UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Engenharia Elétrica e de Computação

Julia Amaya Tavares

Deep-Learning Segmentation for Skin Lesion Images

Aprendizado Profundo em Segmentação para Imagens de Lesão de Pele

Campinas

2018

Julia Amaya Tavares

Deep-Learning Segmentation for Skin Lesion Images

Aprendizado Profundo em Segmentação para Imagens de Lesão de Pele

Dissertation presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Master, in the area of Computer Engineering

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Engenharia Elétrica, na área de Engenharia da Computação

Supervisor: Prof. Dr. Eduardo Alves do Valle Jr

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA JULIA AMAYA TAVARES E ORIENTADA PELO PROF. DR. EDUARDO ALVES DO VALLE JR

Campinas

2018

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

T197d

Amaya Tavares, Julia, 1991-
Deep-learning segmentation for skin lesion images / Julia Amaya Tavares. – Campinas, SP : [s.n.], 2018.

Orientador: Eduardo Alves do Valle Júnior.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Melanoma. 2. Segmentação de imagens. 3. Segmentação de imagens médicas. 4. Redes neurais (Computação). I. Valle Júnior, Eduardo Alves do. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Segmentação de imagens de lesão de pele com aprendizado profundo
Palavras-chave em inglês:
Melanoma
Image segmentation
Image segmentation of medical images
Neural networks (Computer Science)
Área de concentração: Engenharia de Computação
Titulação: Mestra em Engenharia Elétrica
Banca examinadora:
Eduardo Alves do Valle Júnior [Orientador]
Leticia Rittner
Gerberth Adín Ramírez Rivera
Data de defesa: 13-06-2018
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidata: Julia Amaya Tavares **RA:** 091741
Data da Defesa: 13 de junho de 2018
Título da Tese: "Aprendizado Profundo em Segmentação para Imagens de Lesão de Pele"

Prof. Dr. Eduardo Alves do Valle Jr (Presidente)
Prof. Dr. Gerberth Adín Ramírez Rivera
Prof. Dr. Leticia Rittner

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de PósGraduação da Faculdade de Engenharia Elétrica e de Computação.

Abstract

Skin cancer has a high cure rate during its early stages, but can become fatal if left untreated. A skin lesion may range from a benign type of lesion such as nevus and actinic keratosis, to a cancerous type such as carcinoma or melanoma. The latter is the most aggressive type of skin cancer, and causes the most deaths.

In the context of medical image analysis, obtaining the location of the region of interest is an important step for physicians, which assists, for instance, in the preparation of medical records and the monitoring of the progression of diseases. In those cases, the use of automatic tools can be very useful. As in other medical domains, the location of skin lesions in the images is also relevant because aspects such as irregular growth of the lesion and the presence of some tumor markers may occur in the peripheral regions or in its borders.

In recent years, many new solutions have been proposed to tackle the task of image segmentation, leveraging the use of deep neural networks. Most of those solutions focus on the more general task of segmenting objects of different types. The use of these approaches are relatively new in the skin lesion domain.

In this work, we provide an overview of this literature, explore a subset of the main ideas with a series of experiments on a dataset of dermoscopic images and propose an open-sourced implementation that achieves a promising Jaccard score of 0.754.

Keywords: Deep Learning, Skin lesion, Melanoma, Dermoscopy, Automated Screening, Image Segmentation, Neural Networks

Resumo

Câncer de pele tem um índice alto de cura durante os seus primeiros estágios, mas pode se tornar fatal se não for tratado. Uma lesão de pele pode variar desde tipos benignos como nevos e queratose actínica, até tipos malignos como carcinoma ou melanoma. Este último é o tipo mais agressivo de câncer de pele, e o que causa o maior número de mortes.

No contexto de análise de imagens médicas, a localização da região de interesse é um passo importante para os médicos, que auxilia, por exemplo, a elaboração de laudos e acompanhamento da progressão de lesões, e com isso a utilização de ferramentas automático pode ser muito útil. Assim como em outros domínios médicos, a localização de lesões de pele nas imagens também é relevante pois aspectos como crescimento irregular da lesão e a presença de alguns marcadores típicos de sua malignidade podem ocorrer nas regiões periféricas ou nas suas bordas.

Nos últimos anos, muitas novas soluções foram propostas para a tarefa de segmentação de imagem, potencializando o uso de redes neurais profundas. Muitas destas soluções focam na tarefa mais geral de segmentar objetos de diferentes tipos. O uso dessas abordagens é relativamente novo no domínio de lesões de pele.

Neste trabalho, apresentamos uma visão geral dessa literatura, exploramos um subgrupo das principais ideias em uma série de experimentos em um dataset de imagens dermoscópicas, e propomos uma implementação open-source que atinge um resultado promissor, com coeficiente Jaccard de 0.754.

Palavras-chave: Aprendizado Profundo; Lesão de pele; Melanoma; Dermoscopia; Triagem Automática; Segmentação de Imagens; Redes neurais.

Table of Contents

1 Introduction

Melanoma is the most dangerous form of skin cancer — an uncontrolled growth of abnormal skin cells. It results from the cumulative damage to the cells' DNA, mainly from ultraviolet radiation. If treated in its early stages, it is highly curable with non-invasive surgeries; after it has metastasized (spread to other parts of the body) treatment is extremely difficult. To better illustrate the importance of early diagnosis, consider that the 5-year survival rate is 98% for the localized stage, 62% for regional stage, and 18% for distant metastasis-stage (AMERICAN CANCER SOCIETY, 2017).

However, determining if a skin lesion may be cancerous is not trivial to the untrained eye, and requires special training. Therefore, automatic tools to assist the decision-making process of which patients have a higher risk case than others can be very helpful, especially in areas where a specialist doctor may not be readily available for those initial consultations. Some of the information extraction that has the potential to be automated and help health professionals make an informed decision — such as lesion size, shape and progression over time — can be accomplished through the use of image segmentation.

Image segmentation consists of splitting an image into meaningful subparts. Segmentation may identify common objects, such as cars, people, dogs or chairs — or, in medical imaging, highly specialized structures such as cells, organs, lesions, etc. It is a difficult task, that besides recognizing the object, also requires determining the object's boundaries.

In the classical approach to automated segmentation, features were typically handcrafted, with the use of multiple low-level pixel processing and mathematical models, as well as complex rule-based systems. We present an overview of those ideas in Section 2.3. In recent years, supervised techniques have become the standard, and segmentation is typically learned from a much larger number of labeled samples.

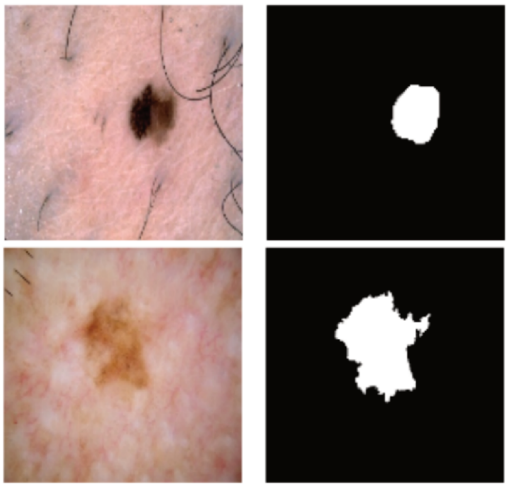
Although there are general computer vision techniques for image segmentation, they often need to be adapted for medical images, which usually contain images with a different structure (oftentimes 3D or grayscale) and are obtained from specialized equipment. For many applications, manual segmentations need to be obtained from a clinical expert, as a non-trained person cannot accurately identify the parts of interest. Computer vision techniques try to reproduce this manual expert segmentation to new, unseen images.

1.1 Motivation

In the context of medical images, segmentation is an area of interest that can help achieve a number of different goals. One possible application is the evaluation of temporal disease progression on a given patient, for instance, if a lesion is growing in size or changing shape, as this information can be important to make decisions regarding the next steps for the patient's treatment.

Other possible applications include detection of morphology, volume estimation, visualization in medical equipment for surgical planning or administering treatment. The segmentation can provide invaluable information that can help quantify tissue volumes, localize pathologies, identify anatomical structures and assist in computer-integrated surgery.

As an effort to improve the automatic understanding of medical images, this work focuses specifically on the segmentation task of skin lesion images, using a dataset of images that have undergone annotation by skin cancer experts (Fig. 1), to segment background normal skin and other structures from the lesions (cancerous or not).



(a) Original images (b) Ground-truth masks

Figure 1 – Skin lesion segmentation[1] (resized)

1.2 Objectives

Our main goal is to explore a modern approach for automated skin lesion segmentation and investigate image segmentation techniques that can benefit that and other tasks dealing with small and specialized datasets. The primary objectives of this work are:

[1] Source: `https://challenge.kitware.com/#challenge/583f126bcad3a51cc66c8d9a`

- Provide a critical review of existing approaches to the semantic segmentation task present in the literature, specifically the more modern deep-learning approaches.
- Improve the accuracy of automated skin lesion segmentation.
- Provide an open-source implementation of a working model for this task, reducing the barrier of entry for others who are interested.

A secondary objective is to enable the investigation of possible benefits of using the predictions of a segmentation network to improve the performance of the classification task based on types of diagnoses of skin lesions, such as melanoma, nevus, and seborrheic keratosis. These can be treated as binary classification tasks, for example, classifying images as melanoma or not melanoma, or alternatively, as keratosis or not keratosis.

1.3 Contributions and Achievements

- Co-author of "RECOD Titans at ISIC Challenge 2017" (MENEGOLA et al., 2017), an extended abstract that describes our participation in the ISIC Challenge 2017: Skin Lesion Analysis Towards Melanoma Detection (IEEE International Symposium on Biomedical Imaging), both in the classification task and in the lesion segmentation tasks.

  In this challenge, the leaderboard and results are independent for each task and teams are not required to participate in all tasks. There is also another task, "lesion dermoscopic feature extraction", for which our team did not make any submissions.

  Our team placed 3rd in the overall classification task (1st in the melanoma classification results) out of 23 submissions, and 5th in the segmentation task out of 21 submissions.

  Although our team already had a long history of experience with melanoma classification, this challenge was the very first time we worked on skin-lesion segmentation.

- Co-author of "Data, Depth, and Design: Learning Reliable Models for Melanoma Screening" (VALLE et al., 2017) (currently under review).

  The main goal of this paper is to investigate methodological issues for designing deep learning models for melanoma classification, exploring the influence of factors such as transfer learning, architecture, image resolution, types of data augmentation and use of segmentation. Two full factorial experiment were performed for five different test datasets for a total of 2560 exhaustive trials in the main experiment, which is analyzed with multi-way ANOVA.

Contributed by training a segmentation network in two datasets and generating predicted images for previously unseen datasets, with the objective of evaluating the possibility that this additional information would improve the classification performance. Also contributed code to aggregate the segmentation mask information to the original images, for instance, cropping the original image based on the corresponding mask or overlaying the mask and changing the background color.

In the context that was tested in the corresponding classification task, which was done by adding the segmentation mask as a fourth input channel (alongside the usual RGB channels), we found that there were no performance gains for the classification task, in fact, the overall score decreased. Future work includes evaluating the use of segmentation masks for cropping the original images and using the resulting RGB images as input, as this has been previously found to have a positive effect.

The specifics of possible approaches to leverage the skin lesion segmentation information to improve classification performance are beyond the scope of this work, and we refer the reader to the previously mentioned paper and to Codella et al. (2015).

- Open-source implementation was made available on a public **github repository**[2].

  This repository contains all of the working code that is needed to reproduce the models used in the segmentation part of the two previously mentioned papers. At the time the code was published, it was — as far as we know —, the only available open implementation of a segmentation network for skin lesion images.

  The code also contains all of the pre-processing steps needed to re-train the model, as well as online data augmentation. This training pipeline is able to reach good results using only publicly available data, and can be trained from scratch in a few hours using a consumer-level build (with 8 GiB RAM and a single 8 GiB GPU).

- Second author of "Adversarial Images for Variational Autoencoders" (TABACOF et al., 2016), presented at NIPS 2016 Workshop on Adversarial Training as a spotlight presentation.

  Contributed to the research and coding of the experiments, which is available at the first authors' **github repository**[3]. While this is not directly related to this dissertation, autoencoder networks have many similarities to segmentation networks, and a summary is given in the appendix B.1.

[2] Available at: https://github.com/juliafeec/isic-2017-segmentation
[3] Available at: https://github.com/tabacof/adv_vae

2 Literature Review

We briefly expose in Section 2.1 the present state in Deep Neural Networks research and in Section 2.2 we exemplify their usefulness for medical imaging. We summarize for the sake of completeness the goals and approaches to the image segmentation task over the years in Section 2.3.

Section 2.4 contains the main part of this literature review, where we delve into details some of the most relevant contributions to the area of image segmentation neural networks and recent state-of-the-art methods. In Section 2.5 we expose the current state-of-the-art model in the more specific case of image segmentation of skin lesion images.

2.1 Deep neural networks

Since the introduction of AlexNet (KRIZHEVSKY et al., 2012) — which set a new bar for results in the ImageNet Challenge (DENG et al., 2009) — deep neural networks became the state of the art in image classification.

Since then, new applications of deep learning have been introduced and refined, leveraging the extraction of relevant features in structured and unstructured data, without the need for hand-crafted or rule-based feature extractors. Examples are speech recognition (HANNUN et al., 2014); biometric identification and verification by facial images (SCHROFF et al., 2015) and speech audio (LEI et al., 2014); strategy games and video games (SILVER et al., 2016), (OH et al., 2015) and object detection such as R-CNN (GIRSHICK et al., 2014), Fast-R-CNN (GIRSHICK, 2015), Faster-R-CNN (REN et al., 2015), and YOLO (REDMON et al., 2016).

Innovations have appeared in state-of-the-art networks in fast pace: ReLU activations (GLOROT et al., 2011), batch-normalization layers (IOFFE; SZEGEDY, 2015), dropout layers (HINTON et al., 2012), Adam optimizer (KINGMA; BA, 2014), and residual learning (HE et al., 2016). Transfer learning (HINTON, 2007), (YOSINSKI et al., 2014) has also proved to be very useful, especially in tasks with limited training data.

Many implementation deep learning frameworks surfaced: Theano (BERGSTRA et al., 2010), Torch7 (COLLOBERT et al., 2011), Caffe (JIA et al., 2014), MXNet (CHEN et al., 2015), TensorFlow (ABADI et al., 2016), significantly reducing the amount of code needed to define and train networks. Those frameworks have worked in synergy to the GP-GPU revolution (General Processing in Graphical Processing Units) allowing to exploit the

massively parallel capacity GPUs without requiring the technical skills of writing GPU code by hand. The use of frameworks and GPUs has enabled training in days or weeks for networks that formerly would have required months or years to converge.

2.2 Medical Imaging

One particular area of interest is the analysis of medical data, and automatic systems capable of extracting information from various forms of data could assist doctors, allowing for faster and more precise diagnostics, especially in disadvantaged areas where the patients/doctor ratio is much higher.

Due to the sensitive nature of the data and the fact that annotation requires medical experts' opinions or diagnostics, the availability of datasets of medical images is much more limited than of datasets of general images, which can also be annotated, for example, by crowd-sourcing.

There is an increasing number of areas where the use of neural networks has proven helpful to work with medical imaging, such as retinopathy, glaucoma, and tumors, as described in the works of Shi et al. (2009) and Litjens et al. (2017).

2.3 Image Segmentation

Image segmentation is the task of partitioning an image into meaningful subregions. Many different algorithms have been developed for this task, and their goals may also differ. In some cases, the goal of segmentation is to obtain a simpler representation that is easier to analyze, that is, segmentation is used as an intermediate step in a computer vision pipeline (FORSYTH; PONCE, 2002). The end goal may be, for instance, to classify the image.

Over the years, different types of segmentation techniques have emerged. Some of the classical approaches include:

- Edge detection or boundary techniques
- Thresholding
- Region-based techniques
- Normalized cut—a graph theoretic approach
- Morphological Watershedding

Edge-based segmentation (SUMENGEN; MANJUNATH, 2005), (TABB; AHUJA, 1997) uses edge detection methods — such as classical edge detectors, color edge detectors and zero crossing — to define closed region boundaries.

Threshold-based segmentation (PAL; PAL, 1993) (BRINK, 1995) is based on a threshold (or multiple level thresholds) that turn the input image into a binary image, based on whether a pixel exceeds a given threshold or not.

Region-based segmentation (GONZALEZ; WOODS, 2012) (HOJJATOLESLAMI; KITTLER, 1998) uses a comparison between a pixel and its neighbors and if a similarity criterion — which may be based on color and intensity — is satisfied, it is considered to be part of a cluster.

In the normalized cut graph method (SHI; MALIK, 2000),(YANG et al., 2007),(COMANICIU; MEER, 2002), the image segmentation is treated as a graph partitioning problem. The normalized cut criterion is based on two metrics: the total similarity within groups — which should be minimized — and the total dissimilarity between the different groups — which should be maximized.

Morphological Watershedding (GONZALEZ; WOODS, 2012) (SERRA, 1983) segments the image by splitting it on the basis of image topology. The algorithm's name refers to a geological watershed — that is, a water drainage divide. In this method, an image is treated like a topographic map, with the pixel value corresponding to an elevation. Different approaches and criteria may be employed to segment the image using this principle, such as growing, merging and saturation.

Most of those classical approaches require some pre-existing knowledge regarding the specific type of image data that is being analyzed, for instance, which parameters should be operated upon in terms or the underlying intensity distribution.

The non-classical approaches, which are based on soft computing methods, can typically be applied without this a priori domain-specific knowledge.

- Fuzzy Sets and Fuzzy Logic
- Genetic Algorithms
- Neural Networks

Fuzzy set theory and fuzzy logic (ROSS et al., 2002) (DUBOIS; PRADE, 1982) are very common in the field of image processing and image classification. One of the most popular methods is the fuzzy c-means (FCM) clustering algorithm (BEZDEK, 1981), which splits a

group of pixels into c fuzzy groups and finds a cluster center in each group by minimizing a cost function of dissimilarity. We refer the reader to the works of Bhattacharyya (2011) and Cheng et al. (2001) for a comprehensive survey of image segmentation approaches using fuzzy logic.

Genetic algorithms (GOLDBERG; HOLLAND, 1988) are commonly used to find solutions to optimization and search problems where it is difficult to obtain the global optimum, by relying on operations such as selection, crossover and mutation. They can be effective in finding solutions in large, complex search spaces, and as the name implies, they were originally inspired in genetics and evolutionary processes. Bhanu et al. (1995) and Alander (2000) exemplify the use of GAs in image segmentation.

Some other examples of the aforementioned techniques specifically in the medical imaging field are fuzzy-based (UDUPA; SAMARASEKERA, 1996), graph-based (FELZENSZWALB; HUTTENLOCHER, 2004), and atlas-based algorithms (PHAM et al., 2000).

For a more complete background on these techniques, we refer the reader to Gonzalez and Woods (2012) and De et al. (2016). Neural networks approaches will be covered in Section 2.4.

In this work, we will focus on the semantic segmentation, the task of obtaining pixel-wise segmentations indicating the object class present at each pixel, as shown in Fig. 2a. It is more specific than object detection, Fig. 2b, which often only requires indicating the rectangle around the objects. We will focus on the deep-learning based approaches developed on the past few years.

Some of the most commonly used large-scale datasets are Pascal VOC (EVERINGHAM et al., 2010) and Microsoft COCO (LIN et al., 2014). The Pascal VOC dataset was created as the backbone of the PASCAL VOC challenge, a yearly competition that was established in 2005 and set the precedent for standardized evaluation of recognition algorithms. The ImageNet Challenge (RUSSAKOVSKY et al., 2015), which began in 2010 and enabled breakthroughs computer vision with its large-scale dataset, also followed the same structure containing a public dataset and annual competition.

(a) Segmentation (b) Object detection

Figure 2 – Pascal VOC annotation samples

The Microsoft COCO dataset was released in 2015, with the goal of advancing the state-of-the-art in object recognition. The large-scale dataset contains everyday scenes with common objects, including multiple instances where those objects are in the background or partially occluded, which aims to reflect the composition of actual real-word scenes.

Figure 3 – MS COCO annotation samples

2.4 Segmentation networks

One of the first works to successfully introduce a deep learning approach to the segmentation problem was proposed by Ciresan et al. (2012), who used a patch around each pixel as the input to a neural network in order to separately classify each pixel.

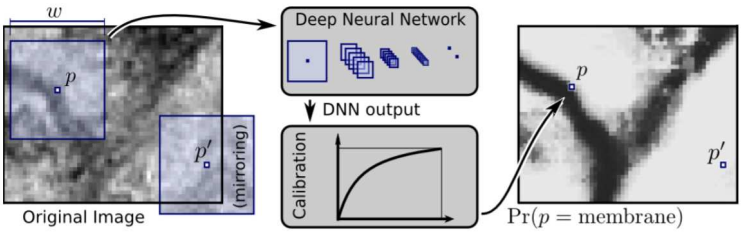
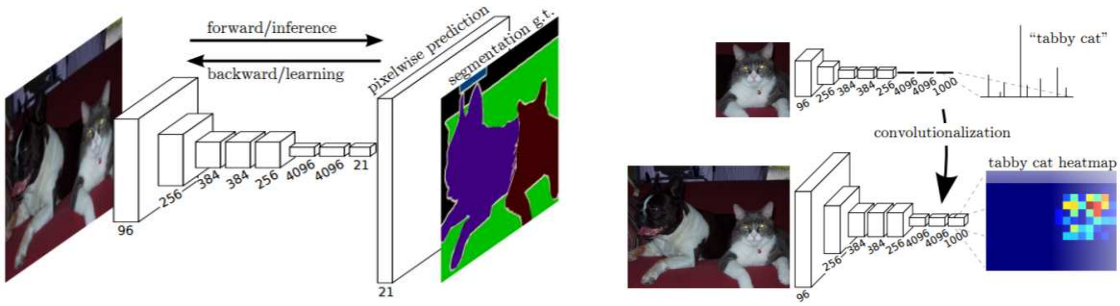


Figure 4 – DNN for electron microscopy images, (CIRESAN et al., 2012)

With the goal of segmenting (biological) neuron membranes in images, a deep neural network was used to classify each pixel as membrane or non-membrane and generate a binary mask. That approach essentially worked as a classification network, with a succession of convolutional and max-pooling layers and a single output with the probability of a membrane label. The input was a square window of raw pixel values, centered around the pixel that was being classified.

A major breakthrough in the area was the introduction of fully convolutional networks (LONG et al., 2014), which built upon a classification network. It did not use any fully

connected layers, replacing them with convolutional layers (Fig. 5a) instead.



(a) Network (b) Heat-map example

Figure 5 – FCN (LONG et al., 2014)

This allowed what was previously a classification network to output a heat-map (Fig. 5b), which was then upsampled with fractionally strided convolutions (Fig. 6). These are also sometimes referred to as transposed convolutions or deconvolutions, and have since become widely used in multiple architectures for segmentation. This approach is much faster than doing segmentations patch by patch, and it allowed the CNN training to be done end-to-end.

A transposed convolution is not the actual mathematical inverse operation of a convolution, and for this reason it is preferable to avoid naming it a deconvolution layer. But the goal of the transposed convolution is to revert the convolution operation in terms of spatial resolution, without the need of a separate upscaling step.

The transposed convolution layer operation is similar to a regular convolution but with some additional spacing in the input, allowing the output of this operation to restore the original dimensions. Fig 6 shows a transposed 2D convolution with kernel of 3, stride of 2 and no padding, resulting in a 5×5 output after performing the operation on a 3×3 input.

Another widely used building block for segmentation architectures is the dilated convolution, which is also referred to as atrous convolution. This operation was proposed by Yu and Koltun (2015), with the objective of aggregating contextual information without the loss of resolution, by expanding the receptive field of kernels. Using a larger kernel size would be an alternative way of obtaining a larger context, but dilated convolutions are able to achieve this without increasing the number of required parameters.

Dilated convolutions require an additional parameter, the dilation rate. It determines the spacing used by the kernel, and a regular 2D convolution is equivalent to a dilated convolution with a dilation rate of one.

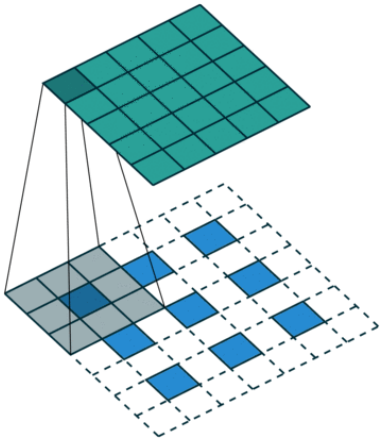


Figure 6 – Fractionally strided convolution[1]

In Fig. 7, a 3×3 kernel with a dilation rate of 2 is used. This results in a field of view as large as that of a 5×5 kernel in a 2D convolution, but only using 9 parameters. The output is equivalent to using a 5×5 kernel and removing the second and fourth columns and rows, resulting in a wide field of view without the need of larger kernels or more convolutions.

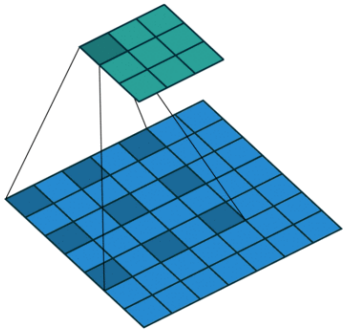


Figure 7 – Dilated convolution[1]

Ronneberger et al. (2015) proposed U-net (Fig. 8), a convolutional network intended for the segmentation of biomedical images.

The U-shaped network has an encoder–decoder structure. Its first half is a contracting path, which captures context, and its second half is an expanding path, symmetric to the first, aiming at precise localization. Such structure has many similarities to autoencoder networks B.1. However, while autoencoders output a reconstruction of the input, U-nets output the segmentation mask (with two or more object classes).

[1] Source: `http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html`

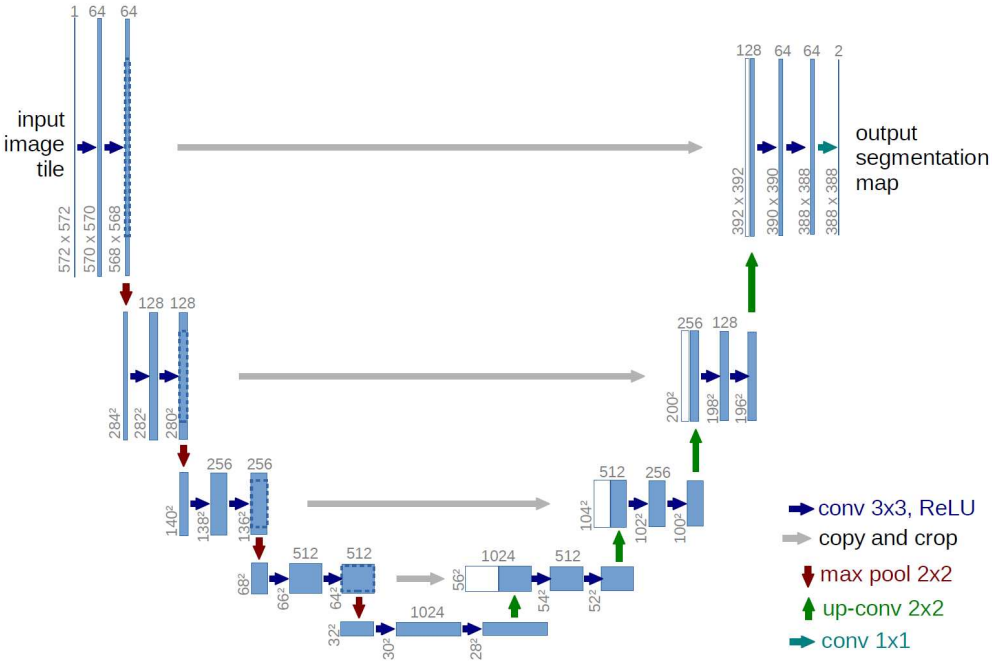


Figure 8 – U-net (RONNEBERGER et al., 2015)

Those networks are trained end-to-end, by back-propagation. They add skip-connections that allow additional information to be transferred from the encoding half to the decoding half directly.

Another useful characteristic of those architectures is their fully-convolutional design, which allows inputs of any size. That also allows relatively small model sizes, which are simpler to train. That is especially interesting in the context of biomedical images due to the limited training data. U-nets are designed to perform well with small datasets helped with data augmentation.

DeepLab (CHEN et al., 2016) further increased the state-of-the-art performance, using parallel atrous convolutions (Figure 9) and, as a post-processing step, conditional random fields (CRFs) (KRÄHENBÜHL; KOLTUN, 2011) as shown in Figure 10a.

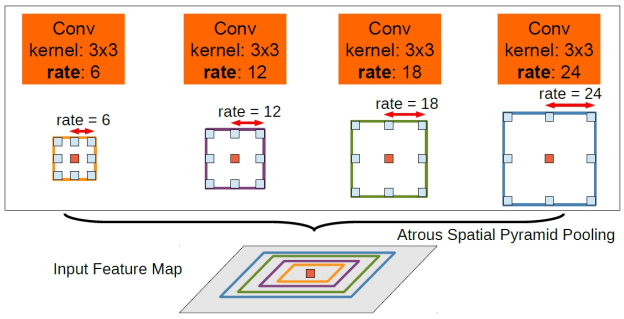
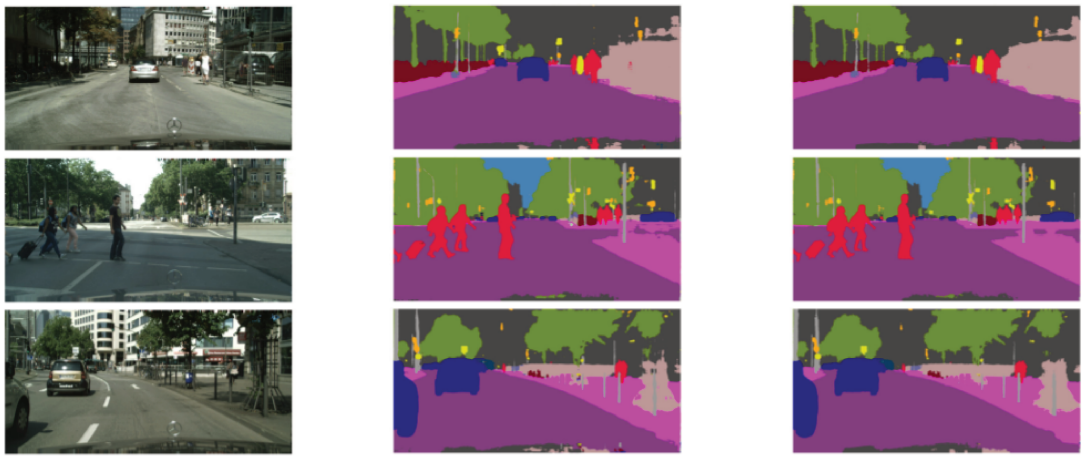


Figure 9 – DeepLab Parallel Atrous Convolutions[2]



(a) Input image (b) Before CRF (c) After CRF

Figure 10 – DeepLab CRF Post-processing (CHEN et al., 2016)

The parallel atrous convolution block is shown in Figure 9. Considering the orange central pixel, multiple atrous convolutions are applied, using the same input but with different dilation rates. As a result, multiple field-of-views are applied, which helps capture multi-scale features.

Large Kernel Matters (PENG et al., 2017) also focused on the issue of capturing contextual information, and achieved state-of-the-art results at the time. In current classification architectures such as the ones proposed by Simonyan and Zisserman (2014) and He et al. (2016), filters with small kernels (such as 3×3) are stacked, because this is more efficient than fewer large kernels, given the same complexity. However, in order to generate pixel predictions, large kernels and their corresponding large receptive field are important to capture context information. Additionally, the requirements of classification and localization problems are different in the following sense: for the classification task, the model should be invariant to transformation such as shifts, rotations or re-scalings. But for the localization task, the model needs to be sensitive to the positions of the input. The paper has two main contributions: a Global Convolutional Network to address both classification and localization tasks simultaneously, and a Boundary Refinement block that improves the localization performance near the object boundaries. This BR block has a residual structure, as shown in Fig. 11, and is integrated into the network and trained end-to-end.

[2] Image from `http://liangchiehchen.com/projects/DeepLab.html`

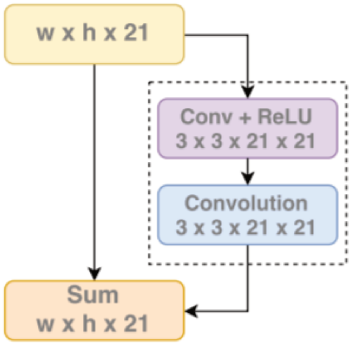


Figure 11 – The Boundary Refinement (BR) block, containing a residual structure, is trained end-to-end along with the network and is used for performance improvement on objects' boundaries.

Large kernels contain many parameters and are computationally expensive, and the key idea behind the GCN module is to provide an approximation that has fewer parameters and less complexity. A k×k convolution is approximated with the sum of a sequence of 1×k and k×1 convolutions, as shown in Fig. 12.

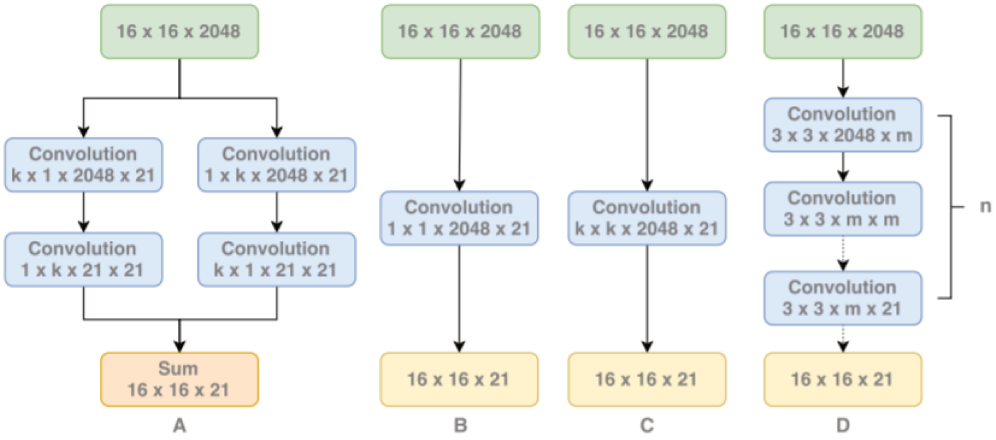


Figure 12 – (A) Global Convolutional Network (B) 1×1 convolution (C) k×k convolution (D) stacked 3×3 convolutions.
The GCN block (A) is used to approximate the k×k convolution (C) with a sum of stacked convolutions. The advantage of this approximation is that it is less complex and has fewer parameters.

Many other successful architectures leveraging CNNs have been recently proposed, such as SegNet (BADRINARAYANAN et al., 2017), RefineNet (LIN et al., 2017) and PSPNet (ZHAO et al., 2017). At the time of writing, the DeepLab v3 network (CHEN et al., 2017)

reaches state-of-the-art segmentation results[3] in the Pascal VOC 2012 dataset. As the title suggests, the paper focuses on atrous convolution, proposing improvements to the atrous module and employing it in cascade or in parallel to capture multi-scale context. It also no longer uses CRF post-processing.

Another very recent paper that is worth mentioning is SegCaps (LALONDE; BAGCI, 2018), the first work in literature to tackle the task of object segmentation using Capsule Networks (SABOUR et al., 2017).

Capsule Networks are motivated by a drawback in the widely used Convolutional Neural Networks: in a CNN's internal data representation, important hierarchies between simple and complex objects are not explicitly taken into account, such as how parts of an object are spatially oriented relative to each other. In Capsule Networks, each capsule is a group of neurons whose activities represent properties of an entity. The entity may be an object or part of an object that is present in the image, and its properties may represent different parameters, such as position, orientation, texture, velocity, etc. The method that is proposed by the authors to achieve this is an iterative routing-by-agreement mechanism, where a lower-level capsule decides where to send its output by choosing the higher level capsules that "agree" with it the most (measured by a scalar product between the lower-level capsule's own prediction and the higher-level capsules' activity vectors). One of the interesting benefits of this architecture that is reported in the paper is that it is considerably better at recognizing highly overlapping digits in the MNIST dataset than a regular convolutional network.

SegCaps extends the idea of convolutional capsules and proposes the concept of deconvolutional capsules. A modification to the original dynamic routing algorithm is also proposed, in which transformation matrices are shared within a capsule type and children are only routed to parents within a defined spatially-local window. These modifications are helpful to train the network in larger images sizes. The authors compared results in the pathological lung segmentation task, and found that SegCaps provided a better segmentation metric (98.479% dice coefficient) than the U-Net network (98.449%), while simultaneously reducing the number of parameters of architecture by 95.4%.

2.5 Skin Lesion Segmentation

In this section we focus on the specific case of skin lesion images. For a more general overview of medical images segmentation applications, we refer the reader to Litjens et al. (2017).

[3] `http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6`

Until recently, it was common to use a multi-step approach, such as hair removal, contrast increase, saliency mapping and post-processing. And at times, a semi-automatic approach would require an additional step performed by a person. We refer the reader to Maglogiannis and Doukas (2009), Celebi et al. (2015), and Ma, Tavares, et al. (2015) for a comprehensive survey of state-of-the-art methods, many of which are based on the methods we briefly exposed in Section 2.2. Pathan et al. (2018) also reviews commonly used methods for dermoscopic images segmentation, and exemplifies some of their evaluation metrics results (although they should not be directly compared, as the training datasets differ).

In the past few years, deep-neural-network-based solutions have become more prevalent in the skin lesion segmentation task, and they usually encompass much fewer (if any) pre-processing and post-processing steps.

The ISBI Challenge (CODELLA et al., 2017) has the goal of helping participants develop image analysis tools to enable the automated diagnosis of melanoma from dermoscopic images. The challenge has three parts, the first of which is lesion segmentation.

This challenge provides training data (though the use of additional external data is allowed), a separate public validation dataset and a blind held-out test dataset, and therefore enables a direct comparison of performance between different models.

Codella et al. (2016) achieved state-of-the-art performance at the time in the melanoma classification task. In their solution, the segmentation masks were also used as features for the classification network. They found that using the ground truth segmentation masks showed a higher but very similar performance to automatically generated segmentation masks (0.649 *vs.* 0.645). The authors considered that a possible cause would be the automated segmentation performance falling in the range of human performance and variability. Their segmentation network is largely based on the U-net network, with a few changes. Besides the usual RGB channels, the authors also used the HSV representation, for a total of 6 channels for the images. Additionally, Gaussian noise was introduced in some of the initial layers, and a 8192-dimensional fully connected layer.

Another interesting experiment from this work provides a measure of agreement between human experts. Three clinical experts were asked to generate ground truth segmentations for a subset of 100 lesion images. Then, the Jaccard metric was calculated for the 3 pairs of experts, resulting in measurements of 0.743, 0.754, and 0.861, and their average of 0.786.

Due to the dataset changes, the performance obtained in different challenges cannot be directly compared. At the time of writing, the last challenge was in March 2017 and the best score in the lesion segmentation task reached a Jaccard Index of 0.765, with the

method described in Yuan et al. (2017) and Yuan and Lo (2017). The authors used deep fully convolutional-deconvolutional neural networks, with 29 layers and about 5M trainable parameters.

The network used ReLU activations for each convolutional or deconvolutional layer, and the sigmoid as the activation function in the last layer. Batch normalization was added to the output of every convolutional or deconvolutional layer. And finally, the loss used was based on the Jaccard distance.

An additional post-processing step with dual-thresholds is used as well. A threshold of 0.8 is applied to determine the lesion center, which is calculated as the centroid of the region that has the largest mass after applying the threshold. Afterwards, a lower threshold of 0.5 is applied to the output map. In sequence, small holes are filled with morphological dilation and the final lesion mask is determined as the region that embraces the lesion center. An ensemble strategy is implemented to combine outputs of 6 similar networks.

The work of Berseth (2017) achieved the second place. Images were resized to 192×192 pixels and a small U-net was used. The network contained only three down-sampling layers, a fully connected layer, and two up-sampling layers, along with some Dropout layers. A lot of the effort was focused on the preprocessing step, in order to generate additional images. All training images were elastically distorted in four ways. Additionally, each training image was also rotated 90 degrees and additional elastic distortions were applied, before resizing. With this pipeline, the total number of training images was brought up to 20,000, and an additional set of transformations were done online in the batches, such as flipping, rotation and zoom.

The third and fourth place were submissions by the same team (BI et al., 2017). Their network was based on the FCN architecture, based on a ResNet. Aside from the challenge's 2,000 training images, the team used additional 8,000 images annotated by experts from the international skin imaging collaboration (ISIC) archive.

Images were downscaled to 500 pixels. The ResNet model was pretrained on ImageNet dataset, and then fine-tuned with the lesion images. Data augmentation such as random crops and flips were also used. Test images were augmented by resizing and flipping left-right, upside-down and both flips at the same time. Once those masks were unflipped, the final output was produced by integrating the outputs.

Li and Shen (2018) proposed a network that simultaneously performs lesion segmentation and classification, with a framework that consists of a multi-scale fully-convolutional residual network and a lesion index calculation unit (LICU). Their reported Jaccard score is 0.718 in the ISIC 2017 test set (equivalent to the ninth position in the segmentation rank).

Mishra and Daescu (2017) also proposed a U-net based network and compared the results to the Otsu segmentation method. For the post-processing step, it is assumed that there is a single lesion per image, and its center is defined to be the largest connected component. Afterwards, a hole-filling morphological operation is applied to complete that region as a single structure.

Codella et al. (2018) provides a comparison between two architectures: U-Net and Dense Residual U-Net. Dense Residual U-Net makes use of ideas from classification networks such as the ones proposed by He et al. (2015) and Huang et al. (2017) and have the following changes relative to regular U-Nets: residual layers are added to the end of the U-Net (in order to model contextual awareness) and convolutions are replaced with dense convolutions (in order to increase the reuse of features).

For pre-training networks, a larger set of images (over 2500) from public image segmentation datasets for healthy skin is used. A smaller specialized dataset of 400 images contains images with various pathological states of diseased skin.

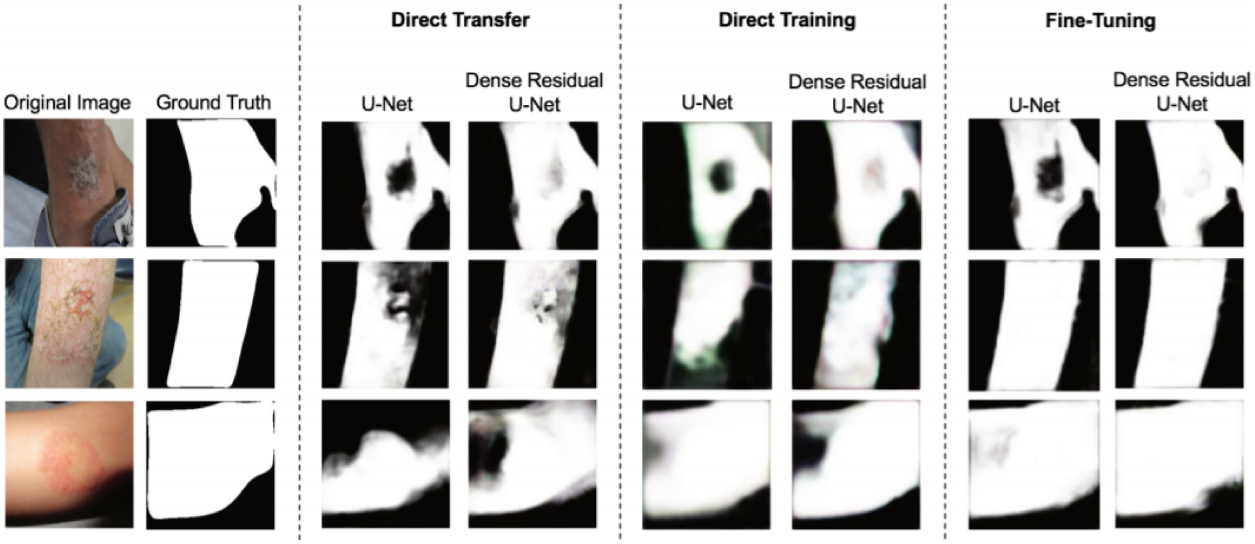


Figure 13 – U-Net and Dense Residual U-Net outputs on clinical photos. (CODELLA et al., 2018)

The authors found that for direct transfer, without the use of the specialized dataset, Dense Residual U-Nets outperformed regular U-Nets (0.55 *vs.* 0.51 Jaccard coefficients). However, U-Nets outperformed Dense Residual U-Nets for both direct training (0.83 *vs.* 0.80) and pre-training followed by fine-tuning (0.89 *vs.* 0.88).

Our approach to this problem is heavily based in the previously mentioned works of (RONNEBERGER et al., 2015) and (CODELLA et al., 2016) and is described in Section 3.3.

2.6 Literature Conclusion

In this section, we covered the main recent findings for deep neural network architectures for the segmentation of general datasets, as well as specific approaches to the segmentation of dermoscopic images.

Many of the proposed segmentation networks require a large amount of computational power and consequently, a long training time, due to the number of parameters and number of operations.

The U-net architecture is one of the options with a relatively smaller number of layers and parameters, and is often found to produce results similar to or better than larger networks for problems with a limited amount of training samples.

One drawback, especially for the specialized dermoscopic solutions covered in Section 2.5, is that implementations are seldom provided, and hence require a significantly larger amount of time to re-implement and reproduce. In the cases where an architecture is pre-trained on a larger general dataset, there is an even larger potential gap for experiment reproductions.

Capsule nets are an interesting new idea and the SegCaps results are promising due to the significantly smaller parameter numbers, despite only providing comparisons for a single dataset, that could already be solved with similarly high scores with convolutional neural networks. However, at the time of writing, there are still very few available implementations for classification and none for segmentation.

Section 3 describes our solution, where we explore some of the ideas from this literature review that we found to be the most promising.

3 Experiments

In this work, we decided to focus on U-net-like networks. The reason for this is that besides being specifically targeted at biomedical images, networks based on it demonstrated good performance in similar tasks, while having a much smaller number of layers and parameters than for instance, ResNet-based FCN networks. Smaller networks are typically faster to train to convergence, and are less likely to overfit on small datasets.

In Sections 3.1 and 3.2 we present the dataset and the metrics that are used for evaluation. In Section 3.3 we present an overview of the variations that were tested and each of them is then discussed in the subsequent sections.

3.1 Datasets

3.1.1 ISIC Archive Dataset

The International Skin Imaging Collaboration (ISIC) is an international effort to improve melanoma diagnosis and the ISIC Archive[1] contains the largest publicly available collection of quality controlled skin lesion images.

The ISIC Archive contains over 13,000 dermoscopic images, which were obtained from several clinical centers internationally, using a variety of devices. The images come in a variety of different sizes and resolutions, and contain associated clinical meta-data, and some also contain one or more associated binary masks.

Dermoscopy is an imaging technique used to visualize the skin unobstructed by skin surface reflections. This allows for an enhanced inspection of the skin, which has been shown to improve diagnostic accuracy by dermatologists when compared to standard unedited photos.

Some of the images contain multiple masks, and as shown in Figure 14. Because masks are obtained in many different ways and by different people, it is common for the Jaccard coefficient between pairs to be relatively high. In some extreme cases, there is complete disagreement between two masks. In Figure 14, the first four rows contain Jaccard indexes between 0.66 and 0.68, while the last row has a Jaccard index of 0.

[1] https://isic-archive.com

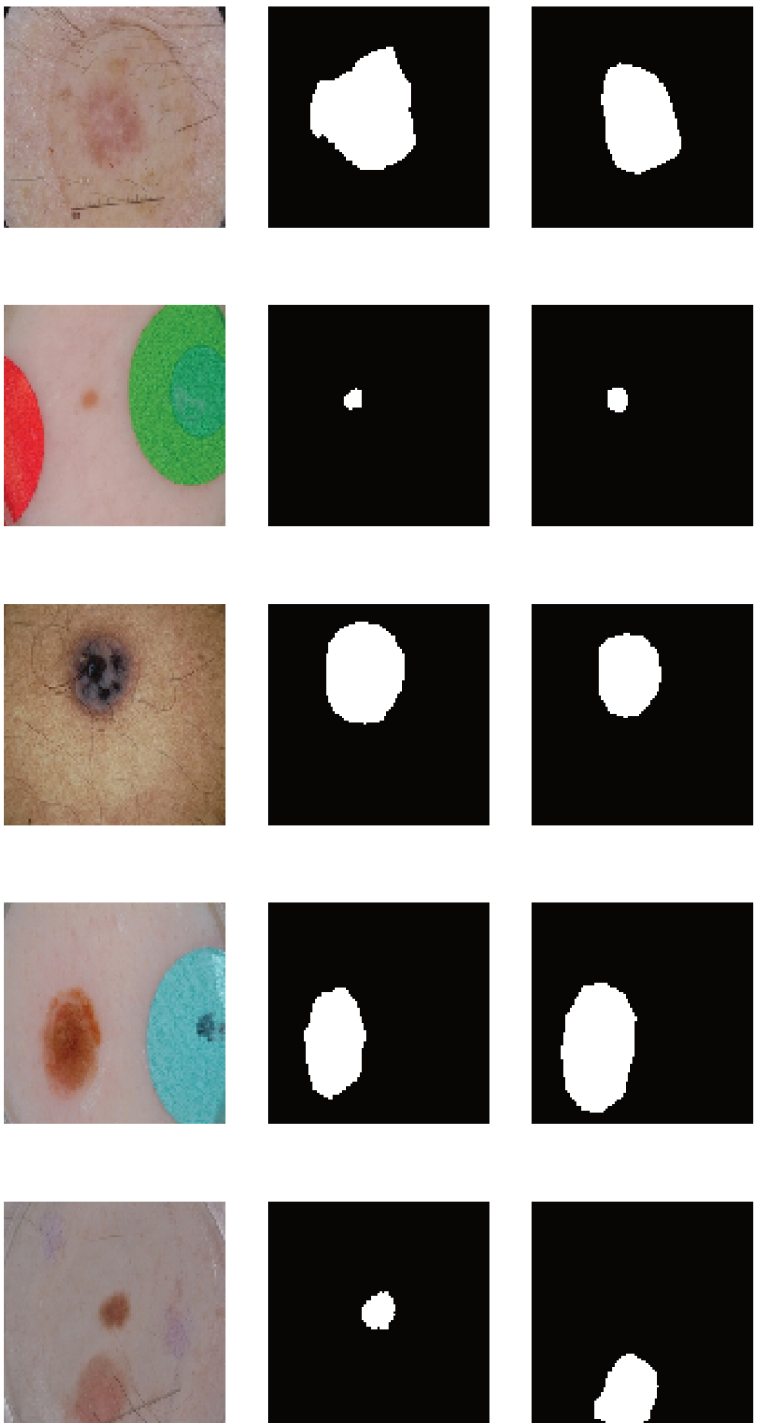


Figure 14 – ISIC Archive examples, with two different masks for the same image in each row.

3.1.2 ISIC Challenge 2017 Dataset

We used the ISIC 2017 Challenge (CODELLA et al., 2017) official train dataset (dermoscopic images).

The lesion segmentation training data contains the ground-truth masks (only one

Table 1 – Dataset

Dataset	**Train Description**	**# of images**
ISIC Challenge 2017	374 melanomas 254 seborrheic keratoses 1,372 benign nevi	Train: 2,000 Validation: 150 Test: 600

mask per image), which are binary masks based on the expert manual tracing of the lesion boundaries also in the form of a binary mask. White pixels are considered inside the area of the lesion and black pixels are considered outside.

A separate public validation dataset (150 images) and test dataset (600 images) are also available (but without the ground-truth masks for the duration of the challenge).

All ISIC Challenge images also contained the gold standard definitive diagnosis for use in the classification task. The gold standard signifies that the diagnosis data is obtained from pathology report information and expert consensus.

3.2 Metrics

$$Jaccard\,(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.1}$$

$$Dice\,(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{3.2}$$

The Jaccard and Dice coefficients are used to compare the similarity of two samples. In the equations above, **A** represents the first mask (which can be the ground-truth labeled mask) and **B** represents the second mask (which can be the mask predicted by the network). In the binary mask case we are considering, each pixel can have only two values after being normalized: 0 or 1. Each pixel in mask **A** is compared only to the corresponding pixel in the same position in mask **B**. Their intersection corresponds to the number of pixels that are simultaneously equal to 1 in both mask **A** and in mask **B**.

They are often used in image segmentation tasks and have the advantage of not giving too much weight to outliers (compared to the Euclidean distance, for example) and not being biased against unbalanced sets (compared to accuracy).

Both are equivalent in the sense that the Jaccard coefficient can be obtained directly from the Dice coefficient and vice versa.

3.3 Experiments Overview

Table 2 shows in bold the cases that resulted in the best validation scores.

Table 2 – Experiments performed, as detailed in the following subsections.

Name	**Options tested**
Architecture	A1) VGG-based U-net-like model A2) VGG-based U-net-like model (initialized with pre-training on ImageNet) B1) **U-net based model with fully connected layers** B2) U-net based model without fully connected layers
Dataset	**ISIC Challenge Train** Full ISIC-Archive
Input Channels	**RGB (3 channels)** RGB and HSV (6 channels)
Input Resolution	**128×128** 256×256
Loss	Binary Cross-Entropy **Dice** Jaccard Mean Squared Error
Normalization	**Subtract Imagenet Mean** Subtract Sample-wise Mean Subtract Train Dataset Mean Subtract Train Dataset Mean and Divide by Std
Optimizer	**Adam** Stochastic Gradient Descent

3.3.1 Architecture

3.3.1.1 Pre-trained VGG-based model

The model identified as **A1** in 2 consisted of the VGG-16 layers except the last max-pooling layer and the following fully-connected layers. Additionally, it included the U-shape according to the U-net model, with convolutional and up-sampling layers and without any fully-connected layer. Dropout layers were added after each of the five layers with concatenated features. For a visualization of the graph of this model, refer to 20.

A transfer-learning routine was followed for model **A2**, which shares the same architecture but is initialized with pre-trained weights:

- the original layers were initialized with pre-trained weights and **frozen** (meaning their weights would not be updated in subsequent iterations)
- the network was trained for 100 epochs, and during this time, only the weights of the new, unfrozen layers were updated
- all original layers were unfrozen
- the network was trained for another 100 epochs, and during this time, all layers' weights were updated

We initialized the layers from the original classification VGG-16 model with the weights of the pre-trained network on the ImageNet dataset on the classification task.

For this network, the official validation Jaccard score was 0.753.

3.3.1.2 U-net-based model

The best results were obtained with a network based on the work of (CODELLA et al., 2016) for the segmentation task.

We found that the train and validation scores decreased when using a smaller number of neurons on the first fully connected layer. The best individual network, **B1**, used 8192 neurons and 0.5 dropout and reached an official validation score of 0.783 after training for 220 epochs. This model architecture is shown in 15.

By removing the fully-connected layers (as in the original U-net paper) and adding batch normalization layers, in model **B2**, we obtained close but lower scores, with a validation score of 0.774 after the same 220 epochs. This second model was faster to train and much smaller due to having fewer parameters, however, we ultimately did not use it in the challenge submission.

Below are some of the main changes in our approach relative to the work it was based on.

- Dice coefficient as the loss function.
- Adam as the optimizer.
- ReLU activations on all layers except the last, where the sigmoid activation was used.

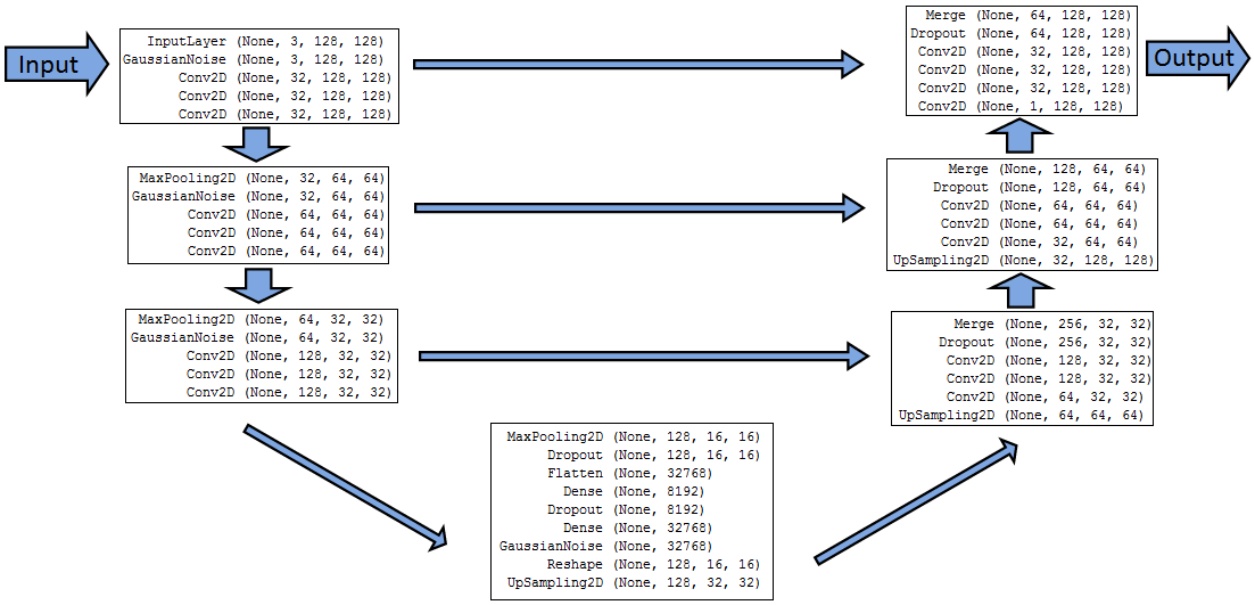


Figure 15 – Best performing model (U-net based model with fully connected layers).

- Ground-truth masks were rescaled by dividing by 255 (maximum pixel value).
- Subtracting the mean RGB value of the ImageNet training set from input RGB values.

3.3.2 Data Augmentation

We used various forms of online image augmentation. For each batch, transformations were done on the CPU while the previous batch executed on the GPU. Transformations were selected between a set of options using the Keras pipeline, and by using a different seed for the random number generator, a different sequence of transformations per batch would happen in each training sequence. So for each batch, a vertical shift value would be randomly chosen and applied, then a horizontal shift value would be chosen and applied on the already transformed image, and so on.

The transformations used are shown in Table 3.

Table 3 – Transformations

Type	**Possible values to sample**
Vertical Shifts	0 to 10% of image height
Horizontal Shifts	0 to 10% of image width
Zoom	0 to 20%
Rotation	0 to 270°
Fill Mode	Reflection (input image) Nearest pixel (mask)

When transformations are applied to images, it is common to fill the missing parts (such as when an image is rotated) with a reflection instead of with a constant color background, as that produces more natural-looking images, which tend to have more similarities to other images in the dataset. This is the fill mode we use for the input images. However, in preliminary testing, we found that using reflection in the mask images was generally increasing error rates. Because in our particular dataset, lesions are almost always centered and there is typically exactly one lesion per image, this data augmentation was causing the network to predict lesions near the border more often as well as multiple lesions, which was in many cases, a wrong prediction. For this reason, we chose not to reflect the mask images during data augmentation, opting instead to fill the missing parts with the nearest pixel.

3.3.3 Dataset

We attempted to use only the ISIC Challenge 2017 dataset (Subsection 3.1.2), as well as the full ISIC-Archive (Subsection 3.1.1) images containing corresponding masks.

The attempt using additional images from the ISIC Archive resulted in lower validation scores (obtained in the validation leaderboard of the challenge website).

One possible explanation to getting lower scores with the larger dataset is that, while ISIC Archive contains a larger number of images and corresponding masks (over 10,000 dermoscopic images), the ISIC Challenge dataset is more selective, and contains ground-truth annotations generated by a panel of dermoscopic experts.

Segmentation does not have a perfect agreement between medical experts, and various tools are utilized, so using inaccurate masks as the ground-truth makes it more difficult to train the network.

3.3.4 Framework and code

The models were implemented using Keras (CHOLLET et al., 2015) with the Theano back-end (BERGSTRA et al., 2010).

The code is publicly available at the **github repository**[2].

3.3.5 Input Channels

We attempted adding three additional HSV (hue, saturation, value) channels in the input alongside the RGB channels for a total of six channels.

[2] Available at: https://github.com/juliafeec/isic-2017-segmentation

HSV is an alternative representations of the RGB color model The idea behind this is that the HSV channels could make it easier to extract certain attributes, as some information is more directly accessible, such as the saturation and lightness of a color. And as mentioned in Section 2.5, some authors report that this improves performance.

In our implementation, we use OpenCV (BRADSKI, 2000) for the color-space conversion, where Hue range is [0,179], Saturation range is [0,255] and Value range is [0,255].

3.3.6 Input Resolution

All images were resized to 128×128 pixels. We also tried with 256×256 pixels for a few runs at first, but the performance was similar and training was slower.

Transforming the images before resizing them was also briefly attempted, but did not improve the results and was much slower.

3.3.7 Loss

As the evaluation metric being used is the Jaccard Index, as introduced in Section 3.2, it should perhaps come as no surprise that loss functions that directly minimize this result in better performance than others. As previously mentioned, the Jaccard and Dice metrics can be directly converted into one another, and in our experiments, the loss functions based on those yielded the best results, with a slight edge to a dice-based loss.

One thing to note is that both the Dice and Jaccard metrics are not differentiable, and as such, the loss functions that are based on them need to be changed slightly. In order to calculate the metrics, each pixel assumes a discrete value of either 0 or 1, and to calculate the intersections and unions, a simple comparison is made to determine if pixels are equal or different.

However, when the loss is being calculated, each pixel has a continuous value, which is given by the output of a sigmoid activation, and rounding is also not an option as it is also not a differentiable operation.

So the way we implement this, as shown in 3.3, is that for each batch, the intersection of pixels is calculated as the sum of the element-wise multiplication of each pixel, and the union of pixels is calculated as the sum of the ground-truth (which are already binary) and predicted pixels summations. A smoothing constant is also added to prevent a division by zero in case all pixels in a given batch turn out to be equal to zero.

$$DiceLoss = \frac{2(y_{true} \cdot y_{pred}) + \alpha}{sum\{y_{true}\} + sum\{y_{pred}\} + \alpha} \tag{3.3}$$

where y_{true} and y_{pred} are the flattened arrays of, respectively, the ground-truth output and the predicted output of a given batch. α is the smoothing constant.

3.3.8 Normalization

Normalization is an important pre-processing step when training neural networks (SOLA; SEVILLA, 1997) to obtain good results. Normalizations are linear transformations of the original data, and a commonly used type is the standard norm, which assumes a normal distribution, such as in 3.4.

$$z = \frac{x - \mu}{\sigma} \tag{3.4}$$

where μ is the mean and σ is the standard deviation of the population.

We attempted taking the mean of each sample, or the mean and standard deviation of all the training dataset, as well as simply the global average of the ImageNet training set (which could be thought of as a more general population when considering the mean pixel values in each channel).

3.3.9 Optimizer

Our best results were obtained using the Adam optimizer (KINGMA; BA, 2014), an adaptive optimization algorithm that uses the history of iterations.

One of the main advantages of the Adam optimizer is that the hyper-parameters typically require little tuning and the convergence is typically reached faster than with the stochastic gradient descent (SGD) optimizer. However, as shown in (WILSON et al., 2017), adaptive optimizers often generalize worse than SGD, even when obtaining a better training performance.

A possible explanation of Adam reaching better results than SGD in our case is that further tuning of the SGD hyper-parameters would have been required to obtain better results.

With these shortcomings in mind, we briefly evaluated a third optimizer: SWATS (KESKAR; SOCHER, 2017).

3.3.9.1 SWATS Optimizer

The main idea of Keskar and Socher (2017) is that since Adam typically converges faster and outperforms SGD in the initial epochs of training, while SGD typically finds a better final solution, a hybrid strategy is proposed: training begins with an adaptive method and switches to SGD. The authors also found that switching early in the training process results in testing accuracy comparable to SGD but switching too late results in a generalization gap comparable to using only the adaptive method.

The proposed triggering condition relates to the projection of Adam steps on the gradient subspace and obtains the hyper-parameters for SGD, as shown in Figure 16.

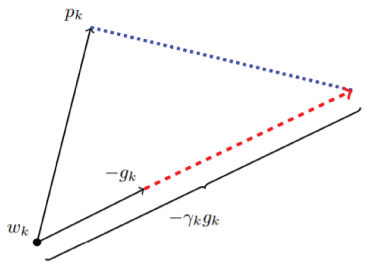
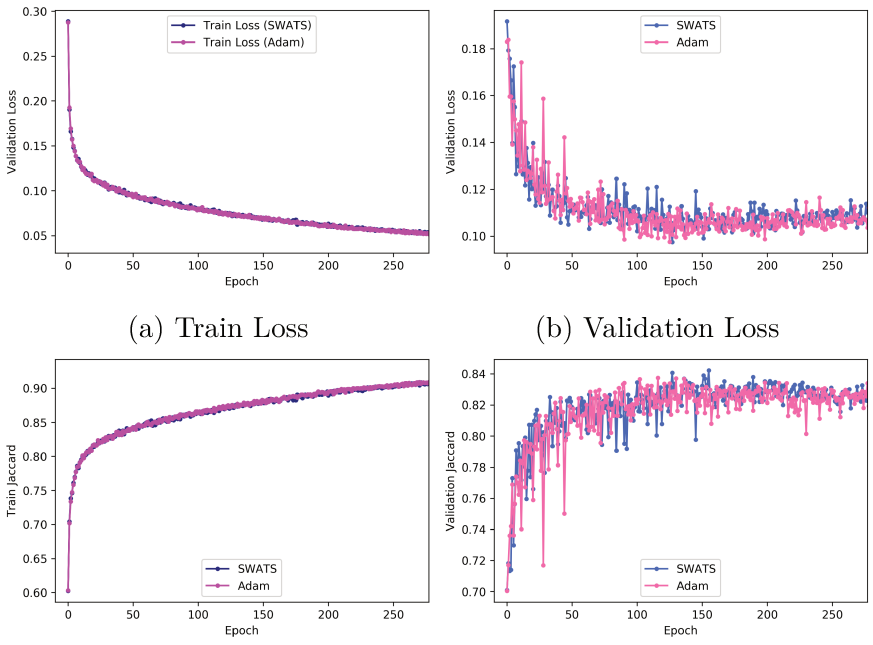


Figure 16 – SWATS proposed projection given an iterate w_k, stochastic gradient p_k and Adam step p_k, obtaining the learning rate γ_k for SGD (KESKAR; SOCHER, 2017)



(a) Train Loss (b) Validation Loss

(c) Train Jaccard Coefficient (d) Validation Jaccard Coefficient

Figure 17 – Comparison of SWATS and Adam optimizers. For SWATS, the switch point between Adam and SGD was triggered in epoch 50.

The paper does not make the code available, so a re-implementation was done using the Keras framework with the TensorFlow back-end.

We compared the progression of the loss function, as well as the Jaccard coefficient, between the SWATS and Adam optimizers for the U-net model with the ISIC Challenge 2017 dataset, as shown in Figure 17. We found the results to be very similar for both optimizers.

3.4 Results

We found that the train and validation scores decreased when using a smaller number of neurons on the first fully connected layer. The best individual network used 8192 neurons and 0.5 dropout and reached an official validation score of 0.783 after training for 220 epochs.

By removing the fully-connected layers (as in the original U-net paper) and adding batch normalization layers, we obtained close but lower scores, with a validation score of 0.774 after the same 220 epochs. This second model was faster to train and much smaller due to having fewer parameters, however we ultimately did not use it in the challenge submission. Our best result was obtained by using an ensemble of four models:

- Two networks were trained with all 2,000 samples, without a validation split, for 250 and 500 epochs respectively.
- Two networks were trained and validated with two different 1600/400 splits, for 220 epochs each.

These four models scored, individually, between 0.780 and 0.783 in the validation set. The ensemble, which was obtained by taking the average of the four models, achieved a Jaccard score of 0.793 in the validation set. For the final submission (ensemble), the test set score was 0.754. This segmentation approach as well as a classification approach were described in our abstract (MENEGOLA et al., 2017) for the ISIC 2017 Challenge.

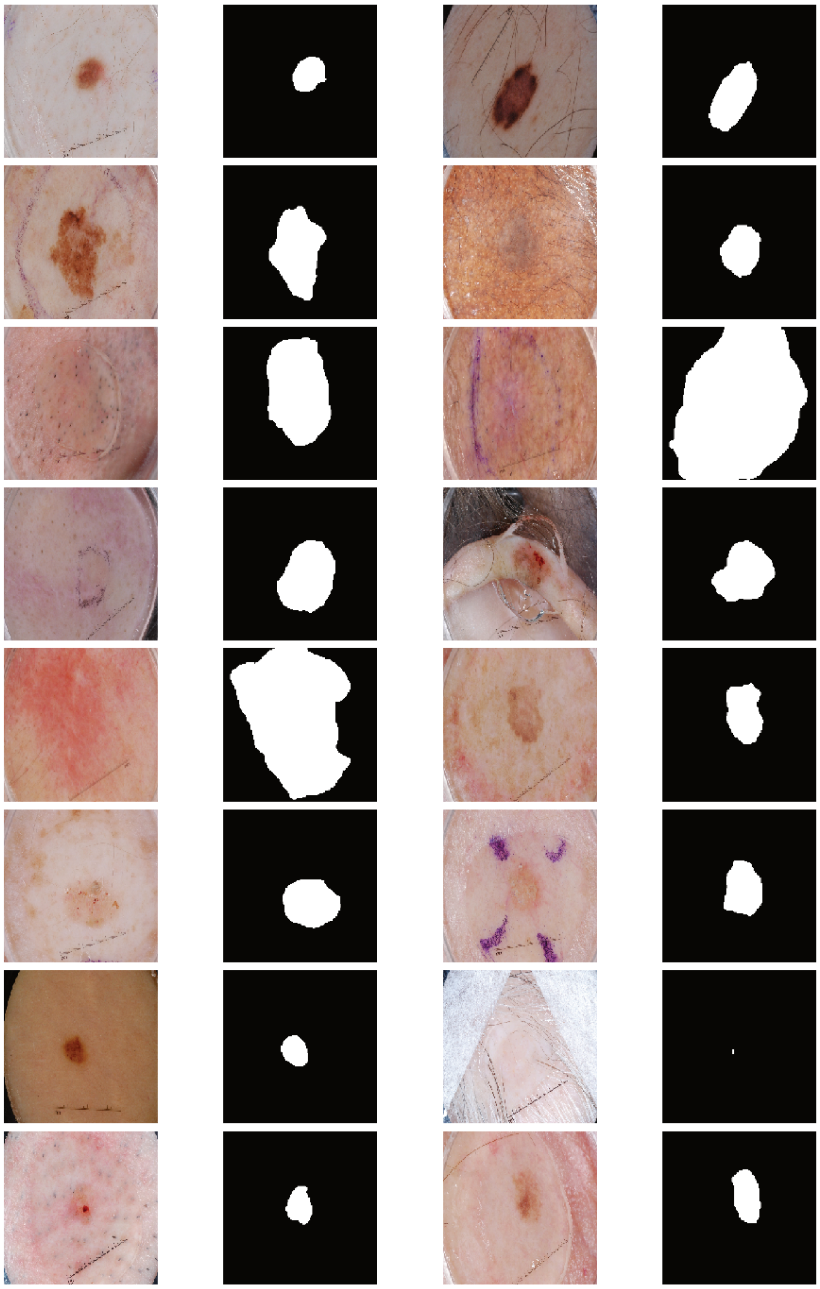
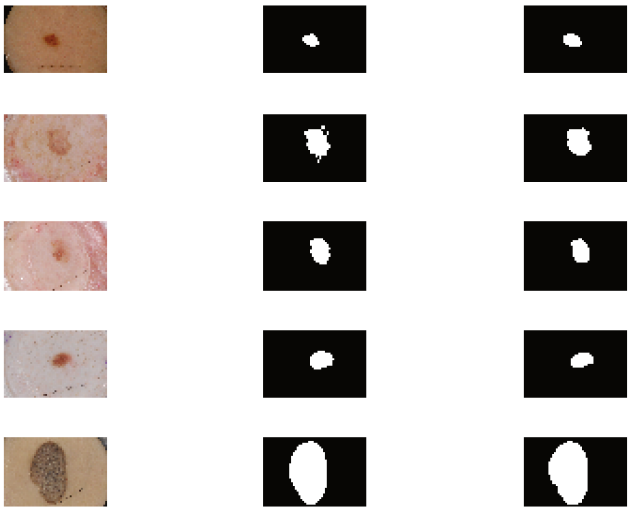
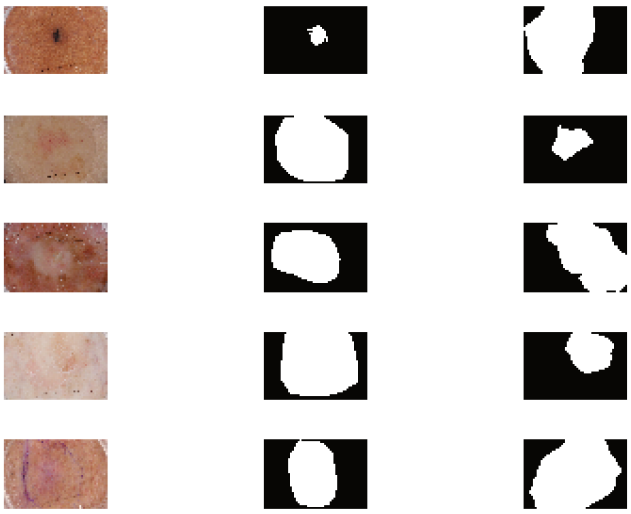


Figure 18 – Samples of test images and their corresponding predictions with the best performing model ensemble.

3.4.1 Prediction Examples



(a) Good prediction examples (high Jaccard score)



(b) Bad prediction examples (low Jaccard score)

Figure 19 – Test image examples. First column is the original image, second column is the ground-truth mask and the third column is the predicted mask.

3.5 Exploring the use of segmentation in the classification task

Deep learning has enabled many improvements in the area of automated melanoma screening in recent years. In (VALLE et al., 2017), we investigate methodological issues for the design and evaluation of deep learning models in the context of melanoma detection.

Two full factorial experiments were performed for five different test datasets, for a total of 2560 exhaustive trials in the main experiment, and 1280 trials in the transfer learning

experiment.

The evaluated factors were:

- Use of transfer learning
- Model architecture
- Train dataset
- Image resolution
- Type of data augmentation
- Input normalization
- Use of segmentation
- Duration of training
- Additional use of SVM
- Test data augmentation

In this work, we focus on the evaluation of the impact of using segmentation to attempt to improve classification, and we refer the reader to the paper for more details on the remaining factors.

It is common to use segmentation as a preprocessing step in on deep-learning models for melanoma detection (NASR-ESFAHANI et al., 2016) (YANG et al., 2017) and improved accuracies are usually reported in such approaches.

In some cases, a dedicated network segments the lesion image, and subsequently forwards it to the classification network (YU et al., 2017) (CODELLA et al., 2016) (DÍAZ, 2017).

3.5.1 Train Dataset

Table 4 – Train Dataset

Dataset	**Melanoma**	**Nevus**	**Keratosis**
ISIC Challenge 2017 (train split)	374	1372	254
Full train (composition of datasets)	1227	10124	710

The Full train split contains data from different sources:

- the ISIC Archive[3], with over 13,000 dermoscopic images
- the Dermofit Image Library (BALLERINI et al., 2013), with 1,300 clinical images (76 melanomas, 257 seborrheic keratoses)
- the PH2 Dataset (MENDONÇA et al., 2013), with 200 dermoscopic images (40 melanomas)

For this work, we streamlined the best model introduced in Section 3.3, reducing the number of parameters, removing the fully-connected and Gaussian-noise layers, and adding batch-normalization and dropout layers. This had the effect of occupying much less disk space and being much faster to train, an important point as these experiments introduced a much larger dataset, as shown in Table 4.

The segmentation models and their corresponding classification models were trained on the same images.

Because of the lack of literature consensus on the optimal approach for using segmentation for melanoma classification, we opted for schemes with minimal changes to both data and networks.

Before starting the full factorial analysis, we did a preliminary evaluation of two candidate approaches:

- Pixel-wise multiplication of the input RGB images by the segmentation masks
- Pre-encoding the four dimensions (R, G, B, and binary mask) into three dimensions, keeping the rest of the networks unchanged

The latter appeared to be more promising in our preliminary tests, so only this approach was considered for the full analysis.

Pre-encoding the masks required an adaptation of the ResNet and Inception networks. For both models, we added three convolutional layers before the input, two layers with 32 filters, and a third with 3 filters.

All convolutional layers used 3×3 kernels and stride of 1. Due to ResNet-101-v2 and Inception-v4 models using input dimensions of 299×299 pixels, the pre-encoding adapter layer used 305×305 pixel images, to account for the two border pixels lost at each convolutional layer.

[3] https://isic-archive.com/

Our results showed that the use of segmentation caused the classification scores to become worse, compared to using the original unaltered images without any mask information.

The negative result on segmentation needs further exploration, due to the fact that works in the literature report many different ways to incorporate segmentation information to classification (NASR-ESFAHANI et al., 2016) (YANG et al., 2017), often with improved performances.

Prior work (CODELLA et al., 2016) has found that a combination of classification from the original unaltered image with classification from a bounding box cropped input (based on the corresponding mask) achieved the best results, rather than using only the latter input option.

There are many possible approaches and as previously mentioned, no consensus or comprehensive direct comparisons of different approaches for integrating segmentation information in the context of melanoma classification. Future work is required in this subject, and some of the possible relevant factors include: the use of bounding boxes *vs.* pixel-wise masks, binary *vs.* smooth integration when using predicted masks, fusion at lower layer *vs.* higher layers.

4 Conclusion

In this work, we provided a critical review of the subject of semantic image segmentation, especially in the context of limited datasets, such as in the case of medical imaging. We reviewed the recent literature in deep neural network approaches to image segmentation in general, and focused in the skin lesion problem, providing an analytical review of the automated skin lesion segmentation literature. We also covered the challenges of dealing with these medical datasets, due to the limited number of images, as well as the variance of ground-truth masks caused by medical experts' disagreement and different devices being used.

We also evaluated a number of recent techniques, focusing on the U-net architecture (RONNEBERGER et al., 2015), which is widely used for biomedical and other small, specific datasets. Our results reinforce the findings of Codella et al. (2017) that adding fully-connected layers to the U-net network (which in its original form, only contains convolutional layers) results in better performance. We also found that using a loss based on the Jaccard metric leads to better results, a similar finding to that of (YUAN et al., 2017) in the same ISIC 2017 Challenge.

Another objective of this work is the accuracy improvement of automated skin lesion segmentation. Towards this goal, the implementation of the segmentation network that we propose, which is heavily based in the existing literature that is reviewed in this work, is able to achieve a reasonable performance with the use of only a small dataset and a consumer-level single PC. Since our open-source implementation of the code is available for general use, our findings are reproducible and can provide a simpler starting point for future research. Our final Jaccard score in the closed test set of the ISIC Challenge 2017 was 0.754, which was the fifth best out of 21 submissions, while the overall best result reached in the challenge was 0.765.

As a secondary goal, we also experimented with using the predictions of the segmentation network to improve the performance of the classification task, although our approach did not achieve a better classification performance.

Future work includes exploring some of the ideas mentioned in the literature we reviewed that we did not explore, for instance, evaluating deeper networks such as ResNet or Inception-based FCNs, other types of data augmentation such as elastic augmentations and additional post-processing steps and transfer-learning from larger general datasets such as Pascal VOC. It would be interesting to observe if the images that are the most difficult for

the U-net network to predict are the same ones as for the ResNet-based network. Should the networks have different strengths in terms of what types of images they can better predict, an ensemble of different types of architectures might significantly improve overall performance.

One of the problems with found with using the extended ISIC Archive dataset was that the ground-truth masks were much more unreliable than in the reduced ISIC Challenge dataset, and the same image would at times have multiple, disagreeing masks, as we showed in this work. One idea for future work is to only keep images for training when the masks agree by more than a chosen Jaccard threshold, or to give a sample weight during training that is proportional to this agreement.

Another adjacent line of work involves exploring the possible contributions of predicted segmentation masks to a classification network, and in which cases this information transfer may be beneficial.

In conclusion, due to the promising results of this research, achieving a Jaccard performance that is on par with that of trained specialists with a relatively small dataset, we support the idea that automated segmentation can become a useful tool in a clinical setting, and assist physicians to make informed health-care decisions.

References

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. **arXiv preprint arXiv:1603.04467**, 2016.

ALANDER, J. T. **Indexed bibliography of genetic algorithms in optics and image processing**. [sinenomine], 2000.

AMERICAN CANCER SOCIETY. Cancer Facts and Figures, 2017. Address: <http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-048738.pdf>.

BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, vol. 39, no. 12, pp. 2481–2495, 2017.

BALLERINI, L.; FISHER, R. B.; ALDRIDGE, B.; REES, J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: [sinenomine]: Springer, 2013. pp. 63–86.

BERGSTRA, J.; BREULEUX, O.; BASTIEN, F.; LAMBLIN, P.; PASCANU, R.; DESJARDINS, G.; TURIAN, J.; WARDE-FARLEY, D.; BENGIO, Y. Theano: a CPU and GPU Math Expression Compiler. **Proceedings of the Python for Scientific Computing Conference (SciPy)**, 2010.

BERSETH, M. ISIC 2017 - Skin Lesion Analysis Towards Melanoma Detection. **CoRR**, abs/1703.00523, 2017. arXiv: 1703.00523. Address: <http://arxiv.org/abs/1703.00523>.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. Norwell, MA, USA: Kluwer Academic Publishers, 1981. ISBN 0306406713.

BHANU, B.; LEE, S.; MING, J. Adaptive image segmentation using a genetic algorithm. **IEEE Transactions on systems, man, and cybernetics**, IEEE, vol. 25, no. 12, pp. 1543–1567, 1995.

BHATTACHARYYA, S. A brief survey of color image preprocessing and segmentation techniques. **Journal of Pattern Recognition Research**, vol. 1, no. 1, pp. 120–129, 2011.

BI, L.; KIM, J.; AHN, E.; FENG, D. Automatic Skin Lesion Analysis using Large-scale Dermoscopy Images and Deep Residual Networks. **CoRR**, abs/1703.04197, 2017. arXiv: 1703.04197. Address: <http://arxiv.org/abs/1703.04197>.

BRADSKI, G. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, 2000.

BRINK, A. Minimum spatial entropy threshold selection. **IEE Proceedings-Vision, Image and Signal Processing**, IET, vol. 142, no. 3, pp. 128–132, 1995.

CELEBI, M. E.; WEN, Q.; IYATOMI, H.; SHIMIZU, K.; ZHOU, H.; SCHAEFER, G. A state-of-the-art survey on lesion border detection in dermoscopy images. **Dermoscopy Image Analysis**, Boca Raton, CRC Press, pp. 97–129, 2015.

CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **arXiv preprint arXiv:1606.00915**, 2016.

CHEN, L.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. Rethinking Atrous Convolution for Semantic Image Segmentation. **CoRR**, abs/1706.05587, 2017. arXiv: `1706.05587`. Address: <`http://arxiv.org/abs/1706.05587`>.

CHEN, T.; LI, M.; LI, Y.; LIN, M.; WANG, N.; WANG, M.; XIAO, T.; XU, B.; ZHANG, C.; ZHANG, Z. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. **arXiv preprint arXiv:1512.01274**, 2015.

CHENG, H.-D.; JIANG, X. H.; SUN, Y.; WANG, J. Color image segmentation: advances and prospects. **Pattern recognition**, Elsevier, vol. 34, no. 12, pp. 2259–2281, 2001.

CHOLLET, F. et al. Keras. GitHub, 2015. Address: <`https://github.com/fchollet/keras`>.

CIRESAN, D.; GIUSTI, A.; GAMBARDELLA, L. M.; SCHMIDHUBER, J. Deep neural networks segment neuronal membranes in electron microscopy images. **Advances in neural information processing systems**, pp. 2843–2851, 2012.

CODELLA, N. C. F.; NGUYEN, Q.; PANKANTI, S.; GUTMAN, D.; HELBA, B.; HALPERN, A.; SMITH, J. R. Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. **CoRR**, abs/1610.04662, 2016. Address: <`http://arxiv.org/abs/1610.04662`>.

CODELLA, N. C.; ANDERSON, D.; PHILIPS, T.; PORTO, A.; MASSEY, K.; SNOWDON, J.; FERIS, R.; SMITH, J. Segmentation of both Diseased and Healthy Skin from Clinical Photographs in a Primary Care Setting. **arXiv preprint arXiv:1804.05944**, 2018.

CODELLA, N. C.; GUTMAN, D.; CELEBI, M. E.; HELBA, B.; MARCHETTI, M. A.; DUSZA, S. W.; KALLOO, A.; LIOPYRIS, K.; MISHRA, N.; KITTLER, H., et al. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). **arXiv preprint arXiv:1710.05006**, 2017.

CODELLA, N.; CAI, J.; ABEDINI, M.; GARNAVI, R.; HALPERN, A.; SMITH, J. R. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. **International Workshop on Machine Learning in Medical Imaging**, pp. 118–126, 2015.

COLLOBERT, R.; KAVUKCUOGLU, K.; FARABET, C. Torch7: A matlab-like environment for machine learning. **BigLearn, NIPS Workshop**, EPFL-CONF-192376, 2011.

COMANICIU, D.; MEER, P. Mean shift: A robust approach toward feature space analysis. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, vol. 24, no. 5, pp. 603–619, 2002.

DE, S.; BHATTACHARYYA, S.; CHAKRABORTY, S.; DUTTA, P. Image Segmentation: A Review. In: [sinenomine]: Springer, 2016. pp. 29–40.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 248–255, 2009.

DÍAZ, I. G. Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. **arXiv preprint arXiv:1703.01976**, 2017.

DUBOIS, D.; PRADE, H. Fuzzy sets and systems: Theory and applications. **American Mathematical society**, vol. 7, no. 3, pp. 603–612, 1982.

EVERINGHAM, M.; VAN GOOL, L.; WILLIAMS, C. K. I.; WINN, J.; ZISSERMAN, A. The Pascal Visual Object Classes (VOC) Challenge. **International Journal of Computer Vision**, vol. 88, no. 2, pp. 303–338, June 2010.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Efficient graph-based image segmentation. **International journal of computer vision**, Springer, vol. 59, no. 2, pp. 167–181, 2004.

FORSYTH, D. A.; PONCE, J. **Computer Vision: A Modern Approach**. [sinenomine]: Prentice Hall Professional Technical Reference, 2002. ISBN 0130851981.

GIRSHICK, R. Fast R-CNN. **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**, IEEE Computer Society, Washington, DC, USA, pp. 1440–1448, 2015. DOI: 10.1109/ICCV.2015.169. Address: <http://dx.doi.org/10.1109/ICCV.2015.169>.

GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, IEEE Computer Society, Washington, DC, USA, pp. 580–587, 2014. DOI: 10.1109/CVPR.2014.81. Address: <http://dx.doi.org/10.1109/CVPR.2014.81>.

GLOROT, X.; BORDES, A.; BENGIO, Y. Deep sparse rectifier neural networks. **Proceedings of the 14th International Conference on Artificial Intelligence and Statistics**, pp. 315–323, 2011.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. **Machine learning**, Springer, vol. 3, no. 2, pp. 95–99, 1988.

GONZALEZ, R. C.; WOODS, R. E. **Digital image processing**. [sinenomine]: Upper Saddle River, NJ: Prentice Hall, 2012.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. **arXiv preprint arXiv:1412.6572**, 2014.

HANNUN, A.; CASE, C.; CASPER, J.; CATANZARO, B.; DIAMOS, G.; ELSEN, E.; PRENGER, R.; SATHEESH, S.; SENGUPTA, S.; COATES, A., et al. Deep speech: Scaling up end-to-end speech recognition. **arXiv preprint arXiv:1412.5567**, 2014.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep Residual Learning for Image Recognition. **arXiv preprint arXiv:1512.03385**, 2015.

— Deep residual learning for image recognition. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 770–778, 2016.

HINTON, G. E. Learning multiple layers of representation. **Trends in Cognitive Sciences**, vol. 11, pp. 428–434, 2007.

HINTON, G. E.; SRIVASTAVA, N.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. R. Improving neural networks by preventing co-adaptation of feature detectors. **arXiv preprint arXiv:1207.0580**, 2012.

HOJJATOLESLAMI, S.; KITTLER, J. Region growing: a new approach. **IEEE Transactions on Image processing**, IEEE, vol. 7, no. 7, pp. 1079–1084, 1998.

HUANG, G.; LIU, Z.; WEINBERGER, K. Q.; MAATEN, L. van der. Densely connected convolutional networks. In: 2. vol. 1, p. 3.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. **International Conference on Machine Learning**, pp. 448–456, 2015.

JIA, Y.; SHELHAMER, E.; DONAHUE, J.; KARAYEV, S.; LONG, J.; GIRSHICK, R.; GUADARRAMA, S.; DARRELL, T. Caffe: Convolutional architecture for fast feature embedding. **Proceedings of the 22nd ACM international conference on Multimedia**, pp. 675–678, 2014.

KESKAR, N. S.; SOCHER, R. Improving Generalization Performance by Switching from Adam to SGD. **arXiv preprint arXiv:1712.07628**, 2017.

KINGMA, D.; BA, J. Adam: A method for stochastic optimization, 2014. arXiv: 1412.6980.

KRÄHENBÜHL, P.; KOLTUN, V. Efficient inference in fully connected crfs with gaussian edge potentials. **Advances in neural information processing systems**, pp. 109–117, 2011.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, pp. 1097–1105, 2012.

LALONDE, R.; BAGCI, U. Capsules for Object Segmentation, 2018. arXiv: 1804.04241.

LEI, Y.; SCHEFFER, N.; FERRER, L.; MCLAREN, M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 1695–1699, 2014.

LI, Y.; SHEN, L. Skin lesion analysis towards melanoma detection using deep learning network. **Sensors**, Multidisciplinary Digital Publishing Institute, vol. 18, no. 2, p. 556, 2018.

LIN, G.; MILAN, A.; SHEN, C.; REID, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In:

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft COCO: Common Objects in Context. **European Conference on Computer Vision (ECCV)**, 2014. Address: <http://mscoco.org/se3/wp-content/uploads/2014/09/coco_eccv.pdf>.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. A. van der; GINNEKEN, B. van; SÁNCHEZ, C. I. A survey on deep learning in medical image analysis. **Medical image analysis**, Elsevier, vol. 42, pp. 60–88, 2017.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully Convolutional Networks for Semantic Segmentation. **CoRR**, abs/1411.4038, 2014. arXiv: 1411.4038. Address: <http://arxiv.org/abs/1411.4038>.

MA, Z.; TAVARES, J. M. R., et al. A review of the quantification and classification of pigmented skin lesions: from dedicated to hand-held devices. **Journal of medical systems**, Springer, vol. 39, no. 11, p. 177, 2015.

MAGLOGIANNIS, I.; DOUKAS, C. N. Overview of advanced computer vision systems for skin lesions characterization. **IEEE transactions on information technology in biomedicine**, IEEE, vol. 13, no. 5, pp. 721–733, 2009.

MENDONÇA, T.; FERREIRA, P. M.; MARQUES, J. S.; MARCAL, A. R.; ROZEIRA, J. PH 2-A dermoscopic image database for research and benchmarking. In: IEEE, pp. 5437–5440.

MENEGOLA, A.; TAVARES, J.; FORNACIALI, M.; LI, L. T.; AVILA, S.; VALLE, E. RECOD Titans at ISIC Challenge 2017. **arXiv preprint arXiv:1703.04819**, 2017.

MISHRA, R.; DAESCU, O. Deep learning for skin lesion segmentation. In: IEEE, pp. 1189–1194.

NASR-ESFAHANI, E.; SAMAVI, S.; KARIMI, N.; SOROUSHMEHR, S. M. R.; JAFARI, M. H.; WARD, K.; NAJARIAN, K. Melanoma detection by analysis of clinical images using convolutional neural network. In: IEEE, pp. 1373–1376.

OH, J.; GUO, X.; LEE, H.; LEWIS, R. L.; SINGH, S. Action-conditional video prediction using deep networks in atari games. **Advances in Neural Information Processing Systems**, pp. 2863–2871, 2015.

PAL, N. R.; PAL, S. K. A review on image segmentation techniques. **Pattern recognition**, Elsevier, vol. 26, no. 9, pp. 1277–1294, 1993.

PATHAN, S.; PRABHU, K. G.; SIDDALINGASWAMY, P. Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review. **Biomedical Signal Processing and Control**, Elsevier, vol. 39, pp. 237–262, 2018.

PENG, C.; ZHANG, X.; YU, G.; LUO, G.; SUN, J. Large Kernel Matters–Improve Semantic Segmentation by Global Convolutional Network. **arXiv preprint arXiv:1703.02719**, 2017.

PHAM, D. L.; XU, C.; PRINCE, J. L. Current methods in medical image segmentation. **Annual review of biomedical engineering**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, vol. 2, no. 1, pp. 315–337, 2000.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You only look once: Unified, real-time object detection. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 779–788, 2016.

REN, S.; HE, K.; GIRSHICK, R.; SUN, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. **Proceedings of the 28th International Conference on Neural Information Processing Systems**, MIT Press, Montreal, Canada, pp. 91–99, 2015. Address: <`http://dl.acm.org/citation.cfm?id=2969239.2969250`>.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. **CoRR**, abs/1505.04597, 2015. Address: <`http://arxiv.org/abs/1505.04597`>.

ROSS, T. J.; BOOKER, J. M.; PARKINSON, W. J. **Fuzzy logic and probability applications: bridging the gap**. [sinenomine]: SIAM, 2002.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M., et al. Imagenet large scale visual recognition challenge. **International Journal of Computer Vision**, Springer, vol. 115, no. 3, pp. 211–252, 2015.

SABOUR, S.; CAO, Y.; FAGHRI, F.; FLEET, D. J. Adversarial Manipulation of Deep Representations. **arXiv preprint arXiv:1511.05122**, 2015.

SABOUR, S.; FROSST, N.; HINTON, G. E. Dynamic routing between capsules. In: pp. 3859–3869.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pp. 815–823, 2015.

SERRA, J. **Image analysis and mathematical morphology**. [sinenomine]: Academic Press, Inc., 1983.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. **IEEE Transactions on pattern analysis and machine intelligence**, Ieee, vol. 22, no. 8, pp. 888–905, 2000.

SHI, Z.; HE, L.; SUZUKI, K.; NAKAMURA, T.; ITOH, H. Survey on neural networks used for medical image processing. **International journal of computational science**, NIH Public Access, vol. 3, no. 1, p. 86, 2009.

SILVER, D.; HUANG, A.; MADDISON, C. J.; GUEZ, A.; SIFRE, L.; VAN DEN DRIESSCHE, G.; SCHRITTWIESER, J.; ANTONOGLOU, I.; PANNEERSHELVAM, V.; LANCTOT, M., et al. Mastering the game of Go with deep neural networks and tree search. **Nature**, Nature Publishing Group, vol. 529, no. 7587, pp. 484–489, 2016.

SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. **CoRR**, abs/1409.1556, 2014. Address: <`http://arxiv.org/abs/1409.1556`>.

SOLA, J.; SEVILLA, J. Importance of input data normalization for the application of neural networks to complex industrial problems. **IEEE Transactions on nuclear science**, IEEE, vol. 44, no. 3, pp. 1464–1468, 1997.

SUMENGEN, B.; MANJUNATH, B. Multi-scale edge detection and image segmentation. In: IEEE, pp. 1–4.

SZEGEDY, C.; ZAREMBA, W.; SUTSKEVER, I.; BRUNA, J.; ERHAN, D.; GOODFELLOW, I.; FERGUS, R. Intriguing properties of neural networks, 2013. arXiv: `1312.6199`.

TABACOF, P.; TAVARES, J.; VALLE, E. Adversarial Images for Variational Autoencoders. **NIPS Workshop**, 2016. arXiv: `1612.00155`. Address: <`http://arxiv.org/abs/1612.00155`>.

TABB, M.; AHUJA, N. Multiscale image segmentation by integrated edge and region detection. **IEEE Transactions on image processing**, IEEE, vol. 6, no. 5, pp. 642–655, 1997.

UDUPA, J. K.; SAMARASEKERA, S. Fuzzy connectedness and object definition: theory, algorithms, and applications in image segmentation. **Graphical models and image processing**, Elsevier, vol. 58, no. 3, pp. 246–261, 1996.

VALLE, E.; FORNACIALI, M.; MENEGOLA, A.; TAVARES, J.; BITTENCOURT, F. V.; LI, L. T.; AVILA, S. Data, Depth, and Design: Learning Reliable Models for Melanoma Screening. **CoRR**, abs/1711.00441, 2017. arXiv: 1711.00441. Address: <http://arxiv.org/abs/1711.00441>.

WILSON, A. C.; ROELOFS, R.; STERN, M.; SREBRO, N.; RECHT, B. The marginal value of adaptive gradient methods in machine learning. In: pp. 4151–4161.

YANG, W.; GUO, L.; ZHAO, T.; XIAO, G. Improving watersheds image segmentation method with graph theory. In: IEEE, pp. 2550–2553.

YANG, X.; ZENG, Z.; YEO, S. Y.; TAN, C.; TEY, H. L.; SU, Y. A novel multi-task deep learning model for skin lesion segmentation and classification, 2017. arXiv: 1703.01025.

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. How transferable are features in deep neural networks? **CoRR**, abs/1411.1792, 2014. arXiv: 1411.1792. Address: <http://arxiv.org/abs/1411.1792>.

YU, F.; KOLTUN, V. Multi-Scale Context Aggregation by Dilated Convolutions. **CoRR**, abs/1511.07122, 2015. arXiv: 1511.07122. Address: <http://arxiv.org/abs/1511.07122>.

YU, L.; CHEN, H.; DOU, Q.; QIN, J.; HENG, P.-A. Automated melanoma recognition in dermoscopy images via very deep residual networks. **IEEE transactions on medical imaging**, IEEE, vol. 36, no. 4, pp. 994–1004, 2017.

YUAN, Y.; CHAO, M.; LO, Y. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. **CoRR**, abs/1703.05165, 2017. arXiv: 1703.05165. Address: <http://arxiv.org/abs/1703.05165>.

YUAN, Y.; LO, Y.-C. Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. **IEEE Journal of Biomedical and Health Informatics**, IEEE, 2017.

ZHAO, H.; SHI, J.; QI, X.; WANG, X.; JIA, J. Pyramid scene parsing network. In: pp. 2881–2890.

Appendices

A Additional Graph

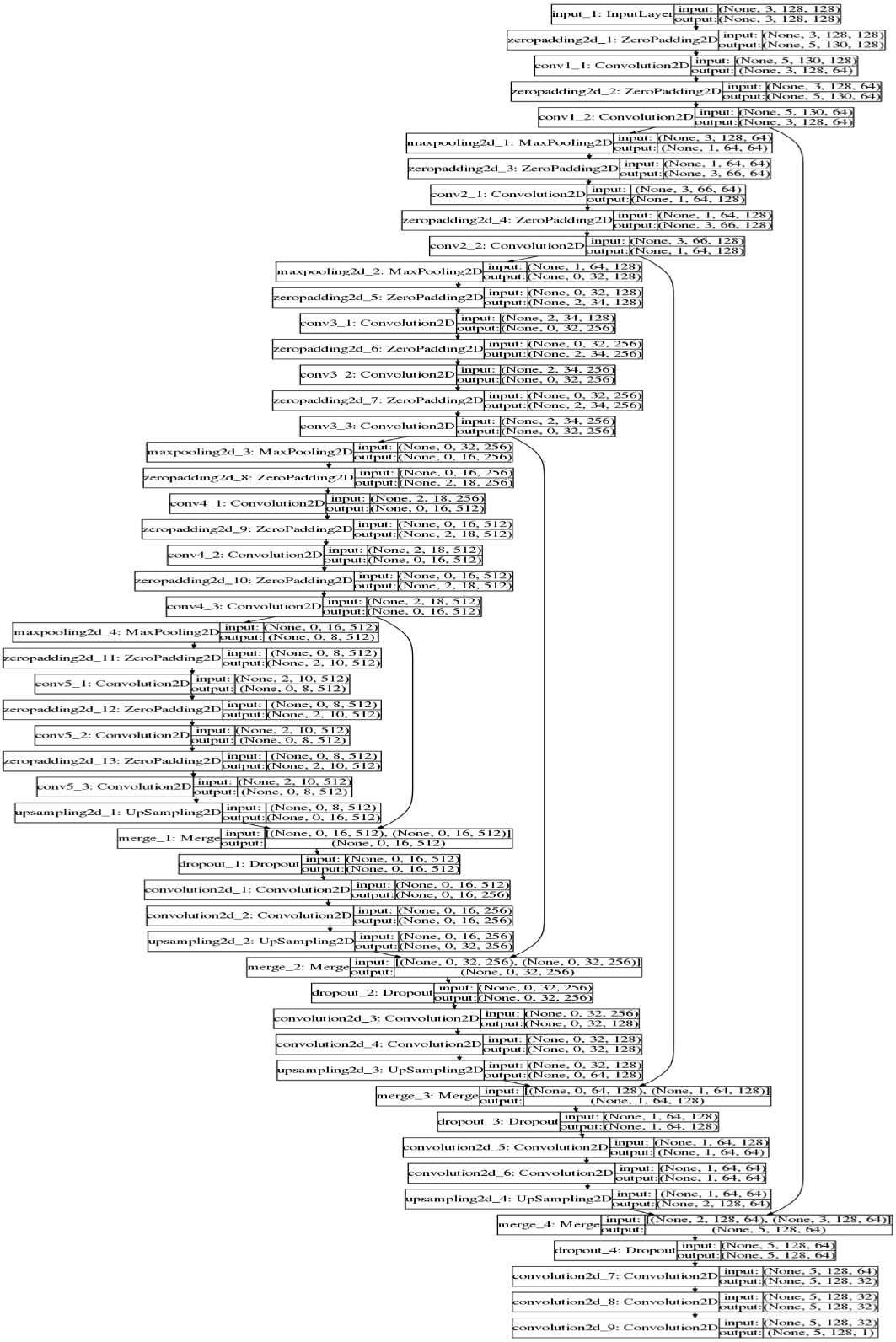


Figure 20 – VGG-based U-net Model

B Interesting properties of encoder-decoder networks

B.1 Adversarial Autoencoders

A type of network similar to what we explored in this segmentation work is the autoencoder, which aims to reproduce the input image, such that the output is as similar as possible to the input. This is made challenging by the fact that in the middle of the network, the number of learned parameters is much smaller than the number of input (and output) pixels. This causes the network to learn an intermediate compressed representation of the input.

(SZEGEDY et al., 2013) introduced adversarial images and (GOODFELLOW et al., 2014) proposed a method for generating images that seemed unchanged to the human eye but were capable of fooling a classification network into labeling it as a completely different class. (SABOUR et al., 2015) further investigated adversarial images and how they are changed throughout the network so that their internal representation is increasingly similar to the target class.

In our work (TABACOF et al., 2016) we investigated the behavior of adversarial images in the context of autoencoders, as shown in Fig. 21, using the optimization described in B.1.

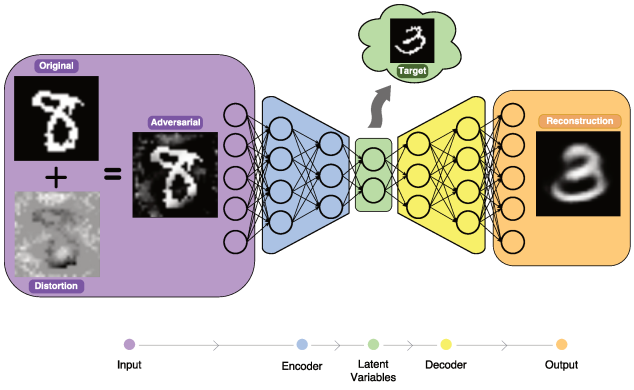


Figure 21 – Adversarial Autoencoder

$$\begin{aligned} \min_{\boldsymbol{d}} \quad & \Delta(\boldsymbol{z_a}, \boldsymbol{z_t}) + C\|\boldsymbol{d}\| \\ \text{s.t.} \quad & L \leq \boldsymbol{x} + \boldsymbol{d} \leq U \\ & \boldsymbol{z_a} = \text{encoder}(\boldsymbol{x} + \boldsymbol{d}) \end{aligned} \tag{B.1}$$

where: $\boldsymbol{z_a}$ and $\boldsymbol{z_t}$ are the latent representations of the adversarial and target images, respectively; C is the regularizing constant; $\boldsymbol{x}$ is the original image; $\boldsymbol{d}$ is the adversarial distortion; $\boldsymbol{x} + \boldsymbol{d}$ is the adversarial image; L and U are the bounds on the input space;

We found that autoencoders are more robust to adversarial attacks and that the trade-off between being similar to the original image and having latent features similar to the target image is almost linear. An example for the MNIST dataset is shown in Fig. 22

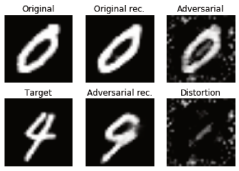


Figure 22 – MNIST example. Distortion is added to the original image of a character 0, creating an adversarial image so that its reconstruction becomes approximately like the image of the character 4.

The code is publicly available at the **github repository**[1].

[1] Available at: https://github.com/tabacof/adv_vae