Surface loss for medical image segmentation

by

Jihene BOUCHTIBA

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER’S DEGREE
WITH THESIS IN AUTOMATED MANUFACTURING ENGINEERING
M.A.Sc.

MONTREAL, MAY 27, 2019

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

Jihene Bouchtiba, 2019
This Creative Commons license allows readers to download this work and share it with others as long as the author is credited. The content of this work cannot be modified in any way or used commercially.
BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

M. Ismail Ben Ayed, Thesis Supervisor
Department of Automated Manufacturing Engineering, École de technologie Supérieure

M. Christian Desrosiers, Co-supervisor
Department of Software and IT Engineering, École de technologie Supérieure

M. Hervé Lombaert, President of the Board of Examiners
Department of Software and IT Engineering, École de technologie Supérieure

M. Matthew Toews, Member of the jury
Department of Systems Engineering, École de technologie Supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MAY 14, 2019

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
ACKNOWLEDGEMENTS

I would like to address special thanks to my supervisor, Professor Ismail Ben Ayed, for his trust, guidance and support during the two last years. Thank you for contributing to my academic and personal development.

I also express my deepest thanks to my co-supervisor Professor Christian Desrosiers for his valuable comments, feedback and supervision.

Many thanks also to Professor Jose Dolz for his help and guidance during my learning process.

I thank the members of the jury Professor Hervé Lombaert and Professor Matthew Toews that kindly accepted to evaluate my master’s thesis.

Special thanks to all my friends from LIVIA and LINCS laboratories for the constant support and the nice moments we shared.

I wish to thank my parents for their endless love and daily encouragement.

I can not forget to express my very deep gratitude to Mr. Gino Belanger, a research agent at the student recruitment and international coordination office. Thank you for your help and kindness.

To finish with, I would like to express my love and gratitude to my husband for his trust, love and endless belief in my success. This accomplishment would not have been possible without you. Thank you.
Une fonction de coût surfacique pour la segmentation des images médicales

Jihene BOUCHTIBA

RÉSUMÉ

Ces dernières décennies ont été témoins d’une expansion sans précédent des données médicales dans divers systèmes complexes et à grande échelle. Malgré toutes les réussites dans la résolution de nombreux problèmes médicaux complexes, il reste encore des défis à relever. Le déséquilibre des classes est l’un des problèmes répandus pour la segmentation des images médicales. Il se produit lorsqu’il y a une distribution de classe sévèrement inégale, par exemple, lorsque la taille de la région cible est beaucoup plus petite (parfois négligeable) que la taille de la région de fond. Dans de tels problèmes, les fonctions de coût typiques utilisées pour la segmentation avec des réseaux neuronaux convolutionnels (CNN) ne parviennent pas à produire de bonnes performances.

Les fonctions de coût couramment utilisées sont basées sur des termes régionaux et supposent que toutes les classes sont réparties de façon équitable. Ainsi, elles tendent à favoriser la classe majoritaire et à mal classifier la classe cible. Pour résoudre ce problème, le principal objectif de ce travail est de construire une fonction de coût surfacique, basée sur la distance entre les frontières et non sur les régions. Nous soutenons qu’une fonction de coût surfacique peut atténuer les problèmes des pertes régionales en introduisant une information complémentaire basée sur la distance. Notre fonction de coût est inspirée des techniques d’optimisation discrètes (basées sur des graphes) pour calculer les flux de gradient de l’évolution des courbes.

Suivant une approche intégrale pour calculer les variations des frontières, nous exprimons une distance non symétrique $L_2$ sur l’espace des formes en tant qu’intégrale régionale, ce qui évite les calculs différentiels complètement locaux. Notre fonction de coût est la somme des fonctions linéaires des sorties régionales de probabilité Softmax du réseau. Par conséquent, elle peut être facilement combinée avec des fonctions de coût régionales et intégrée dans n’importe quelle architecture existante de réseau profond pour la segmentation à N dimensions (N-D).

Nos expérimentations sont menées sur trois bases de données de référence correspondant à des problèmes de segmentation de plus en plus déséquilibrés : segmentation multimodale de la tumeur cérébrale (BRATS17), lésion ischémique cérébrale (ISLES) et hyperintensité de la matière blanche (WMH). Utilisée en conjonction avec la fonction de Dice généralisé (GDL) basée sur la région, notre fonction de coût améliore considérablement les performances par rapport à GDL seul, atteignant jusqu’à 8 % d’amélioration du score de Dice et 10 % d’amélioration du score de Hausdorff. Elle a également aidé à mieux stabiliser le processus d’apprentissage.

Mots-clés: Fonction de coût surfacique, segmentation déséquilibrée, apprentissage profond, techniques d’optimisation discrètes
Surface loss for medical image segmentation

Jihene BOUCHTIBA

ABSTRACT

Last decades have witnessed an unprecedented expansion of medical data in various large-scale and complex systems. While achieving a lot of successes in many complex medical problems, there are still some challenges to deal with. Class imbalance is one of the common problems of medical image segmentation. It occurs mostly when there is a severely unequal class distribution, for instance, when the size of target foreground region is several orders of magnitude less than the background region size. In such problems, typical loss functions used for convolutional neural networks (CNN) segmentation fail to deliver good performances.

Widely used losses, e.g., Dice or cross-entropy, are based on regional terms. They assume that all classes are equally distributed. Thus, they tend to favor the majority class and misclassify the target class. To address this issue, the main objective of this work is to build a boundary loss, a distance based measure on the space of contours and not regions. We argue that a boundary loss can mitigate the problems of regional losses via introducing a complementary distance-based information. Our loss is inspired by discrete (graph-based) optimization techniques for computing gradient flows of curve evolution.

Following an integral approach for computing boundary variations, we express a non-symmetric $L_2$ distance on the space of shapes as a regional integral, which avoids completely local differential computations. Our boundary loss is the sum of linear functions of the regional softmax probability outputs of the network. Therefore, it can easily be combined with standard regional losses and implemented with any existing deep network architecture for N-dimensional segmentation (N-D).

Experiments were carried on three benchmark datasets corresponding to increasingly unbalanced segmentation problems: Multi modal brain tumor segmentation (BRATS17), the ischemic stroke lesion (ISLES) and white matter hyperintensities (WMH). Used in conjunction with the region-based generalized Dice loss (GDL), our boundary loss improves performance significantly compared to GDL alone, reaching up to 8% improvement in Dice score and 10% improvement in Hausdorff score. It also yielded a more stable learning process.

Keywords: Boundary loss, imbalanced segmentation, deep learning, discrete optimization techniques
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 1 LITERATURE REVIEW</td>
<td>5</td>
</tr>
<tr>
<td>1.1 Traditional approaches for medical image segmentation</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Convolution neural networks (CNNs): basic concepts</td>
<td>7</td>
</tr>
<tr>
<td>1.2.1 Discrete convolution</td>
<td>8</td>
</tr>
<tr>
<td>1.2.2 Distinguishing features of CNNs</td>
<td>9</td>
</tr>
<tr>
<td>1.2.3 Layers</td>
<td>10</td>
</tr>
<tr>
<td>1.2.4 Activation functions</td>
<td>12</td>
</tr>
<tr>
<td>1.2.5 Basic Loss functions</td>
<td>15</td>
</tr>
<tr>
<td>1.3 Deep learning for medical image segmentation: Overview, successes and challenges</td>
<td>16</td>
</tr>
<tr>
<td>1.3.1 Deep learning: from natural to medical imaging</td>
<td>16</td>
</tr>
<tr>
<td>1.3.2 Lack of annotated data</td>
<td>19</td>
</tr>
<tr>
<td>1.3.3 Class unbalanced problems</td>
<td>20</td>
</tr>
<tr>
<td>CHAPTER 2 BOUNDARY LOSS FOR HIGHLY UNBALANCED SEGMENTATION</td>
<td>23</td>
</tr>
<tr>
<td>2.1 Motivation</td>
<td>23</td>
</tr>
<tr>
<td>2.2 Methodology</td>
<td>25</td>
</tr>
<tr>
<td>2.2.1 Proposed boundary loss</td>
<td>25</td>
</tr>
<tr>
<td>2.2.2 Scheduling strategy</td>
<td>28</td>
</tr>
<tr>
<td>2.2.3 Network architecture</td>
<td>29</td>
</tr>
<tr>
<td>CHAPTER 3 EXPERIMENTS AND RESULTS</td>
<td>31</td>
</tr>
<tr>
<td>3.1 Datasets</td>
<td>31</td>
</tr>
<tr>
<td>3.2 Experimental protocol</td>
<td>32</td>
</tr>
<tr>
<td>3.2.1 Data Pre-processing</td>
<td>32</td>
</tr>
<tr>
<td>3.2.2 Implementation details</td>
<td>33</td>
</tr>
<tr>
<td>3.3 Results for binary problems</td>
<td>33</td>
</tr>
<tr>
<td>3.4 Results for multi-class problem</td>
<td>34</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>36</td>
</tr>
<tr>
<td>CONCLUSION AND RECOMMENDATIONS</td>
<td>41</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>43</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3.1  DSC and HD values achieved on ISLES and WMH validation subsets. The values represent the mean performance (and standard deviation) of 2 runs for each setting. .......................................... 33

Table 3.2  DSC and HD values achieved on BRATS validation subset. The values represent the mean performance of 2 runs. ................................. 35
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 0.1</td>
<td>A visual comparison that shows the positive effect of our boundary loss on a validation data from the WMH dataset. Our boundary loss helped recovering small regions that were otherwise missed by the generalized Dice loss (GDL). Best viewed in colors.</td>
<td>2</td>
</tr>
<tr>
<td>Figure 0.2</td>
<td>Thesis outline</td>
<td>3</td>
</tr>
<tr>
<td>Figure 1.1</td>
<td>Illustration of a feature map (Taken from: <a href="http://https://www.quora.com/What-is-meant-by-feature-maps-in-convolutional-neural-networks">http://https://www.quora.com/What-is-meant-by-feature-maps-in-convolutional-neural-networks</a>)</td>
<td>7</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Discrete convolution</td>
<td>8</td>
</tr>
<tr>
<td>Figure 1.3</td>
<td>Local connectivity in CNN</td>
<td>10</td>
</tr>
<tr>
<td>Figure 1.4</td>
<td>Max-Pooling with a 2x2 filter</td>
<td>11</td>
</tr>
<tr>
<td>Figure 1.5</td>
<td>Droptout method as introduced by N.Srivastava et al. (2014) (Taken from the same paper)</td>
<td>13</td>
</tr>
<tr>
<td>Figure 1.6</td>
<td>Activation functions</td>
<td>14</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Case of zero Dice</td>
<td>24</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Same Dice with spatial variance</td>
<td>25</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>The relationship between <em>differential</em> and <em>integral</em> approaches for evaluating boundary variation.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Original UNet architecture (taken from ( Ronneberger et al. (2015b))</td>
<td>29</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Evolution of DSC and HD values on WMH/ISLES validation subsets. The blue curve shows the performance of the network trained using the GDL loss, while the red curve represents the optimization process with the GDL combined with our proposed boundary loss.</td>
<td>34</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Evolution of DSC and HD values on BRATS validation subset. The blue curve shows the performance of the network trained using the GDL loss, while the red curve represents the optimization process with the GDL combined with our proposed boundary loss.</td>
<td>35</td>
</tr>
</tbody>
</table>
Figure 3.3  Qualitative segmentation results on BRATS validation set. Colors correspond to different tumour regions: Red for edema (ED), green for Non enhancing tumor (NET). Best viewed in colors. ...................... 36

Figure 3.4  Effect of our boundary loss on false positives/negatives. Best viewed in colors. ......................................................... 37

Figure 3.5  Visual results from WMH and ISLES validation sets. ......................... 38

Figure 3.6  Visual results for our boundary loss as a stand alone loss trained only on positive samples on WMH dataset .................................. 39
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>Cross Entropy</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Network</td>
</tr>
<tr>
<td>ETS</td>
<td>École de Technologie Supérieure</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphical Processing Unit</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi Layer Perceptron</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified Linear Unit</td>
</tr>
<tr>
<td>SGD</td>
<td>Stochastic Gradient Descent</td>
</tr>
<tr>
<td>WCE</td>
<td>Weighted Cross Entropy</td>
</tr>
</tbody>
</table>
INTRODUCTION

Problem statement

Driven by the revolution of big data and computing power, deep learning techniques have shown their ability to deliver outstanding performances in different fields where machines become more suitable than humans dealing with a huge amount of data in a comprehensive way. The medical field is one of the fields where deep learning had the largest impact.

Last years have witnessed a revolution in medical imaging that has broken paradigms and seems to open new possibilities in the field. Recent studies estimate that medical images currently account for at least 90% of medical data, which is an overwhelming amount on human scale. The detection, classification and segmentation of specific medical patterns, e.g. organs, tissues, tumors are the main key steps for diagnosis, treatment or monitoring of different diseases progression. For instance, an accurate delineation of one or multiple structures is essential to derive important clinical parameters such as ventricular volume, ejection fraction, and myocardial mass for cardiac images or lesion volume for brain images. However, the manual segmentation of these structures is complex, time consuming and requires multiple expert opinions. In some cases, the annotation may not be possible due to the absence of experts or due to the rarity of the disease or lesion to be segmented. To address this problem, deep neural networks have shown their ability to deliver outstanding performances in such tasks.

Recently, a substantial growth in the number of deep learning techniques has been witnessed. However, while achieving a lot of successes in many complex medical problems, researchers have realized that there are still some challenges to deal with. One of them is the data unbalance problem, which occurs for instance when the size of target foreground region is several orders of magnitude less than the background region size. Such problem occurs mostly in cases of abnormalities detection/segmentation, e.g. tumors (M.Havaei et al. (2015); Pereira et al. (2016);
Menze et al. (2014); Hwang & Kim (2016); Bellver et al. (2017)), the breast (Kallenberg et al. (2016); Wang et al. (2016); Hassanien et al. (2014)), the lungs (Kalinovsky & Kovalev (2016); Mi et al. (2014); Skourt et al. (2018)), etc. For example, in the characterization of white matter hyperintensities (WMH) of presumed vascular origin, the foreground composed of target regions may be 500 times smaller than the background (see the typical example in Fig 0.1). In such case, common approaches, assuming that all the patterns are evenly distributed, fail to deliver good segmentation performance. The issue is that they tend to favour the majority class and, then, misclassify the target class.

Figure 0.1 A visual comparison that shows the positive effect of our boundary loss on a validation data from the WMH dataset. Our boundary loss helped recovering small regions that were otherwise missed by the generalized Dice loss (GDL). Best viewed in colors.

**Research hypotheses and objectives**

All common proposed strategies to re-balance class prior distributions are region-based. In this work, our objective is to propose a boundary loss that takes the form of a distance metric on the space of contours (or shapes), not regions. We argue that a boundary loss can mitigate the issues related to regional losses in highly unbalanced segmentation problems. Rather than using unbalanced integrals over the regions, a boundary loss uses integrals over the boundary (interface)
between the regions. Furthermore, a boundary loss provides information that is complimentary to regional losses. It is, however, challenging to represent the boundary points corresponding to the regional softmax outputs of a CNN. This difficulty may explain why boundary losses have been mostly avoided in the context of deep segmentation networks. Our boundary loss is inspired by techniques in discrete (graph-based) optimization for computing gradient flows of curve evolution (Boykov et al., 2006). Following an integral approach for computing boundary variations, we express a non-symmetric L2 distance on the space of shapes as a regional integral, which avoids completely local differential computations involving contour points. Our boundary loss is the sum of linear functions of the regional softmax probability outputs of the network. Therefore, it can be easily combined with standard regional losses and implemented with any existing deep network architecture for N-D segmentation.

Figure 0.2  Thesis outline
Thesis outline

This document is organized as shown in Fig.0.2. Chapter one introduces the literature review where we report some basics of CNNs and discuss their major impact on medical image analysis, showing both successes and current challenges. In the second chapter, we propose a new boundary loss for highly unbalanced problems, and explain in details the methodology that we followed. The next chapter is dedicated to the experimental protocol, where we show our results on different datasets. Finally, we conclude with a summary of the work and a discussion of some possible future directions.
CHAPTER 1

LITERATURE REVIEW

In this chapter, we introduce the literature review. First, we discuss some traditional techniques used for medical image segmentation. Then, we report some basics of convolutional neural networks. In the following section, we give an overview of the fast-growing place of deep learning techniques in medical image analysis, showing various successes and also discussing two main challenges faced by current research: lack of annotations and unbalanced problems.

1.1 Traditional approaches for medical image segmentation

Image segmentation is the process of delineating one or multiple semantic structures in the image. It is one of the most active research tasks in medical image processing where it facilitates the extraction of useful informations for diagnostic medicine and biomedical imaging applications such as pathology localization, analysis of the anatomical structure and computer-assisted surgery. Several traditional approaches have been introduced in the literature and can be divided into four main categories: thresholding-based methods, atlas-guided methods, statistical approaches and deformable models.

Based on the assumption that target objects can be identified by their intensity values, thresholding, as a region-based technique, has been largely used to segment medical images. The idea is to apply a threshold value on pixels intensities to extract the target. Typically, this value corresponds to a peak in the histogram that divides an image into two different parts: pixels with intensities greater than a given threshold \( T \) correspond to a first class while remaining pixels belong to a second class.

The main advantage of thresholding is its simplicity and rapidity. However, global thresholding assumes that the image has a bimodal histogram, which is not the case in many medical images, especially when the background is not homogeneous or the object presents a lot of intensity variation.
To address that, local thresholding has been introduced (N & S (2016); Aja-Fernández et al. (2015)). Authors in (Wu et al. (1995); Manikandan et al. (2014); Xie & Bovik (2013)) proposed the so-called Otsu’s method for an automated selection of the threshold. One of the common weaknesses of thresholding approach is its strong dependence to the database. Also, thresholding does not take into account the spatial information, which leads to poor segmentation results especially with noisy images. Another popular approach to segment medical images is the use of atlases. The idea is to build a reference model from some training data and then register the unseen image to the reference model. Registration is the process of aligning the image so that corresponding features can be easily related and distances can be computed. Although they often deliver good performances, atlas-based methods suffer from complex registration steps. Furthermore, it is not easy to cover all the anatomical variability of target objects e.g., rare diseases or anatomical abnormalities. Thus, atlas-based methods can fail to deliver good performances when dealing with huge deformations.

Over the last decades, the explosion of medical datasets called for new solutions. In that context, statistical approaches propose a parametric model, which describes the variability of the target object (shape, texture, size, etc). Training data is used during the learning process. Despite their success in various medical applications, such techniques can lead to an over fitting problem when the size of data is limited comparing to the number of parameters. Unlike statistical techniques, deformable models do not need a training data. They are defined as dynamic contours that evolve to fit a target shape via an energy-minimization strategy. Deformable models can be divided into two main categories; parametric and geometric models. In the first category, we find the well known active contours widely used for medical image segmentation (McInerney & Terzopoulos (1996); Yushkevich et al. (2006); Wang et al. (2009); He et al. (2008). In a different strategy, geometric models use a distance transformation to describe the target shape. In other words, they rely on the so-called level-set based shape representation (Zhang et al. (2008); Droske et al. (2001); Lin et al. (2003); Li et al. (2011); Mitiche & Ben Ayed (2011)). The most straightforward advantage of deformable models is their flexibility and their usefulness in specific tasks, e.g., object tracking. However, they are
still sensitive not only to the initialization but also to the presence of noise, when the initial boundary is far from the target, they might get stuck in a local minima.

1.2 Convolutional neural networks (CNNs): basic concepts

In image analysis, one common idea between traditional approaches is that they all need to extract features in a separated way before applying another method for classification. A logical next step is to let the machine learn these features in an automatic fashion. This is the key idea behind most deep learning techniques. In simple words, deep neural networks (DNNs) are models composed of many successive layers in order to transform an input data to an output while learning increasingly higher level features. Convolutional neural networks (CNNs) are among the most successful models for image analysis, in general, and for image segmentation particularly.

A CNN is a multilayer perceptron (mlp) designed to use a fewer amount of parameters than a fully connected network by taking advantage of the spatial relationships between neighboring pixels in an image. Typically, the network is a set of convolutional and fully connected layers. It can also include the so-called pooling layers.
Each convolutional layer is composed of small neuron collections (called receptive fields) which process portions of the input image and produce a set of outputs called feature maps (see Fig. 1.1). A feature map is obtained by a local convolution operation using a linear filter \( K \), adding a bias term and then applying an activation function.

After introducing discrete convolution, we discuss the basic components of a CNN.

### 1.2.1 Discrete convolution

Discrete convolution is a powerful mathematical tool used in many areas of math and engineering, including computer vision. Applied to image analysis, the idea is to use a kernel (or filter) \( K \) and slide it over the image \( I \). For each position of the kernel, we multiply the overlapping values of the kernel and image together, and add up the results. This sum of products will be the value of the output image (see Fig. 1.2). The discrete convolution of an image \( I \) with a kernel \( K \) of size \((2h_1 + 1) \times (2h_2 + 1)\) can be given by:

\[
(I * K)_{p,q} := \sum_{u=-h_1}^{h_1} \sum_{v=-h_2}^{h_2} K_{u,v} I_{p+u,q+v} \quad (1.1)
\]
Where the filter $K$ is given by:

$$K = \begin{pmatrix}
K_{-h1,-h2} & \cdots & K_{-h1,h2} \\
\vdots & \ddots & \vdots \\
K_{h1,-h2} & \cdots & K_{h1,h2}
\end{pmatrix}$$  \hspace{1cm} (1.2)

1.2.2 Distinguishing features of CNNs

The first question that can be asked about CNNs is the following: **What is the difference between a CNN and an MLP?**

A CNN is an MLP with fewer parameters and a special structure that aims to reduce computational complexity and ensure a spatial invariance, i.e., the network performs the input regardless of its spatial location in the image. In other words, a car is a car regardless of its position in the image. Convolutional neural networks have two main distinguishing features.

**Local connectivity:** In the case of high-dimensional inputs such as images, a fully connected hidden layer would have unmanageable parameters. Also, computing activation functions in respective neurons would be very expensive. In order to avoid these problems, CNNs take advantage of the local spatial coherence property of images and enforce a local connectivity pattern between neurons of adjacent layers. In other words, neurons in one layer are only connected to neurons in the next layer that are spatially close to them (see Fig. 1.3). Basically, we assume that two distant pixels have less significant relationship that close pixels.

**Parameter sharing:** The idea behind weights sharing is the convolutional concept of layers. The idea behind this property is to control and reduce the number of parameters by making an assumption: certain neurons are aggregated together and share the parameters matrix (weighted connections) forming, then, a feature map. This concept allows features to be detected regardless of their position in the image. Parameter sharing reduces the number of free parameters, achieving better generalization and computation performance.
1.2.3 Layers

Convolutional layer

The convolutional layer is the core building block of a CNN. It is designed to learn hierarchical features from the image. In general, the input is a feature map. For the first layer, it would be the initial image. Let layer $l$ be a convolutional layer. Then, the input layer contains $m^{(l-1)}_1$ feature maps, each of size $m^{(l-1)}_2 \times m^{(l-1)}_3$. The output of layer $l$ consists of $m^{(l)}_1$ feature maps, each of size $m^{(l)}_2 \times m^{(l)}_3$. The $i^{th}$ feature map in layer $l$, denoted $Y^{(l)}_i$, is computed as:

$$Y^{(l)}_i = B^{(l)}_i + \sum_{j=1}^{m^{(l-1)}_1} K^{(l)}_{i,j} * Y^{(l-1)}_i$$  \hspace{1cm} (1.3)

Where $B^{(l)}_i$ is the bias matrix introduced as an external input. $K^{(l)}_{i,j}$ is the filter of size $(2h^{(l)}_1 + 1) \times (2h^{(l)}_2 + 1)$ used to connect the $j^{th}$ feature map in layer $l - 1$ with the $i^{th}$ feature map in layer $l$.

The size of output feature maps is influenced by borders effect. In other words, the discrete convolution can be applied only on pixels where the sum of equation (1.3) is defined properly.
Thus, output feature maps have size

$$m_2^{(l)} = m_2^{(l-1)} - 2h_1 + 2p$$

and

$$m_3^{(l)} = m_3^{(l-1)} - 2h_1 + 2p$$

(1.4)

Where $p$ denotes the amount of zero padding used.

**Pooling and sub-sampling layer**

Another important feature in CNNs is the pooling operation. It aims at reducing the size of extracted feature maps via a non linear subsampling function. Max-pooling is the most common used function. As illustrated in Fig. 1.4, max-pooling consists in partitioning the input feature map into a set of non-overlapping rectangles. For each rectangle, the maximum value is retained as output. Max pooling is then, useful for two reasons; first, it reduces computation for the next layers. It assures also a form of local translation invariance by eliminating non-maximal values. In addition to this technique, pooling can also be performed with other functions such as average pooling which takes the average value in each rectangle.

![Figure 1.4 Max-Pooling with a 2x2 filter](image)

**ReLu layer**

In order to improve processing efficiency, a ReLu (Rectified linear Unit) layer is typically used. It is a layer of neurons with the non-saturating activation function $f(z) = \max(0,z)$. This
rectifier has been argued to be more practical than many other functions since it adds a non linearity to the network while allowing faster and effective training for deep neural architectures on large and complex datasets.

**Fully connected layer**

After alternating between several convolutional and pooling layers, the final classification is typically done via fully connected layers. Neurons in a fully connected layer \( l \) have connections with all neurons in the previous layer \( l - 1 \), as seen in multi-layer perceptrons. Generally, one or more fully connected layers are used for classification purposes based on computed features in previous convolutional layers.

**Dropout layer**

Neural networks contain many non linear layers and a lot of parameters to adjust during training. With a limited training data, an overfitting problem can occur. It consists in a poor predictive performance; the system overreacts to small fluctuations in the training data. Thus, the system is not able to perform well on unseen data. To solve this problem, the dropout technique was proposed by (N.Srivastava et al. (2014)) as a simple method to prevent neural networks from overfitting. The idea is illustrated in Fig. 1.5. In a given layer, some neurons are dropped out, that is set to zero. For each training example, a different set of neurons to drop is randomly chosen. This technique is used only in the training step. At test time, all units are used.

**1.2.4 Activation functions**

There is a variety of common activation functions used in neural networks. They can be divided into two main categories:

- Linear functions, e.g, the identity:

\[
\begin{align*}
  f(z) &= z
\end{align*}
\] (1.5)
When used in a multilayer architecture, linear functions make the network equivalent to a single-layer model and fail to solve complex non-linear problems.

- Non linear functions: For classification purposes, Sigmoid functions are often used. In fact, a sigmoid function is a bounded differentiable function defined for all real input values and has an 'S' shaped curve. Thus, a large variety of sigmoid functions have been used as activation functions including the logistic and hyperbolic tangent functions which are linearly-related by:

\[
2f(z) = 1 + \tanh\left(\frac{z}{2}\right) \tag{1.6}
\]

where \( f \) is the logistic function defined as:

\[
f(z) = \frac{1}{1 + \exp(-z)} \tag{1.7}
\]
and tanh is the hyperbolic function:

\[
\tanh(z) = \frac{1 - \exp(-2z)}{1 + \exp(-2z)}
\]  

(1.8)

Softsign is another activation function:

\[
f(z) = \frac{z}{1 + |z|}
\]  

(1.9)

Figure 1.6 Activation functions

In deep learning, the choice of the activation function is a challenging task since it affects the training quality and the hole system performance. A comparative study between activation
functions in deep architectures is presented in (Glorot & Bengio (2010)). Experiments show that logistic and hyperbolic functions are less suited for deep learning due to their fast saturation and their slow convergence. Better performance is reported with softsign function.

The network seems to be more robust to the initialization procedure. Although these smooth non linear functions were mathematically inspired by the biological neuron behavior, most of current methods are using the ReLU function $f$ defined by:

$$f(z) = \max(0, z)$$

Authors in (Skourt et al. (2018); Paszke et al. (2016)) argue that it achieves better results in deep networks as it makes the activation sparse and more efficient. However, the horizontal line of the function (where $z < 0$) causes some problems when the activation function stacks in this region. In such case, the gradient of corresponding neurons will be zero and corresponding weights would not be adjusted during descent. This is called the Dying ReLU problem. To mitigate that, a modified version called Leaky RELU (LReLU) was proposed by Xu et al. (2015) introducing a light slope in the negative part of the function.

### 1.2.5 Basic Loss functions

Network training is an important task in supervised learning. It consists in approximating the network parameters (weights) using a training dataset composed of $N$ labelled samples:

$$T_N := (\mathbf{i}_n, \mathbf{g}_n) : 1 \leq n \leq N$$

where $\mathbf{i}_n$ is the network input and $\mathbf{g}_n$ is the desired output.

The objective is to adjust the network parameters $\theta$ in order to minimize the so-called loss function that evaluates the gap between the network output $\mathbf{s}_{\theta,n}$ and the desired output $\mathbf{g}_n$ for each training sample $\mathbf{i}_n$. The loss function is used to compute the network errors. Commonly used
mathematical expressions include the mean squared loss \((MSL)\), the cross entropy\((CE)\) and more recently the Dice loss \((DL)\).

The mean squared error measure is given by :

\[
MSL(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} (s_{\theta,n,k} - g_{n,k})^2 
\]  

(1.12)

where \(d\) is the number of labels per sample. In practice, the mean squared (or quadratic) measure gives good results. However, if the sigmoid activation function is used, the quadratic loss would converge slowly to the final output. To resolve this problem, the \(CE\) was proposed:

\[
CE(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{d} g_{n,k} \log(s_{\theta,n,k}) 
\]  

(1.13)

More recently (Milletari et al. (2016)), the Dice measure, initially used as a similarity metric to evaluate the segmentation quality, was introduced as a loss function for two-class segmentation problem. It is defined as:

\[
DL = 1 - \frac{\sum_{n=1}^{N} s_{\theta,n} g_{n} + \varepsilon}{\sum_{n=1}^{N} s_{\theta,n} + g_{n} + \varepsilon} - \frac{\sum_{n=1}^{N} (1 - s_{\theta,n})(1 - g_{n}) + \varepsilon}{\sum_{n=1}^{N} 2 - s_{\theta,n} - g_{n} + \varepsilon} 
\]  

(1.14)

Where \(\varepsilon\) is a small constant to avoid the problem of dividing by zero.

1.3 Deep learning for medical image segmentation: Overview, successes and challenges

1.3.1 Deep learning: from natural to medical imaging

The few last decades have witnessed a digital revolution and fast technical advances in medical imaging with the explosion of new imaging modalities, e.g, X ray computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound (US). Hence, experts are no longer able to manually process the huge amount of data and extract useful informations. Searching for new paths, the medical image community has been attracted by the astonishing successes of
deep learning techniques in natural imaging and have applied a lot of these findings to improve medical image analysis for different tasks including classification, detection, segmentation and registration. Image segmentation is one of the most important tasks in medical image analysis. It consists of partitioning the image into different segments that often correspond to organs, tissue categories, pathologies or other biological structures. An accurate delineation allows efficient quantitative analysis of important clinical measures related to volume or shape as in cardiac or cerebral image analysis. Such measures are very useful for diagnostic medicine and clinical interventions.

Recently, deep learning methods have shown their ability to deliver excellent performances in such task. Many methods have been proposed for segmentation of specific organs or other substructures, e.g., brain (Pereira et al. (2016); Brébisson & G. Montana (2015); M. Havaei et al. (2015); Chen et al. (2018); Ganaye et al. (2018)), heart cavities and muscles (M. Havaei et al. (2015); Yu et al. (2017a); Avendi et al. (2016)), vertebrae (Chen et al. (2015)), bone (Minnema et al. (2018); Zhang et al. (2010); Klein et al. (2018)) or abnormalities such as tumors (M. Havaei et al. (2015); Pereira et al. (2016); Menze et al. (2014)) or lesions (Hwang & Kim (2016); Bellver et al. (2017)). During the last years, thousands of new papers have been published. However, it is worth to note that most of them suffer from reproducibility problems. This can be either a problem of methodology or also lack of details in the published paper as the use of hyper-parameters or heuristics specific to a particular dataset. In some cases, researchers can validate their method only on some selected scans and omit the rest which biases their results. One way to guarantee a fair and direct comparison of proposed methods on the same dataset and evaluation procedure is to organize public challenges. From 2007, around 36 public challenges\(^1\) and 118 hidden ones have been organized to deal with specific medical image analysis problems, e.g., brain lesion segmentation (BRATS\(^2\) (2012-2018), ISLES\(^3\), WMH\(^4\),

---

\(^1\) [https://grand-challenge.org/stats](https://grand-challenge.org/stats)
\(^2\) [http://braintumorsegmentation.org/](http://braintumorsegmentation.org/)
\(^4\) [http://wmh.isi.uu.nl](http://wmh.isi.uu.nl)
MSSEG\(^5\), cardiac image segmentation (ACDC\(^6\), LVSC\(^7\), MM-WHS 2017\(^8\), CETUS2014\(^9\)), liver segmentation (CHAOS\(^10\)), etc. Deep neural networks have been the winners of most of the challenges.

The typical idea of a DNN-based approach is to train one or more deep networks to assign each pixel (2d input) or voxel (3D input) to its corresponding anatomical region based on different input features. As the architectural choice is an important key for a good segmentation, dozens of different designs have been introduced. Initially, the common used strategy was patch-based (Avendi et al. (2016); Pereira et al. (2016); Brébisson & G.Montana (2015)). It consists in the use of the neighborhood of each pixel in the image as input (for example: 3x3 or 5x5 window centered on the pixel, called a patch). A variety of features was proposed: 2D and 3D intensity patches (Brébisson & G.Montana (2015); Lai (2015); Kleesiek et al. (2016)), orthogonal views of MRI scans (A.Prasoon et al. (2013)). One straightforward drawback of a sliding-window strategy is that it doesn’t take into account the global context of the image which is needed to segment medical images. By changing the last fully connected layers by convolutional ones, the DNN becomes able to process the entire image in a single forward pass and produce a probability map. This was the idea of a fully convolutional neural network proposed by (J.Long et al. (2015)) for natural image segmentation. Influenced by the FCNN idea, Ronneberger et al. (Ronneberger et al. (2015a)) have proposed the famous U-net architecture to segment biomedical images. Authors combined a regular FCNN with an up-sampling path where "deconvolutions" are applied to build an output of the same size than the input. The particularity of a U-net is the insertion of "skip connections" to connect opposing; convolutional and deconvolutional, layers between the down-sampling and up-sampling paths. To exploit the volumetric context of medical data, an extension to 3D U-Net was suggested (Çiçek et al.

\(^5\) https://portal.fli-iam.irisa.fr/msseg-challenge/overview
\(^6\) https://www.creatis.insa-lyon.fr/Challenge/acdc/index.html
\(^7\) http://www.cardiacatlas.org/challenges/lv-segmentation-challenge/
\(^8\) http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/
\(^9\) https://www.creatis.insa-lyon.fr/Challenge/CETUS/index.html
\(^10\) https://chaos.grand-challenge.org/
Milletari et al. (Milletari et al. (2016)) have proposed an other variant of U-Net for 3D prostate segmentation. In this study, authors used a loss function based on the Dice coefficient to mitigate the problem of imbalanced data.

1.3.2 Lack of annotated data

Fully supervised approaches (Çiçek et al. (2016b); J.Long et al. (2015); Avendi et al. (2016); Brébisson & G.Montana (2015); M.Havaei et al. (2015); LC.Chen & G.Papandreou (2015)) rely on available ground truth sets derived generally from comprehensive manual annotations by experts. Initially, the majority of researches have focused on supervised deep learning. However, they require huge training dataset, which is not easy to meet in the medical field. The non-availability of large and high-quality labeled data still be the biggest barrier for the performance of deep learning in medical imaging. In fact, an accurate annotation is a difficult and time consuming task. An expert may spend 3 or 4 days to segment a full brain. On the other hand, manual annotations require multiple expert opinions to get over the human error. In some cases, annotation may not be possible due to the absence of experts or due to the rarity of the target disease or lesion. To address this problem, different methods and techniques have been proposed, e.g. transfer learning (Tran (2016); Shin et al. (2016); Yu et al. (2017a); Ghafoorian et al. (2017); Tajbakhsh et al. (2019)) and data augmentation (Asperti & Mastronardo (2018); Vasconcelos & Vasconcelos (2017); Tran (2016); Pereira et al. (2016); Vasconcelos & Vasconcelos (2017)).

Transfer learning is a common strategy that has shown its usefulness especially in cases of limited labeled data. The basic idea is to use a pre-trained network on a large dataset and re-apply it on another dataset. Typically, a fine-tuning step is performed to adapt the network to the given task of interest. In (Yu et al. (2017a)), authors use dynamic CNNs to segment fetal left ventricle in echocardiographic sequences. Initially, a pre-trained model is fine-tuned by a deep tuning using the first annotated frame then a shallow tuning for the remaining frames. Conducted on 51 echocardiographic sequences, dynamic networks achieve good results in terms of Dice accuracy of 94.5%. As another way to mitigate the lack of annotations, recent methods
have focused more on semi and weak supervision. A variety of weak labels such as image tags (Pathak et al. (2015); Papandreou et al. (2015)), partial labels, bounding boxes (Shah et al. (2018); Dai et al. (2015); Papandreou et al. (2015)) and scribbles (Wang et al. (2017a,b); Kervadec et al. (2019)) have been used. In a very recent work, Kervadec et al. (2019) exploit weak annotations, e.g., scribbles and dots for semantic segmentation. They have proposed new constrained losses based on anatomical priors (shape, size) to guide the learning process during training. Validated on three medical applications, the method showed promising results close to full supervision performances.

1.3.3 Class unbalanced problems

While achieving a lot of successes in many complex medical problems, researchers have realized that there are still some challenges with deep learning techniques to deal with. In addition to lack of annotations discussed previously, another major issue is the data unbalance that occurs when the target is underrepresented in the image. This is very common in health field, for instance, when the task consists in segmenting abnormalities, rare diseases or small regions. For example, to segment a brain cancer, the frequency of the target class (cancer) can be 1000 times less than the majority class (healthy patient). Assuming that all the classes are equally distributed, classic classifiers are highly biased toward the majority class and fail to deliver good segmentation performances. To combat this problem, three main strategies have been suggested; methods operating directly on the data via down-sampling or over-sampling, methods operating on the learning process and methods introducing specific loss functions. A common solution; which falls into the first category, is to re-balance the data by down-sampling frequent labels (Havaei et al. (2017); Valverde et al. (2017)). Others perform a data augmentation on target samples (Kamnitsas et al. (2017); Litjens et al. (2016); Pereira et al. (2016)).

From a different point of view, authors in (M.Havaei et al. (2015)) propose a two-phase training for brain tumor segmentation; the first is performed using equi-probable patches to adjust lower layers parameters. Then, in a second step, they re-train only the output layer to take into
account the unbalanced nature of real patches. The same strategy was used in (Minnema et al. (2018)) to segment bone in CT scans. A considerable number of recent works has focused on the adaptation of the loss function to alleviate data imbalance issue. A common practice is to assign weights to the different classes, inversely proportional to the frequency of the corresponding labels (Brosch et al. (2015); Ronneberger et al. (2015b); Kamnitsas et al. (2017); Yu et al. (2017b)). In (Ronneberger et al. (2015b)), the standard cross entropy (CE) loss is modified by adding weights that assign more importance to the rare class than to the majority class. In a distinguished way, Sudre et al. (Sudre et al. (2017)) generalized the Dice loss (GDL) initially proposed by Milletari et al. (Milletari et al. (2016)). To have better trade-off between precision and recall, some recent works (Salehi et al. (2017)) investigated losses based on the Tversky similarity index. Inspired by the concept of focal loss (Lin et al., 2018), Dice and Tversky losses have been extended to integrate a focal term, which is parameterized by a value that controls the importance between easy and hard training samples (Abraham & Khan (2018); Wong et al. (2018); Hashemi et al. (2018)).
CHAPTER 2

BOUNDARY LOSS FOR HIGHLY UNBALANCED SEGMENTATION

In the present chapter, we propose a new method that tackle imbalanced medical segmentation problems. First, we describe the motivation behind the idea of our method. Then, we explain the proposed methodology via a detailed mathematical background.

2.1 Motivation

During the last decades, fully deep neural networks have shown their great potential to deliver the best performances in medical image analysis. However, one of the major challenges encountered is the data unbalance problem where the size of target region is several orders of magnitude less than the background size. Such problem occurs mostly in cases of abnormalities detection/segmentation, e.g. tumors segmentation (brain (M.Havaei et al. (2015); Pereira et al. (2016); Menze et al. (2014); Hwang & Kim (2016); Bellver et al. (2017)), breast (Kallenberg et al. (2016); Wang et al. (2016); Hassani et al. (2014)), lung (Kalinovsky & Kovalev (2016); Mi et al. (2014); Skourt et al. (2018)), etc). For example, to segment a brain cancer, the frequency of the target class (cancer) can be 1000 times less than the majority class (healthy patient). In such cases, common approaches, assuming that all patterns are evenly distributed, fail to deliver good segmentation performance. The issue is that they tend to favour the majority class and, then, misclassify the target class.

As discussed previously in section (1.3.3), similarity loss functions have been proposed to mitigate the unbalance problem, e.g., weighted cross entropy (Ronneberger et al. (2015b)) and the generalized dice loss (GDL) (Sudre et al. (2017)). Although effective for some unbalanced problems, weighted cross entropy may undergo serious difficulties when dealing with highly unbalanced datasets. The CE gradient computed over the few pixels of infrequent labels is typically noisy, and amplifying this noise with a high class weight may lead to instability. To address this problem, many recent researches (Sudre et al. (2017); Milletari et al. (2016) relied on GDL as an alternative. Despite the improvements over CE, the dice loss has some
limitations in a highly unbalanced problem. Initially defined as an overlap measure to evaluate the network performance, the dice coefficient accounts only the correctly classified pixels/voxels without taking into account the spatial distribution. As shown in Fig.2.1 and Fig.2.2, Dice coefficient gives same values in both cases regardless the distance that separates segments. Thus, when used as a loss function, the information of how far is the prediction from the ground truth is missing. Moreover, regional dice losses still be not convenient for small region segmentation such as lesions, tumors or small structures. In fact, the probability of low overlap between two small regions is high. In such case, dice loss would give a small value regardless of how far regions are from each other.

Based on all the above-mentioned limitations of region-based losses, we propose a new boundary loss that takes the form of a distance metric on the space of contours (or shapes), not regions. We argue that a boundary loss can mitigate the issues related to regional losses in highly unbalanced segmentation problems. Moreover, it provides information that is complimentary to regional losses.

Figure 2.1 Case of zero Dice
2.2 Methodology

2.2.1 Proposed boundary loss

Here, we introduce the formulation of the new proposed loss. Let $I : \Omega \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$ be an input image with spatial domain $\Omega$, and $g : \Omega \rightarrow \{0,1\}$ a binary ground-truth segmentation of the image: $g(p) = 1$ if pixel/voxel $p$ belongs to the target region $G \subset \Omega$ (foreground region) and 0 otherwise. Let $s_\theta : \Omega \rightarrow [0,1]$ denotes the softmax probability output of a deep segmentation network, and $S_\theta \subset \Omega$ the corresponding segmentation region: $S_\theta = \{p \in \Omega | s_\theta(p) \geq \delta\}$ for some threshold $\delta$. Widely used segmentation loss functions involve a regional integral for each segmentation region in $\Omega$, which measures some similarity (or overlap) between the region defined by the probability outputs of the network and the corresponding ground-truth. In the two-region case, we have an integral of the general form $\int_\Omega g(p)f(s_\theta(p))dp$ for the foreground, and of the form $\int_\Omega (1-g(p))f(1-s_\theta(p))dp$ for the background. The generalized Dice loss (GDL) (Sudre et al. (2017)) involves regional integrals with $f = 1$, subject to some normalization, and is given as follows for the two-region case:

$$\mathcal{L}_{GD}(\theta) = 1 - 2 \frac{w_G \int_{p \in \Omega} g(p)s_\theta(p)dp + w_B \int_{p \in \Omega} (1-g(p))(1-s_\theta(p))dp}{w_G \int_{\Omega}[s_\theta(p) + g(p)]dp + w_B \int_{\Omega}[2-s_\theta(p) - g(p)]dp} \quad (2.1)$$
where coefficients \( w_G = 1/\left( \int_{p \in \Omega} g(p) \, dp \right)^2 \) and \( w_B = 1/\left( \int_{\Omega} (1 - g(p)) \, dp \right)^2 \) are introduced to reduce the well-known correlation between the Dice overlap and region size.

Now, our objective is to build a boundary loss \( \text{Dist}(\partial G, \partial S_\theta) \), which takes the form of a distance metric on the space of contours (or region boundaries) in \( \Omega \), with \( \partial G \) denoting a representation of the boundary of ground-truth region \( G \) (e.g., the set of points of \( G \), which have a spatial neighbor in background \( \Omega \setminus G \)) and \( \partial S_\theta \) denoting the boundary of the segmentation region defined by the network output. On the one hand, a boundary loss should be able to mitigate the above-mentioned difficulties for unbalanced segmentations: rather than using unbalanced integrals within the regions, it uses integrals over the boundary (interface) between the regions.

The main challenge here is how to represent this surface between the two boundaries \( \partial S_\theta \) and \( \partial G \) as a differentiable function that can accommodate standard stochastic optimizers such as SGD. This challenge might explain why distance-based losses have been, to the best of our knowledge, largely avoided in the context of deep segmentation networks.

![Figure 2.3](image_url)  
Figure 2.3 The relationship between differential and integral approaches for evaluating boundary variation.
Our boundary loss is inspired from discrete (graph-based) optimization techniques for computing gradient flows of curve evolution (Boykov et al. (2006)). Similarly to our problem, curve evolution methods aim to evaluate boundary variations. The typical solution to compute the distance between contours is based on a variational method where we should compute each pixel’s local differential motion (see illustration in Fig. 2.3). However, similarly to any contour distance invoking directly points on boundary $\partial S$, the computed distance can not be used as a loss function. For that, we propose to use an approximation that expresses the differential boundary variation as a regional integral:

$$\text{Dist}(\partial G, \partial S) = 2 \int_{\Delta S} |\phi_G(p)| dp$$

(2.2)

where $\Delta S$ denotes the region between the two contours and $\phi_G : \Omega \to \mathbb{R}$ is a level set representation of boundary $\partial G$: $\phi_G(p)$ evaluates a signed distance between point $p \in \Omega$ and the nearest point $z_{\partial G}(p)$ on contour $\partial G$: $\phi_G(p) = -\|p - z_{\partial G}(p)\|$ if $p \in G$ and $\phi_G(p) = \|p - z_{\partial G}(p)\|$ otherwise. Fig. 2.3.b illustrates this integral framework for evaluating the boundary distance.

Thus, the non-symmetric $L_2$ distance between contours can be expressed as a sum of regional integrals:

$$\int_S \phi_G(p) dp - \int_G \phi_G(p) dp = \int_\Omega \phi_G(p)s(p) dp - \int_\Omega \phi_G(p)g(p) dp$$

(2.3)

where $s : \Omega \to \{0, 1\}$ is binary indicator function of region $S$: $s(p) = 1$ if $p \in S$ belongs to the target and 0 otherwise. Now, for $S = S_\theta$, i.e., replacing binary variables $s(p)$ in Eq. (2.3) by the softmax probability outputs of the network $s_\theta(p)$, we obtain the following boundary loss which, up to a constant independent of $\theta$, approximates boundary distance $\text{Dist}(\partial G, \partial S_\theta)$:

$$\mathcal{L}_B(\theta) = \int_\Omega \phi_G(p)s_\theta(p) dp$$

(2.4)

Notice that we omitted the last term in Eq. (2.3) as it is independent of network parameters. The level set function $\phi_G$ is pre-computed directly from the ground-truth region $G$. In practice, our boundary loss in Eq. (2.4) is the sum of linear functions of the regional softmax probability...
outputs of the network. Therefore, it can be easily combined with standard regional losses and implemented with any existing deep network architecture for N-D segmentation. In the experiments, we tested our boundary loss as a stand alone loss in a first step. Unfortunately, we did not get competitive results. We believe that this is due to the following technical facts. In theory, the global optimum of our boundary loss corresponds to a negative value (when the softmax probabilities correspond to a non-empty foreground). However, an empty foreground (null values of the softmax probabilities almost everywhere) corresponds to low gradients. Therefore, this trivial solution is close a local minimum or a saddle point. To avoid getting stuck in such trivial solution, we propose to use our loss in conjunction with the regional Generalized Dice loss:

\[ \alpha L_{GD}(\theta) + (1 - \alpha) L_B(\theta) \]  

Finally, it is worth noting that our boundary loss uses ground-truth boundary information via pre-computed level-set function \( \phi_G(p) \), which encodes the distance between each point \( p \) and \( \partial G \). In Eq. (2.4), the softmax for each point \( p \) is weighted by the distance function. Such distance-to-boundary information is omitted in widely used regional losses, where all the points within a given region are treated equally, independently of their distances from the boundary.

### 2.2.2 Scheduling strategy

In our work, we propose to combine the new boundary-loss with the GDL as it has shown good results when dealing with unbalanced problems (Nie et al. (2018); Shen et al. (2018); Horváth et al. (2018); Górriz et al. (2018)). The task of finding the ideal combination between the two losses is not obvious. We argue that it is not very convincing to use an empirical fixed weight. For that, we propose to use a simple scheduling strategy. During training, the value of \( \alpha \) in Eq. (2.5) is initialized to 1, then decreased by 0.01 after each epoch until reaching 0.01. Thus, we give more importance to the GDL in first epochs then we gradually increase the impact of our boundary loss.
2.2.3 Network architecture

It is worth to note that the choice of the network architecture is important for medical image image segmentation. To validate our method, we employed UNet (Ronneberger et al. (2015b)) as a deep learning architecture. In fact, the network is initially designed as a combination of a regular FCNN with an up-sampling path where "deconvolutions" are applied to build an output of the same size than the input. As shown in Fig. 2.4, the particularity of a U-net is the insertion of "skip connections" to connect opposing; convolutional and deconvolutional, layers between the down-sampling and up-sampling paths.

Figure 2.4 Original UNet architecture (taken from (Ronneberger et al. (2015b)))
CHAPTER 3

EXPERIMENTS AND RESULTS

3.1 Datasets

To evaluate our proposed method, we selected three challenging public datasets for brain lesion segmentation where classes are highly unbalanced: the ischemic stroke lesion (ISLES), white matter hyperintensities (WMH) and Multi modal brain tumor segmentation (BRATS17). WMH and ISLES segmentation correspond to binary problems while BRATS segmentation is a multi-class problem.

WMH

The public dataset of White Matter Hyperintensities (WMH) ¹ MICCAI 2017 challenge is composed of 60 subjects for training and 110 subjects for testing. Images are acquired from five different scanners belonging to three different vendors in three various hospitals. Each subject has 3D T1-weighted scan and 2D multi-slice FLAIR. The ground truth of the 60 subjects of training set is defined only on FLAIR images. From a total of 60 subjects, training was performed on 48 subjects while validation on the 12 remaining ones. Patients were chosen randomly from three scanners and with equal proportions in order to have representative sets. The percentage of lesions in the validation fold ranges from 0.01 % to 0.43% of voxels, with a mean of 0.25% between patients, std of 0.32% and median of 0.04%.

ISLES

The training dataset provided by the ISLES organizers is composed of 94 ischemic stroke lesion multi-modal scans. In our experiments, we split this dataset into training and validation sets containing 74 and 20 examples, respectively. Each scan contains Diffusion maps (DWI)

¹ http://wmh.isi.uu.nl
and Perfusion maps (CBF, MTT, CBV, Tmax and CTP source data), as well as the manual ground-truth segmentation. More details about the data can be found in the ISLES website\(^2\).

**BRATS**

The public dataset for the Multimodal Brain Tumor Segmentation Challenge 2017 (BRATS17)\(^3\) is composed of multi-modal scans acquired from different 19 institutions. Provided scans are divided into two categories; subjects with *high grade Glioblastoma (HGG)* and subjects with *Low Grade Glioblastoma (LGG)*. Manual annotations cover three different regions; non enhancing tumor (NET), Peritumoral Edema (ED) and Enhacing Tumor (ET). Our experiments were performed on the training set of a total of 285 cases that were split into 80\% for training and 20\% for validation.

### 3.2 Experimental protocol

#### 3.2.1 Data Pre-processing

Before training the network, a pre-processing step was performed. While the scans are provided as 3D images, we process them as a stack of independent 2D images, which are fed into the network. During the pre-processing, all scans are normalized between 0 and 1 and saved as a set of 2D matrices. As mentioned before, WMH scans are from different scanners and with different sizes. Thus, all scans are re-scaled to a uniform size 200×200 pixels via cropping/-padding operations. On the other hand, BRATS scans are normalized and saved as 2D matrices of 240×240 pixels. To test the impact of our loss on a very basic setting, We did not use any data augmentation during experiments.

---

\(^2\) [http://www.isles-challenge.org](http://www.isles-challenge.org)

\(^3\) [https://www.med.upenn.edu/sbia/brats2017/data.html](https://www.med.upenn.edu/sbia/brats2017/data.html)
3.2.2 Implementation details

In our experiments, we employed Adaptive Moment Estimation (Adam) (Kingma & Ba (2014)) optimizer with a batch size of 10 and an initial learning rate equal to 0.001. We trained a UNet model over 200 epochs for WMH and ISLES datasets and 100 epochs for BRATS without an early stopping. Distance maps or level-set functions $\phi_G$ in Eq. (2.4) were computed using standard SciPy functions \(^4\). In case of slices with only background (no target class), we used a zero-distance map, assuming that the GDL, used alone, is more suitable in those cases.

3.3 Results for binary problems

We trained and validated our network on WMH and ISLES training datasets. Table 3.1 reports the Dice score (DSC) and the Hausdorff distance (HD) performance for our experiments using GDL alone and the loss we proposed in Eq. (2.5). We note that adding our boundary loss jointly with GDL ameliorates efficiently the segmentation performance which is reflected in significant better values of DSC and HD. On the WMH dataset, we had an improvement of around 2% of the DSC while it achieves 8% on the ISLES segmentation task. The same behaviour is observed for the HD metric with a larger gain on ISLES dataset.

<table>
<thead>
<tr>
<th>Loss</th>
<th>ISLES</th>
<th>WMH</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{L}_{GD}$</td>
<td>0.575 (0.028)</td>
<td>0.727 (0.006)</td>
</tr>
<tr>
<td>$\mathcal{L}<em>{GD} + \mathcal{L}</em>{B}$</td>
<td>0.656 (0.023)</td>
<td>0.748 (0.005)</td>
</tr>
<tr>
<td>DSC</td>
<td>HD (mm)</td>
<td>DSC</td>
</tr>
<tr>
<td>4.009 (0.016)</td>
<td>1.045 (0.014)</td>
<td>3.562 (0.009)</td>
</tr>
</tbody>
</table>

Table 3.1 DSC and HD values achieved on ISLES and WMH validation subsets. The values represent the mean performance (and standard deviation) of 2 runs for each setting.

Figure 3.1 shows the learning process behaviour of validation metrics (DSC and HD) over 200 epochs. The learning curves show the gap in performance of the GDL with respect to the

\(^4\) https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.ndimage.morphology.distance_transform_edt.html
GDL with boundary loss, where the difference becomes significant at convergence. In addition to outperforming GDL, we can also observe that the boundary loss term helps stabilizing the training process, yielding a much smoother curve as the network training converges. This behaviour is consistent for both metrics and both datasets, which clearly shows the efficiency of our proposed boundary term.

3.4 Results for multi-class problem

To investigate the behaviour of our boundary loss in a multi-class imbalanced problem, we trained and validated the same network on BRATS database for brain tumour segmentation. Quantitative results are reported in table 3.2. We note that, in a multi-class problem, our
boundary loss jointly with GDL was also able to ameliorate the segmentation performance. Significant better values of the Dice score and the Hausdorff distance are achieved. The mean DSC over classes improves of around 9%. The same behaviour is noted for the Hausdorff distance.

Table 3.2 DSC and HD values achieved on BRATS validation subset. The values represent the mean performance of 2 runs.

<table>
<thead>
<tr>
<th>Loss</th>
<th>N.enh.</th>
<th>ED</th>
<th>ET</th>
<th>N.enh.</th>
<th>ED</th>
<th>ET</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{GD}$</td>
<td>0.73</td>
<td>0.66</td>
<td>0.66</td>
<td>1.32</td>
<td>2.23</td>
<td>1.69</td>
</tr>
<tr>
<td>$L_{GD} + L_B$</td>
<td>0.75</td>
<td>0.75</td>
<td>0.78</td>
<td>1.43</td>
<td>1.98</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Figure 3.2 Evolution of DSC and HD values on BRATS validation subset. The blue curve shows the performance of the network trained using the GDL loss, while the red curve represents the optimization process with the GDL combined with our proposed boundary loss.
Figure 3.3  Qualitative segmentation results on BRATS validation set. Colors correspond to different tumour regions: Red for edema (ED), green for Non enhancing tumor (NET). Best viewed in colors.

3.5 Discussion

Qualitative evaluation

Inspecting qualitative results (for both binary and multi-class problems) illustrated by Fig. 3.5 and Fig.3.3, we note that there are two main types of improvements when using our boundary loss. First, as mentioned previously, region-based losses (e.g. GDL) do not include the information of how far is the prediction from the ground truth. Thus, a close detected object is treated equally than a totally missed one. However, as our boundary loss is based on the dis-
Figure 3.4 Effect of our boundary loss on false positives/ negatives. Best viewed in colors.

Distance map from the ground-truth boundary $\partial G$, it will penalize much more such cases, helping to recover small and far regions. This effect is best illustrated in Fig. 0.1, Fig.3.5 (third row) and Fig 3.4. False positives (first row in Fig. 3.5, second and third row in 3.4) will be far away from the closest foreground, getting a much higher penalty than with the GDL alone. This helps in reducing the number of false positives.
Figure 3.5 Visual results from WMH and ISLES validation sets.

**Behaviour of boundary loss alone**

As mentioned in 2.2.1, the boundary loss (used alone) is not able to give competitive results in a highly imbalanced problem. To investigate its behaviour on a less imbalance level, we trained the network only on positive samples (where the target is present). In that way, we decreased the imbalance rate of the whole dataset. Preliminary results, depicted in Fig. 3.6 show that, when trained only on positive samples, our boundary loss gives smooth contours. We might explain this behaviour with the following fact. In theory, distance maps used in our boundary loss give, on the contrary of typical ground truth images, non binary values but real ones. In
other words, pixels which are close to the foreground boundary are less penalized than far ones. This results in a more regular segmentation.

Figure 3.6 Visual results for our boundary loss as a stand alone loss trained only on positive samples on WMH dataset
CONCLUSION AND RECOMMENDATIONS

Summary of contributions

In this work, we tackled a common issue in medical image processing, the class imbalance problem. This problem occurs mostly when there is an unequal class distribution. In chapter 1, we briefly presented the main classic approaches used for medical image segmentation. We then reported some basics of deep CNNs and discussed both their successes and current challenges. In chapter 2, we introduced our methodology. We proposed a new boundary loss that takes the form of a distance metric on the space of contours, not regions. We argue that a distance-based loss can mitigate the issues related to regional losses in highly unbalanced segmentation problems. Rather than using unbalanced integrals over the regions, a boundary loss uses integrals over the boundary (interface) between the regions. Also, our loss provides a new information which is missing in regional losses.

Conducted on both binary and multi-class highly imbalanced problems, our experiments showed the great impact of including a distance based loss. Results demonstrated a good improvement not only on segmentation performance but also on the training stability.

Limitations and recommendations

Even though experiments have shown good results on three challenging medical problems, we argue that there are still some limitations and recommendations left for a future work. Thus, we suggest the following recommendations for the next steps in our research problem:

- In our main experiments, we used the boundary loss only as a joint loss. Motivated by preliminary results illustrated by Fig.3.6, it would be interesting to investigate the loss behaviour, when used alone, in other problems with different levels of imbalance and with careful initialisations.
- Although our experiments were limited to 2-D segmentation problems, the proposed framework can be trivially extended to 3-D, which could further improve the performance of deep networks, as more context is analyzed. We just need to compute distance transform functions in 3D and use it with any choice of 3-D framework.

- In order to compare to other state-of-the-art methods for these particular applications, future works will evaluate on testing sets of the challenges.
BIBLIOGRAPHY


