Universidade Estadual de Campinas
Instituto de Computação

Victória Pedrazzoli Ferreira

O Futuro do Monitoramento Automático de Publicidades: Como Redes Transformers Impactam na Classificação de Publicidades Alimentícias

CAMPINAS
2024

Victória Pedrazzoli Ferreira

O Futuro do Monitoramento Automático de Publicidades: Como Redes Transformers Impactam na Classificação de Publicidades Alimentícias

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestra em Ciência da Computação.

Orientadora: Profa. Dra. Sandra Eliza Fontes de Avila
Coorientadora: Profa. Dra. Paula Martins Horta
Coorientadora: Profa. Dra. Leo Sampaio Ferraz Ribeiro

Este exemplar corresponde à versão final da Dissertação defendida por Victória Pedrazzoli Ferreira e orientada pela Profa. Dra. Sandra Eliza Fontes de Avila.

CAMPINAS
2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Ferreira, Victória Pedrazzoli, 1998-
F413f O futuro do monitoramento automático de publicidades : como redes transformers impactam na classificação de publicidades alimentícias / Victória Pedrazzoli Ferreira. – Campinas, SP : [s.n.], 2024.

Orientador(es): Sandra Eliza Fontes de Avila.
Coorientador(es): Paula Martins Horta, Leo Sampaio Ferraz Ribeiro.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Aprendizado de máquina. 2. Aprendizado profundo. 3. Transformers (Arquitetura de computador). 4. Classificação multimodal. 5. Publicidade de alimentos. I. Avila, Sandra Eliza Fontes de, 1982-. II. Horta, Paula Martins. III. Ribeiro, Leo Sampaio Ferraz, 1995-. IV. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. V. Título.

Informações complementares

Título em outro idioma: The future of automated advertising monitoring : how transformers networks impact the classification of food advertising
Palavras-chave em inglês:
Machine learning
Deep learning
Transformers (Computer architecture)
Multimodal classification
Food publicity
Área de concentração: Ciência da Computação
Titulação: Mestra em Ciência da Computação
Banca examinadora:
Sandra Eliza Fontes de Avila [Orientador]
Paula Dornhofer Paro Costa
Laís Amaral Mais
Data de defesa: 25-10-2024
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0002-4865-9886
- Currículo Lattes do autor: http://lattes.cnpq.br/5193523403814592

Universidade Estadual de Campinas
Instituto de Computação

Victória Pedrazzoli Ferreira

O Futuro do Monitoramento Automático de Publicidades: Como Redes Transformers Impactam na Classificação de Publicidades Alimentícias

Banca Examinadora:

- Profa. Dra. Sandra Eliza Fontes de Avila (Orientadora)
  Universidade Estadual de Campinas (UNICAMP)
- Profa. Dra. Paula Dornhofer Paro Costa
  Universidade Estadual de Campinas (UNICAMP)
- Dra. Laís Amaral Mais
  Instituto Brasileiro de Defesa do Consumidor (IDEC)

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 25 de outubro de 2024

Agradecimentos

Primeiramente, gostaria de expressar minha mais profunda gratidão à Professora Sandra Avila, minha orientadora, por toda a inspiração, dedicação, apoio, carinho e amizade ao longo desta jornada. Seu compromisso foi essencial para o sucesso desta dissertação e meu maior incentivo para me aventurar pela área acadêmica.

Agradeço também à minha co-orientadora Paula Martins e à minha parceira Michele Rodrigues pela oportunidade de participar deste projeto e pelos valiosos conhecimentos compartilhados. Sua disposição em ensinar e orientar foi imprescindível, especialmente em um campo que eu ainda conhecia tão pouco.

Um obrigada especial à minha outra co-orientadora, Leo, que, apesar de ter se juntado mais tarde à jornada, desempenhou um papel essencial, redirecionando completamente os rumos desta dissertação. Sem sua ajuda, eu não teria chegado até aqui!

À UNICAMP, ao SAE e a todo o corpo docente, sou profundamente grata por terem sido pilares fundamentais na minha formação acadêmica, proporcionando aprendizagens valiosas ao longo dos muitos anos em que fiz parte desta universidade.

Minha gratidão também vai ao Becas Santander, cujo financiamento me permitiu dedicar-me a este projeto.

Aos amigos e familiares, incluindo as adições não mais tão recentes, Loraine e Valter, agradeço por todo o apoio, incentivo e compreensão ao longo deste processo. Sua presença e suporte foram imprescindíveis para a conclusão desta dissertação.

Sou especialmente grata aos meus pais, Carla e Vitório, que sempre me apoiaram incondicionalmente. Obrigada por todo o incentivo e por compreenderem minha ausência em tantos momentos. Seu amor e suporte foram fundamentais para que eu pudesse me dedicar plenamente a esta jornada. À minha avó Silvandira, meu avô Ary e minha madrinha e tia Ana, envio um obrigada especial, onde quer que estejam, por continuarem a me inspirar e iluminar o meu caminho.

A Maria, ao Eduardo e ao André, sou grata pelos momentos de lazer e descanso que tanto me ajudaram a recarregar as energias ao longo dessa trajetória e por escutarem as milhares de reclamações quando eu achava que tudo daria errado, sem nunca se queixarem.

Por fim, a todos que, de alguma forma, direta ou indiretamente, contribuíram para a realização desta dissertação, deixo aqui os meus mais sinceros agradecimentos. Obrigada por enriquecerem meu processo de aprendizado.

Resumo

A obesidade é uma doença crônica de escala global que representa um grave problema de saúde pública, sendo um dos principais fatores de risco para doenças não transmissíveis. Esse cenário está associado a mudanças nos hábitos alimentares, que incluem um aumento do consumo de alimentos ultraprocessados. A publicidade de alimentos exerce um papel central nessa transformação, ao influenciar negativamente os padrões de consumo.

Em resposta a esse reconhecimento, a Organização Mundial da Saúde tem destacado a importância de monitorar e restringir esse tipo de publicidade. Diversos países já adotaram medidas legais para limitar as publicidades de alimentos, com a criação de projetos de lei que regulamentam a publicidade televisiva.

No entanto, estudos de monitoramento de publicidades televisivas mostram que ainda prevalecem publicidades de alimentos com baixo valor nutricional, frequentemente associadas às estratégias de marketing persuasivas. Diante desse contexto, torna-se necessário implementar um sistema automático para monitorar publicidades de alimentos e bebidas não alcoólicas que seja capaz de classificar os tipos de publicidade e suas principais estratégias de marketing.

Embora bases de dados contendo publicidades alimentícias sejam limitados, a rede INFORMAS Brasil, com nosso apoio, tem desenvolvido um novo banco de publicidades televisivas brasileiras para essa finalidade. Ainda assim, a escassez de dados e a complexidade da classificação dessas publicidades, representam grandes desafios. Isso porque a base ainda apresenta problemas de distribuição, o que significa que algumas classes estão desbalanceadas, com uma quantidade desigual de exemplos. Esse desequilíbrio pode afetar o desempenho dos modelos, tornando mais difícil para eles aprenderem a identificar corretamente classes menos representadas.

Apesar de nossos esforços para mitigar esse problema, ele persiste e requer atenção. Para a classificação binária, distinguindo entre publicidades alimentícias e não alimentícias, o melhor modelo (EviT, baseado em Transformers) alcançou uma acurácia balanceada de 92,4%. Já ao expandir para quatro classes (*fast-food*, supermercado, alimento/bebida e não alimentícia), o modelo mais eficaz (EfficientNet, baseado em redes neurais convolucionais) obteve uma acurácia balanceada de 87,4%.

O modelo baseado na EviT também demonstrou capacidade de classificar publicidades alimentícias de maneira binária provinda de outras mídias (*YouTube*) sem maiores adaptações. Mas, apresentou dificuldades com publicidades não alimentícias, resultando em falsos negativos, especialmente em vídeos infantis. A semelhança visual entre publicidades infantis e de alimentos, com cores vibrantes e personagens lúdicos, mostrou ser um complicador na diferenciação. Além disso, as mudanças nos padrões visuais de publicidades recentes (2022-2024) em relação à base de dados original (2018-2020) prejudicaram a generalização do modelo.

Resumidamente, ao avaliar o desempenho considerando o conjunto de dados coletados em sua totalidade, as Transformers se destacaram por sua maior capacidade de generali-

zação, evidenciando o potencial dos modelos baseados em atenção para capturar padrões mais complexos e contextuais em dados visuais. Essas técnicas não só podem acelerar como também aumentar a precisão no monitoramento de publicidades alimentares, complementando o trabalho manual realizado por especialistas e oferecendo grandes benefícios à saúde pública.

Abstract

Obesity is a chronic disease worldwide and represents a severe public health issue. It is one of the main risk factors for non-communicable diseases. This situation is linked to changes in eating habits, including increased consumption of ultra-processed foods. Food advertising plays a central role in this shift, as it negatively influences consumption patterns.

In response to this concern, the World Health Organization has emphasized the importance of monitoring and restricting such advertising. Several countries have already adopted legal measures to limit food advertisements by creating legal measures that regulate television advertising.

However, studies monitoring TV ads show that advertisements for low-nutritional-value foods still prevail, often associated with persuasive marketing strategies. Given this context, it becomes necessary to implement an automated system to monitor non-alcoholic food and beverage advertisements capable of classifying the types of ads and their main marketing strategies.

Although databases containing food ads are limited, the INFORMAS Brazil network, with our support, has developed a new database of Brazilian TV ads for this purpose. Still, the scarcity of data and the complexity of classifying these ads posed significant challenges. This is because the database still presents distribution problems, meaning that some classes are unbalanced, with an unequal number of examples.

This imbalance can affect model performance, making it more difficult for them to correctly identify underrepresented classes. Despite our efforts to mitigate this problem, it persists and requires attention. For binary classification, distinguishing between food and non-food advertisements, the best model (EviT) achieved a balanced accuracy of 92.4%. When expanding to four classes (*fast-food*, supermarket, food/beverage, and non-food), the most effective model (EfficientNet) achieved a balanced accuracy of 87.4%.

EviT also showed some ability to classify food ads in a binary manner from other media (*YouTube*) without major adaptations but faced difficulties with non-food ads, resulting in false negatives, especially in children's videos. The visual similarity between children's and food ads, with vibrant colors and playful characters, complicated differentiation. Additionally, changes in visual patterns of recent ads (2022–2024) compared to the original database (2018–2020) hindered the model's generalization.

In summary, when evaluating the performance across the entire dataset, Transformers stood out for their greater generalization capability, highlighting the potential of attention-based models to capture more complex and contextual patterns in visual data. These techniques can not only speed up but also increase the accuracy of monitoring food ads, complementing the manual work done by experts and offering significant benefits to *public health*.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Descrição do Problema

A obesidade é um problema de saúde pública que atinge, de maneira crescente, vários países, sendo essa uma doença crônica e um dos principais fatores de risco para várias doenças não transmissíveis (por exemplo, hipertensão, doenças cardiovasculares, diabetes, depressão e ansiedade) que estão entre as principais causas de morte globalmente [33, 57].

Em 2022, a Organização Mundial de Saúde (OMS) estimou que mais de 2,5 bilhões de adultos estavam acima do peso, sendo mais de 890 milhões de obesos. Em relação às crianças e adolescentes de 5 a 19 anos, cerca de 390 milhões apresentavam excesso de peso, dos quais 160 milhões viviam com obesidade [59]. No Brasil, a prevalência de obesidade na população adulta também aumentou consideravelmente em menos de 20 anos [18].

O crescente aumento do consumo de produtos alimentícios ultraprocessados [7] e a redução na inclusão de alimentos in natura ou minimamente processados na dieta dos brasileiros é uma das principais causas do aumento das taxas de excesso de peso. Esses produtos, caracterizados por seu baixo teor de fibras e elevada densidade energética, além de concentrações significativas de gorduras saturadas, açúcar e sódio, têm impacto direto na saúde da população [45, 46, 49]. No Brasil, estimativas de 2019 indicam que mais de 10% das mortes de indivíduos entre 30 e 69 anos foram atribuídas ao consumo desses alimentos ultraprocessados [53]. [53].

A publicidade tem grande participação nessa mudança de dieta pois ela pode influenciar negativamente o consumo alimentar dos indivíduos [9]. Isso ocorre porque, normalmente, é centrada em alimentos não saudáveis. Além disso, a publicidade pode utilizar estratégias persuasivas com o objetivo de impactar o comportamento do consumidor, acarretando na fidelização pela marca desde a mais tenra idade [52, 68, 69].

Tal fato é extremamente preocupante porque grande parte dos hábitos alimentares se formam na infância, tornando o público infantil o mais vulnerável por esse tipo de publicidade, visto que, por ainda estarem em processo de desenvolvimento cognitivo, não conseguem identificar o caráter persuasivo da mensagem publicitária [8, 13, 37].

A recomendação da OMS em 2010 para reduzir a exposição das crianças à publicidade de alimentos, sobretudo de produtos ultraprocessados [58], reflete a crescente preocupação com o impacto desses alimentos na saúde infantil. Logo, o monitoramento e a fiscalização são cruciais para mitigar o impacto dessas publicidades na saúde das crianças.

No Brasil, assim como em muitos outros países, apesar das regulamentações estabelecidas no Código de Defesa do Consumidor e no Estatuto da Criança e do Adolescente (ECA), os órgãos de fiscalização, como o Conselho Nacional dos Direitos da Criança e do Adolescente (CONANDA), ainda enfrentam limitações que impedem a aplicação efetiva dessas normas, incluindo a análise de grandes volumes de conteúdo de mídia, resultando em processamento lento, caro e propenso a erros humanos, o que retarda a resposta das autoridades. Publicidades potencialmente abusivas podem alcançar e assim afetar vastos públicos infantis [8, 37] antes de serem formalmente identificadas e retiradas de circulação.

Portanto, a implementação de tecnologias de monitoramento automatizado pode ser uma solução importante para melhorar o monitoramento e a fiscalização dessas publicidades. Ferramentas baseadas em Inteligência Artificial (IA) podem auxiliar na triagem de grandes volumes de conteúdo de mídia (por exemplo, televisão, internet e outras plataformas digitais), identificando automaticamente características de publicidades de alimentos e bebidas não alcoólicas que infringem as regulamentações.

Ainda assim, a relevância desta dissertação de Mestrado extrapola a fiscalização de publicidades, pois fortalece a base acadêmica atual, e assim contribui diretamente para a geração de jurisprudências, o que desempenha um papel essencial na formulação de políticas públicas.

Diante dos dados disponíveis atualmente, concentramos especificamente em uma solução voltada para a classificação de quatro tipos de publicidades: publicidades de supermercado, de *fast-food*, de alimento ou bebida, e não alimentícias. Todas as publicidades foram coletadas e classificadas por pessoas especialistas da área de Nutrição.

Destacamos que, apesar do público infantil ser um motivador relevante para este trabalho, os modelos desenvolvidos não lidam diretamente com a classificação baseada em conteúdos direcionados a crianças. A proposta atual busca contribuir indiretamente para o monitoramento dessas publicidades.

Neste âmbito, os métodos de aprendizado de máquina e IA já se provaram efetivos nas mais diversas tarefas, incluindo a classificação de imagens e vídeos [5, 14, 34, 80], tarefa principal desta dissertação. Dentre os métodos de aprendizado de máquina, as redes neurais Transformers (ViTs, *Vision Transformers* [20]) representam o estado da arte, superando as Redes Neurais Convolucionais (CNNs, *Convolutional Neural Networks* [41]) ao utilizar módulos de atenção, especialmente em grandes conjuntos de dados [20]. No entanto, o uso dessas ViTs em conjuntos de dados pequenos e médios ainda apresenta desafios [12], como é o caso dos vídeos de publicidades de alimentos e bebidas não alcoólicas.

1.2 Motivações e Desafios

Atualmente, a quantidade de bases de dados contendo publicidades alimentícias é quase nula. Um dos principais fatores responsáveis pela escassez era o desconhecimento do tema e a inexistência de protocolos padronizados de coleta. Dessa maneira, os estudos focados em classificação automática de publicidades alimentícias são escassos, e diminuem ainda mais quando tratamos de publicidades brasileiras, sendo estes inexistentes.

Para abordar essa questão, em 2018, um grupo de pesquisadores da rede INFORMAS

(*International Network for Food and Obesity/Non-communicable Disease Research, Monitoring and Action Support*) [74] Brasil iniciou a construção de uma base de dados de publicidades com uma ou mais estratégias persuasivas de marketing (veja a Tabela 4.3, a qual é utilizada nesta dissertação) com o intuito de assim monitorar as publicidades alimentícias e a partir disso facilitar o processo de fiscalização e tomada de decisões. Apesar dos avanços obtidos, a base de dados ainda é limitada devido à sua recente construção, e depende de diversas etapas que tornam o processo de construção mais lento, como a coleta e a anotação de dados de forma manual.

A limitação de dados tem um grande impacto nesta pesquisa, uma vez que a base de dados em questão é pequena (com algumas centenas de vídeos) quando comparada a outras mais comuns na literatura dentro do campo da Visão Computacional [40, 67, 82]. Por exemplo, bases populares de imagens e vídeos, como a ImageNet-21K [64] e o YouTube-8M [1], possuem milhões de imagens e vídeos, o que destaca ainda mais a limitação em termos de tamanho da nossa base.

Quando analisamos as CNNs em respeito ao tamanho da base de dados utilizada, elas costumam performar bem em bases menores devido ao seu forte viés indutivo [30], fazendo com que sejam particularmente eficazes na codificação de informações locais por meio de convoluções, o que, além de se adequar às tarefas visuais, reduz a quantidade de dados necessários.

Em contraste, as Transformers, que dependem de mecanismos de auto-atenção, requerem uma quantidade significativamente maior de dados (pois não possuem esse viés local implícito). Para contornar essa limitação, trabalhos recentes têm adotado novas estratégias, mesclando técnicas comuns às CNNs. Entre essas abordagens, destacam-se técnicas como a *Teacher-Student*, em que redes CNNs pré-treinadas são utilizadas como mentoras para guiar o aprendizado de Transformers [79], o desenvolvimento de arquiteturas híbridas [25] e adaptações diretas nos blocos de uma Transformer para incorporar convoluções [21, 62, 85].

Especificamente, o uso de arquiteturas que empregam estratégias de atenção para classificar publicidades, que, por sua própria natureza, também aplicam técnicas para captar a atenção do público, parece especialmente promissor. No entanto, vale destacar que, apesar dos resultados encorajadores das arquiteturas Transformers, a maioria dos trabalhos pressupõe uma quantidade de dados muito maior do que a que temos disponível. Consequentemente, sua performance geralmente decai consideravelmente quando poucas amostras estão disponíveis.

Além disso, o próprio teor dos dados coletados para essa dissertação é complexo e dificulta sua classificação, visto que o fato de uma publicidade ser alimentícia ou não, não necessariamente garante a presença ou a falta de alimentos durante a publicidade, ainda mais quando consideramos trabalhar com quatro classes (publicidades de supermercado, de *fast-food*, de alimento ou bebida, e não alimentícias). Quanto às estratégias persuasivas de marketing, algumas têm caráter subjetivo, onde nem mesmo os especialistas da área chegam a um consenso quanto à sua classificação. Por exemplo, a presença de uma celebridade em uma publicidade pode ser interpretada de diferentes maneiras: alguns especialistas podem classificá-la como uma estratégia de apelo à autoridade, enquanto outros podem considerá-la apenas uma tática de reconhecimento de marca. Da mesma

forma, o uso de cores e sons pode ser percebido como uma forma de enfatizar o produto ou criar uma conexão emocional.

Neste sentido, a aplicação de técnicas computacionais para monitorar e fiscalizar publicidades de alimentos no Brasil complementa o trabalho manual atualmente realizado e pode trazer grandes benefícios à saúde pública. Isso ocorre porque essas técnicas facilitam o processo de tomada de decisões por parte das pessoas especialistas, permitindo acelerar a implementação de medidas mais eficazes no âmbito governamental [36].

1.3 Objetivos

O principal objetivo desta dissertação de Mestrado é conduzir uma investigação aprofundada das principais técnicas de aprendizado para Transformers — e das adaptações necessárias dessas técnicas, para que se possa desenvolver um método que auxilie o monitoramento automático de publicidades de alimentos e bebidas não alcoólicas de canais de televisão no Brasil. Mais especificamente, os objetivos desta dissertação são:

O1. Criar uma base de dados com diferentes tipos de publicidades de alimentos e bebidas não alcoólicas de canais de televisão no Brasil, especificamente dividida em quatro classes.

O2. Avaliar a aplicabilidade de Transformers para classificação de publicidades alimentícias.

O3. Desenvolver um modelo que facilite o processo de investigação e classificação de publicidades alimentícias.

1.4 Questões de Pesquisa

Q1. Os resultados obtidos pelas redes neurais profundas baseadas em Transformers superam os resultados obtidos pelas redes neurais convolucionais para o problema de classificação de publicidades alimentícias?

Q2. Como gerar um modelo eficaz de classificação baseado em Transformers, contornando o problema de desbalanceamento e falta de dados, que seja capaz de identificar os diferentes tipos de publicidades alimentícias?

Q3. Como combinar áudio e Transformers para melhorar a classificação de publicidades alimentícias?

1.5 Contribuições

C1. Criação da primeira base de dados com diferentes tipos de publicidades alimentícias brasileiras, divididas em quatro classes.

C2. Implementação e avaliação de diferentes arquiteturas de redes neurais profundas, incluindo *EfficientNet*, *EViT*, e *Everything at Once*, no contexto da classificação de publicidades alimentícias televisivas brasileiras.

C3. Desenvolvimento de modelos de classificação publicidades alimentícias e não alimentícias de maneira binária (alimentícias vs. não alimentícias) e multiclasse (publicidades de supermercado, de *fast-food*, de alimento ou bebida, e não alimentícias).

C4. Demonstração da eficácia do uso de áudio no contexto da classificação de publicidades alimentícias televisivas brasileiras.

1.6 Organização do Texto

Esta dissertação está organizada da seguinte forma. No Capítulo 2, são apresentados os principais conceitos, tanto da área computacional quanto da área da nutrição, para o entendimento deste texto. No Capítulo 3, são descritos e apresentados os principais trabalhos relacionados a esta pesquisa, que foram identificados por uma revisão da literatura focada na aplicação de Transformers para classificação de imagens e vídeos. No Capítulo 4, é descrita a base de dados bem como uma breve descrição das etapas de coleta e preparação da mesma. No Capítulo 5, é apresentada a metodologia proposta e os passos gerais para atingir nossos objetivos e validar nossa proposta. No Capítulo 6, apresentamos os principais resultados desta dissertação. Por fim, no Capítulo 7, é apresentada as conclusões, trabalhos futuros e considerações éticas desta pesquisa.

Capítulo 2

Conceitos Relacionados

Neste capítulo, apresentamos os conceitos relacionados ao desenvolvimento desta dissertação. Especificamente, vamos abordar a rede INFORMAS (Seção 2.1) e seu protocolo de coleta de publicidades alimentícias, as Transformers (Seção 2.2), como as ViTs, EViT, *Everything at Once*, e a EfficientNet (Seção 2.3), uma rede neural convolucional.

2.1 INFORMAS

A rede INFORMAS (*International Network for Food and Obesity/Non-communicable Disease Research, Monitoring and Action Support*) [83] (ou, em português, Rede Internacional para Pesquisa, Monitoramento e Apoio à Ação em Alimentos e Obesidade/Doenças Não Transmissíveis) foi criada em 2013 para monitorar de maneira abrangente os ambientes alimentares e avaliar o impacto das políticas do setor público e privado, a fim de fortalecer os sistemas de responsabilização e prevenir a obesidade e as doenças crônicas não transmissíveis. Atualmente, a rede INFORMAS conta com a colaboração de 41 países[1], incluindo o Brasil, sendo composta por uma ampla gama de instituições acadêmicas e organizações de saúde pública.

Essa rede propõe algumas formas de atuação diferentes, tanto no setor público quanto no privado, sendo o protocolo de monitoramento da publicidade televisiva de alimentos [35] mais importante para esta pesquisa. Especificamente, esse protocolo procura direcionar a coleta de dados e contabilizar a promoção de alimentos supersaturados e bebidas não alcoólicas feitas no meio televisivo. Ele visa estabelecer uma metodologia para avaliar a frequência e o nível de exposição de grupos populacionais — especialmente crianças menores de 18 anos — às promoções de alimentos, o poder de persuasão das técnicas utilizadas nas comunicações promocionais e a composição nutricional dos produtos alimentícios.

Para realizar tal tarefa, além das informações contextuais (referente ao país, regulamentações existentes, ano de coleta, horário), o INFORMAS recomenda a coleta mínima das seguintes variáveis descritas na Tabela 2.1.

A partir dos dados coletados, é possível fazer uma análise através dos indicadores recomendados, divididos em dois grupos: os de monitoramento de publicidade televisiva e os de 'poder de persuasão' do conteúdo dessas publicidades. Ambos são avaliados

[1] `https://www.informas.org/countries`

Tabela 2.1: Variáveis mínimas recomendadas pelo protocolo INFORMAS para a coleta de dados sobre publicidades televisivas.

Variável	**Descrição**
Nome do país	Identificação do país onde a coleta de dados é realizada.
Área de coleta de dados	Região específica de coleta.
Ano de coleta	Ano em que os dados foram coletados.
Nome e número do canal	Identificação do canal onde a publicidade foi exibida.
Percentual de audiência	Percentual da audiência do canal nos horários de maior visualização.
Data da gravação	Data em que a gravação foi feita, no formato dia/mês/ano.
Dia da semana	Dia em que a publicidade foi exibida.
Nome do programa	Nome do programa em que a publicidade apareceu.
Categoria do programa	Classificação do programa (notícias, esportes, novelas, etc.).
Hora de início e término do programa	Horários de início e fim do programa onde a publicidade foi exibida.
Faixa de horário da publicidade	Código que representa a faixa de horário da publicidade conforme uma tabela predefinida.
Hora de início e término da publicidade	Horários específicos de início e fim do publicidade.
Momento da publicidade	Indicação se a publicidade foi exibida durante uma pausa de um programa específico ou entre dois programas.
Tipo de publicidade	Código que classifica a publicidade em categorias como produto alimentício ou bebida, promovido por uma empresa/marca, etc.
Nome da empresa	Nome da empresa responsável pela publicidade.
Nome e descrição do produto	Descrição detalhada do produto anunciado, incluindo variações de sabor ou marca.
Categoria do produto alimentício	Classificação detalhada do produto alimentício (produto alimentar ou bebida — empresa/marca de alimentos, produto alimentar ou bebida promovido em propaganda por marca/empresa/varejista/serviço/evento não alimentar, empresa ou marca de alimentos ou bebidas (sem varejista) sem comida ou bebida, supermercado ou loja de conveniência com alimentos ou bebidas, supermercado ou loja de conveniência sem produtos alimentares ou bebidas, restaurante ou *takeaway* ou *fast-food* com alimentos ou bebidas,restaurante ou *takeaway* ou *fast-food* sem comida ou bebida, produto não alimentar ou bebida)
Indicador de permissão para marketing para crianças	Informação sobre se o produto é permitido para ser comercializado para crianças pela Organização Mundial da Saúde ou pela Organização Pan-Americana da Saúde.
Indicador de poder da publicidade	Informação sobre se foram usadas estratégias de persuasão na publicidade.
Tipo de estratégia de publicidade	Código que descreve a estratégia de persuasão utilizada, como personagens de desenho animado, celebridades, etc.
Descrição do poder da publicidade	Descrição das estratégias de persuasão utilizadas na publicidade.
Indicador de ofertas promocionais presentes	Informação sobre a presença de ofertas promocionais na publicidade.
Descrição das ofertas promocionais	Detalhamento das ofertas promocionais presentes, como jogos, concursos, descontos, etc.
Sistema de categorização alimentar específico do país	Sistema opcional para categorizar alimentos conforme as especificidades do país.
Estratos	Definição dos estratos com base em se o dia da coleta foi um dia de semana ou fim de semana.
Peso	Variável derivada para análise adicional.

considerando cada hora ao longo do dia, horários de pico e não pico para crianças de 5 a 12 anos, horários regulados para crianças (manhã e tarde) e programas populares para esse público.

Para os indicadores de **monitoramento da publicidade televisiva**, a intenção é avaliar a frequência e a composição das publicidades exibidas, ou seja, o nível de publicidade, através dos seguintes dados:

- Taxa média ou frequência de anúncios por canal por hora.
- Taxa média ou frequência de anúncios de alimentos versus anúncios não alimentares por canal por hora.
- Taxa média ou frequência de anúncios de alimentos não saudáveis versus anúncios de alimentos saudáveis por canal por hora.
- Taxa média ou frequência de grupos de alimentos não saudáveis (não essenciais) por canal por hora.
- Taxa média ou frequência de publicidades que utilizam técnicas promocionais persuasivas, diferenciando entre alimentos saudáveis e não saudáveis (essenciais e não essenciais).
- Razão entre publicidades de alimentos saudáveis e não saudáveis.
- Número total de publicidades de alimentos saudáveis versus não saudáveis.
- Proporção de publicidades de alimentos por principais categorias alimentares.

Já para os indicadores de **'poder de persuasão'**, o conteúdo dessas promoções é avaliado pelas seguintes taxas:

- Taxa média de personagens promocionais (alimentos vs. não alimentos; alimentos não saudáveis versus alimentos saudáveis).
- Taxa média de prêmios oferecidos (alimentos vs. não alimentos; alimentos não saudáveis versus alimentos saudáveis).
- Proporção de alegações nutricionais e de saúde associadas a alimentos não saudáveis vs. alimentos saudáveis.
- Proporção de publicidades de alimentos por principais categorias alimentares.

Esses indicadores são essenciais para entender a influência do conteúdo publicitário sobre o público, especialmente crianças e adolescentes, e assim poder pensar em diferentes estratégias de intervenção.

O protocolo também cobre o controle de qualidade sobre os dados e estudos derivados deste, incluindo até mesmo testes de confiabilidade para garantir comparabilidade dos resultados. Assim, todas as pessoas assistentes de pesquisa envolvidas na análise dos dados (também denominadas de codificadores), primeiro recebem um treinamento e depois são avaliadas por duas notas, uma calculada com base nas outras pessoas pesquisadoras de seu país de origem e outra com base nas pessoas pesquisadoras dos outros países. Espera-se uma alta confiabilidade entre codificadores (de 90% a 100%).

Historicamente, havia uma falta de protocolos padronizados que considerassem todas as etapas, desde a coleta até a análise dos dados. Por esse motivo, o INFORMAS tornou-se uma referência mundial no monitoramento da publicidade televisiva. Além de estar alinhado com as recomendações da OMS, este estabelece critérios detalhados para amostragem, definição do período de coleta de dados e variáveis a serem coletadas, incluindo tipos de alimentos anunciados e estratégias de marketing utilizadas.

2.2 Transformers

As Transformers [84] são redes neurais profundas propostas em 2017 para tarefas de tradução automática, que rapidamente se tornaram uma ferramenta essencial no processamento de linguagem natural (NLP, *Natural Language Processing*) e outras áreas da IA. A principal inovação dessas é o uso de mecanismos de auto-atenção (*self-attention*), que refletem o conceito lúdico do que é a atenção humana. Elas têm causado bastante "barulho" por causa da ferramenta ChatGPT (e variações); o T da sigla GPT significa Transformer.

O conceito de 'atenção' foi primeiramente introduzido no contexto da tradução automática em 2014 por Bahdanau [4], através do mecanismo de atenção cruzada (*cross-attention*), permitindo que o modelo focasse em diferentes partes de uma sentença de entrada. A equação que rege esse processo é dada por:

$$\alpha_{ij} = \frac{\exp(f(s_{i-1}, h_j))}{\sum_{k=0}^{K} \exp(f(s_{i-1}, h_k))}, c_i = \sum_{k=0}^{K} \alpha_{ik} h_k,$$

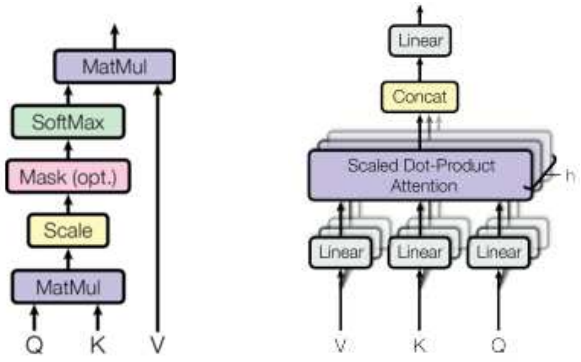
onde s_{i-1} é o estado atual do decodificador, $h_0, \ldots, h_k$ são os estados ocultos do codificador, e f é a função aprendida pelo modelo, que são usados para calcular o contexto c, que é a entrada para o próximo passo do decodificador.

A entrada refere-se as unidades básicas de processamento chamadas *tokens*. Para as tarefas textuais, estes podem ser palavras, sílabas ou até mesmo caracteres, dependendo do nível de granularidade escolhido. Essa abordagem permite que o modelo analise e manipule o texto em partes, facilitando o aprendizado das relações semânticas.

Em 2017, foi proposta uma evolução para um esquema de auto-atenção, permitindo com que o modelo avalie a relação entre todas as palavras de uma sentença simultaneamente, sem depender de um processamento sequencial. A partir do contexto, a auto-atenção procura aproximar os *tokens* mais semelhantes e distanciar os menos semelhantes.

Esse esquema, ilustrado na Figura 2.1a, é descrito matematicamente como uma função denominada "atenção de produto escalado". O processo começa convertendo os tokens de entrada em vetores (*embeddings*), que são transformados linearmente, por meio de multiplicações matriciais, em três matrizes: *query* (Q), *key* (K) e *value* (V).

A saída desta função é calculada por uma soma ponderada. O peso de cada entrada é determinado pela similaridade entre os vetores *query* e *key*, que é normalizada pela dimensão dos vetores (d_k) e transformada em probabilidades por meio de uma função *softmax*. Essa técnica permite que o modelo ajuste a relevância das diferentes partes da entrada, atenuando ou amplificando-as conforme necessário.



(a) *Scaled dot-product attention* (b) *Multi-head self-attention*

Figura 2.1: Esquema de atenção: à esquerda, *Scaled dot-product attention*; à direita, *Multi-head self-attention*. Figura reproduzida de Vaswani et al. [84].

Na prática, a função de atenção é aplicada simultaneamente a toda a sequência, utilizando as matrizes Q, K, e V, conforme descrito na Equação 2.1.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V. \tag{2.1}$$

Para que o modelo possa se concentrar em várias posições relevantes ao mesmo tempo, os autores propuseram o *multi-head self-attention* (Figura 2.1b). Este mecanismo consiste em múltiplas camadas de auto-atenção operando em paralelo. Cada camada (*head* ou "cabeça") produz uma saída, de forma independente, que é então combinada e normalizada por uma normalização em camada [3], permitindo ao modelo capturar diferentes aspectos das relações entre os *tokens* de entrada. Essa combinação é formalizada da seguinte forma:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V),$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O,$$

sendo W_i^Q, W_i^K, W_i^V as matrizes de projeção para a i-ésima *head*, e W^O a matriz de projeção final após a concatenação das saídas das h cabeças de atenção.

A Figura 2.2 ilustra a arquitetura final de uma Transformer, que é fundamentada na estrutura de codificador-decodificador proposta por Bahdanau [4], sendo esta composta por uma pilha de camadas de auto-atenção (o *multi-head self-attention*), organizadas em módulos de codificadores e decodificadores. Cada etapa do modelo é auto-regressiva, utilizando os símbolos gerados anteriormente como entrada para a próxima camada [26].

A eficácia das Transformers em NLP tem sido demonstrada por vários trabalhos subsequentes [19, 63]. Foi devido ao sucesso de Transformers que pesquisadores começaram recentemente a explorar o seu uso em outras áreas.

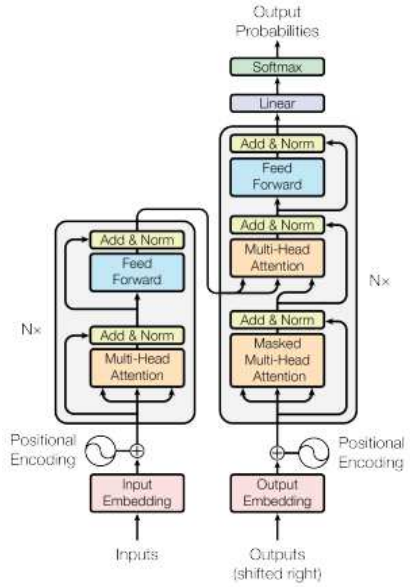


Figura 2.2: Arquitetura de uma Transformer. Figura reproduzida de Vaswani et al. [84].

Transformers Visuais

O primeiro trabalho a aplicar com sucesso Transformers em Visão Computacional foram as *Vision Transformers* (ViTs) (ou Transformers Visuais) de Dosovitskiy et al. [20], que propõe a utilização de Transformers nas tarefas de visão computacional.

A arquitetura das ViTs divide uma imagem em pequenos *patches*. Cada *patch* é composto por um vetor de dimensão fixa que representa parte da imagem original. Este é combinado com outro *embedding* posicional, de mesmo tamanho, a fim de manter a informação espacial (Figura 2.3).

Isso é necessário visto que as Transformers originalmente em sua proposição não têm uma noção intrínseca da sequência dos dados. Assim, as ViTs adicionam a entrada um vetor posicional, que neste caso tem uma dimensão, e é inicializado aleatoriamente e treinado juntamente com o modelo. Esses *embeddings* posicionais são somados aos *patches* formando os *tokens* de entrada.

Além disso, as ViTs precisam adaptar parte da arquitetura tradicional de uma Transformer pois foram especificamente projetadas para tarefas de classificação de imagens. Dessa maneira, como outros modelos de NLP utilizados para classificação e se baseando no estado da arte da época para codificadores de Transformers (o BERT [19]), os decodificadores são removidos do fluxograma.

Ademais, um *token* especial de classificação é adicionado à primeira posição da sequência de *embeddings* dos *patches* de maneira com que ele possa representar a informação global da imagem após a passagem pelo codificador e facilitar a tarefa de classificação. Este *token*, um vetor de dimensão fixa (o mesmo tamanho que os *embeddings* dos *patches*), é inicializado com valores aleatórios e é treinado e aprendido juntamente com o modelo.

Com isso, a sequência de entrada de uma ViT está formada, sendo essa os *patches* e os *tokens* de classe e posicionais. É essa entrada que será aplicada ao codificador, e será tratada de maneira semelhantes às palavras de um texto.

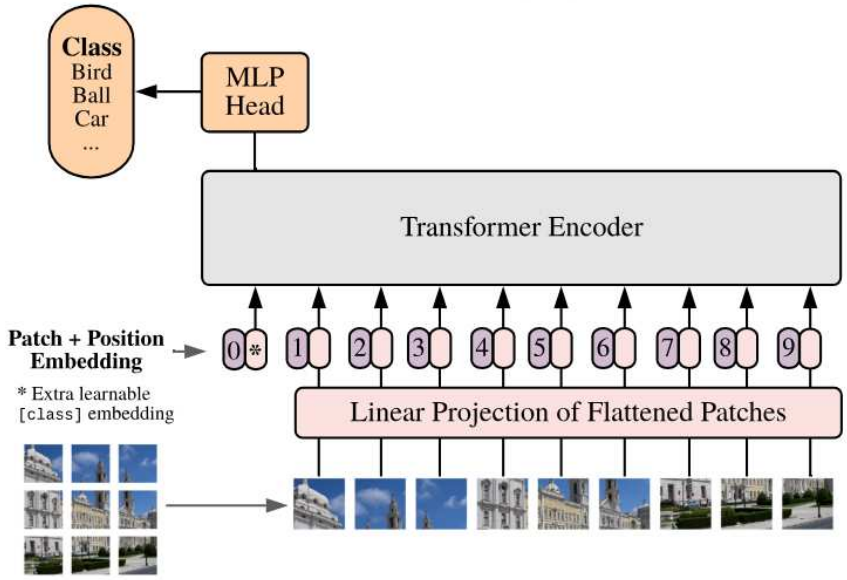


Figura 2.3: Geração dos *token* de entrada de uma ViT e sua arquitetura. Figura reproduzida de Dosovitskiy et al. [20].

Dessa maneira, a estrutura tradicional de uma Transformer pode ser adaptada para tarefas envolvendo imagens, permitindo com que modelo considere simultaneamente a dependência entre cada *patch*, da mesma maneira com que as Transformers tradicionais agem sobre as tarefas de NLP.

Os resultados, apesar de serem promissores e representarem uma mudança de paradigma na visão computacional, dependiam de um vasto conjunto de dados e por isso tinham um alto custo computacional.

Rapidamente, o conceito de Transformers Visuais se popularizou, e diversos outros trabalhos surgiram na área [38]. O primeiro modelo a reduzir consideravelmente a quantidade de dados necessários foi a *Training Data-Efficient Image Transformers* (DeiT) [78], que baseado no seu predecessor usa uma ViT em um modelo de *Teacher-Student*, incluindo um *token* de destilação a fim de inserir suavemente os vieses aprendidos pelo *teacher* (uma rede neural convolucional) para o *student* (a Transformer Visual).

Ao contrário das ViTs tradicionais, que normalmente exigem grandes quantidades de dados (na ordem de centenas de milhões de imagens, como a ImageNet-21K), o DeiT é muito mais eficiente, alcançando desempenho competitivo com um volume consideravelmente menor de dados, como no caso do ImageNet-1K, que possui aproximadamente 1,2 milhão de imagens.

Os resultados do modelo, quando pré-treinados ou não na ImageNet [67], foram competitivos para as base de dados de imagens mais relevantes da literatura, como ImageNet [67], iNaturalist 2019 [82].

A literatura recente, que será discutida na Seção 3.3, demonstra que mecanismos de auto-atenção podem ser benéficos em Visão Computacional.

2.2.1 EViT

A EviT, publicada no artigo "*Not All Patches Are What You Need: Expediting Vision Transformers via Token Reorganizations*" [48], trata de uma abordagem diferente, também baseada nas ViTs. O intuito é aumentar sua eficiência atacando o problema do alto custo computacional, e por consequência também acelerando o processo de treinamento, mas sem perder a precisão.

Isso é feito através de uma proposta de reorganização dos *patches*, que parte da ideia de que, muitas vezes, nem todos os eles contribuem igualmente para a decisão final do modelo. Sendo assim, a ideia é cortar, ou ao menos reduzir, a presença desses *tokens* que contém informações redundantes ou irrelevantes no decorrer do processo o treinamento, fazendo com que o modelo foque apenas naqueles mais informativos e desconsidere ou processe superficialmente os menos relevantes. Como o mecanismo de atenção possui um custo quadrático em relação ao número de *patches*, a redução de quantidade resulta em economias proporcionais no custo computacional.

A EviT utiliza o mesmo conceito de atenção aplicado nas Transformers para identificar quais *patches* de uma imagem são mais relevantes, avaliando a importância de cada *token* na contribuição para a saída do modelo, e com base nessa importância eles são reorganizados dinamicamente. *Patches* informativos são priorizados, enquanto os menos informativos colapsados, processo ilustrado Figura 2.4.

Esse processo é denominado abordagem de fusão de *tokens* inatentos. A fusão é realizada utilizando uma operação de média ponderada, representada como $x_{\text{fused}} = \sum_{i \in N} a_i x_i$, onde N é o conjunto de índices dos *tokens* e a_i são os pesos atribuídos a cada *token*. O *token* fusionado x_{fused} é então anexado aos demais e processado nas camadas subsequentes.

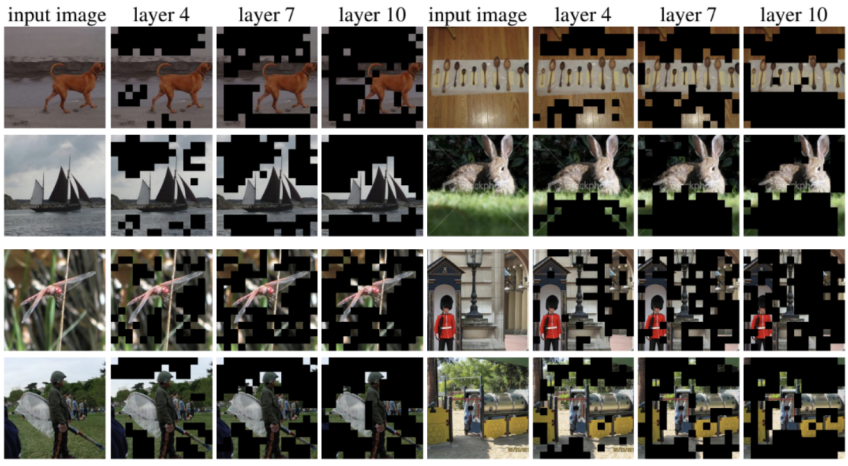


Figura 2.4: Visualização de *patches* inativos em uma EViT. Figura reproduzida de Liang et al. [48].

O método foi proposto para que fosse versátil e pudesse ser aplicado a diferentes arquiteturas, mas garantindo com que a qualidade das predições não fosse comprometida e a precisão permanece comparável a outros modelos de ViTs em conjuntos de dados padrão como ImageNet [67]. Por exemplo, a velocidade de inferência do DeiT-S é aumentado em

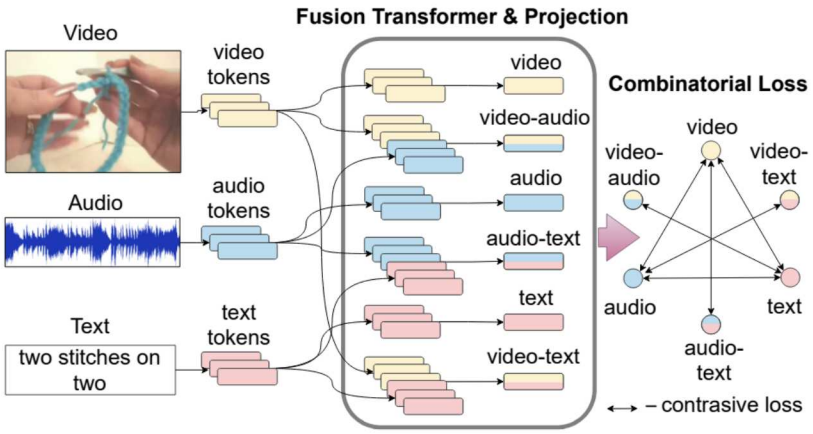


Figura 2.5: Visão geral da abordagem de fusão multimodal da *Everything at Once*. Figura reproduzida de Shvetsova et al. [71].

50%, enquanto sua precisão diminui apenas 0,3% para a ImageNet.

2.2.2 *Everything at Once*

A proposta da *Everything at Once* (EaO), publicada no artigo "*Everything at Once — Multi-modal Fusion Transformer for Video Retrieval*" [71], é abordar o problema da recuperação de vídeos por meio de uma abordagem multimodal, integrando simultaneamente texto, áudio e vídeo durante todo o treinamento. A arquitetura baseia-se em auto-supervisão, onde as três modalidades formam triplas que representam uma mesma instância, aproveitando o fato de que essas fontes de dados estão frequentemente associadas a uma única entidade de informação. Dessa forma, o modelo pode aprender as correspondências entre as modalidades sem a necessidade de rótulos.

Métodos tradicionais de recuperação de vídeos geralmente se limitam a uma ou duas modalidades, restringindo a capacidade de compreensão dos modelos. A EaO supera essa limitação ao utilizar o mecanismo de atenção das Transformers para processar todas as modalidades simultaneamente desde o início do treinamento, com o objetivo de aprimorar a precisão e a eficiência na tarefa de recuperação.

Isso é feito através da fusão multimodal, que foi proposta nessa pesquisa, com o intuito de integrar essas informações de várias modalidades (Figura 2.5). Cada modalidade de entrada é codificada usando outros modelos específicos [30, 51, 66] que extraem as *features* (ou recursos) relevantes de cada tipo de dado. Essas *features* são transformadas nos *tokens* de entrada por meio de projeções e camadas de normalização específicos para cada modalidade, sendo ambos aprendíveis. Ao final, são obtidos três conjuntos de *tokens*, um para cada modalidade: (1) $[\tau_{i1}, \ldots, \tau_{ik}]$ do texto t_i, (2) $[\nu_{i1}, \ldots, \nu_{im}]$ do vídeo v_i, e (3) $[\alpha_{i1}, \ldots, \alpha_{in}]$ do áudio a_i.

Por fim, uma Transformer recebe como entrada essa unificação e a processa com a intenção de definir a importância de cada modalidade e também integrar esses dados de forma eficaz, fazendo uso de uma combinação de perda contrastiva (*contrastive loss*, descritas nas Equações 2.2, 2.3, 2.4). Tais equações modelam diferentes comparações entre

os vetores de cada modalidade e têm como finalidade garantir que entradas semanticamente semelhantes sejam mapeadas para pontos próximos no espaço de representação. No contexto da EaO, a ideia é não apenas aproximar as modalidades, mas também forçar a troca de informações entre elas, adicionando perdas específicas para cada combinação de modalidades e utilizando o contraste entre elas para guiar o aprendizado.

Por exemplo, L_{t_v}, se refere à comparação entre *tokens* textuais (t) e *tokens* de vídeo (v) e L_{t_va} refere-se à comparação entre *tokens* textuais (t) e a combinação vídeo-áudio (va).

$$\begin{aligned} L = \lambda_{t\_v} L_{t\_v} + \lambda_{v\_a} L_{v\_a} + \lambda_{t\_a} L_{t\_a} + \\ + \lambda_{t\_va} L_{t\_va} + \lambda_{v\_ta} L_{v\_ta} + \lambda_{a\_tv} L_{a\_tv}, \end{aligned} \tag{2.2}$$

onde $\lambda_{m\_m\hat{}}$ denota um coeficiente de ponderação para a combinação $(m, m\hat{})$.

$$L = \sum_{\mathcal{X},\mathcal{Y} \subset \mathcal{M}; \mathcal{X} \cap \mathcal{Y} = \varnothing} \lambda_{\mathcal{X}\mathcal{Y}} L_{\mathcal{X}\mathcal{Y}}. \tag{2.3}$$

A perda combinatória considera todas as combinações possíveis e disponíveis de modalidades e pode ser generalizada para qualquer conjunto de modalidades $\mathcal{M} = \{m_1, \ldots, m_N\}$. $L_{\mathcal{X}\mathcal{Y}}$ é uma perda contrastiva entre as representações fundidas dos subconjuntos $\mathcal{X}$ e $\mathcal{Y}$, e $\lambda_{\mathcal{X}\mathcal{Y}}$ é um coeficiente de ponderação.

Para calcular as perdas contrastivas para todas as combinações, utiliza-se a *Information Noise Contrastive Estimation* (ou, Estimação Contrastiva do Ruído da Informação) com temperatura τ e tamanho de lote B.

$$\mathrm{NCE}(x, y) = -\log \left(\frac{\exp\left(x^\top y / \tau\right)}{\sum_{i=1}^{B} \exp\left(x_i^\top y_i / \tau\right)} \right). \tag{2.4}$$

Durante o treinamento, o modelo é então aplicado seis vezes para obter seis *embeddings* correspondendo às modalidades de texto, vídeo, áudio, texto-vídeo, texto-áudio e vídeo-áudio e assim calcular a perda combinatória.

O treinamento da EaO depende de grandes volumes de dados para que modelo possa generalizar bem e por isso foi concluído na base de dados HowTo100M [50]. Durante o *fine-tuning*, o modelo é treinado em um novos conjunto de dados menores (YouCook2 [88], MSR-VTT [87], CrossTask [89] e Mining YouTube [42]), mais relevante para a tarefa-alvo, a fim de melhorar sua capacidade de mapear as diferentes modalidades. Para todas as bases em que a EaO foi testada, ela foi capaz de atingir o estado da arte.

Os resultados demostram que o modelo supera significativamente os métodos tradicionais de recuperação de vídeos e que a capacidade da EaO de considerar todas as modalidades ao mesmo tempo resulta em uma compreensão aprimorada dos dados.

2.3 EfficientNet

O reconhecimento de imagem é um problema clássico de classificação, e as redes neurais convolucionais (CNNs, *Convolutional Neural Networks*), especificamente as Efficient-

Nets [76], têm grande destaque nessa área.

CNNs são arquiteturas especialmente projetadas para processar dados que têm uma estrutura em grade, como imagens. Essas redes funcionam aplicando filtros convolucionais em cada camada para a detecção de padrões visuais em diferentes níveis fazendo uso de métodos como convolução, *pooling* e camadas totalmente conectadas.

A arquitetura da EfficientNet foi desenvolvida para abordar o problema de como escalar redes neurais de forma eficaz. Tradicionalmente, aumentar a precisão de uma rede envolvia aumentar uma de suas dimensões, seja a profundidade, a largura (mais filtros por camada) ou a resolução da entrada, o que pode levar a uma ineficiência.

A EfficientNet busca melhorar a precisão escalando uniformemente em todas as direções por meio de algumas observações cruciais: ao aumentar a resolução da imagem de entrada, o modelo se beneficiaria de um aumento proporcional em profundidade e largura e é essencial equilibrar todas as dimensões da rede (Figura 2.6). Sendo assim, a principal motivação do artigo é demonstrar que o investimento em apenas uma dessas dimensões leva a um *plateau* na relação entre desempenho e custo computacional.

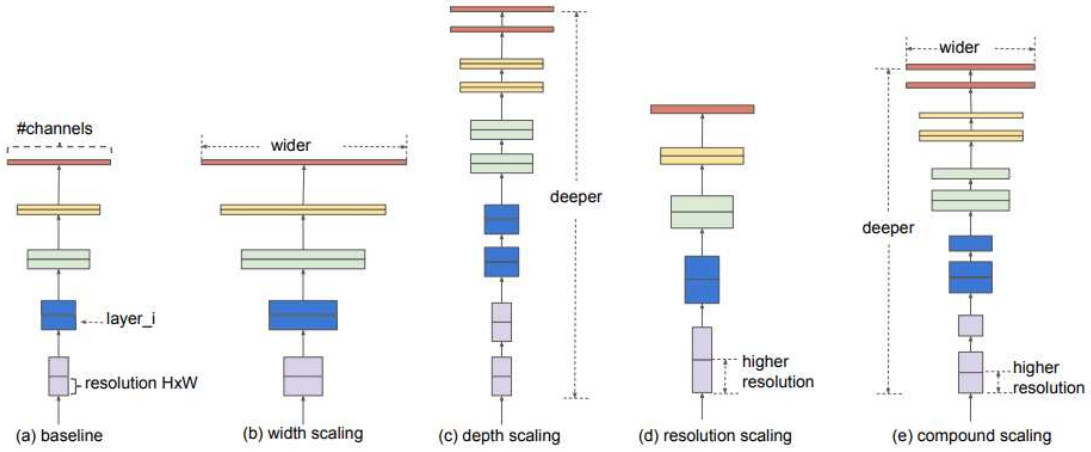


Figura 2.6: Métodos de dimensionamento vs. o escalonamento composto da EfficientNet. Figura reproduzida de Tan e Le [76].

Esse processo é feito por meio de coeficientes, utilizando um conjunto fixo de três coeficientes de dimensionamento (um para cada dimensão). Na etapa de busca de hiperparâmetros para os coeficientes, escolhe-se um valor para o coeficiente composto ϕ que determina a escala geral da rede. Nas próximas etapas do treinamento, esses valores são determinados por constantes α, β e γ que podem ser determinadas por uma pequena busca em grade e usadas para escalar a rede, resultando em uma arquitetura balanceada (Equação 2.5).

$$\begin{aligned} &\textit{Profundidade: } (d = \alpha^{\phi}), \\ &\textit{Largura: } (w = \beta^{\phi}), \\ &\text{Resolução:} \quad r = \gamma^{\phi}, \\ &\text{sujeito a: } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \quad \text{com} \quad \alpha \geq 1,\ \beta \geq 1,\ \gamma \geq 1. \end{aligned} \tag{2.5}$$

Ao escalar uniformemente em todas as direções, a EfficientNet mantém um equilíbrio

entre as três dimensões da rede (profundidade, largura e resolução), evitando os problemas de superdimensionamentos, alcançando uma alta precisão com menos parâmetros e menor custo computacional, demonstrando resultados competitivos em *benchmarks* de reconhecimento de imagem, como ImageNet [67].

Capítulo 3

Trabalhos Relacionados

Neste capítulo, discutimos os principais trabalhos relacionados a esta dissertação. Inicialmente, abordamos a aplicação da Inteligência Artificial (IA) no marketing, com ênfase em técnicas de aprendizado de máquina, como redes neurais profundas (Seção 3.1). Em seguida, exploramos o uso da IA no monitoramento de publicidades (Seção 3.2), ressaltando a carência de soluções automatizadas, especialmente no contexto de publicidades de alimentos voltadas ao público infantil. Por fim, apresentamos os avanços em Transformers para classificação de imagens e vídeos, destacando o papel de arquiteturas híbridas e a relevância de sua aplicação em tarefas multimodais (Seção 3.3).

3.1 Aplicação de Inteligência Artificial no Marketing

Atualmente, diversas pesquisas acadêmicas têm como campo de estudo a IA aplicada em áreas do marketing [16, 28], abordando uma ampla gama de aspectos. Entre as áreas de destaque está a automação de processos repetitivos e manuais que incluem o gerenciamento e análise de grande quantidade de dados [39, 75] a fim de apoiar a tomada de decisões estratégicas, reduzir custos operacionais, mas também acelerar processos, aumentando a eficiência geral das operações.

A aplicação de técnicas de aprendizado de máquina, especificamente redes neurais profundas e IA permite com que essas empresas possam extrair informação dos dados de seus consumidores e assim serem capazes de resolverem tarefas de segmentação de clientes [43], previsão de demanda [31] e detecção de fraude [27].

Esses algoritmos buscam prever tendências de consumo e antecipar mudanças no comportamento do consumidor, permitindo que essas empresas possam ajustar tanto seus estoques quanto suas estratégias de marketing para personalizar a publicidade especificamente para cada cliente, tudo feito através da análise de dados comportamentais e históricos [2, 10]. Também se torna possível identificar padrões atípicos em transações financeiras ajudando na prevenção de fraude.

Apesar das inúmeras vantagens, a implementação da IA no marketing enfrenta diversos desafios, principalmente quanto as questões de Ética, especialmente no que diz respeito à transparência e aos vieses [16, 17].

Ademais, a aplicação de IAs, especialmente nas áreas de marketing, pode influenciar

negativamente o consumo alimentar dos indivíduos. Por exemplo, o histórico comportamental nem sempre reflete as preferências atuais de um indivíduo. Isso significa que clientes que desejam adotar uma dieta mais saudável podem enfrentar ainda mais desafios se a IA identificar padrões antigos e continuar a recomendar alimentos não saudáveis com base em comportamentos passados [16].

A publicidade voltada para alimentos não saudáveis, aliada a estratégias cada vez mais persuasivas, já foi identificada por diversos especialistas como um fator que influencia o comportamento e, por consequência, a dieta do consumidor [9], especialmente durante a infância [37, 52, 68, 69].

3.2 Monitoramento Automático de Publicidades

Até onde sabemos, não existem muitos trabalhos voltados para o monitoramento automático de publicidades de alimentos voltados ao meio televisivo, o que ressalta a proposta original deste mestrado. No contexto das mídias mídias sociais, há alguns estudos incipientes [54, 55, 56]. Assim como na televisão, a indústria aproveita o ambiente digital (mídias sociais, videogames, sites e aplicativos de jogos) para alavancar o consumo de alimentos ultraprocessados, com o diferencial de que as crianças têm ainda mais dificuldade de reconhecer uma publicidade fora do meio televisivo [8, 37]. Dessa forma, o marketing digital também fomenta um envolvimento profundo do consumidor com alimentos e marcas não saudáveis o que acaba trazendo consequências negativas a saúde desses indivíduos.

Embora existam esses poucos estudos, eles se limitam a protocolos de pesquisa ou *comentários*. Sendo assim, nenhum desses apresenta resultados práticos, soluções automatizadas ou protocolos específicos para o monitoramento automático de publicidades. Apesar disso, esses trabalhos destacam o interesse na área, trazendo planejamentos de estudos futuros e discussões, análises críticas e reflexões sobre esse tema e a necessidade de se monitorar automaticamente publicidades alimentícias.

Diante do exposto, acreditamos que os trabalhos mencionados nesta seção destacam a importância de incluir regulamentos que se apliquem em toda a programação, que possam ser apoiados por sistemas habilitados por IA para atuar em grande escala e contribuir na diminuição da exposição de crianças a publicidades alimentícias de baixo teor nutricional. Vale destacar que, atualmente, os pesquisadores precisam avaliar manualmente a extensão e a natureza das táticas aplicadas pela indústria alimentícia nos mais diversos meios (digitais e televisivos), em altíssimo volume (visto a quantidade de publicidades que vão ao ar). Como resultado, eles são capazes de catalogar apenas um pequeno número de mídias quando comparado a quantidade de publicidades transmitidas nesses dois meios.

3.3 Transformers

Como mencionado na Seção 2.2, em 2017, as Transformers [84] foram propostas para tarefas de processamento de linguagem natural, substituindo arquiteturas recorrentes anteriormente populares para estas tarefas. Em 2020, pesquisadores começaram a explorar o uso de Transformers Visuais para classificação de imagens, sendo estas o estado da arte.

Com sua popularização, diversos outros trabalhos surgiram na área. Na Tabela 3.1, sumarizamos as arquiteturas que foram propostas a fim de melhorar a performance das Transformers e Transformers Visuais para as diferentes tarefas voltadas para imagens, vídeos e entradas multimodais.

Tabela 3.1: Publicações na área de Transformers. Em destaque, as arquiteturas avaliadas nesta dissertação.

Rede	**Tarefa**	**Tipo**	**Base de Dados**	**Conf.**	**Código**
ViT [20]	Classificação de imagens	Transformer tradicional	ImageNet, ImageNet-21K, JFT	ICLR'21	✓
DeiT [78]	Classificação de imagens	Teacher-Student com uma CNN	ImageNet, CIFAR-10, CIFAR-100, iNaturalist, Flowers-102, Stanford Cars	PMLR'21	✓
LeViT [25]	Classificação de imagens	Rede híbrida de Transformer e ResNet-50	ImageNet	ICCV'21	✓
CvT [85]	Classificação de imagens	Transformers com adição de convolução	ImageNet, ImageNet-21K	ICCV'21	✓
Perceiver [32]	Classificação multimodal (áudio + imagem)	Atenção cruzada e blocos de Transformers em alternância	ImageNet	ICML'21	✓
ConVit [21]	Classificação de imagens	Transformers com adição de convolução	ImageNet, CIFAR-100	ICML'21	✓
EViT [48]	Classificação de imagens	Transformers com mesclagem e redução de patches	ImageNet	ICLR'22	✓
EaO [71]	Busca multimodal (áudio + imagem + texto)	Abordagem multimodal desde o início baseada em fusão multimodal	HowTo100M [50], YouCook2 [88], MSR-VTT [87], CrossTask [89], Mining YouTube [42]	CVPR'22	✓
DeiT III [79]	Pré-treinamento	Pré-treinamento para as ViTs	–	ECCV'22	✗
TubeVit [62]	Classificação de imagens e vídeos	Transformar um encoder ViT ao amostrar de forma esparsa as entrada	ImageNet, Kinetics-400, Kinetics-600, Kinetics-700, Something-SomethingV2	CVPR'23	✗

Tabela 3.1: Publicações na área de Transformers (continuação). Em destaque, as Transformers avaliadas nesta dissertação.

Rede	**Tarefa**	**Tipo**	**Base de Dados**	**Conf.**	**Código**
Uniformer [47]	Classificação de imagens e vídeos e *dense prediction*	Transformer unificada integrando convolução e auto-atenção com afinidade de token local e global	COCO, ImageNet-1K, ADE20k, Kinetics, Something-Something (SthSth) V1/V2, entre outros	TPAMI'23	✓
UVCOM [86]	Vídeo moment retrieval e detecção de highlights	Transformer multimodal com tokens de bottleneck	Charades-STA [23], TVSum [72], QVHighlights [44], YouTube Highlights [73]	CVPR'24	✓
ViT Registers [15]	Identifica artefatos em mapas de features nas ViTs	Propõe o uso de *tokens register* para solucionar *high-norm tokens* em áreas pouco informativas	ImageNet-22k, ADE20k, NYUd	ICLR'24	✗

A maioria dos Transformers Visuais surgiu como resultado da hibridização de Transformers com redes neurais convolucionais, como as ResNets [30]. Geralmente, essas novas abordagens baseiam-se em modelos anteriores, buscando maneiras de integrar convoluções presentes nas CNNs às etapas do Transformer, substituindo ou modificando blocos padrão para incorporar convoluções. Um exemplo é o LeViT [25], que combina a arquitetura de Transformer Visual com a ResNet-50, enquanto o CvT [85] incorpora convoluções diretamente nos blocos de Transformer, aprimorando a extração de *features* em tarefas de classificação de imagens. De maneira semelhante, a Uniformer [47] une convolução e auto-atenção, oferecendo uma integração unificada da afinidade local e global entre *tokens*.

Outras abordagens focam na adaptação da estrutura tradicional do Transformer para lidar com dados multimodais, como vídeos e áudios, simultaneamente. O Perceiver [32], por exemplo, foi um dos primeiros estudos a utilizar atenção cruzada para integrar diferentes modalidades de forma escalável, alternando entre blocos de atenção e modalidades. Propostas mais recentes, como o UVCON [86], buscam trocar informações entre as modalidades por meio de camadas de atenção cruzada. A EaO [71] é um destaque nesse campo, pois adota uma abordagem direta e eficiente, integrando simultaneamente áudio, vídeo e texto desde o início do treinamento. Essa fusão multimodal é especialmente vantajosa em cenários complexos, como a classificação de publicidades de televisão, onde a combinação de diferentes fontes de informação pode ser crucial para melhorar o desempenho, mesmo com recursos computacionais limitados, uma vez que a proposta da EaO é computacionalmente mais barata que a UVCON.

Adicionalmente, várias pesquisas concentram-se em otimizações específicas. O TubeViT [62], por exemplo, otimiza a amostragem de *frames* de vídeos de forma esparsa, reduzindo o custo computacional sem sacrificar a precisão da classificação. Similarmente,

a EViT [48] utiliza a fusão de *tokens* inatentos para diminuir o custo computacional na classificação de imagens, também sem comprometer o desempenho. Já o ViT Registers [15] propõe o uso de *tokens* de "registro" para mitigar artefatos em mapas de *features*, lidando com os *high-norm tokens* em áreas com pouca informação, o que aprimora a interpretação dos resultados. Por sua vez, o DeiT III [79] foca na redução da quantidade de dados necessários para o treinamento de uma ViT, por meio do aprimoramento do processo de pré-treinamento. Em cenários com limitações computacionais rigorosas, como sistemas de monitoramento em tempo real, a EViT se destaca ao oferecer uma significativa redução da complexidade, ou seja, do custo computacional, sem demandar grandes volumes de dados, podendo ser facilmente adaptada para diferentes arquiteturas.

Embora promissora, destacamos que a pesquisa em Transformers Visuais ainda está em seus estágios iniciais.

Capítulo 4

Base de Dados

Esta dissertação de Mestrado foi realizada em parceria com o Departamento de Nutrição da Universidade Federal de Minas Gerais (UFMG), que é responsável pela criação, seleção e anotação da base de dados, e também pela validação dos resultados obtidos para a proposição de um novo método.

Neste capítulo, detalhamos a base de dados utilizada para o desenvolvimento deste trabalho, incluindo a coleta e construção dos dados (Seção 4.1), conforme o protocolo da Rede INFORMAS. Além disso, descrevemos as variáveis analisadas e as categorias de publicidades, com ênfase nos subtipos de publicidades alimentícias e nas estratégias de marketing persuasivas. Ademais, detalhamos as etapas de preparação (Seção 4.2) e separação (Seção 4.3) da base de dados, etapas essenciais e antecedentes ao treinamento dos modelos escolhidos. No Apêndice A, apresentamos a ficha de especificação da base de dados seguindo o artigo *Datasheets for Datasets* [24].

4.1 Coleta e Construção da Base de Dados

Os procedimentos empregados para a coleta dos dados se basearam no protocolo de monitoramento da publicidade televisiva de alimentos [35], definido pela Rede INFORMAS [83] (*International Network for Food and Obesity/Non-communicable Disease Research, Monitoring and Action Support*, ou em português, Rede Internacional para Pesquisa, Monitoramento e Apoio à Ação em Alimentos e Obesidade/Doenças Não Transmissíveis) (veja a Seção 2.1). Para a construção da base de dados, foram escolhidos para monitoramento três canais da TV aberta (*Globo*, *Record* e *SBT*) examinados nos períodos de Abril de 2018, Maio de 2019 e Junho de 2020, e dois canais da TV fechada (*Discovery Kids* e *Cartoon Network*) monitorados nos períodos de Maio de 2019, Setembro de 2019 e Junho de 2020 (Tabela 4.1). Exemplos das publicidades coletadas podem ser encontrados na Figura 4.1.

Para cada ciclo de coleta de dados, a programação dos canais de TV foi gravada em formato digital, por uma empresa especializada durante oito dias não consecutivos, sorteados aleatoriamente, sendo quatro dias de fins de semana (sábado ou domingo) e quatro dias durante a semana, totalizando 18 horas diárias para cada canal e 2.268 horas nos três anos do estudo. Dessas horas, foram extraídos 4.065 vídeos de publicidades

Tabela 4.1: Detalhamento de cada ciclo de coleta por canal, incluindo canais abertos (Rede Globo, SBT, Rede Record) e canais fechados (*Discovery Kids* e *Cartoon Network*).

Canais	**Ano de Coleta**	**Meses**	**Dias**
Rede Globo	2018	Abril	05, 14, 15, 19, 22, 24, 25, 29
	2019	Maio	05, 07, 10, 11, 14, 18, 26, 27
	2020	Junho	04, 07, 09, 11, 13, 20, 24, 27
SBT	2018	Abril	05, 14, 15, 19, 22, 24, 25, 29
	2019	Maio	05, 07, 10, 11, 14, 18, 26, 27
	2020	Junho	04, 07, 09, 11, 13, 20, 24, 27
Rede Record	2018	Abril	05, 14, 15, 19, 22, 24, 25, 29
	2019	Maio	05, 08, 12, 14, 17, 21, 25, 28
	2020	Junho	04, 07, 09, 11, 13, 20, 24, 27
Discovery Kids	2019	Maio	05, 07, 10, 11, 14, 18, 26, 27
	2019	Junho[1]	04, 07, 09, 20, 24, 27
	2020	Setembro	05, 08, 12, 14, 17, 21, 25, 28
Cartoon Network	2019	Maio	05, 07, 10, 11, 14, 18, 26, 27
	2019	Junho	05, 07, 10, 11, 14, 18, 26, 27
	2020	Setembro	05, 08, 12, 14, 17, 21, 25, 28

[1] Perda de dois dias (11, 13) de gravação na coleta de junho de 2020 para o canal por erro da empresa de *clipping*.

alimentícias e 29.928 não alimentícias.

Entre as informações coletadas estão: nome do canal, a data da gravação, o nome do programa, o horário de início e término do anúncio e o tipo de publicidade. Em seguida, especificamente para as publicidades envolvendo alimentos, também foram divididos em subtipos de anunciantes, além de serem também anotadas o nome da marca ou empresa, nome e descrição do produto e a categoria do alimento (Tabela 4.2). Também foram investigadas e classificadas as suas estratégias de publicidade, divididas entre 28 estratégias, pertencentes a um dos três seguintes grupos: poder das estratégias de publicidade, uso da oferta de prêmios e uso de alegações de benefícios da marca (Tabela 4.3), que apesar de coletadas, não são o foco desta pesquisa.

4.2 Preparação e Limpeza da Base de Dados

A etapa de limpeza da base de dados envolveu a remoção de vídeos duplicados, sendo que esse processo foi realizado de maneira distinta para as publicidades alimentícias e não alimentícias.

Tabela 4.2: Variáveis utilizadas para extração geral dos dados da publicidade televisiva.

Variável	**Descrição**
Nome	O nome do canal avaliado
Data	Data da gravação
Nome do programa	Nome do programa de acordo com a grade do canal
Horário da publicidade	Faixa de horário da publicidade
Horário de início	Horário de início da publicidade (HH:MM:SS)
Horário de término	Horário de termino da publicidade (HH:MM:SS)
Momento da publicidade	Intervalo de um programa específico
	Intervalo entre dois programas
Tipo de publicidade	Alimento ou bebida – empresa/marca de alimentos
	Alimento ou bebida – não anunciado pela empresa/marca do produto
	Empresa/marca de alimentos sem anunciar um produto
	Supermercado/Loja de conveniência anunciando alimentos
	Supermercado/Loja de conveniência sem anunciar alimentos
	Restaurante/*Fast-food* anunciando alimentos
	Restaurante/*Fast-food* sem anunciar alimentos
	Reality show ou Programa culinário
	Suplemento alimentar
	Bebidas alcoólicas
	Produto não alimentício
Código da publicidade	Código copiado da grade de monitoramento

4.2.1 Publicidades Alimentícias

Dos 4.065 vídeos de publicidades alimentícias coletados, cerca de 83% foram removidos manualmente por serem duplicados. Os vídeos alimentícios foram divididos, pelo Tipo de Publicidade (Tabela 4.2). Esse processo foi feito com auxílio da equipe de nutrição da UFMG. Essa subdivisão é parte do processo de aplicação do protocolo INFORMAS [35], e surgiu da necessidade de categorizar de maneira mais precisa as estratégias de marketing utilizadas por diferentes setores da indústria alimentícia.

No entanto, devido às limitações na quantidade de dados coletados, tivemos que focar em apenas três tipos de publicidade para as categorias alimentícias, que serão discutidos a seguir. Isso se deve ao fato de que não haviam exemplos suficientes para algumas classes (sendo algumas até mesmo inexistentes, por exemplo a categoria 'Empresa/marca de alimentos sem anunciar um produto'). Dessa forma, foi necessária uma adaptação, onde classes sem dados foram removidas, e não mais subdividimos os tipos de publicidade pela presença ou não de alimento. Agrupamos algumas classes com poucos dados (por exemplo, a classe de suplemento alimentares, alimento ou bebida não alcoólica e bebidas alcoólicas).

Do ponto de vista computacional, essa adaptação foi essencial para mitigar os desafios impostos por um conjunto de dados desbalanceado e de tamanho reduzido. Classes com

Tabela 4.3: Estratégias persuasivas de marketing de alimentos de acordo com o protocolo da rede INFORMAS.

Tipo	**Estratégia**	**Descrição**
Poder da publicidade	Personagem próprio da marca	Personagem criado pela própria marca.
	Personagem licenciado	Empresa usa de outro personagem de desenho/filmes.
	Personagem esportista amador	Pessoa praticando esportes desde que não é atleta famosa.
	Celebridade não esportiva	Inclui uma pessoa famosa que não é atleta.
	Filme	Publicidades relacionadas a filmes.
	Esportista famoso/time	Inclui atletas famosos ou times famosos.
	Eventos comemorativos/festivos	Inclui datas ou eventos comemorativos.
	'Para crianças'	Produtos destinados especificamente a crianças.
Oferta de prêmios	Prêmios da marca	Expor premiações conquistada pela marca.
	Edição limitada	Produto é oferecido em um período limitado.
	Brinde ou colecionável	Ganhar um bride colecionável junto a compra.
	Concursos	Marca oferece concursos para consumidores.
	Desconto no preço	Ofertar desconto no preço do produto.
	Programas de fidelidade	Incentivar a compra para obter algo (desconto ou brinde).
	Pague 2, leve 3 ou outros	Na compra de x produtos um sai como brinde.
	Downloads de jogos e aplicativos	Apresenta links para downloads para interação dos consumidores.
	20% extra ou outros	Oferta de porcentagem a mais do produto.
Alegações de benefícios	Características sensoriais	Apresentar alegações por sabor, textura, aparência e aroma.
	Uso sugerido para crianças	Quando o produto é sugerido a algum público específico.
	Novos desenvolvimentos	Lançamento de um novo produto da marca.
	Preço	Quando a publicidade usa o preço como vantagem.
	Inovador	Quando frente a outra marca, o produto seja inovador.
	Uso sugerido	Apresentar consumo/preparo junto a outro produto.
	Conveniência	Produto fácil de ser adquirido ou consumido.
	Alegações emotivas	Quando a marca apresenta sensações ao consumo do produto.
	Parceria com outra marca	Publicidade de duas marcas distintas parceiras
	Ingredientes relacionados a saúde	Alegações sobre ingredientes relacionado a saúde.
	Conteúdo nutricional	Alegações sobre o conteúdo nutricional do produto.

(a) Publicidades alimento ou bebida

(b) Publicidades supermercado

(c) Publicidades *fast-food*

(d) Publicidades não alimentícias

Figura 4.1: Exemplos de publicidades coletadas e veiculadas nos canais abertos da TV brasileira.

poucos exemplos tendem a ser mal representadas no treinamento de modelos de aprendizagem, o que prejudica a capacidade de generalização do modelo, resultando em baixa precisão para essas categorias e em uma tendência a priorizar as classes majoritárias. Essas adaptações foram cruciais para viabilizar o uso do aprendizado de máquina para a classificação de publicidades.

A separação final incluiu:

1. Publicidades de *fast-food*/restaurante, que são voltadas para redes de alimentação rápida, independentemente da presença explicita ou não de alimentos. Vamos denominar essas publicidades como ***fast-food***.
2. Publicidades de supermercado/loja de conveniência que promovem produtos alimentícios disponíveis nas prateleiras de grandes redes varejistas, independentemente da presença explicita ou não. Vamos denominar essas publicidades como **supermercado**.
3. Publicidades de alimentos ou bebidas (alcóolicas ou não) que não se encaixam nos grupos anteriores, como cereais, refrigerantes, cervejas, suplementos alimentares e

outros produtos consumidos em várias ocasiões, necessariamente não vinculados ao contexto de supermercados ou *fast-food*. Vamos denominar essas publicidades como **alimento/bebida**.

No total, obtivemos 690 vídeos alimentícios distribuídos pelas três classes, sendo 105 de *fast-food*, 149 de supermercado e 436 vídeos de alimento/bebida.

4.2.2 Publicidades Não Alimentícias

Como os vídeos coletados pela equipe de nutrição da UFMG não estavam separados para as publicidades não alimentícias, foi necessário automatizar o processo de corte dos vídeos da programação completa de cada canal, através das *timestamps* anotadas[1]. Para tanto, foi criado um código em Python que utiliza a ferramenta `ffmpeg` [77] para auxiliar esse processo.

Dos 29.928 vídeos de publicidades não alimentícias cortados, identificamos diversas publicidades duplicadas, correspondendo cerca de 80% dos vídeos coletados. Ao final desse processo, obtivemos 6.139 publicidades não alimentícias. As duplicatas consistiam em dois tipos: algumas eram cópias perfeitas, enquanto outras seriam versões reduzidas de uma publicidade original. Para resolver esse problema, foi proposta uma maneira simples (descrita a seguir) de remover esses dois tipos de duplicatas.

Para remover as duplicatas, aplicamos funções de *average hash*[2], visto que se trata de uma abordagem barata, rápida e eficiente. A abordagem consiste em calcular um *hash* para cada *frame* do vídeo. Quando duas publicidade tiverem um número suficiente de *hashs* iguais, ou seja, quando esse valor fosse maior que um certo limiar, a maior delas será considerada a publicidade 'original' e a menor será considerada uma 'cópia' e será excluída. Uma vez que uma publicidade seja considerada original, ela não pode mais ser considerado duplicata de nenhuma outra. Portanto, em cada nova remoção, apenas as publicidades que não foram classificadas como "originais" anteriormente podem ser categorizadas como cópias.

Realizamos uma série de experimentos para determinar um limiar eficaz para a remoção de duplicatas de publicidade. Observamos que a natureza das publicidade nos canais abertos e fechados difere, sendo que as duplicatas são mais comuns nos canais abertos. Assim, adotamos o seguinte fluxograma: primeiramente, removemos as cópias existentes em um mesmo canal, repetindo o processo de remoção três vezes, de forma sequencial. Só depois de remover as duplicatas de um canal individualmente juntamos as publicidades de todos os canais e as removemos, também de maneira sequencial.

As taxas foram definidas empiricamente com base na natureza do canal (aberto ou fechado) e tornam-se progressivamente menos restritivas, como apresentado a seguir:

1. Canais Abertos (Rede Globo, SBT, Rede Record): as taxas são de 40, 20 e 10.
2. Canais Fechados (*Discovery Kids* e *Cartoon Network*): as taxas são de 50, 25 e 13.
3. Entre canais (Rede Globo, SBT, Rede Record, *Discovery Kids* e *Cartoon Network*): as taxas são de 60, 30, 25.

[1]Esta etapa foi realizada durante o Projeto Final de Graduação [22] da autora da dissertação.
[2]`https://github.com/gklc811/duplicate_video_finder`

Assim, ao final do processo, a grande maioria dos vídeos duplicados são removidos. Essa divisão em duas etapas (uma por canal e só depois a coleta completa) foi necessária porque, dado a natureza de cada canal, tentar fazer a remoção em uma única etapa se provou (de forma empírica) menos eficaz através dos nossos experimentos.

Essa é a primeira base de dados desse tipo no Brasil e pretendemos disponibilizá-la publicamente. Para isso, seguimos as recomendações feitas no artigo *Datasheets for Datasets* [24]. No Apêndice A, apresentamos a ficha de especificação da base de dados de acordo com Gebru et al. [24].

Destacamos que apesar dos nossos esforços, a bases de dados ainda sim é desbalanceada (cerca de 90% das publicidades são da classe não alimentícia), pela própria natureza das publicidades expostas no meio televisivo, e pequena, contendo 6.829 vídeos no total. Para contextualizar, outras bases populares na área de vídeo, como o YouTube-8M [1] e a ImageNet-21K [64] possuem milhões de vídeos e imagens, o que ressalta ainda mais a limitação em termos de tamanho da nossa base.

4.3 Separação da Base de Dados

Após a preparação da base de dados, foi iniciada a divisão da base de dados para o treinamento dos modelos. Para tanto, a "base inicial de dados" foi dividida em três novas: uma foi destinada para treinar os algoritmos (Base 1, que contém grande parte dos vídeos de 2020 e 2019), uma foi destinada para validação dos resultados (Base 2, que contém uma pequena porcentagem dos vídeos de 2018 e 2019) e a última foi destinada ao teste final do modelo (Base 3, que contém os vídeos de 2018 e 2019). Essa divisão é necessária para evitar o enviesamento dos modelos ao serem executados em novos dados. Ela foi feita de maneira que a grande parte dos dados foram destinados ao treinamento do modelo, mas de forma que tanto a Base 2, quanto a Base 3 ainda sejam capazes de representar a totalidade destes (Figura 4.2).

Destacamos que esse trabalho de separação iniciou durante a pandemia, o que trouxe desafios adicionais. Durante esse período, não tínhamos acesso a todos os dados coletados, o que impactou diretamente a forma como as bases foram divididas. Essas limitações no acesso e na disponibilidade, forçou uma adaptação na estratégia de divisão, para garantir que cada conjunto de dados fosse balanceado o suficiente e representasse adequadamente as classes analisadas.

Na Tabela 4.4, são apresentadas os números de vídeos por tipo de publicidades e pela separação da base de dados, sendo que a Base 1 contém 4.760 vídeos (70%), a Base 2 1.349 (20%), e a Base 3 720 vídeos (10%), totalizando 6.829 vídeos.

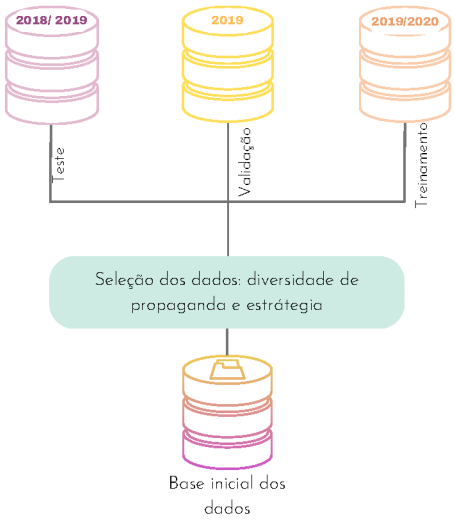


Figura 4.2: Síntese da metodologia que compreende a etapa de divisão da base de dados.

Tabela 4.4: Número de vídeos por tipo de publicidades (alimentícias: *fast-food*, supermercado, alimento/bebida, vs. não alimentícias) e pela separação da base de dados.

		Base 1 (Treino)	**Base 2 (Valid.)**	**Base 3 (Teste)**	**Total**
Alim.	*Fast-Food*	79	28	11	118
	Supermercado	111	25	13	149
	Alimento/Bebida	299	88	49	436
	Não Alimentícia	4.281	1.211	647	6.139

Capítulo 5

Metodologia

Neste capítulo, detalhamos as estratégias utilizadas para responder às perguntas de pesquisa, com foco na reprodutibilidade da pesquisa, aprofundando um estudo em arquiteturas baseadas em Transformers para a classificação de publicidades alimentícias.

5.1 Tipos de Classificação

A tarefa central desta dissertação consiste na classificação de vídeos publicitários de maneira supervisionada, onde o objetivo principal é que o algoritmo aprenda a mapear corretamente as entradas (vídeos ou imagens) para as saídas (categorias associadas a esses vídeos). Ou seja, buscamos desenvolver um modelo que seja capaz de classificar cada vídeo publicitário em uma ou mais classes predeterminadas por especialistas da área, segundo o protocolo INFORMAS.

Essa tarefa pode ser dividida em dois problemas principais: um de classificação binária e outro de classificação multiclasse, que devem ser abordados de maneiras distintas.

Para a **classificação binária**, o objetivo é determinar se um vídeo publicitário pertence a uma de duas categorias específicas: alimentícia ou não alimentícia. Nesse cenário, o desafio é diferenciar ambas classes com base nas características extraídas dos vídeos sendo que uma delas (a não alimentícia) é muito mais frequente que a outra.

No problema de **classificação multiclasse**, o objetivo é categorizar um vídeo publicitário em apenas uma de várias classes possíveis. Para o escopo desse estudo, foram definidas quatro categorias, sendo três (*fast-food*, supermercado, alimento/bebida) minoritárias, associadas a publicidades alimentícias, e uma classe majoritária (a não alimentícia). Neste caso, o modelo final deve ser capaz de distinguir entre essas múltiplas categorias e fazer uma única predição para cada vídeo ou imagem. Com a adição de mais categorias, a complexidade do problema cresce pois existe uma maior variedade de padrões, sendo que as três classes de interesse possuem características visuais e contextuais semelhantes.

5.2 Metodologia Proposta

Como discutido na seção anterior, identificamos duas tarefas de classificação que guiaram nossa fase de experimentação: binária e multiclasse. Uma das principais propostas

deste estudo foi a avaliação de modelos baseados em Transformers; portanto, testamos diversas abordagens diferentes. A escolha dos modelos baseados em Transformers foi feita após uma extensa revisão da literatura, com destaque para arquiteturas como o EViT (Seção 2.2.1 e o EaO (Seção 2.2.2). Para comparar as Transformers com as CNNs, a escolha da CNN EfficientNet (Seção 2.3), especificamente, baseou-se no fato de que, no início desta dissertação, ela representava o estado da arte para tarefas de classificação de imagens.

As tarefas abordadas pelas arquiteturas também variam em natureza. A EfficientNet e a EViT foram desenvolvidas para tarefas de classificação de imagens, enquanto a EaO foi originalmente projetada para tarefas de busca. Portanto, algumas adaptações no fluxograma foram necessárias, como a adição ou não de um classificador, a depender da abordagem. Em todos os métodos avaliados, utilizamos modelos pré-treinados (nas bases ImageNet e HowTo100M). Além disso, mantivemos consistentes as configurações básicas que determinam a estrutura e os hiperparâmetros dos modelos em todas as abordagens.

A Figura 5.1 representa uma visão geral do fluxograma completo. A imagem apresenta um fluxo de trabalho para o processo de classificação de vídeos de publicidades alimentícias. Ele começa com a coleta de dados de fontes, como a Nutrição da UFMG (*passo 1*), seguida pela seleção e preparação dos dados (*passos 2 e 3*), incluindo a extração de *features* de vídeo e áudio (*passo 4*). Destacamos que esse é um passo opcional, a depender da arquitetura. Em seguida, a base de dados é dividida (*passo 5*) para ser utilizada nas etapas de treinamento (*passos 6, 7, 8 e 9*) e validação (*passo 10*). Após a avaliação de todos os modelos treinados, chegamos a um modelo final (*passo 11*) que é testado (*passo 12*), com os resultados finais sendo analisados (*passo 13*).

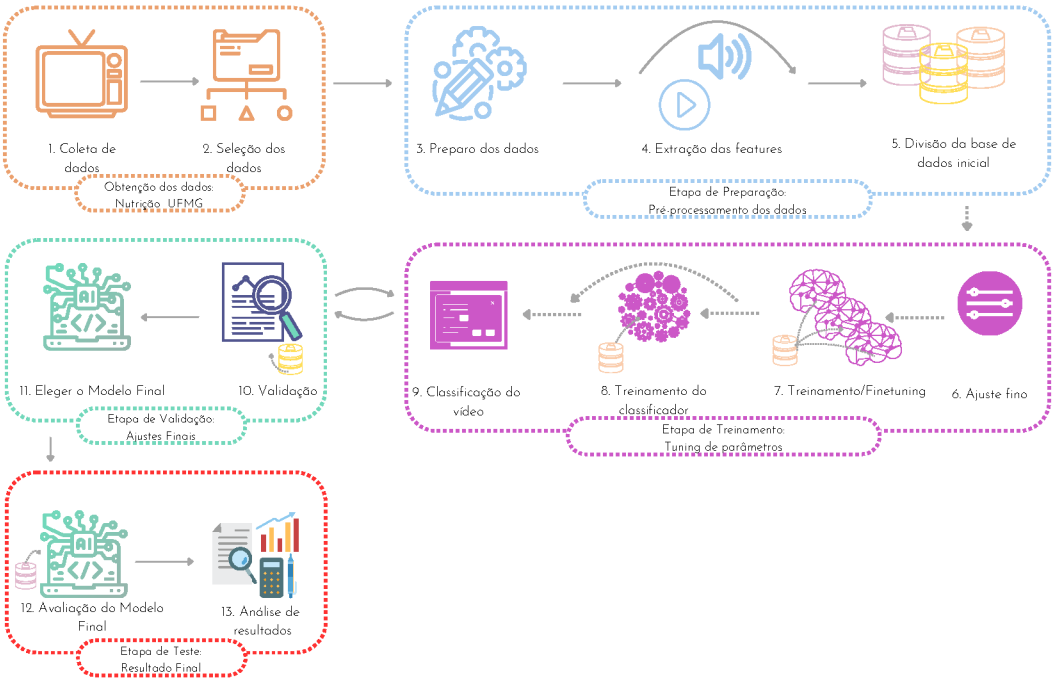


Figura 5.1: Fluxograma da metodologia proposta. As setas da linha pontilhada representam as possibilidades para os experimentos, e as setas da linha sólida representam etapas fixas.

Esses experimentos comparativos iniciais nos permitiram responder a questão de pesquisa Q1, que avalia o desempenho das CNNs em comparação aos modelos baseados em Transformers. Após essa análise inicial, focamos em aprimorar os resultados obtidos com o modelo EaO. Este modelo apresentou o melhor desempenho na tarefa de classificação multiclasse, devido à quantidade de dados disponíveis, que impedia a resolução do problema de classificação das estratégias de marketing.

Ao adaptar a EaO, uma arquitetura baseada em Transformers Multimodais (áudio, texto e vídeo), também abordamos a questão de pesquisa Q2, além de simultaneamente explorar a Q3 e as questões relacionadas ao uso de áudio. Isso nos permitiu focar a metodologia proposta com base nos resultados obtidos e nas questões de pesquisa levantadas:

Q1. **Os resultados obtidos pelas redes neurais profundas baseadas em Transformers superam os resultados obtidos pelas redes neurais convolucionais para o problema de classificação de publicidades alimentícias?** Além de comparar os métodos descritos acima, exploramos esta questão tanto para o problema de classificação binária quanto o problema de classificação multiclasse a fim de avaliar que tipo de rede pode generalizar melhor para a tarefa alvo de maneira menos custosa.

Q2. **Como gerar um modelo eficaz de classificação baseado em Transformers, contornando o problema de desbalanceamento e falta de dados, que seja capaz de identificar os diferentes tipos de publicidades alimentícias?** Com o melhor modelo escolhido, podemos avaliar sua performance na nossa base de testes e contar com o auxílio de especialistas da área de Nutrição para validar seus resultados.

   Além disso, como discutido anteriormente, a disponibilidade limitada de dados, especialmente para as classes alimentícias, exigiu o uso de técnicas adicionais para mitigar o desbalanceamento, pois nenhuma das abordagens performou bem sem o auxílio das mesmas. Testamos técnicas como *data augmentation*, *synthetic minority over-sampling technique*[1] (SMOTE) [11], atribuição de peso por classe, *batches* balanceados, entre outras, para melhorar o desempenho dos modelos a depender o tipo de arquitetura testada.

Q3. **Como combinar áudio e Transformers para melhorar a classificação de publicidades alimentícias?** Optamos por uma rede multimodal, o que nos permitiu explorar o uso do áudio para auxiliar na tarefa de classificação. Diversos experimentos foram realizados para determinar a eficácia da utilização do áudio e a melhor forma de integrá-lo ao fluxograma.

5.3 Treinamento e Validação

Destacamos que todos os modelos selecionados foram pré-treinados em bases de dados maiores, visto que essa abordagem permite explorar melhor os padrões aprendidos previamente. Portanto, durante o treinamento, realizamos um *finetuning*, utilizando os pesos

[1] Técnica que cria novas amostras sintéticas da classe minoritária.

pré-treinados nessas bases de dados maiores no início do processo e continuamos o treinamento desses modelos na nossa Base 1. Dessa forma, ajustamos os modelos para a tarefa específica de classificação das publicidades brasileiras.

Para cada época, validamos os resultados obtidos na Base 2, para verificar se a taxa de acerto ao longo do treinamento. Isso é feito através do cálculo da perda e da acurácia, utilizando classificadores apropriados, e da plotagem de curvas de desempenho (tanto para perda quanto para acurácia). Também empregamos métodos de visualização, como matrizes de confusão, para entender melhor o comportamento dos modelos e gráficos de *embeddings* gerados pelo modelo, utilizando técnicas como o *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [81], quando fosse necessário.

Além disso, ao final da validação de cada modelo, nos concentramos em identificar as fontes de erros preditivos, que serão discutidas posteriormente. Essa análise é crucial para entender as limitações do modelo e para orientar melhorias nas iterações seguintes.

Após a seleção e definição do que seria melhor modelo para a tarefa de classificação, realizamos uma etapa adicional de treinamento na Base 2 (conjunto de validação) por algumas épocas, antes de finalmente testá-lo na Base 3 (conjunto de teste). Os resultados finais obtidos nessa fase serão apresentados e discutidos em detalhes na Seção 6.

5.4 Métricas

As métricas de avaliação são essenciais para entender a qualidade de um modelo de classificação e sua capacidade de generalização, ajudando a identificar áreas de melhoria e a tomar decisões informadas sobre ajustes futuros no modelo.

Acurácia Balanceada

Acurácia é uma métrica de desempenho amplamente utilizada em tarefas de classificação. É a razão entre o número de previsões corretas (os verdadeiros positivos (VP) e verdadeiros negativos (VN)) e o número total de exemplos avaliados (Equação 5.1), ou seja, a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo. Embora a acurácia seja uma medida intuitiva, ela pode ser enganosa em cenários onde há um desbalanceamento, pois um modelo pode apresentar alta acurácia por sempre prever a classe majoritária corretamente, errando as demais classes, por isso a utilização de outras métricas foi necessária.

$$\text{Acurácia} = \frac{VP + VN}{\text{Total de Predições}}. \tag{5.1}$$

Acurácia Balanceada é uma métrica de desempenho que busca corrigir as limitações da acurácia tradicional, especialmente em cenários de desbalanceamento de classes. Para mitigar esse problema, calcula-se a média da taxa de acertos (sensibilidade) de cada classe, levando em conta tanto os verdadeiros positivos (VP) quanto os falsos negativos (FN), além dos verdadeiros negativos (VN) e falsos positivos (FP).

Para cada classe, calcula-se a taxa de acerto para a classe, e então tira-se a média dessas taxas em todas as classes, assim garantindo que o desempenho em cada uma delas

tenha o mesmo peso, independentemente do tamanho da classe (Equação 5.2).

$$\text{Acurácia Balanceada} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{VP_i}{VP_i + FN_i} + \frac{VN_i}{VN_i + FP_i}\right), \tag{5.2}$$

onde VP_i representa os verdadeiros positivos da classe i, VN_i representa os verdadeiros negativos da classe i, FP_i representa os falsos positivos da classe i e FN_i representa os falsos negativos da classe i.

Loss

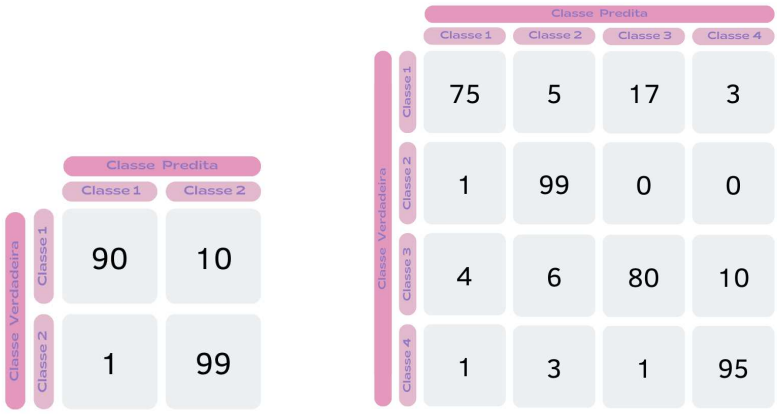
A *loss* ou função de perda é uma medida que quantifica o quão bem o modelo está performando em relação ao seu objetivo (neste caso, a classificação de vídeos). Durante o treinamento, o modelo busca minimizar essa perda, ajustando seus parâmetros para melhorar suas previsões. Existem diversos tipos de funções de perda, dependendo da tarefa em questão. A função de perda mais comum para tarefas de classificação é a entropia cruzada, que mede a diferença entre as distribuições de probabilidade previstas e as reais (Equação 5.3). Quanto menor a perda, melhor o modelo se ajusta aos dados de treinamento.

$$\text{Loss(p,q)} = -\sum_{i}^{i \in classes} p(x)\log(q(x)), \tag{5.3}$$

onde $p(x)$ é a distribuição verdadeira das classes e $q(x)$ é a distribuição predita pelo modelo.

Matriz de Confusão

A matriz de confusão nada mais é que uma tabela, onde as linhas representam as classes reais e as colunas representam as classes previstas pelo modelo. Cada célula da matriz indica o número de instâncias correspondentes a uma combinação particular de classe real e classe prevista. A diagonal principal da matriz contém as previsões corretas, enquanto as células fora da diagonal mostram os erros de classificação (Figura 5.2). A análise de uma matriz de confusão pode revelar se um modelo tende a confundir certas classes específicas, a fim de facilitar a interpretação e melhoria do modelo.



(a) *Problema Binário* (b) *Problema Multiclasse*

Figura 5.2: Exemplo de matriz de confusão: à esquerda, matriz de um *problema binário*; à direita, matriz de um *problema multiclasse* contendo quatro classes.

Capítulo 6

Resultados Experimentais

Neste capítulo, apresentamos e discutimos os principais resultados relacionados a esta dissertação. Inicialmente, abordamos a etapa de pré-processamento dos dados (Seção 6.1) e alguns detalhes de implementação (Seção 6.2). Em seguida, apresentamos os resultados (Seção 6.3), tanto para o problema de classificação binária quanto o multiclasse, incluindo uma análise qualitativa dos resultados.

6.1 Pré-processamento da Base de Dados

6.1.1 Extração de *Frames*/Imagens

Algumas das arquiteturas analisadas trabalham apenas com imagens. Para criar uma esse tipo de base de dados, utilizamos os *frames* (quadros) brutos dos vídeos obtidos em etapas anteriores com auxílio da ferramenta `ffmpeg` [77]. Para os vídeos de publicidades alimentícias, estes foram obtidos em uma amostra de um *frame* por segundo, enquanto que para os vídeos de publicidades não-alimentícias, os *frames* foram retirados da seguinte forma, onde T é a duração do vídeo em segundos:

- $1/2$ *frames* por segundos, se $T \leq 10$;
- $1/4$ *frames* por segundos, se $10 < T \leq 25$;
- $1/6$ *frames* por segundos, se $25 < T \leq 35$;
- T$/15$ *frames* por segundos, se $35 < T \leq 60$;
- T$/30$ *frames* por segundos, se $60 < T \leq 120$;
- T$/45$ *frames* por segundos, se $120 < T$.

Essa decisão foi tomada a fim de minimizar o desbalanceamento notado. Todos os *frames* foram padronizados e centralizados de acordo com a métrica compatível ao tamanho de entrada do modelo escolhido.

6.1.2 Extração de *Features*

Algumas das arquiteturas escolhidas avaliadas nesta dissertação têm como entradas *features* de vídeos e áudio. Portanto, uma etapa de extração e preparação de dados foi

necessária, seguindo os procedimentos descritos a seguir, baseados nas recomendações feitas pela rede *Everything at Once* [71] (EaO).

Extração de *Features* Visuais

Com auxílio do extrator de *features* visuais disponível no repositório `Video Feature Extractor`[1], que utiliza o modelo ResNext-101 [29], foram extraídos as *features* visuais (usando uma rede convolucional 2D e 3D), para todos os vídeos disponíveis na nossa base de dados.

Extração de Espectrogramas de Áudio

Para a extração das *features* de áudio, seguimos o processo disponível no repositório `AVLnet`[2], que envolve três etapas: a extração do áudio, a conversão para mono e a criação de espectrogramas Mel, que são representações visuais das frequências de um sinal de áudio (Figura 6.1).

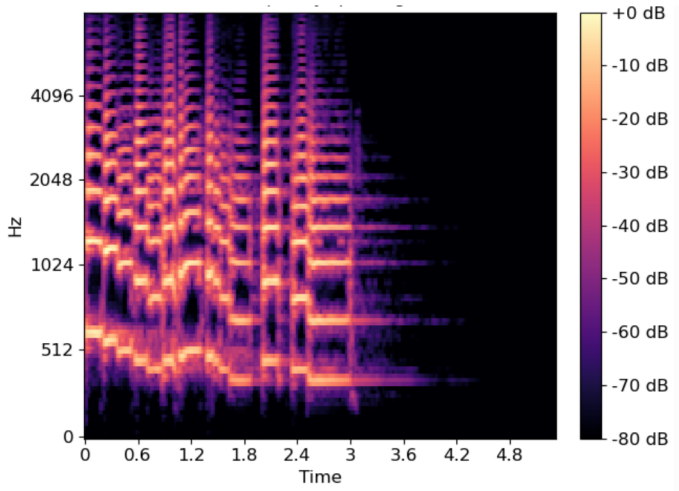


Figura 6.1: Exemplo de um espectrograma Mel. Figura reproduzida de `librosa`[3].

Resumidamente, extraímos o áudio do vídeo na sua taxa de amostragem nativa para um arquivo WAV separado (utilizando o `ffmpeg`). Em seguida, o áudio é convertido de estéreo para mono e é feito *downsampling* para uma taxa de amostragem de 16 kHz (utilizando a ferramenta `sox`). Finalmente, o áudio é convertido para um espectrograma Mel, vale destacar que a EaO tem como parte do seu fluxograma uma CNN treinável que será a extratora de características desses espectrogramas.

Criação do Arquivo Pickle

Após a extração das *features*, foi necessária a criação de um arquivo pickle[4] para cada base (treinamento, validação e teste), pois essa é a entrada esperada pela EaO. Este arquivo

[1] `https://github.com/roudimit/video_feature_extractor`

[2] `https://github.com/roudimit/AVLnet`

[4] Pickle é um módulo que permite serializar e deserializar objetos Python. Serializar um objeto significa convertê-lo em um fluxo de bytes que pode ser salvo em um arquivo. Deserializar é o processo inverso, onde a partir do fluxo de bytes reconstruímos o objeto original.

contém uma lista de dicionários, onde cada elemento representa um clipe de vídeo. As chaves de cada dicionário incluem '2d' (o vetor de características 2D pré extraído), '3d' (vetor de características 3D pré extraído), '*audio*' (o espectrograma do áudio), '*label*' (o rótulo/*label* da classe), '*class*' (o nome da classe) e '*caption*' (neste caso, consideramos uma das quatro opções '*aliment drink*', '*fast food*', '*supermarket*' e '*other*', de acordo com a classe do vídeo, cuidadosamente escolhidas para facilitar a diferenciação dos *embeddings*. As palavras foram definidas em inglês porque a rede *Everything at Once* [71] foi treinada em inglês).

6.2 Detalhes de Implementação

Para todos os experimentos, seguimos o fluxograma apresentado no Capítulo 5. Para encontrar o melhor modelo para classificação de publicidades alimentícias, realizamos experimentos com modelos pré-treinados e ajustados para a tarefa alvo. A Tabela 6.1 sumariza as principais configurações dos experimentos.

EfficientNet

Optamos por utilizar a EfficientNet-B7 [76] como a arquitetura para as redes convolucionais, visto que esta é uma boa referência comparativa para abordagens mais recentes devido a sua popularidade. Por se tratar de uma arquitetura para classificação de imagens, cada *frame* extraído das publicidades foi classificado de maneira independentemente. Para determinar a categoria do vídeo, empregamos um *pooling* (agregação) com média ponderada de cada *frame* para combinar as previsões de cada imagem para uma previsão final do vídeo.

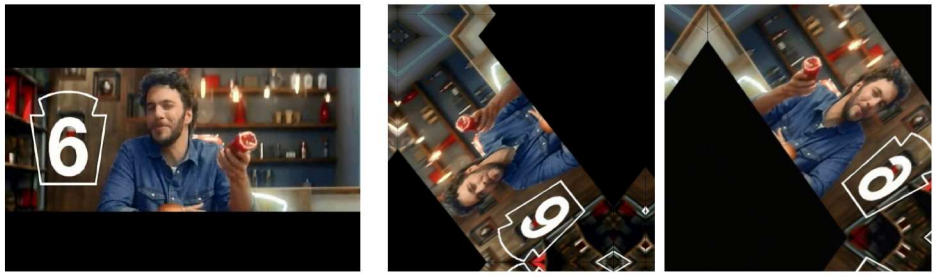
No decorrer desta pesquisa, percebemos que apenas esse tipo de arquitetura não foi suficiente para contornar o desbalanceamento de dados. Para solucionar o problema, aplicamos três estratégias distintas a fim de encontrar o melhor resultado: *data augmentation* [61], *batches* balanceados e adição de peso por classes. Vale destacar que o código foi desenvolvido de maneira a poder aplicar qualquer combinação dessas técnicas como for desejado.

Para o processo de *data augmentation*, aplicamos as seguintes transformações, por se tratarem de transformações simples e baratas, para os *frames* extraídos dos vídeos de publicidades alimentícias: rotação, translação, *zoom*, adição ou remoção de contraste e giros no eixo vertical e/ou horizontal (Figura 6.2).

Para obter *batches* balanceados, criamos o nosso próprio *data generator*, ou seja, ao definir o tamanho do *batch* como X, a quantidade de imagens para cada classe estaria próxima ou exatamente igual a $X/(\textit{número de classes})$. Para tanto, seguimos exemplos de abordagens populares populares em códigos públicos de outros métodos da literatura, entre elas destacamos o `BalancedDataGenerator`[5].

Também decidimos adicionar pesos para cada classe durante o processo de classificação de uma imagem. Assim, podemos contornar o desbalanceamento adicionando um peso maior paras a classes minoritárias (que possuem menos dados), e um peso menor para

[5] `https://gist.github.com/arnaldog12/16efc663c869b35e2479bd607d56c1da`



(a) *Frame* original de uma publicidade alimentícia

(b) Mesmo *frame* após transformações

Figura 6.2: Exemplos das transformações de *data augmentation* realizadas.

a as majoritárias (visto que ela tem mais amostras). Isso é feito durante o treinamento do modelo.

EviT

Assim como na abordagem com a EfficientNet, cada *frame* das publicidades foi tratado de forma independente, sendo classificado individualmente. Posteriormente, aplicamos uma média ponderada das previsões desses *frames* para determinar a categoria final do vídeo. Apesar das vantagens inerentes à arquitetura Transformer, a EviT também enfrentou desafios devido ao desbalanceamento dos dados. A fim de mitigar esses desafios, aplicamos as mesmas três estratégias que se mostraram eficazes com a EfficientNet: *data augmentation*, *batches* balanceados e ponderação de classes, com algumas alterações a fim de adaptar estas estratégias para a arquitetura em questão.

Entre as alterações está a adição de mais três técnicas de aumentação de dados [79], são essas: solarização, desfoque gaussiano e escala de cinza (Figura 6.3). Estas técnicas foram aplicadas em todas as classes, em conjunto com outras técnicas discutidas anteriormente.

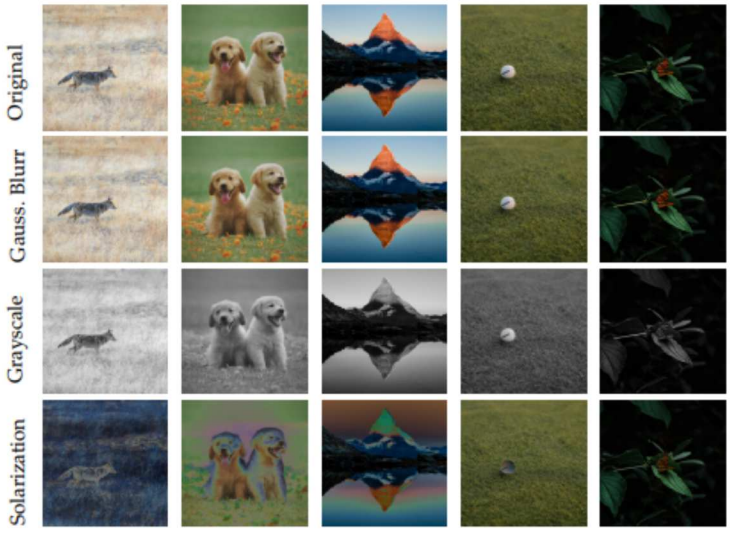


Figura 6.3: Outras técnicas de aumentação de dados. Figura reproduzida de Touvron et al. [79].

Tabela 6.1: Principais configurações dos experimentos para classificação de publicidades alimentícias.

Método	**Parâmetros de Treino**								**Estratégias de Balanceamento de Dados**				
	model	*classificador*	*learning rate*	*scheduler*	*otimizador*	*dropout*	*paciência*	*épocas*	*ruído*	*SMOTE*	*batches balanceados*	*peso por classe*	*data augmentation*
EfficientNet	B7	–	0.00001	–	Adam	0,2	10	150	✗	✗	✓	–	–
EfficientNet	B7	–	0,00001	–	Adam	0,2	10	150	✗	✗	✗	balanceado	–
EfficientNet	B7	–	0,00001	–	Adam	0,2	10	150	✗	✗	✗	–	rotação, translação, zoom, contraste e giros
EviT	deit base patch16 224	–	0,0001	Cosine	Adam	0,3	10	300	✗	✗	✓	–	–
EviT	deit base patch16 224	–	0,0001	Cosine	Adam	0,3	10	300	✗	✗	✗	balanceado	–
EviT	deit base patch16 224	–	0,0001	Cosine	Adam	0,3	10	300	✗	✗	✗	–	rotação, translação, zoom, contraste, solarização, *Gaussian blur*, escala de cinza e giros
EaO	HowTo100M	KNN(cosine, NB=50)	0,0001	ReduceOnPlateau	Adam	0,3	10	150	✓	✓	✗	–	–
EaO	HowTo100M	SVM (rbf, C=0,1, balanced, ovo, gamma=1)	0,0001	ReduceOnPlateau	Adam	0,3	10	150	✓	✓	✗	–	–
EaO	HowTo100M	LR(lbfgs, balanced)	0,0001	ReduceOnPlateau	Adam	0,3	10	150	✓	✓	✗	–	–
EaO	Youcook	KNN(cosine, NB=50)	0,0001	ReduceOnPlateau	Adam	0,3	10	150	✓	✓	✗	–	–
EaO	Youcook	SVM (rbf, C=0,1, balanced, ovo, gamma=1)	0,0001	ReduceOnPlateau	Adam	0,3	10	150	✓	✓	✗	–	–
EaO	Youcook	LR(lbfgs, balanced)	0,0001	ReduceOnPlateau	Adam	0,3	10	150	✓	✓	✗	–	–

Além disso, também testamos duas maneiras de balancear os *batches*, uma balanceando o próprio *sampler* que alimenta o modelo, de maneira semelhante ao feito para a EfficientNet e outra igualando a quantidade de exemplos em cada classe.

Por fim, replicamos todos os testes com e sem a fusão de *tokens* inatentos, a fim de realmente avaliar a performance da mesma.

Everything at Once

Durante a fase de experimentação a EaO[6] foi a arquitetura escolhida. As características modificadas serão detalhadas a seguir, além das técnicas aplicadas para contornar o desbalanceamento de dados na metodologia final desta dissertação. As configurações principais da arquitetura, incluindo dimensões de *embeddings*, configurações de projeção de *tokens*, pesos de perda combinatória, entre outros, foram mantidas conforme as configurações originais da EaO, para facilitar a adaptação do modelo, embora outras combinações tenham sido testadas.

Durante o treinamento, configuramos parte da arquitetura para garantir que o modelo aprenda representações mais robustas e generalizáveis com os dados disponíveis, visto que a pequena quantidade de vídeos poderia facilmente levar ao *overfitting*.

O *scheduler* de taxa de aprendizado com melhor desempenho foi ReduceLROnPlateau[7], configurado para reduzir a taxa de aprendizado pela metade se a métrica monitorada não melhorar após 3 épocas, ajudando a ajustar dinamicamente a taxa de aprendizado para melhorar o treinamento. Mantivemos o otimizador proposto (Adam), com a adição de regularização (*weight decay*) para prevenir o *overfitting*.

Incluímos o uso de *embeddings* posicionais e da estratégia *max pool* como *pooling* de áudio, além de adicionarmos ruído gaussiano sobre a entrada de vídeo em uma taxa de 20% a 35% da nossa base de dados a fim de evitar um *overfitting*.

Especificamente para os parâmetros de fusão do modelo, configuramos três tipos de *dropout*, que desativa aleatoriamente 30% das unidades durante a fase de treinamento para manter o modelo robusto e generalizável. Aplicamos *dropout* tanto nos neurônios quanto nas camadas de atenção e nas camadas do modelo (em profundidade estocástica).

Por fim, para interromper o treinamento caso a performance do modelo não melhorasse, utilizamos uma paciência de 10, que pode ser aplicada sobre a *loss* ou acurácia calculada durante a classificação.

A arquitetura da EaO foi proposta para resolver um problema de busca. Portanto, foi necessária algumas modificações na implementação para que obtivéssemos a classificação dos vídeos. Para isso, usamos os *embeddings* de saída do modelo e os passamos por classificadores, obtendo uma classificação final para o vídeo. Três classificadores diferentes foram implementados e testados: *k-nearest neighbors* (kNN)[8] com métrica de distância cosseno, *support vector machines* (SVM)[9] com kernel RBF (*radial basis function*) e regressão lo-

[6] `https://github.com/ninatu/everything_at_once/`

[7] `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html`

[8] `https://scikit-learn.org/dev/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`

[9] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`

gística[10] com *solver* lbfgs (*Limited-memory Broyden–Fletcher–Goldfarb–Shanno*), todos balanceados por classes (quando possível). Vale destacar que dois treinamentos estão sendo feitos paralelamente, o da tarefa de busca e o do classificador. Portanto, ambos passaram por um processo de ajuste fino.

Para reduzir o desbalanceamento dos dados, aplicamos a técnica de *oversampling* SMOTE [11], criando novas instâncias sintéticas das classes minoritárias até que elas atingissem, individualmente, 40% da quantidade de amostras da classe majoritária. Isso melhora significativamente o desempenho do modelo, já que as classes minoritárias são de interesse, sem adicionar tanto custo computacional, pois não estamos de fato equilibrando as quatro classes. O SMOTE é aplicado apenas durante o treinamento do classificador, nas *features* de saída da arquitetura principal. Durante a validação e a inferência, o classificador é usado diretamente para fazer previsões, sem aplicar qualquer técnica de *oversampling*.

Para cada época, treinamos o modelo e calculamos a *loss* e a acurácia tanto no conjunto de treinamento quanto no de validação com técnicas de precisão mista para otimizar o uso de memória e acelerar o treinamento. Além de calcular a *loss* média e a acurácia no conjunto de validação, e também salvar *checkpoints*. Por fim, também salva-se o melhor modelo caso a *loss* média fosse minimizada, ou a acurácia fosse maximizada, a depender da métrica escolhida.

6.3 Resultados

Para responder as perguntas de pesquisa propostas nesta dissertação (Seção 1.4), conduzimos uma série de experimentos. Comparamos os modelos avaliados e selecionamos os melhores resultados tanto para a tarefa de classificação binária (Seção 6.3.1) quanto para a tarefa de classificação multiclasse (Seção 6.3.2). A Figura 6.4 detalha como cada experimento feito está relacionado com perguntas de pesquisas. A Tabela 6.2 sumariza todos os experimentos.

Tabela 6.2: Sumário de todos os resultados obtidos na base de teste (Base 3) para classificação (binária e multiclasse) de publicidades alimentícias.

	Acurácia Balanceada (%)	
	Binária	**Multiclasse**
EfficientNet	90,5	87,4
EViT	92,4	83,8
EaO	85,9	82,8

[10] `https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html`

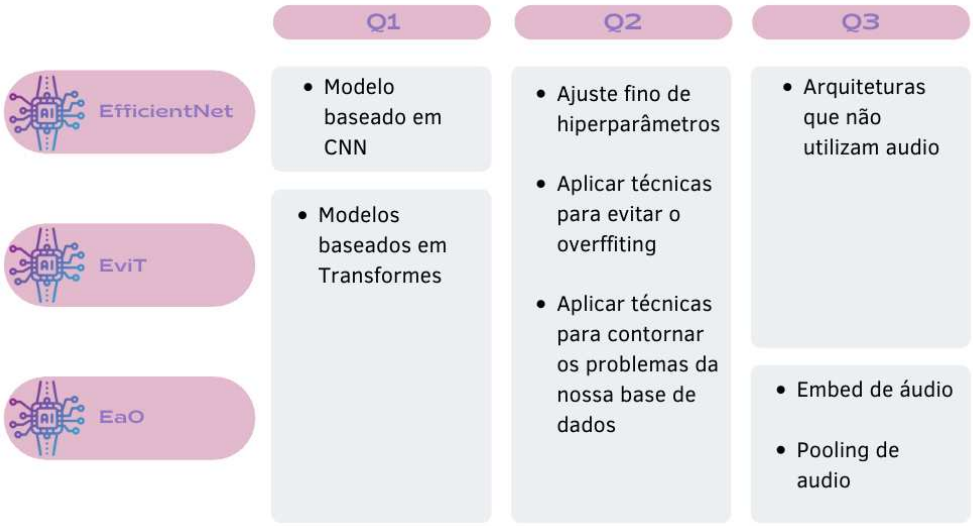


Figura 6.4: Síntese da relação entre os experimentos feitos e as perguntas de pesquisa.

6.3.1 Classificação Binária

EfficientNet

As matrizes de confusão a seguir retratam os principais resultados por vídeo. Estão retratados a porcentagem de número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos, onde a classe positiva é a alimentícia (*fast-food*, *supermercado* e *alimento/bebida*) e a classe negativa é a não alimentícia. A Tabela 6.3 representa o melhor resultado dos experimentos realizados, de acordo com o conjunto de validação (Base 2). A técnica com melhor performance foi a aplicação de *batches* balanceados com adição de peso dois no cálculo da classificação do vídeo.

Tabela 6.3: Matriz de confusão da classificação binária de publicidades alimentícias vs. publicidades não alimentícias. EfficientNet-B7 na base de treinamento (Base 1), na base de validação (Base 2) e na base de teste (Base 3).

Treinamento na Base 1	**Alimentícia**	**Não Alimentícia**
Alimentícia	100% (489/489)	0% (0/489)
Não Alimentícia	0,4% (15/4281)	99,6% (4266/4281)
Validação na Base 2	**Alimentícia**	**Não Alimentícia**
Alimentícia	90,1% (127/141)	9,9% (14/141)
Não Alimentícia	9,6% (115/1198)	90,4% (1083/1198)
Teste na Base 3	**Alimentícia**	**Não Alimentícia**
Alimentícia	90,4% (66/73)	9,6% (7/73)
Não Alimentícia	9,3% (60/647)	90,7% (587/647)

Esta etapa da dissertação resultou no artigo "*Revolutionizing food advertising monitoring: A machine learning-based method for automated classification of food videos*" [65], publicado no periódico *Public Health Nutrition*, em novembro de 2023.

EviT

A Tabela 6.4 representa o melhor resultado dos experimentos realizados, de acordo com o conjunto de validação (Base 2). A melhor técnica foi a adição de peso por classe, aplicando a fusão de *tokens* inativos.

Tabela 6.4: Matriz de confusão da classificação binária de publicidades alimentícias vs. publicidades não alimentícias. EviT na base de treinamento (Base 1), na base de validação (Base 2) e na base de teste (Base 3).

Treinamento na Base 1	**Alimentícia**	**Não Alimentícia**
Alimentícia	98,8% (483/489)	1,2% (6/489)
Não Alimentícia	4,6% (197/4281)	95,4% (4084/4281)
Validação na Base 2	**Alimentícia**	**Não Alimentícia**
Alimentícia	91,5% (129/141)	8,5% (12/141)
Não Alimentícia	11,1% (133/1198)	88,9% (1065/1198)
Teste na Base 3	**Alimentícia**	**Não Alimentícia**
Alimentícia	94,5% (69/73)	5,5% (4/73)
Não Alimentícia	9,7% (63/647)	90,3% (584/647)

Everything at Once

A Tabela 6.5 representa o melhor resultado dos experimentos realizados. A melhor configuração inclui o uso de *embeddings* de texto, áudio e vídeo com KNN considerando 50 vizinhos.

Tabela 6.5: Matriz de confusão da classificação binária de publicidades alimentícias vs. publicidades não alimentícias. EaO na base de treinamento (Base 1), na base de validação (Base 2) e na base de teste (Base 3).

Treinamento na Base 1	**Alimentícia**	**Não Alimentícia**
Alimentícia	100% (479/479)	0% (0/479)
Não Alimentícia	0% (1/4281)	100% (4280/4281)
Validação na Base 2	**Alimentícia**	**Não Alimentícia**
Alimentícia	85,5% (118/138)	14,5% (20/138)
Não Alimentícia	4,0% (49/1211)	96,0% (1162/1211)
Teste na Base 3	**Alimentícia**	**Não Alimentícia**
Alimentícia	72,6% (53/73)	27,4% (20/73)
Não Alimentícia	0,8% (5/647)	99,2% (642/647)

6.3.2 Classificação Multiclasse

De maneira semelhante, os principais resultados dos modelos treinados foram retratados pelas matrizes de confusão por vídeo, de acordo com o conjunto de validação (Base 2). As classes de interesse são *fast-food*, *supermercado*, *alimento/bebida* e a classe negativa (*não alimentícia*).

EfficientNet

Continuamos com a aplicação da EfficientNet-B7 [76] de maneira semelhante a classificação binária, empregando um *pooling* com média ponderada de cada *frame* para combinar as previsões de cada imagem para uma previsão final do vídeo. Ainda assim, aplicamos três estratégias distintas a fim de encontrar o melhor resultado: *data augmentation* [61], *batches* balanceados e adição de peso por classes.

A Tabela 6.6 representa o melhor resultado dos experimentos realizados, de acordo com o conjunto de validação (Base 2). A técnica com melhor performance foi a aplicação de *batches* balanceados.

Tabela 6.6: Matriz de confusão da classificação multiclasse de publicidades alimentícias (*fast-food*, supermercado e alimento/bebida) e publicidades não alimentícias. EfficientNet-B7 na base de treinamento (Base 1), na base de validação (Base 2) e na base de teste (Base 3).

Treinamento na Base 1	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	100,0% (79/79)	0,0% (0/79)	0,0% (0/79)	0,0% (0/79)
Supermercado	0,0% (0/111)	100,0% (111/111)	0,0% (0/111)	0,0%(0/111)
Alimento/Bebida	0,0% (0/299)	0,0% (0/299)	100,0% (299/299)	0,0% (0/299)
Não Alimentícia	0,0% (1/4281)	0,3% (11/4281)	0,1% (7/4281)	99,6% (4262/4281)
Validação na Base 2	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	70,4% (19/27)	7,4% (2/27)	14,8% (4/27)	7,4% (2/27)
Supermercado	0,0% (/25)	88,0% (22/25)	0,0% (0/25)	12,0% (3/25)
Alimento/Bebida	4,5% (4/88)	3,4% (3/88)	81,8% (72/88)	10,2% (9/88)
Não Alimentícia	1,1% (13/1198)	4,5% (54/1198)	9,3% (111/1198)	85,1% (1020/1198)
Teste na Base 3	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	100,0% (11/11)	0,0% (0/11)	0,0% (0/11)	% 0,0% (0/11)
Supermercado	0,0% (0/13)	76,9% (10/13)	7,7% (1/13)	15,4% (2/13)
Alimento/Bebida	6,1% (3/49)	0,0% (0/49)	85,7% (42/49)	8,2% (4/49)
Não Alimentícia	1,7% (11/647)	5,6% (36/647)	5,6% (36/647)	87,2% (564/647)

EviT

As mesmas técnicas para contornar o desbalanceamento foram aplicadas. Dessa vez, a melhor abordagem foi o uso de peso por classe com *pooling* balanceado, também aplicando a fusão de *tokens* inativos. A Tabela 6.7 representa o melhor resultado dos experimentos realizados.

Tabela 6.7: Matriz de confusão da classificação multiclasse de publicidades alimentícias (*fast-food*, supermercado e alimento/bebida) e publicidades não alimentícias. EviT na base de treinamento (Base 1) e na base de validação (Base 2) e na base de teste (Base 3).

Treinamento na Base 1	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	93,7% (74/79)	5,1% (4/79)	0,0% (0/79)	1,3% (1/79)
Supermercado	0,0% (0/111)	96,4% (107/111)	0,0% (0/111)	3,6% (4/111)
Alimento/Bebida	1,3% (4/299)	1,3% (4/299)	95,7% (286/299)	1,7% (5/299)
Não Alimentícia	1,3% (54/4281)	2,6% (113/4281)	3,6% (154/4281)	92,5% (3960/4281)
Validação na Base 2	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	77,8% (21/27)	7,4% (2/27)	7,4% (2/27)	7,4% (2/27)
Supermercado	0,0% (0/25)	76,0% (19/25)	8,0% (2/25)	16,0% (4/25)
Alimento/Bebida	3,4% (3/88)	1,1% (1/88)	83,0% (73/88)	12,5% (11/88)
Não Alimentícia	2,5% (30/1198)	4,7% (56/1198)	6,5% (78/1198)	86,3% (1034/1198)
Teste na Base 3	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	81,8% (9/11)	0,0% (0/11)	9,1% (1/11)	% 9,1% (1/11)
Supermercado	0,0% (0/13)	84,6% (11/13)	7,7% (1/13)	7,7% (1/13)
Alimento/Bebida	0,0% (0/49)	2,0% (1/49)	91,8% (45/49)	6,1% (3/49)
Não Alimentícia	2,3% (15/647)	9,0% (58/647)	11,9% (77/647)	76,8% (487/647)

Everything at Once

A Tabela 6.8 resume o resultado dos principais experimentos utilizando esta arquitetura. Testamos dois modelos que foram pré-treinados em duas bases distintas, classificadores diferentes, distintos *embeddings* para alimentar estes classificadores e mais quatro técnicas para diminuir o desbalanceamento dos dados e evitar o *overfitting*. Por fim, reportamos a melhor acurácia de acordo com o conjunto de validação (Base 2).

A Tabela 6.9 representa o melhor resultado dos experimentos realizados.

6.3.3 Análise dos Resultados

Os resultados apresentados destacam a eficácia e as limitações das três redes neurais analisadas no contexto da classificação de publicidades televisivas brasileiras. A seguir, analisamos qualitativamente cada arquitetura.

Um estudo detalhado das matrizes de confusão e das métricas obtidas nos ajuda a responder a questão de pesquisa Q1. Embora a EfficientNet tenha apresentado resultados sólidos, destacando sua robustez em tarefas de classificação de imagens, quando consideramos os resultados nas bases de validação, as redes baseadas em Transformers se destacaram, superando os resultados da rede convolucional em ambos os problemas de classificação, tanto binária (EaO em 0,6 pontos percentuais) quanto multiclasse (EaO em 4,2 pontos percentuais). Quando analisamos os resultados na base de teste, a EViT (Transformers) superam a EfficientNet (CNNs) na tarefa de classificação binária, com uma vantagem de 1,9 pontos percentuais. No entanto, esse cenário se inverte na classificação multiclasse, onde a EfficientNet se destaca, superando as Transformers por 3,6 pontos percentuais. Esse é um indicativo de que existe uma limitação dos dados, já que a base de validação não parece ser suficientemente representativa da distribuição dos dados

Tabela 6.8: Principais testes usando a EaO. Resultados na base de validação (Base 2).

Pré-Treino	**Classificador**	**Embedding**	**Ruído**	**SMOTE**	**Dropout**	**Audio Pooling**	**Best Acc@1**
HowTo100M	KNN	texto, vídeo e áudio	✓	✓	✓	max pool	85,6%
HowTo100M	Matriz de Similaridade	texto, vídeo e áudio	✓	✓	✓	max pool	85,1%
HowTo100M	KNN	texto, vídeo e áudio	✗	✓	✓	–	79,5%
HowTo100M	SVM	texto, vídeo e áudio	✓	✓	✓	max pool	78,5%
HowTo100M	KNN	vídeo e áudio	✗	✗	✗	–	73,9%
HowTo100M	KNN	texto, vídeo e áudio	✗	✗	✗	–	71,6%
HowTo100M	LR	texto, vídeo e áudio	✓	✓	✓	max pool	71,1%
HowTo100M	KNN	video	✗	✗	✗	–	63,4%
HowTo100M	LR	texto, vídeo e áudio	✗	✓	✓	–	40,1%
HowTo100M	KNN	áudio	✗	✗	✗	–	38,8%
HowTo100M	SVM	texto, vídeo e áudio	✗	✓	✓	–	38,4%
Youcook	KNN	texto, vídeo e áudio	✗	✓	✓	–	79,4%
Youcook	KNN	texto, vídeo e áudio	✗	✗	✗	–	72,1%
Youcook	KNN	video	✗	✗	✗	–	72,0%
Youcook	KNN	vídeo e áudio	✗	✗	✗	–	66,6%
Youcook	Matriz de Similaridade	texto, vídeo e áudio	✓	✓	✓	max pool	40,6%
Youcook	KNN	áudio	✗	✗	✗	–	40,2%
Youcook	KNN	texto, vídeo e áudio	✓	✓	✓	max pool	42,0%
Youcook	SVM	texto, vídeo e áudio	✓	✓	✓	max pool	25,7%

de teste, o que impacta a performance.

Quanto a questão Q2, o uso de estratégias simples, como aumentação de dados, *batches* balanceados, ponderação de classes, SMOTE e adição de ruído, provaram-se extremamente eficazes para mitigar os efeitos do desbalanceamento e da escassez de dados. Essas técnicas foram cruciais para melhorar a generalização dos modelos, levando em conta a discrepância no número de amostras entre as classes.

Em relação à questão Q3, nota-se que a introdução de uma Transformer multimodal no fluxograma de classificação mostrou-se eficaz no contexto multiclasse para diferenciar os tipos publicidades alimentícias (*fast-food*, supermercado e alimento/bebida). O uso de áudio em comparação com os experimentos que não o incluíam, mostra que esse se destacou ao fornecer novas informações contextuais que complementaram a análise visual, especialmente em casos em que os vídeos por si só não eram suficientes para uma classificação precisa, possibilitando o modelo de captar nuances adicionais que corrigem falhas de interpretação. Por outro lado, no contexto de classificação binária, o áudio se mostrou menos relevante, com os modelos baseados apenas em vídeo obtendo resultados superiores. Isso indica que, enquanto o áudio oferece um ganho substancial na distinção entre subcategorias de publicidades alimentícias, sua contribuição é mais modesta quando

Tabela 6.9: Matriz de confusão da classificação multiclasse de publicidades alimentícias (*fast-food*, supermercado e alimento/bebida) e publicidades não alimentícias. EaO na base de treinamento (Base 1) e na base de validação (Base 2) e na base de teste (Base 3).

Treinamento na Base 1	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	100% (69/69)	0,0% (0/69)	0,0% (0/69)	0,0% (0/69)
Supermercado	0,0% (0/111)	100% (111/111)	0,0% (0/111)	0,0% (0/111)
Alimento/Bebida	0,0% (0/299)	0,0% (0/299)	99,0% (296/299)	1,0% (3/299)
Não Alimentícia	0,0% (0/4281)	0,0% (0/4281)	0,1% (4/4281)	99,9% (4277/4281)
Validação na Base 2	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	88,0% (22/25)	8,0% (2/25)	4,0% (1/25)	0,0% (0/25)
Supermercado	0,0% (0/25)	80,0% (20/25)	8,0% (2/25)	11,0% (3/25)
Alimento/Bebida	3,4% (3/88)	1,1% (1/88)	79,5% (70/88)	15,9% (14/88)
Não Alimentícia	0,8% (10/1211)	1,2% (14/1211)	3,4% (41/1211)	94,6% (1146/1211)
Teste na Base 3	**Fast-Food**	**Supermercado**	**Alimento/Bebida**	**Não Alimentícia**
Fast-Food	77,7% (8/11)	18,2% (2/11)	9,1% (1/11)	0,0% (0/11)
Supermercado	0,0% (0/13)	76,9% (10/13)	0,0% (0/13)	23,1% (3/13)
Alimento/Bebida	4,1% (2/49)	0,0% (0/49)	77,5 % (37/49)	20,4% (10/49)
Não Alimentícia	0,0% (0/647)	0,2% (1/647)	0,6% (4/647)	99,2% (642/647)

a tarefa é identificar se o conteúdo pertence ou não a uma categoria alimentar. Resumidamente, as Transformers multimodais se provaram uma boa estratégia para introduzir o uso de áudio ao fluxograma, que também se provou, empiricamente, uma boa alternativa para aumentar a precisão em tarefas de classificação mais complexas.

EfficientNet

Após a aplicação de técnicas de balanceamento, sendo a preterida *batches* balanceados, a CNN demonstrou um bom desempenho no problema de classificação binária, conforme esperado. A arquitetura também apresentou bons resultados na classificação multiclasse, ainda assim, mesmo que os resultados tenham sido um pouco superiores no conjunto de teste (Base 3), a matriz de confusão na base de validação (Base 2) sugere que o modelo final pode ter dificuldades para generalizar em dados desconhecidos. Além disso, a EfficientNet-B7 tem alto custo computacional e de tempo de treinamento.

Resumidamente, a CNN mostrou-se eficaz, mas limitada em cenários com maior complexidade e severo desbalanceamento de dados, sendo necessária a aplicação de técnicas que contornam o problema de desbalanceamento e falta de dados para alcançar maior eficácia. No geral, os resultados sugerem que arquiteturas Transformers podem ser mais adequadas a resolver a tarefa de classificação de publicidades.

EviT

O modelo EviT teve seu desempenho próximo ao da EfficientNet na tarefa binária, com uma melhora no cenário de teste e um resultado semelhante para a tarefa multiclasse. O modelo conseguiu reduzir falsos positivos e falsos negativos de maneira mais eficaz em comparação com a CNN, especialmente no problema de classificação binária, sugerindo que a arquitetura Transformer tem uma maior capacidade de generalização.

A introdução do processo de fusão de *tokens* inativos não só otimizou o custo computacional como se provou, empiricamente, mais eficaz para o problema de classificação de publicidades alimentícias.

Everything at Once

O modelo EaO apresentou o melhor desempenho quando pré-treinado na base de dados HowTo100M, o que era esperado devido ao fato de essa ser uma base de dados bem maior. Durante o treinamento, foram utilizados *embeddings* multimodais, combinando informações de texto (o *label* da classe), vídeo e áudio, o que proporcionou uma melhor representação dos dados.

Após o ajuste de hiperparâmetros, o melhor classificador identificado foi o KNN, utilizando a distância de cosseno com 50 vizinhos. Isso sugere que tanto abordagens mais simples quanto mais complexas do que essa poderiam não oferecer bons resultados sem modificações significativas no fluxograma. Técnicas como adição de ruído, SMOTE e *dropout* foram aplicadas e, empiricamente, demonstraram trazer melhorias em comparação com outras configurações, a fim de evitar o *overfitting*.

Entre as três arquiteturas estudadas, a EaO apresentou o melhor desempenho na base de validação, para ambas as tarefas, mas o pior desempenho na base de testes. Esse é o maior indicativo das limitações da nossa base de dados e da dificuldade em garantir que ela seja representativa da distribuição dos dados de treinamento, validação e teste.

Embora a precisão final na base de teste tenha sido inferior as demais, o resultado ainda foi promissor. A EaO demonstra ser o modelo treino com maior capacidade de generalização, especialmente considerando que o conjunto de validação contém, em sua maioria, dados mais recentes em comparação ao conjunto de teste.

Em resumo, esses resultados indicam que apesar dos erros graves serem mais raros (os erros serão detalhados na próxima seção), o modelo EaO enfrenta desafios específicos devido à falta de dados e à subjetividade das publicidades alimentícias, em comparação com as não alimentícias, tendo algumas vezes alta probabilidade associada às classificações, não garantindo precisão, especialmente em vídeos com conteúdo ambíguo.

6.3.4 Análise Qualitativa dos Erros

Essa análise foca nas previsões erradas, suas características e o impacto delas na avaliação final do modelo (EaO). Para isso, a base de teste (Base 3) foi utilizada. O modelo apresentou um conjunto de erros que refletem a subjetividade e ambiguidade dos vídeos publicitários analisados. De 720 vídeos, 23 foram classificados incorretamente. A seguir, detalhamos as principais confusões entre as classes:

1. **Confusão entre *supermercado* e *não alimentícia*:** Alguns vídeos foram classificados incorretamente como *não alimentícia*, embora fossem *supermercado*. Nesses casos, a ausência de referências explícitas a alimentos parece ter confundido o modelo. Esses vídeos apenas mencionam a rede de supermercado, anunciando outros produtos, quase sem fazer referência direta a alimentos.

2. **Confusão entre *fast-food* e outras classes:** Vídeos nos quais uma rede de *fast-food* aparece, mas sem nenhuma referência visual ou verbal clara ao contexto alimentício, também são incorretamente classificados como *não alimentício*, inclusive com alta probabilidade. Além disso, o modelo ainda cometeu erros ao não ser capaz de diferenciar publicidades de *fast-food* de outras publicidades de alimento/bebida, especialmente quando os vídeos eram curtos, evidenciando uma necessidade de propagandas mais longas para ser capaz de fazer essa diferenciação entre classes alimentícias.
3. **Confusão entre *alimento/bebida* e *não alimentícia*:** Alguns vídeos foram classificados incorretamente como *não alimentícios*, embora fossem de *alimento/bebida*. Nestes vídeos, o modelo foi confundido pelo contexto visual, onde, apesar de haver referência a alimentos, o foco da propaganda está em outros elementos, como celebridades ou atividades, a maior parte do tempo.
4. **Confusão entre suplementos alimentares e a classe *não alimentícia*:** Um padrão recorrente foi a dificuldade do modelo em lidar com publicidades de medicamentos ou suplementos alimentares. Vídeos que fazem referência a vitaminas e suplementos (que deveriam ser considerados *alimento/bebida*) e remédios (*não alimentícia*), confundiram o modelo. O problema maior se deve ao fato de essa ser uma área cinzenta na classificação, visto que apesar de esse produtos serem ingeríveis, possuem características que os distinguem de *alimento/bebida* convencionais. Idealmente, segundo o protocolo INFORMAS [35], suplementos nutricionais deveriam ter sua própria categoria. No entanto, pela baixa quantidade de dados, os poucos vídeos existentes foram considerados como publicidades de *alimento/bebida*.
5. **Erros graves:** Nenhum dos erros pode ser explicitamente classificado como "grave", pois embora o modelo tenha cometido erros de classificação, nenhum deles levou a confusões críticas no contexto geral das publicidades. Mesmo assim, vale destacar que a presença de erros com alta confiança entre algumas classes foi foi observada.

Em suma, o modelo EaO apresentou dificuldades em classificar vídeos que envolvem referências indiretas a alimentos, produtos alimentares com foco secundário, ou quando a publicidade se desvia para temas como saúde ou interação social (veja a Tabela 6.10). A presença de produtos por um breve período ou em contextos altamente subjetivos parece ser um dos fatores mais impactantes nos erros de classificação, sugerindo que o modelo ainda enfrenta dificuldades em distinguir entre classes alimentícias e não alimentícias. Ajustes futuros no fluxograma de treinamento, como um maior refinamento nos *embeddings* multimodais ou o uso de técnicas que permitam ao modelo focar melhor no conteúdo visual relevante, poderiam mitigar alguns desses problemas.

6.3.5 Avaliação em Vídeos do YouTube

Para analisar o comportamento em outras mídias, coletamos manualmente 1.158 vídeos publicitários do YouTube, provenientes de canais brasileiros de diversas marcas, como

Tabela 6.10: Síntese dos erros de classificação do modelo EaO. Em 'Nome do Vídeo', os vídeos listados foram mantidos com os nomes originais por uma questão de referência.

Categoria de Erro	**Nome do Vídeo**	**Descrição do Erro**
Supermercado	*121036_2018*, *130203_2018*, *133747_2019*	Classificados incorretamente devido à ausência de referências (visuais e verbais) explícitas a alimentos
Fast-Food	*133849_2019*, *133703_2019*, *122317_2018*	Classificados incorretamente devido à ausência de referências (visuais e verbais) explícitas a alimentos ou à curta duração do vídeo
Alimento/bebida	*133661_2019*, *122280_2018*, *130105_2018*, *133661_2019*, *120821_2018*	Publicidades que mencionam alimentos de forma indireta ou com foco em outros elementos (por exemplo, atividades ou celebridades), ou com características visuais muito distintas (por exemplo, produto representado em forma de desenho)
Suplementos e Medicamentos	*117261_2018*, *22404203_2018*, *0050480_2018*, *12904130_2018*, *121344_2018*	Produtos como vitaminas, medicamentos e suplementos confundiram o modelo, sendo algumas vezes classificados como *alimento/bebida* e outras vezes como *não alimentícia*
Graves	–	Embora existam erros, nenhum foi explicitamente classificado como grave, visto os contextos subjetivos associados aos erros

McDonald's, *Omo*, *Burger King*, *Barbie*, *Hot Wheels*, *Extra* e *Magazine Luiza*. Essas marcas foram escolhidas porque estavam muito presentes em nossa base de dados inicial, o que indicava sua relevância.

A coleta abrangeu todos os vídeos disponíveis nesses canais, com foco especial na categoria *shorts* (vídeos curtos), excluindo aqueles com mais de 5 minutos de duração. Embora nem todos os vídeos fossem necessariamente publicitários, a maioria se enquadrava nessa categoria ou apresentava conteúdo bastante similar.

Adicionalmente, a anotação das classes foi realizada de forma semi-automática, sem a participação direta de especialistas. Nesse processo, classificamos como publicidades alimentícias todos os conteúdos associados a marcas amplamente reconhecidas no setor alimentício ou que apresentaram um número elevado de publicidades alimentícias na base de dados inicial. Os demais casos foram classificados como não alimentícios. Por exemplo, anúncios de marcas como *McDonald's*, *Burger King*, *Extra* e *Magazine Luiza* foram considerados alimentícios, enquanto publicidades de marcas como *Omo*, *Barbie* e *Hot Wheels* foram categorizadas como não alimentícias.

Seguindo um protocolo semelhante ao utilizado anteriormente, buscamos avaliar o

Tabela 6.11: Matriz de confusão da classificação binária de publicidades alimentícias vs. publicidades não alimentícias. EviT na base de vídeos do Youtube.

Teste no YouTube	**Alimentícia**	**Não Alimentícia**
Alimentícia	71,5% (535/712)	24,9% (177/712)
Não Alimentícia	39,5% (176/446)	60,5% (270/446)

desempenho da EViT (o melhor modelo) na tarefa de classificação binária, por sua menor complexidade. Não foi realizado novo treinamento; em vez disso, extraímos *frames* desses vídeos a uma taxa de um quadro por segundo para todas as categorias. Para aumentar a robustez do modelo, aplicamos uma técnica de aumentação nos *frames* coletados [60, 70]. A ideia principal é gerar versões modificadas desses *frames*, utilizando as técnicas de aumentação previamente discutidas, como solarização, escala de cinza, *Gaussian blur*, rotações e translações. Isso contribui para melhorar a capacidade do modelo de generalizar e lidar com pequenas variações ou ruídos no ambiente real.

Realizamos cinco classificações por *frames*: uma para o original e outras quatro utilizando as técnicas de aumentação. A classificação final de cada *frame* foi calculada pela média das cinco predições, e a classificação final de cada vídeo foi obtida a partir da média ponderada das classificações de seus *frames*. A Tabela 6.11 apresenta o resultado do experimento.

Embora os resultados sejam promissores, visto que o modelo apresentou bom desempenho sem treinamento específico nos dados do YouTube, ainda há desafios a serem superados, especialmente com publicidades mais recentes e aquelas direcionadas ao público infantil, conforme será discutido a seguir.

O modelo obteve uma boa precisão na classificação de publicidades alimentícias. No entanto, a taxa de falsos negativos para publicidades não alimentícias revela dificuldades em distinguir corretamente certos tipos de vídeos; uma análise mais detalhada mostra que a maioria era voltada ao público infantil. Muitos desses vídeos foram erroneamente classificados como alimentícios, o que pode ocorrer devido à semelhança visual e temática entre as publicidades infantis e as de alimentos, que frequentemente utilizam cores vibrantes e personagens lúdicos, dificultando a diferenciação pelo modelo. Mesmo que esse desempenho não seja ideal para o monitoramento estrito de publicidades alimentícias, ele demonstra que, do ponto de vista da fiscalização, a ferramenta pode ser útil para identificar quaisquer publicidades direcionadas a crianças, dado que toda publicidade infantil tem caráter abusivo, independentemente de seu conteúdo específico.

Outro aspecto a ser considerado é a natureza dinâmica das publicidades mais recentes (anos de 2022, 2023 e 2024), que podem apresentar elementos visuais diferentes dos padrões da base de dados original (Base 1, anos de 2020 e 2019), utilizada para o treinamento do modelo. Essa mudança pode prejudicar a capacidade do modelo de generalizar corretamente para novos tipos de vídeos, indicando a necessidade de um retreinamento com dados mais recentes, não obrigatoriamente vindos do meio televisivo, podendo assim incluir outras fontes como as mídias sociais.

Vale ressaltar que, diferentemente dos dados usados no treinamento original, os vídeos

do YouTube não foram coletados seguindo o protocolo do INFORMAS [35], devido a limitações de tempo. Isso pode ter introduzido inconsistências na base, como a inclusão de vídeos que não são estritamente publicitários, ou que fogem ao escopo de classificação esperado, contribuindo para as dificuldades do modelo em classificar adequadamente entre publicidades alimentícias e não alimentícias.

Esses fatores explicam porque, apesar dos resultados promissores, o modelo enfrenta dificuldades. Ainda assim, o desempenho da EViT sem ajustes específicos para essa nova base de vídeos é positivo, apontando para futuras melhorias, como uma adaptação mais robusta aos diferentes estilos de publicidade.

Capítulo 7

Conclusão

Neste capítulo, apresentamos as conclusões finais desta dissertação (Seção 7.1), além de discutirmos prospectivas futuras sobre as aplicações dessa pesquisa (Seção 7.2), e as principais consequências éticas (Seção 7.3).

7.1 Considerações Finais

Esta dissertação oferece uma contribuição pioneira ao disponibilizar a primeira base de dados brasileira dedicada à classificação de diferentes tipos de publicidades veiculadas na televisão. A criação dessa base representa um marco importante, pois fornece um recurso inédito para o estudo do monitoramento automático e fiscalização, especialmente de publicidades de alimentos, beneficiando pesquisadores, formuladores de políticas públicas e agentes reguladores.

Além disso, aplicamos técnicas de inteligência artificial e aprendizado de máquina de forma original e inovadora para automatizar as etapas iniciais do protocolo INFORMAS [35]. A automatização dessa parte do protocolo se concentra na classificação de tipos de publicidades, uma tarefa crucial para o monitoramento e fiscalização de publicidades alimentícias. A automatização facilita a análise de um grande volume de publicidades em um curto espaço de tempo, também facilitando futuras coletas de dados pois representa uma economia substancial de recursos financeiros e humanos.

Ainda que o foco principal desta dissertação seja a publicidade brasileira na televisão, as implicações deste estudo se estendem a outras mídias, como as plataformas digitais, que têm um papel cada vez mais central nas campanhas publicitárias no Brasil e em outras partes do mundo. A metodologia desenvolvida nesta dissertação pode ser aplicada, com poucas ou nenhuma adaptação, para treinar modelos capazes de classificar publicidades brasileira provindas de outras fontes. Em contextos internacionais, onde diferentes países enfrentam desafios semelhantes, o método pode ser ajustado com base em uma coleta de dados local, de modo a respeitar as particularidades culturais de cada região, tornando-o versátil e adaptável.

Atualmente, apesar dos poucos trabalhos voltados para o monitoramento automático de publicidades de alimentos [37], os resultados destacam a necessidade de se avançar na agenda do cumprimento da fiscalização de publicidades alimentícias que promovem ali-

mentos não saudáveis, especialmente aquelas direcionadas às crianças [8, 37]. O aumento das taxas de obesidade e outras doenças relacionadas à má alimentação [18, 59] também reforça a necessidade de desenvolver e implementar tecnologias que possam auxiliar na regulação dessas publicidades, indicando que esta é uma uma área de interesse, tanto de pesquisadores [54, 55, 56] quanto do governo e da sociedade como um todo. Portanto, a área de monitoramento de publicidades de alimentos é não apenas de grande interesse acadêmico, mas também de importância estratégica para políticas de saúde pública, contando com o envolvimento de governos, agências reguladoras, ONGs e a sociedade como um todo.

Mais especificamente, este estudo revelou que o modelo final utilizando *Everything at Once* (EaO), tem maior dificuldade em diferenciar referências a marcas, especialmente na ausência de elementos visuais claros dos produtos. Isso é particularmente desafiador quando não há imagens que conectem diretamente a marca a alimentos ou atos de consumo, dificultando o reconhecimento da categoria correta. Ou seja, a presença ou ausência de elementos visuais relacionados a alimentos e a subjetividade das imagens influenciam significativamente os resultados do classificador. Vídeos que não exibem alimentos de maneira clara ou que o fazem por um tempo muito curto tendem a ser classificados incorretamente, especialmente quando o foco está em outra atividade ou produto. Nesses casos, a classificação depende fortemente do conhecimento prévio sobre os produtos que a marca costuma oferecer, algo que o classificador tem dificuldade em aprender devido à quantidade limitada de dados disponíveis.

Por esse motivo, os tipos de publicidades mais desafiadores incluem produtos como suplementos e cervejas, além de marcas alimentícias (como a Sadia). Essa complexidade se deve às características específicas dessas publicidades: suplementos podem ser considerados medicamentos, bebidas alcoólicas não podem anunciar diretamente seus produtos em parte da grade horária, e marcas como a Sadia oferecem uma variedade de itens típicos de *fast-food*, gerando confusões. Esses erros poderiam ser corrigidos com a adição de mais exemplos semelhantes à base de dados e o subsequente retreinamento do modelo, dada que quantidade de exemplos disponíveis para esse tipo de publicidade mais desafiadora é ainda mais limitada.

Resumidamente, embora grandes mudanças nas arquiteturas de redes neurais não tenham sido propostas, a pesquisa avança na aplicação prática dessas tecnologias em um contexto específico, especialmente no que diz respeito ao monitoramento automático de publicidades, apresentando uma contribuição significativa para o campo de estudo da saúde coletiva. Por fim, os achados deste estudo não só corroboram a necessidade de avanço na fiscalização automatizada de publicidades alimentícias, mas também abrem caminho para novas investigações acadêmicas.

7.2 Trabalhos Futuros

Os resultados desta dissertação abrem diversas possibilidades para pesquisas futuras. Uma das principais recomendações é a ampliação da base de dados, que pode incluir a coleta de novos anos de dados televisivos e a inclusão de publicidades de diferentes plataformas,

como redes sociais. Essa ampliação proporcionaria uma visão mais abrangente sobre as estratégias de marketing de alimentos. Além disso, a incorporação de dados de áudio e transcrições de diálogos poderia enriquecer a análise, permitindo uma avaliação mais completa das estratégias persuasivas utilizadas nas campanhas publicitárias e facilitando o treinamento de um modelo mais aprimorado. A coleta de dados adicionais não apenas pode ser facilitada pelos resultados obtidos neste trabalho, pela aplicação de uma abordagem de rotulação assistida [6], mas também é essencial para aprimorar a performance dos modelos e explorar novas abordagens, considerando que o conteúdo das publicidades muda significativamente ao longo dos anos.

A cada ano, novos modelos baseados em Transformers são disponibilizados. Portanto, experimentos com abordagens mais atuais, utilizando modelos pré-treinados em grandes bases de dados de vídeos, podem resultar em um aumento significativo na precisão da classificação, especialmente em categorias desafiadoras. Essa atualização nas arquiteturas é crucial para criar-se um modelo cada vez mais capaz de 'entender' a subjetividade e a ambiguidade presentes nas publicidades.

Além disso, a aplicação de uma abordagem interdisciplinar que envolva áreas como psicologia, sociologia e nutrição pode enriquecer a compreensão do impacto das publicidades de alimentos na saúde pública e no comportamento do consumidor. Essa colaboração pode ajudar a decifrar como diferentes contextos afetam a eficácia das campanhas publicitárias e, por consequência, influenciam as estratégias utilizadas. Assim, uma nova fase de pesquisa pode se concentrar no reconhecimento e na análise dos principais padrões de estratégias persuasivas empregadas nas publicidades de alimentos, completando, assim, o processo de fiscalização e facilitando o processo de automatização.

Por fim, a criação de ferramentas automatizadas para o monitoramento em tempo real das publicidades alimentícias na TV e em outras plataformas permitiria que organizações fiscalizassem continuamente as práticas de marketing e respondessem rapidamente a campanhas que possam ser prejudiciais à saúde pública. Isso possibilitaria uma avaliação ágil das políticas públicas atuais que restringem a publicidade de alimentos não saudáveis, especialmente aquelas direcionadas a crianças.

7.3 Considerações Éticas

Embora o foco principal da pesquisa não tenha sido abordar diretamente as consequências éticas deste trabalho, é crucial considerar as implicações do uso de classificadores automatizados, especialmente no monitoramento de conteúdos voltados para o público infantil. A precisão e a interpretação das classificações impactam diretamente as ações regulatórias e as políticas de proteção ao consumidor. Portanto, quaisquer erros ou injustiças cometidas nesse processo podem ter não só um impacto comercial para as empresas envolvidas, como também afetar a vida cotidiana da população brasileira e ter impactos diretos na saúde pública.

Além disso, se o modelo for utilizado de maneira mal-intencionada, as consequências podem ser severas. Uma possível exploração da automação pode incluir o direcionamento intencional de publicidade enganosa para públicos vulneráveis, especialmente para

crianças, utilizando de IA para explorar vulnerabilidades emocionais ou psicológicas, e direcionando mais publicidades de alimentos não saudáveis a esse público, contribuindo para normalização de comportamentos alimentares prejudiciais e a perpetuação de problemas como a obesidade.

Ainda, a falta de supervisão e a transparência nos algoritmos de classificação podem resultar em discriminação e viés. Se os dados utilizados para treinar os modelos forem enviesados, as classificações poderão refletir esses preconceitos, perpetuando injustiças, especialmente quando consideramos que as publicidades diferem de acordo com o público e suas condições sociais. Isso pode criar um cenário onde certos grupos são injustamente alvos de regulamentações mais rigorosas ou, inversamente, onde produtos potencialmente prejudiciais possam passar despercebidos, pois a base de dados utilizada para o treinamento continha algum tipo de viés.

Adicionalmente, é fundamental que as tecnologias de IA sejam utilizadas de maneira responsável e monitorada por órgãos responsáveis independentes, garantindo assim que a coleta, análise e uso de dados respeitem a privacidade dos indivíduos. Finalmente, promover a alfabetização digital é essencial para garantir que a sociedade possa responder criticamente sobre as implicações do uso da IA, para que tomem decisões informadas a seu respeito.

Referências Bibliográficas

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan e S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 16, 42

[2] H. A. Alawaad. The role of artificial intelligence (ai) in public relations and product marketing in modern organizations. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(14):3180–3187, 2021. 31

[3] J. Ba. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 23

[4] D. Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 22, 23

[5] M. Belić, V. Bobić, M. Badža, N. Šolaja, M. Đurić-Jovičić e V. S. Kostić. Artificial intelligence for assisting diagnostics and assessment of parkinson's disease—a review. *Clinical neurology and neurosurgery*, 184:105442, 2019. 15

[6] B. C. Benato, C. Grosu, A. X. Falcão e A. C. Telea. Human-in-the-loop: Using classifier decision boundary maps to improve pseudo labels. *Computers & Graphics*, 124:104062, 2024. 70

[7] R. M. Bielemann, J. V. S. Motta, G. C. Minten, B. L. Horta e D. P. Gigante. Consumption of ultra-processed foods and their impact on the diet of young adults. *Revista de saude publica*, 49:28, 2015. 14

[8] M. Blades, C. Oates e S. Li. Children's recognition of advertisements on television and on web pages. *Appetite*, 62:190–193, 2013. 14, 15, 32, 69

[9] E. J. Boyland, S. Nolan, B. Kelly, C. Tudur-Smith, A. Jones, J. C. Halford e E. Robinson. Advertising as a cue to consume: a systematic review and meta-analysis of the effects of acute exposure to unhealthy food and nonalcoholic beverage advertising on intake in children and adults, 2. *The American journal of clinical nutrition*, 103 (2):519–533, 2016. 14, 32

[10] H. Boz e U. Kose. Emotion extraction from facial expressions by using artificial intelligence techniques. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 9(1):5–16, 2018. 31

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 46, 56

[12] X. Chen, C.-J. Hsieh e B. Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representations*, 2022. 15

[13] P. C. Coleman, P. Hanson, T. van Rens e O. Oyebode. A rapid review of the evidence for children's tv and online advertisement restrictions to fight obesity. *Preventive Medicine Reports*, page 101717, 2022. 14

[14] I. Contreras e J. Vehi. Artificial intelligence for diabetes management and decision support: literature review. *Journal of medical Internet research*, 20(5):e10775, 2018. 15

[15] T. Darcet, M. Oquab, J. Mairal e P. Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 34, 35

[16] T. Davenport, A. Guha, D. Grewal e T. Bressgott. How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48: 24–42, 2020. 31, 32

[17] A. De Bruyn, V. Viswanathan, Y. S. Beh, J. K.-U. Brock e F. Von Wangenheim. Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, 51(1):91–105, 2020. 31

[18] I. B. de Geografia e Estatística. Pesquisa nacional de saúde - 2019. percepção do estado de saúde, estilos de vida e doenças crônicas: Brasil, grandes regiões e unidades da federação, 2020. 14, 69

[19] J. Devlin, M.-W. Chang, K. Lee e K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 23, 24

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Machine Learning*, 2021. 15, 24, 25, 33

[21] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli e L. Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296, 2021. 16, 33

[22] V. P. Ferreira, P. M. Horta e S. Avila. EfficientNet para o monitoramento automático de publicidades de alimentos. 2024. Relatório Técnico–IC-PFG de Projeto Final de Graduação. 41

[23] J. Gao, C. Sun, Z. Yang e R. Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 34

[24] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii e K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 36, 42, 80

[25] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou e M. Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *International Conference on Computer Vision*, pages 12259–12269, 2021. 16, 33, 34

[26] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 23

[27] K. Gupta, K. Singh, G. V. Singh, M. Hassan, U. Sharma et al. Machine learning based credit card fraud detection-a review. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 362–368. IEEE, 2022. 31

[28] A. Haleem, M. Javaid, M. A. Qadri, R. P. Singh e R. Suman. Artificial intelligence (ai) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3:119–132, 2022. 31

[29] K. Hara, H. Kataoka e Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 51

[30] K. He, X. Zhang, S. Ren e J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016. 16, 27, 34

[31] C. Ingle, D. Bakliwal, J. Jain, P. Singh, P. Kale e V. Chhajed. Demand forecasting: Literature review on various methodologies. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2021. 31

[32] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman e J. Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pages 4651–4664, 2021. 33, 34

[33] G. Jenkin, N. Madhvani, L. Signal e S. Bowers. A systematic review of persuasive marketing techniques to promote food to children on television. *Obesity reviews*, 15 (4):281–293, 2014. 14

[34] P. Jin, X. Ji, W. Kang, Y. Li, H. Liu, F. Ma, S. Ma, H. Hu, W. Li e Y. Tian. Artificial intelligence in gastric cancer: a systematic review. *Journal of cancer research and clinical oncology*, 146:2339–2350, 2020. 15

[35] B. Kelly. Informas protocol: Food promotion module: Food marketing-television protocol. 2017. 19, 36, 38, 64, 67, 68

[36] B. Kelly, L. King, L. Baur, M. Rayner, T. Lobstein, C. Monteiro, J. Macmullan, S. Mohan, S. Barquera, S. Friel et al. Monitoring food and non-alcoholic beverage promotions to children. *Obesity reviews*, 14:59–69, 2013. 17

[37] B. Kelly, K. Backholer, E. Boyland, M. P. Kent, M. A. Bragg, T. Karupaiah e S. Ng. Contemporary approaches for monitoring food marketing to children to progress policy actions. *Current Nutrition Reports*, 12(1):14–25, 2023. 14, 15, 32, 68, 69

[38] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan e M. Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021. 25

[39] F. Kitsios e M. Kamariotou. Artificial intelligence and business strategy towards digital transformation: A research agenda. *Sustainability*, 13(4):2025, 2021. 31

[40] A. Krizhevsky, G. Hinton et al. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 16

[41] A. Krizhevsky, I. Sutskever e G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 15

[42] H. Kuehne, A. Iqbal, A. Richard e J. Gall. Mining youtube-a dataset for learning fine-grained action concepts from webly supervised video data. *arXiv preprint arXiv:1906.01012*, 2019. 28, 33

[43] V. Kumar, B. Rajan, R. Venkatesan e J. Lecinski. Understanding the role of artificial intelligence in personalized engagement marketing. *California management review*, 61(4):135–155, 2019. 31

[44] J. Lei, T. L. Berg e M. Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 34

[45] R. B. Levy, R. M. Claro e C. A. Monteiro. Sugar and overall macronutrient profile in the brazilian family diet (2002-2003). *Cadernos de saude publica*, 26(3):472–480, 2010. 14

[46] R. B. Levy, R. M. Claro, D. H. Bandoni, L. Mondini e C. A. Monteiro. Disponibilidade de"açúcares de adição"no brasil: distribuição, fontes alimentares e tendência temporal. *Revista Brasileira de Epidemiologia*, 15:3–12, 2012. 14

[47] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li e Y. Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600, 2023. 34

[48] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang e P. Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 26, 33, 35

[49] M. L. d. C. Louzada, A. P. B. Martins, D. S. Canella, L. G. Baraldi, R. B. Levy, R. M. Claro, J.-C. Moubarac, G. Cannon e C. A. Monteiro. Ultra-processed foods and the nutritional dietary profile in brazil. *Revista de Saúde Pública*, 49, 2015. 14

[50] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev e J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 28, 33

[51] T. Mikolov, K. Chen, G. Corrado e J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 27

[52] S. Mills, L. Tanner e J. Adams. Systematic literature review of the effects of food and drink advertising on food and drink-related behaviour, attitudes and beliefs in adult populations. *Obesity Reviews*, 14(4):303–314, 2013. 14, 32

[53] E. A. Nilson, G. Ferrari, M. L. C. Louzada, R. B. Levy, C. A. Monteiro e L. F. Rezende. Premature deaths attributable to the consumption of ultraprocessed foods in brazil. *American Journal of Preventive Medicine*, 64(1):129–136, 2023. 14

[54] D. L. Olstad e E. Boyland. Towards effective restriction of unhealthy food marketing to children: unlocking the potential of artificial intelligence. *International Journal of Behavioral Nutrition and Physical Activity*, 20(1):61, 2023. 32, 69

[55] D. L. Olstad e J. Lee. Leveraging artificial intelligence to monitor unhealthy food and brand marketing to children on digital media. *The Lancet Child & Adolescent Health*, 4(6):418–420, 2020. 32, 69

[56] D. L. Olstad, M. Raman, C. Valderrama, Z. S. H. Abad, A. B. Cheema, S. Ng, A. Memon e J. Lee. Development of an artificial intelligence system to monitor digital marketing of unhealthy food to children: Research protocol. *Current Developments in Nutrition*, 6:1151–1151, 2022. 32, 69

[57] W. H. Organization. Global status report on noncommunicable diseases 2014, 2014. URL `https://iris.who.int/bitstream/handle/10665/148114/9789241564854_eng.pdf?ua=1`. Last accessed 11 September 2024. 14

[58] W. H. Organization et al. Set of recommendations on the marketing of foods and non-alcoholic beverages to children, 2010. 14

[59] W. H. Organization et al. World health organization obesity and overweight fact sheet, 2022. URL `https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight`. 14, 69

[60] F. Perez, C. Vasconcelos, S. Avila e E. Valle. Data augmentation for skin lesion analysis. In *Skin Image Analysis Workshop, MICCAI*, pages 303–311, 2018. 66

[61] F. Perez, C. Vasconcelos, S. Avila e E. Valle. Data augmentation for skin lesion analysis. In *Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI*, pages 303–311, 2018. 52, 59

[62] A. Piergiovanni, W. Kuo e A. Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2214–2224, 2023. 16, 33, 34

[63] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al. Improving language understanding by generative pre-training. 2018. 23

[64] T. Ridnik, E. Ben-Baruch, A. Noy e L. Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 16, 42

[65] M. B. Rodrigues, V. P. Ferreira, R. M. Claro, A. P. B. Martins, S. Avila e P. M. Horta. Revolutionising food advertising monitoring: a machine learning-based method for automated classification of food videos. *Public Health Nutrition*, 26(12):2717–2727, 2023. 57

[66] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 27

[67] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 16, 25, 26, 30

[68] S. J. Russell, H. Croker e R. M. Viner. The effect of screen advertising on children's dietary intake: A systematic review and meta-analysis. *Obesity reviews*, 20(4):554–568, 2019. 14, 32

[69] B. Sadeghirad, T. Duhaney, S. Motaghipisheh, N. Campbell e B. Johnston. Influence of unhealthy food and beverage marketing on children's dietary intake and preference: a systematic review and meta-analysis of randomized trials. *Obesity Reviews*, 17(10): 945–959, 2016. 14, 32

[70] D. Shanmugam, D. Blalock, G. Balakrishnan e J. Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021. 66

[71] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass e H. Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 20020–20029, 2022. 27, 33, 34, 51, 52

[72] Y. Song, J. Vallmitjana, A. Stent e A. Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 34

[73] M. Sun, A. Farhadi e S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 787–802. Springer, 2014. 34

[74] B. Swinburn, G. Sacks, S. Vandevijvere, S. Kumanyika, T. Lobstein, B. Neal, S. Barquera, S. Friel, C. Hawkes, B. Kelly et al. Informas (i nternational n etwork for f ood and o besity/non-communicable diseases r esearch, m onitoring and a ction s upport): overview and key principles. *Obesity reviews*, 14:1–12, 2013. 16

[75] J. Tan, K. A. Cherkauer e I. Chaubey. Developing a comprehensive spectral-biogeochemical database of midwestern rivers for water quality retrieval using remote sensing data: a case study of the wabash river and its tributary, indiana. *Remote Sensing*, 8(6):517, 2016. 31

[76] M. Tan e Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 29, 52, 59

[77] S. Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006. 41, 50

[78] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles e H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 25, 33

[79] H. Touvron, M. Cord e H. Jégou. DeiT III: Revenge of the ViT. *arXiv preprint arXiv:2204.07118*, 2022. 16, 33, 35, 53

[80] E. Valle, M. Fornaciali, A. Menegola, J. Tavares, F. V. Bittencourt, L. T. Li e S. Avila. Data, depth, and design: Learning reliable models for skin lesion analysis. *Neurocomputing*, 383:303–313, 2020. 15

[81] L. Van der Maaten e G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 47

[82] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona e S. Belongie. The inaturalist species classification and detection dataset. In *Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 16, 25

[83] S. Vandevijvere, G. Sacks e B. Swinburn. International network for food and obesity/ncd research, monitoring and action support: benchmarking food environments towards healthier diets. In *Annals of Nutrition & Metabolism: Abstracts of the 20th International Congress of Nutrition 2013*, pages 865–865, 2013. 19, 36

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser e I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 22, 23, 24, 32

[85] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan e L. Zhang. Cvt: Introducing convolutions to vision transformers. In *International Conference on Computer Vision*, pages 22–31, 2021. 16, 33, 34

[86] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang e X. Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18709–18719, 2024. 34

[87] J. Xu, T. Mei, T. Yao e Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 28, 33

[88] L. Zhou, C. Xu e J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 28, 33

[89] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev e J. Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 28, 33

Apêndice A

Datasheet

Este documento é baseado no artigo *Datasheets for Datasets* [24].

A.1 Motivação

Para qual propósito a base de dados foi criada? Havia alguma tarefa específica em mente? Havia alguma lacuna a ser preenchida? Por favor, forneça uma descrição. A base de dados foi criada com o objetivo de monitorar a publicidade televisiva de alimentos, conforme o protocolo da Rede INFORMAS. O foco é analisar as categorias de publicidade, com ênfase nos subtipos de publicidades alimentícias. Essa iniciativa preencheu uma lacuna no monitoramento das estratégias da publicidade de alimentos, especialmente na televisão aberta e fechada no Brasil.

Quem criou a base de dados (por exemplo, qual equipe, grupo de pesquisa) e em nome de qual entidade (por exemplo, empresa, instituição, organização)? A base de dados foi criada pelo Departamento de Nutrição da Universidade Federal de Minas Gerais (UFMG), em parceria com o Instituto de Defesa do Consumidor(IDEC) e a Rede INFORMAS, que coordena os protocolos de monitoramento de publicidade televisiva de alimentos.

Quem financiou a criação da base de dados? Se houver um número de financiamento associado, forneça o nome e o número do financiamento. A criação da base de dados foi realizada no âmbito de uma parceria acadêmica com financiamento indireto do Conselho Nacional de Desenvolvimento Científico (CNPq) sob número de processo 442789/2019-0 e do The International Development Research Center sob identificação projectId108166.

A.2 Composição

O que representam as instâncias que compõem a base de dados (por exemplo, documentos, fotos, pessoas, países)? Existem diversos tipos de instâncias (por exemplo, filmes, usuários e classificações; pessoas e interações entre eles;

Tabela A.1: Número de vídeos por tipo de publicidades (alimentícias: *fast-food*, supermercado, alimento/bebida, vs. não alimentícias) e pela separação da base de dados.

		Base 1 (Treino)	**Base 2 (Valid.)**	**Base 3 (Teste)**	**Total**
Alim.	*Fast-Food*	79	28	11	118
	Supermercado	111	25	13	149
	Alimento/Bebida	299	88	49	436
	Não Alimentícia	4.281	1.211	647	6.139

nós e bordas)? Por favor, forneça uma descrição. As instâncias da base de dados representam publicidades televisivas brasileiras transmitidas em canais de TV aberta (*Globo*, *Record*, *SBT*) e TV fechada (*Discovery Kids*, *Cartoon Network*). Essas instâncias consistem em vídeos e suas respectivas classificações em publicidades alimentícia e seus subtipos.

Quantas instâncias existem no total (de cada tipo, se for o caso)? Veja a Tabela A.1.

A base de dados contém todas as instâncias possíveis ou é uma amostra (não necessariamente aleatória) de instâncias de uma base maior? Se a base de dados é uma amostra, qual é a base maior? A amostra é representativa da base maior (por exemplo, cobertura geográfica)? Se sim, descreva como essa representatividade foi validada/verificada. Se não for representativo da base de dados maior, descreva por que não (por exemplo, para cobrir uma gama mais diversificada de instâncias, porque as instâncias foram retidas ou indisponíveis). A base de dados é uma amostra das publicidades veiculadas nesses canais, representando uma amostra focada e aleatória de períodos específicos monitorados entre 2018 e 2020.

Em que dados consiste cada instância? Dados "brutos" (por exemplo, texto ou imagens não processados) ou *features*? Em ambos os casos, forneça uma descrição. Cada instância consiste em vídeos de publicidades televisivas e as suas respectivas anotações, que incluem informações a classificação dos tipos de publicidade.

Existe um rótulo associado a cada instância? Se sim, forneça uma descrição. Sim, cada instância (vídeo de publicidade) possui um rótulo associado, indicando o programa, a timestamp, o tipo de programa, o tipo de publicidade, o tipo de produto, a empresa e um id único para o vídeo.

Falta alguma informação em instâncias individuais? Se sim, forneça uma descrição, explicando por que essa informação está ausente (por exemplo, porque

não estava disponível). Isso não inclui informações removidas intencionalmente, mas pode incluir, por exemplo, texto redigido. Especificamente, para as publicidades não alimentícias não temos anotadas: o tipo de produto e a empresa.

Os relacionamentos entre instâncias individuais da base de dados são explícitos (por exemplo, classificações de filmes dos usuários, links de redes sociais)? Se sim, descreva como essas relações são explicitadas. Não, os relacionamentos entre as instâncias individuais da base de dados não são explícitos. Cada instância (publicidade televisiva) é tratada de forma independente, com suas respectivas anotações sobre tipo de publicidade e estratégias de marketing persuasivas, sem conexões explícitas entre elas.

Existem splits de dados recomendados (por exemplo, treinamento, validação, teste)? Se sim, forneça uma descrição desses splits, explicando a lógica por trás delas. Sim, existe uma divisão recomendada dos dados para os tipos de publicidade. Após a preparação da base de dados, ela foi dividida em três partes:

- **Base 1:** Utilizada para o treinamento dos algoritmos, composta principalmente pelos vídeos dos anos de 2020 e 2019.
- **Base 2:** Destinada à validação dos resultados, contendo uma pequena porcentagem de vídeos de 2018 e 2019.
- **Base 3:** Reservada para o teste final do modelo, composta por vídeos de 2018 e 2019.

Essa divisão foi realizada de maneira a evitar o enviesamento dos modelos, garantindo que eles sejam capazes de generalizar bem para novos dados. A maior parte dos dados foi destinada ao treinamento, enquanto as bases de validação e teste foram planejadas para representar a totalidade da base de forma equilibrada, de modo que resultados mais robustos possam ser obtidos.

No entanto, essa divisão pode ser desconsiderada, a depender do objetivo da análise ou modelo que está sendo implementado pelo pesquisador.

Existem erros, fontes de ruído ou redundâncias na base de dados? Se sim, forneça uma descrição. Embora o processo de coleta e anotação tenha sido cuidadoso, é possível que existam fontes de ruído ou redundâncias, como erros de corte, a repetição de publicidades em diferentes períodos de coleta ou variações nas classificações manuais das estratégias de marketing.

A base de dados é autocontida ou está vinculada ou depende de recursos externos (por exemplo, sites, tweets, outros conjuntos de dados)? Se estiver vinculada ou depender de recursos externos, a) existem garantias de que eles existirão e permanecerão constantes ao longo do tempo; b) existem versões de arquivo oficiais na base de dados completa (ou seja, incluindo os recursos externos que existiam no momento em que a base de dados foi criada); c) existem restrições (por exemplo, licenças, taxas) associadas a algum dos recursos

externos que podem se aplicar a um consumidor da base de dados? Forneça descrições de todos os recursos externos e quaisquer restrições associadas a eles, bem como links ou outros pontos de acesso, conforme apropriado. A base de dados é autocontida e não depende de recursos externos, como links para sites ou outros conjuntos de dados. Todos os dados necessários estão incluídos na base, sem vínculos ou dependências de terceiros. Portanto, não há restrições ou licenças adicionais associadas.

A base de dados contém dados que podem ser considerados confidenciais (por exemplo, dados protegidos por privilégio legal ou por confidencialidade médico-paciente, dados que incluem o conteúdo de comunicações não públicas de indivíduos)? Se sim, forneça uma descrição. Não, a base de dados não contém dados confidenciais, são todos de domínio público.

A base de dados contém dados que, se visualizados diretamente, podem ser ofensivos, ofensivos, ameaçadores ou causar ansiedade? Se sim, descreva o motivo. Não, a base de dados não contém conteúdo que possa ser considerado ofensivo, ameaçador ou causar ansiedade.

O conjunto de dados está relacionado a pessoas? (Se não, as perguntas restantes desta seção podem ser desconsideradas.) Não, o conjunto de dados não está diretamente relacionado a pessoas, mas sim a publicidades televisivas.

A.3 Processo de Coleta

Como os dados associados a cada instância foram adquiridos? (Os dados foram diretamente observáveis (por exemplo, texto bruto, classificações de filmes), relatados pelos sujeitos (por exemplo, respostas a questionários) ou inferidos/derivados indiretamente de outros dados (por exemplo, rótulos de partes do discurso, suposições modeladas sobre idade ou idioma)? Se os dados foram relatados pelos sujeitos ou inferidos/derivados indiretamente de outros dados, eles foram validados/verificados? Se sim, descreva como.) Os dados foram coletados através de gravações automáticas dos canais de TV aberta e fechada em horários específicos, definidos pelo protocolo da Rede INFORMAS. Posteriormente, essas gravações foram revisadas manualmente para a identificação de publicidades alimentícias.

Quais mecanismos ou procedimentos foram usados para coletar os dados (por exemplo, algum hardware ou sensor, curadoria manual humana, programa de software, API de software)? (Como esses mecanismos ou procedimentos foram validados?) Os dados foram coletados por meio de um processo de curadoria manual realizado por pesquisadores do projeto. Os vídeos de propagandas televisivas foram capturados a partir de canais de TV brasileiros. A coleta foi conduzida de maneira sistemática para garantir a integridade e a representatividade das publicidades. A validação foi feita

por meio de um processo de revisão pelos próprios pesquisadores, que verificaram a consistência das anotações e classificações das estratégias de marketing utilizadas.

Se o conjunto de dados é uma amostra de um conjunto maior, qual foi a estratégia de amostragem (por exemplo, determinística, probabilística com probabilidades de amostragem específicas)? O conjunto de dados é uma amostra de um universo maior de publicidades veiculadas na televisão brasileira. A estratégia de amostragem foi probabilística, com os dias de coleta sendo sorteados de forma aleatória, garantindo uma distribuição representativa dentro dos canais específicos e períodos considerados.

Quem esteve envolvido no processo de coleta de dados (por exemplo, estudantes, trabalhadores de *crowdsourcing*, contratados) e como foram compensados (por exemplo, quanto os trabalhadores de *crowdsourcing* foram pagos)? Pesquisadores e assistentes de pesquisa estiveram diretamente envolvidos no processo de coleta de dados. Esses indivíduos eram membros do projeto acadêmico e faziam parte de uma equipe de pesquisa em instituições acadêmicas. Alguns foram pagos diretamente pela coleta dos dados.

Durante qual período de tempo os dados foram coletados? (Esse período coincide com o período de criação dos dados associados às instâncias (por exemplo, captura recente de artigos antigos de notícias)? Caso contrário, descreva o período em que os dados associados às instâncias foram criados.) Os dados foram coletados ao longo dos anos de 2018, 2019 e 2020. Esse período coincide com o período em que as instâncias foram criadas e veiculadas, garantindo que os vídeos coletados refletem o período em análise.

Algum processo de revisão ética foi conduzido (por exemplo, por um comitê de ética institucional)? (Se sim, forneça uma descrição desses processos de revisão, incluindo os resultados, bem como um link ou outro ponto de acesso à documentação de suporte.) Nenhuma revisão ética foi conduzida.

O conjunto de dados está relacionado a pessoas? (Se não, as perguntas restantes desta seção podem ser desconsideradas.) O conjunto de dados não está diretamente relacionado a pessoas.

A.4 Pré-processamento/Limpeza/Rotulagem

Foi realizado algum pré-processamento/limpeza/rotulagem dos dados (por exemplo, discretização ou agrupamento, tokenização, rotulagem de partes do discurso, extração de *features*, remoção de instâncias, processamento de valores ausentes)? (Se sim, forneça uma descrição. Se não, as perguntas restantes desta seção podem ser desconsideradas.) Sim, foi realizado um extenso

pré-processamento dos dados. Os vídeos de publicidades não alimentícias foram cortados automaticamente e também tiveram suas duplicatas removidas. Esse processo foi feito manualmente para as publicidades alimentícias

Os dados "brutos" foram salvos, além dos dados pré-processados/limpos/rotulados (por exemplo, para suportar usos futuros não antecipados)? (Se sim, forneça um link ou outro ponto de acesso aos dados "brutos".) Sim, os dados "brutos" foram armazenados para garantir a possibilidade de futuras análises e reprocessamentos. Eles estão disponíveis para acesso interno, mas não são fornecidos publicamente. O acesso pode ser solicitado mediante autorização do grupo de pesquisa responsável.

O software usado para pré-processar/limpar/rotular as instâncias está disponível? (Se sim, forneça um link ou outro ponto de acesso.) Sim, o software usado para pré-processar os dados foi desenvolvido em Python e utiliza bibliotecas como `ffmpeg`, funções de `average hash`[1]. Ele está disponível em um repositório público, sob requisição. Interessados devem entrar em contato com o grupo de pesquisa para obter acesso ao código.

A.5 Usos

O conjunto de dados já foi usado para alguma tarefa? (Se sim, forneça uma descrição.) Sim, o conjunto de dados foi utilizado para realizar a presente dissertação.

Há um repositório que vincula a qualquer ou todos os artigos ou sistemas que utilizam o conjunto de dados? (Se sim, forneça um link ou outro ponto de acesso.) No momento, não há um repositório público que vincule artigos ou sistemas que utilizam o conjunto de dados.

Para quais (outras) tarefas o conjunto de dados poderia ser usado? O conjunto de dados poderia ser usado para uma ampla variedade de tarefas relacionadas à análise de publicidades. Além da classificação de tipo, ele pode ser utilizado para tarefas como detecção de estratégias de marketing, análise de comportamento do consumidor, identificação de padrões nas publicidades, ou até mesmo para estudos sobre a influência de propagandas na saúde pública.

Há algo sobre a composição do conjunto de dados ou a forma como foi coletado e pré-processado/limpo/rotulado que possa impactar usos futuros? (Por exemplo, há algo que um usuário futuro precise saber para evitar usos que possam resultar em tratamento injusto de indivíduos ou grupos (por exemplo, estereotipagem, problemas de qualidade de serviço) ou outros danos indesejáveis (por exemplo, danos financeiros, riscos legais)? Se sim, forneça uma descrição. Há algo que um usuário futuro poderia fazer para mitigar esses

[1] `https://github.com/gklc811/duplicate_video_finder`

danos indesejáveis?) Sim, uma consideração importante é o desequilíbrio de classes presente na base de dados, que pode impactar negativamente o desempenho dos modelos em classes minoritárias, por exemplo: a classe de *fast-food*, *supermercado*, *alimento/bebida* ou da classe alimentícia como um todo. Além disso, como o conjunto de dados contém publicidades de diferentes tipos de produtos, incluindo produtos alimentícios, seu uso em tarefas que envolvem decisões automatizadas relacionadas a políticas públicas ou recomendações comerciais pode exigir cuidados adicionais para evitar estereotipagem.

Existem tarefas para as quais o conjunto de dados não deve ser usado? (Se sim, forneça uma descrição.) Sim, o conjunto de dados não é adequado para tarefas que envolvem a análise de informações pessoais sensíveis. A base de dados foi coletada com o propósito de analisar padrões de publicidade e não contém dados pessoais diretos, mas seu uso indevido em contextos de monitoramento pessoal poderia levantar preocupações éticas.

A.6 Distribuição

O conjunto de dados será distribuído para terceiros fora da entidade (por exemplo, empresa, instituição, organização) em nome da qual o conjunto de dados foi criado? (Se sim, forneça uma descrição.) Sim, o conjunto de dados pode ser disponibilizado para terceiros, principalmente para fins acadêmicos ou de pesquisa.

Como o conjunto de dados será distribuído (por exemplo, em algum site, API, GitHub)? (O conjunto de dados possui um identificador de objeto digital (DOI)?) O conjunto de dados pode ser acessado através de um repositório GitHub `https://github.com/victoriapf/food-ads-monitoring`. Atualmente, o conjunto de dados não possui um identificador de objeto digital (DOI).

Quando o conjunto de dados será distribuído? O conjunto de dados será distribuído após a conclusão da pesquisa principal e a publicação dos resultados, o que está previsto para o próximo ano.

O conjunto de dados será distribuído sob um direito autoral ou outra licença de propriedade intelectual (PI) e/ou sob os termos de uso aplicáveis (ToU)? (Se sim, descreva essa licença e/ou ToU e forneça um link ou outro ponto de acesso, ou reproduza os termos de licenciamento relevantes ou ToU, bem como quaisquer taxas associadas a essas restrições.) O conjunto de dados será distribuído sob uma licença específica que respeita os direitos autorais e as diretrizes de uso. Ele estará disponível exclusivamente para fins acadêmicos e de pesquisa, sob os termos da Licença de Uso Não Comercial.

Terceiros impuseram restrições baseadas em PI ou outras restrições sobre os dados associados às instâncias? (Se sim, descreva essas restrições e forneça

um link ou outro ponto de acesso, ou reproduza os termos de licenciamento relevantes, bem como quaisquer taxas associadas a essas restrições.) Não, não há restrições impostas por terceiros em relação à propriedade intelectual ou a outros aspectos das instâncias presentes no conjunto de dados. Todos os dados foram coletados em conformidade com as leis de propriedade intelectual vigentes.

Algum controle de exportação ou outras restrições regulatórias se aplicam ao conjunto de dados ou a instâncias individuais? (Se sim, descreva essas restrições e forneça um link ou outro ponto de acesso à documentação de suporte.) Não, o conjunto de dados não está sujeito a controles de exportação ou outras restrições regulatórias específicas.

A.7 Manutenção

Quem está apoiando/hospedando/mantendo o conjunto de dados? O conjunto de dados está sendo mantido pelo grupo de pesquisa responsável por este estudo, associado às instituições de pesquisa que conduziram a coleta, análise e pré-processamento dos dados.

Como o proprietário/curador/gerente do conjunto de dados pode ser contatado (por exemplo, endereço de e-mail)? O responsável pelo conjunto de dados pode ser contatado através do endereço de e-mail: sandra@ic.unicamp.br.

Existe algum errata? (Se sim, forneça um link ou outro ponto de acesso.) Até o momento não foi identificada nenhuma errata.

O conjunto de dados será atualizado (por exemplo, para corrigir erros de rotulagem, adicionar novas instâncias, excluir instâncias)? (Se sim, descreva com que frequência, por quem, e como as atualizações serão comunicadas aos usuários (por exemplo, lista de e-mails, GitHub)?) Sim, o conjunto de dados poderá ser atualizado para corrigir possíveis erros de rotulagem ou adicionar novas instâncias. As atualizações serão conduzidas pelo mesmo grupo de pesquisa responsável por sua criação. Os usuários serão notificados sobre atualizações através de uma lista de e-mails ou via repositório, onde essas mudanças estarão documentadas.

Se o conjunto de dados está relacionado a pessoas, existem limites aplicáveis à retenção dos dados associados às instâncias (por exemplo, as pessoas em questão foram informadas de que seus dados seriam retidos por um período fixo de tempo e depois excluídos)? (Se sim, descreva esses limites e explique como eles serão aplicados.) O conjunto de dados não contém informações que identifiquem diretamente indivíduos. Portanto, não há limites específicos aplicáveis à retenção dos dados.

Versões mais antigas do conjunto de dados continuarão a ser suportadas / hospedadas / mantidas? (Se sim, descreva como. Caso contrário, descreva como sua obsolescência será comunicada aos usuários.) Versões anteriores do conjunto de dados não serão mantidas no repositório institucional.

Se outros quiserem estender/aumentar/construir/contribuir com o conjunto de dados, existe um mecanismo para que isso seja feito? (Se sim, forneça uma descrição. Essas contribuições serão validadas/verificadas? Se sim, descreva como. Caso contrário, por que não? Existe um processo para comunicar/distribuir essas contribuições para outros usuários? Se sim, forneça uma descrição.) Sim, o conjunto de dados poderá aceitar contribuições de outros pesquisadores ou colaboradores, contanto que sigam o protocolo INFORMAS. Essas contribuições serão analisadas e verificadas pela equipe responsável antes de serem integradas ao conjunto principal. Uma vez validadas, as atualizações ou expansões serão comunicadas.