BELIEF FUNCTION-BASED SEMI-SUPERVISED LEARNING FOR BRAIN TUMOR SEGMENTATION

Ling Huang†, Su Ruan⋆, Thierry Denœux†‡

† Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France
⋆ University of Rouen Normandy, Quantif, LITIS, Rouen, France
‡ Institut universitaire de France, Paris, France

ABSTRACT

Precise segmentation of a lesion area is important for optimizing its treatment. Deep learning makes it possible to detect and segment a lesion field using annotated data. However, obtaining precisely annotated data is very challenging in the medical domain. Moreover, labeling uncertainty and imprecision make segmentation results unreliable. In this paper, we address the uncertain boundary problem by a new evidential neural network with an information fusion strategy, and the scarcity of annotated data by semi-supervised learning. Experimental results show that our proposal has better performance than state-of-the-art methods.

Index Terms— belief functions, semi-supervised learning, evidential fusion, brain tumor segmentation

1. INTRODUCTION

Deep learning has achieved great success in many computer vision tasks with abundant labeled training data, such as image recognition, object detection and segmentation, image generation, etc. However, acquiring big labeled training data in the medical domain is particularly challenging, especially for image segmentation. Region labeling in medical image segmentation tasks requires not only careful delineation of boundaries but also qualified expertise. Many methods have been developed to address the scarce annotation problem in deep learning, such as self-training [1], adversarial training [2], co-training [3] and clustering [4], in which only partially labeled or unlabeled data are used. However, training a segmentation model with only unlabeled data cannot always meet high precision requirements. Therefore, researchers now rather focus on semi-supervised learning. Another significant problem for medical segmentation is the uncertainty and imprecision of boundary annotation. There is usually a big contradiction between the increasing demand for accurate radiotherapy and the shortage of accurate annotations. Although probabilistic approaches [5][6] have shown great achievements in medical segmentation, they are limited when it comes to addressing uncertainty and imprecision problems. Belief function theory [7][8], a mathematical framework for modeling, fusing and reasoning with uncertain and/or imprecise information, has been successfully applied in many fields, such as image analysis, information fusion, clustering, etc. Since we only have partially labeled data, the uncertainty is increased. Thus, we propose to use belief function theory for partially supervised image segmentation. The main contributions of this paper can be summarized as follows: (1) An end-to-end evidential deep framework is proposed for brain tumor segmentation; (2) A deep semi-supervised learning method allows us to train deep network with fewer labels; (3) A fusion strategy using probabilities and belief functions is applied to address the uncertainty boundary problem.

2. RELATED WORK

2.1. Semi-supervised medical segmentation

Techniques for semi-supervised medical image segmentation can be divided into three classes: 1) self-learning methods (using image-level labels); 2) graph-constrained methods (using bounding box information) 3) Generative adversarial neural network (GAN)-based methods (using adversarial learning). Baur et al [9] introduce the concept of auxiliary manifold embedding for semi-supervised learning using Fully Convolutional Networks [10] with random feature embedding. The labelled data are employed to optimize the network with a primary loss function and the training batches are further augmented with unlabelled samples and their prior maps. In [11], Xu et al use a rectangle as a soft constraint by transforming it into a Euclidean distance map, they predict object masks with a convolutional neural network. In [12], Mondal et al design a few-shot 3D multi-modality medical image segmentation method with a GAN. A \( K + 1 \)-class prediction method is used to allow the GAN to output plausible predictions for unlabelled true data by restricting its output for fake examples.

2.2. Belief function theory

Belief function theory was first introduced by Dempster [7] and Shafer [8] and was further popularized and developed by
Smets [13]. The great expressive power of belief function theory allows us to represent evidence in a more faithful way than using probabilities alone. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ be a finite set of hypotheses about some question. Evidence about $\Omega$ can be represented by a mapping $m$ from $\Omega$ to $[0, 1]$ such that $\sum_{A \subseteq \Omega} m(A) = 1$, called a mass function. For any hypothesis $A \subseteq \Omega$, the quantity $m(A)$ represents the mass of belief allocated to $A$ and to no more specific proposition. Two mass functions $m_1$ and $m_2$ derived from two independent items of evidence can be combined by Dempster’s rule [8] defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B)m_2(C),$$

for all $A \subseteq \Omega, A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) = 0$. In (1), $\kappa$ represents the degree conflict between $m_1$ and $m_2$ defined as

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B)m_2(C).$$

In [14], Denoeux proposed an evidential neural network classifier based on prototypes. In this model, each prototype $p_i$ is a piece of evidence the class of input $x$; the reliability of this evidence decreases with the Euclidean distance $d_i$ between $x$ and $p_i$. The mass function induced by prototype $p_i$ is defined as

$$m_i(\{\omega_k\}) = \alpha_i u_{ik} \exp(-\gamma_i d_i^2), \quad k = 1, \ldots, K,$$

$$m_i(\Omega) = 1 - \alpha_i \exp(-\gamma_i d_i^2) \quad \text{(3b)}$$

where $u_{ik}$ is the membership degree of prototype $p_i$ to class $\omega_k$, $K$ is the number of classes, $\alpha_i$ and $\gamma_i$ are tuning parameters. The mass functions from all prototypes are combined by the Dempster’s rule (1).

3. MATERIALS AND METHODS

3.1. Dataset and pre-processing

The BraTs2018 brain tumor dataset contains 285 cases of patients for training and 66 for validation. Compared with BraTs2018, BraTs2019 adds 50 additional cases; here we hold the extra 50 cases as a test set. For each patient, we have four kinds of brain MRI volumes: Flair, T1, T2, T1ce, and the tumor segmentation ground truth, as represented in Fig. 1 showing examples of the four modalities. Each volume has $155 \times 240 \times 240$ voxels. In order to facilitate the training process, we first sliced 3D volumes into 2D images and normalized each modality with mean and standard deviation for brain region separately. We then center-cropped images with $160 \times 160$ pixels. Finally, we combined the four modality images into 3D tensors as network inputs.

There are four labels: GD-enhancing tumor (ET-label 4), peritumoral edema (ED-label 2), necrotic and non-enhancing tumor core (NCR/NET-label 1), and background (label 0).

3.2. Network Structure

Fig 2 shows the proposed network architecture composed of a light UNet (LUNet), an evidential neural network (ENN) [14] and an evidential fusion layer (EF). Firstly we use a modified UNet [15] with two additional down-sampling layers: $4 \rightarrow 16$, $16 \rightarrow 32$, and two additional up-sampling layers: $32 \rightarrow 16$, $16 \rightarrow 4$, and we shrink the number of neuron nodes to get a light UNet. Details can be found in Fig 2. The output $p_U$ from LUNet is the vector of probabilities for pixels belongs to four class 0, 1, 2, 4. Secondly, we transfer high-level features into the ENN to better represent uncertainty. A probabilistic network may give pixels a very high probability of belonging to one class, even for the boundary pixels. However, a pixel at the boundary between two regions of generally should have similar probabilities of belonging to both regions due to the low contrast in the boundary. To deal with this problem, and inspired by the ENN introduced in [14], we use an ENN module composed of three layers: 1) an activation layer in orange (Fig 2), which calculates distance maps between features and cluster centers; 2) a basic assignment layer in yellow, which converts activated feature values into a mass function based on distance maps; and 3) a mass fusion layer in blue for combining the contribution of each feature to reach a final decision. This ENN module outputs for each pixel five mass values: four masses corresponding to the four classes 0, 1, 2, 4 and an additional mass assigned to the “don’t know” class $\Omega$. Pixels that cannot be classified into any of the four classes are assigned to $\Omega$.

3.3. Multi-source fusion strategy

Since only part of the data are labeled, the uncertainty is higher than it is in the fully supervised case. We propose to use belief functions to deal with this problem. Probabilistic classifiers assign a probability to each class, which is not reliable for pixels close to region boundaries. In contrast, belief function classifiers compute a mass function that assigns a mass to each class and an uncommitted mass to the full set $\Omega$. Here we propose an additional evidential fusion layer to combine both segmentation results from the UNet and the
ENN using Dempster’s rule (1). Denoting the pixel-wise segmentation outputs of LUNet and ENN as, respectively, \(p_U\) and \(m_E\), the combined mass function is

\[(p_U \oplus m_E)(a_i) = \frac{1}{1 - \kappa} \sum_{b \cap c = a_i} p_U(b)m_E(c),\]

for all \(a_i \subseteq \Omega = \{0, 1, 2, 4\}\), \(a_i \neq \emptyset\), and \((p_U \oplus m_E)(\emptyset) = 0\), where \(\kappa\) is the degree of conflict defined in Eq. (2). Here, a high degree of conflict means that the two classifiers assign a pixel to different classes, which indicates high segmentation uncertainty.

### 3.4. Semi-supervised learning

Assuming that only part of the database has been labeled, we propose a semi-supervised learning method that aims to obtain a segmentation accuracy close to that of a fully-supervised learning method. For each input image \(x\), we use several transformations, such as the addition of Gaussian noise, to get new images, noted \(x_t\). Similar images are expected to produce similar segmentation maps even if some transformations have been performed on them. Here we train the network with 50% of the labeled samples using the following \(\text{loss}_1\) function, which measures the difference between the output \(s_t\) and the ground truth \(G\):

\[\text{loss}_1 = \sum_{p=1}^{P} \sum_{t=1}^{T} \left(1 - \frac{2 \cdot \sum s_t^p \cap G_p}{\sum s_t^p + \sum G_p} + 0.5 \times (s_t^p - G_p)^2\right),\]

where \(T\) is the numbers of transformed image type, \(P\) is the number of pixels for segmentation outputs, and \(G\) are the segmentation labels. For the 50% of data without labels we use the following \(\text{loss}_2\) criterion, which measures the difference between the original output \(s\) and the transformed output \(s_t\):

\[\text{loss}_2 = \sum_{p=1}^{P} \sum_{t \neq t}^{T} \left(1 - \frac{2 \cdot \sum s_t^p \cap s_i^p}{\sum s_t^p + \sum s_i^p} + 0.5 \times (s_t^p - s_i^p)^2\right).\]

During training, when the iteration counter is even, \(\text{loss}_1\) is used, otherwise \(\text{loss}_2\) is used.

### 4. EXPERIMENT AND RESULTS

#### 4.1. Performance measures

In this work we used Dice score, Specificity, Positive predictive value (PPV) and Hausdorff distance as performance measures. These criteria are defined as follows:

\[
\text{Dice}(P, T) = \frac{2 \times TP}{FP + 2 \times TP + FN},
\]

\[
\text{PPV}(P, T) = \frac{TP}{TP + FP},
\]

\[
\text{Sensitivity}(P, T) = \frac{TP}{TP + FN},
\]

where TP, FP and FN denote, respectively, the numbers of true positive, false positive and false negative samples, and

\[
\text{Hausdorff} = \max \left(\max_{i \in s} \min_{j \in G} d(i, j), \max_{j \in G} \min_{i \in s} d(i, j)\right),
\]

where \(d\) represents the Euclidean distance.

#### 4.2. Results and Analysis

The learning rates for LUNet and ENN were set to 0.001 and 0.01, respectively. A quantitative comparison is reported in Tables 1 and 2. UUnet is the original UNet framework, which is considered as the baseline model. LUNet is our re-designed light UNet. ELUNet is our final proposal. We can observe that semi-supervised learning does not perform well with UNet if only partially labeled training data are used. The reason is that there are more parameters to optimise, which makes the network more difficult to train. Furthermore, LUNet and ELUNet have better performances according to the four criteria compared with the baseline method UUnet. We can also remark that the performance of ELUNet trained with partially annotated samples is comparable to those of fully supervised methods.
Fig. 3. Segmentation results of brain tumor with different methods of two patients corresponding to the first row and the second one, respectively. From left to right, it shows input samples, ground truth (GT), the results of UNet, LUNet and ELUNet, the difference between GT and UNet, GT and LUNet, and GT and ELUNet.

Table 1. Comparison of different attributes in BraTs2019 dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Learning</th>
<th>Parameters</th>
<th>Dice WT</th>
<th>Dice TC</th>
<th>Dice ET</th>
<th>PPV WT</th>
<th>PPV TC</th>
<th>PPV ET</th>
<th>Sensitivity WT</th>
<th>Sensitivity TC</th>
<th>Sensitivity ET</th>
<th>Hausdorff WT</th>
<th>Hausdorff TC</th>
<th>Hausdorff ET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unet (baseline)</td>
<td>supervised</td>
<td>39.40M</td>
<td>77.58</td>
<td>88.27</td>
<td>90.14</td>
<td>83.33</td>
<td>91.01</td>
<td>93.38</td>
<td>89.66</td>
<td>94.20</td>
<td>93.61</td>
<td>1.59</td>
<td>0.90</td>
<td>0.82</td>
</tr>
<tr>
<td>LUnet</td>
<td>supervised</td>
<td>9.876M</td>
<td>75.52</td>
<td>84.14</td>
<td>90.35</td>
<td>78.95</td>
<td>87.56</td>
<td>95.04</td>
<td>90.40</td>
<td>92.53</td>
<td>91.53</td>
<td>1.71</td>
<td>1.07</td>
<td>0.82</td>
</tr>
<tr>
<td>ELUnet</td>
<td>supervised</td>
<td>87.19</td>
<td>91.97</td>
<td>92.15</td>
<td>87.04</td>
<td>93.64</td>
<td>93.71</td>
<td>91.30</td>
<td>95.55</td>
<td>95.32</td>
<td>1.10</td>
<td>0.78</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Unet (baseline)</td>
<td>semi-supervised</td>
<td>39.40M</td>
<td>56.78</td>
<td>75.46</td>
<td>76.46</td>
<td>62.56</td>
<td>79.73</td>
<td>80.73</td>
<td>61.58</td>
<td>81.34</td>
<td>79.18</td>
<td>2.04</td>
<td>1.77</td>
<td>1.70</td>
</tr>
<tr>
<td>LUnet</td>
<td>semi-supervised</td>
<td>9.876M</td>
<td>82.74</td>
<td>86.58</td>
<td>86.99</td>
<td>88.76</td>
<td>91.42</td>
<td>99.84</td>
<td>89.64</td>
<td>91.63</td>
<td>94.96</td>
<td>1.49</td>
<td>0.99</td>
<td>1.14</td>
</tr>
<tr>
<td>ELUnet</td>
<td>semi-supervised</td>
<td>84.01</td>
<td>89.73</td>
<td>89.94</td>
<td>88.51</td>
<td>91.62</td>
<td>91.55</td>
<td>94.17</td>
<td>94.94</td>
<td>94.54</td>
<td>1.40</td>
<td>0.86</td>
<td>0.85</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison with the state-of-the-art methods (Dice score) on the BraTS 2018 validation dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Dice</th>
<th>Dice</th>
<th>Dice</th>
<th>Dice</th>
<th>Dice</th>
<th>Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WT</td>
<td>TC</td>
<td>ET</td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELUnet</td>
<td>86.16</td>
<td>90.27</td>
<td>85.15</td>
<td>87.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No-new-Net [16]</td>
<td>90.62</td>
<td>84.54</td>
<td>80.12</td>
<td>85.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMFNet[17]</td>
<td>91.26</td>
<td>86.34</td>
<td>80.87</td>
<td>86.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVDLMD [18]</td>
<td>90.68</td>
<td>86.02</td>
<td>81.73</td>
<td>86.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-VNet[19]</td>
<td>90.48</td>
<td>83.64</td>
<td>77.68</td>
<td>83.93</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We also compared our method with the state-of-the-art approaches (Table 2). ELUnet achieves scores of 88.16%, 90.27% and 85.15% for WT, TC and ET, respectively. Compared to the best scores achieved by DMFNet [17], ELUnet shows better performance for TC and ET with 4% and 5% increase, respectively. The performance for WT is not as good as that of other methods, which may be explained by the fact that we take background loss into consideration, under the assumption that only partially annotated training samples are available.

Fig 3 shows the semi-supervised results with different methods. As can be seen from Fig 3, LUnet and ELUnet yield better results than UNet with semi-supervised learning. When the segmentation object is clear (yellow and red regions in the first and second rows, respectively), all three methods show good performances. If the region contains uncertain pixels, both UNet and LUnet have difficulties to segment them (columns 6 and 7), while ELUnet yields better results for controversial boundary segmentation thanks to ENN and EF (column 8). Fig 4 shows the uncertain boundary of a brain tumor. The segmentation results of UNet and LUnet are more uncertain than those of ELUnet if the pixels are close to the boundary between two classes.

5. CONCLUSION

In this work, a belief function-based medical image segmentation network, called ELUnet, has been proposed with a semi-supervised learning strategy. We first modified the Unet framework to get a light Unet. Then we transferred feature map extracted from LUnet to an ENN module for high-level semantic evidence fusion and decision. Finally, decisions from LUnet and ENN were combined by Dempster’s rule to reach a final segmentation result. In future work, we plan to further investigate uncertain boundary analysis for medical images and tackle 3D volume segmentation.
6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access from https://www.med.upenn.edu/sbia/brats2018/data.html. Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

This work was supported by the China Scholarship Council (grant 201808331005). It was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program “Investments for the future” managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02)

8. REFERENCES


