Deep Image Segmentation based on Mutual Information: A Study

Tiago Gonçalves[1,2]
tiago.f.goncalves@inesctec.pt
Leonardo Capozzi[1,2]
leonardo.g.capozzi@inesctec.pt
Ana Rebelo[3]
ana.maria.s.rebelo@accenture.com
Jaime S. Cardoso[1,2]
jaime.s.cardoso@inesctec.pt

[1] Faculdade de Engenharia
Universidade do Porto
Porto, Portugal
[2] INESC TEC
Porto, Portugal
[3] Accenture Portugal
Lisboa, Portugal

Abstract

The use of deep neural networks has achieved many advancements in many different areas, namely computer vision. Information theory concepts such as the Kullback-Leibler Divergence is often used in deep learning methodologies as optimisation criterion since it quantifies the difference between two probability distributions. Image segmentation is a computer vision problem where the goal is to classify individual pixels of an image. It has many real-world applications such as self-driving cars, medical imaging and object detection, to name a few. In this paper, we present a comparative study of two methodologies based on mutual information applied to the task of deep image segmentation.

1 Introduction

Image segmentation is a computer vision problem where the goal is to classify individual pixels of an image as belonging to a certain class. This allows to detect the location of certain objects/features in an image. In recent years, deep learning methodologies have brought performance improvements compared to methods that did not use deep learning. Information theory is based on probability and statistics and is often used to study the information of the distributions of random variables. For instance, in deep learning, concepts such as the Kullback-Leibler Divergence (KLD), which quantifies the difference between two probability distributions, are widely used as optimisation criteria. In this paper, we present a comparative study of two methodologies based on mutual information applied to the task of deep image segmentation. The code related to this work is available in a GitHub repository[1].

2 Related Work

Deep Learning for Image Segmentation U-Net was proposed by Ronneberger *et al.* for the segmentation of images of electron microscopy. This architecture has the main advantage of being capable to learn from few annotated images effectively because it also relies on the use of data augmentation. Inspired by the Fully Convolutional Network (FCN), proposed by Long *et al.* [4], the encoder of the U-Net performs down-sampling operations and extracts features through several convolutional layers; the main goal of this module is to capture context. On the other hand, the decoder, which is symmetric to the encoder, performs upsampling operations through transposed convolutions, thus reducing the number of feature maps and increasing the resolution of the outputs. In the end, a final convolution is employed to obtain a pixel-wise segmentation map. Later, Lin *et al.* proposed the Feature Pyramid Network (FPN), an architecture that could deal with multi-scale features [3] and Zhao *et al.* developed the Pyramid Scene Parsing Network (PSPN), a multi-scale network that also addressed the learning of global context representation [8]. Recently, another operation has been introduced in CNNs: the dilated convolution (also known as *atrous* convolution), which introduced a new hyperparameter, the dilation rate. The nature of this operation allows us to enlarge the receptive field of the deep neural networks with no increase in computational cost. For instance, the state-of-the-art DeepLabV3+ model is based on the use of this operation along with an encoder-decoder architecture and the Atrous Spatial Pyramid Pooling (ASPP) operation, which allows us to extract features at multiple sampling rates and to capture context information at multiple scales [1].

Concepts of Information Theory We present the fundamental concepts of information theory, needed to understand the intuition behind our analysis. All the concepts are presented as in [2]. Information is formally defined as:

$$I_X(x) = -log_2[p_X(x)] \tag{1}$$

, where X is a random variable with probability mass function $p_X(x)$. Mutual information refers to the amount of information that can be obtained about one random variable by observing another random variable. The mutual information between two random variables X and Y is given as:

$$I(X;Y) = \sum_{x,y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

The Kullback-Leibler Divergence (D_{KL}) measures how one probability distribution is different from another. Given two discrete probability distributions P and Q, with the same probability space X, the D_{KL} from Q to P is defined as:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)}) \tag{3}$$

Mutual Information and Deep Learning An interesting approach to the application of information theory concepts to the field of deep learning is related to model distillation methodologies. The main assumption behind the application of model distillation is that small networks have the same *representation capacity* as large networks but may be more difficult to optimise. The distillation approach starts with the learning of a complex and powerful network (*i.e.*, the *teacher* network). Then, the objective is to train a smaller network (*i.e.*, the *student* network) to mimic the teacher, which, according to the literature [5], may be easier than learning the target function directly. Another approach that is directly related to model distillation is *mutual learning*. While in model distillation the main objective is that the student can learn from a pre-trained network (one-way knowledge transfer), in mutual learning, both the networks are training and the main goal is that both mimic each other. Recently, Zhao *et al.* proposed a *region mutual information* loss for semantic segmentation [9]. In their work, they state that most segmentation models use pixel-wise loss functions to optimise the weights, which end up ignoring the dependencies between the pixels of an image. To overcome this issue, they model the image as a multi-dimensional distribution that encodes the relationship between pixels. This is achieved by representing a given pixel by itself and its neighbours. The main goal is to maximise the mutual information between the multi-dimensional distributions of the prediction and ground-truth, however, since the real-value of mutual information may be hard to compute, the authors derive a lower bound of the mutual information and maximise this lower bound to maximise the real value of the mutual information [9].

3 Methodology

Data We performed experiments with two benchmark data sets: 1) the Cityscapes Dataset[2], composed of urban street scenes recorded from 50 different cities; and, 2) the PASCAL Visual Object Classes (VOC) 2012[3], composed of several visual object classes in realistic scenes. This data processing strategy is employed in all the use-cases: 1) images and masks

[1] https://github.com/TiagoFilipeSousaGoncalves/image-segmentation-mutual-information

[2] https://www.cityscapes-dataset.com

[3] http://host.robots.ox.ac.uk/pascal/VOC/voc2012

are read and are given to the models in RGB format; 2) the images are then resized into the size 234 × 234; 3) a random-crop is then applied to the images and they reach the final size of 224 × 224.

Models We trained three different models: 1) U-Net Baseline (Model 1); 2) U-Net with Region Mutual Information Loss (Model 2); and, 3) Deep Mutual Learning using DeepLabV3 and U-Net (Model 3). In Model 2, we train a U-Net with the loss proposed by [9], given by $\mathcal{L} = \lambda L_{\text{BCE}} + (1-\lambda)L_{\text{RMI}}$, where L_{BCE} is the binary cross-entropy loss, L_{RMI} is the region mutual information loss and $\lambda = 0.5$ is a weight coefficient applied to the loss terms. In Model 3, we apply the training strategy proposed by [7], where we employ a joint training strategy with a *teacher* model and a *student model*. In this work, we used the DeepLabV3 architecture as the teacher model and U-Net as the student model.

Performance Metrics To assess the predictive performance of the deep learning algorithms we use two performance metrics: 1) intersection over union (IoU) and 2) accuracy (Acc), defined as:

$$\text{IoU} = \frac{TP}{TP+FP+FN} \tag{4}$$

$$\text{Acc} = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

, where TP, TN, FP and FN be the number of true positive pixels, the number of true negative pixels, the number of false-positive pixels and the number of false-negative pixels, respectively.

4 Results and Discussion

Table 1 presents the IoU and Acc values obtained for Model 1, Model 2 and Model 3. Despite the clear differences between the two datasets (see results for Model 1), we note that predictive performance does not significantly change for Model 2 (achives worse results for PASCAL VOC 2012), and Model 3 (although it achieves interesting results on PASCAL VOC 2012). Figures 1–2 show examples of predictions on the PASCAL VOC 2012 dataset.

Table 1: Summary of results (IoU and Acc) for all models and datasets. Best results highlighted in bold.

Dataset	Model 1		Model 2		Model 3	
	IoU	Acc	IoU	Acc	IoU	Acc
Cityscapes Dataset	**0.5931**	**0.9388**	0.5244	0.9345	0.5619	0.9342
PASCAL VOC 2012	0.1965	0.7367	0.0607	0.6893	**0.2203**	**0.7539**

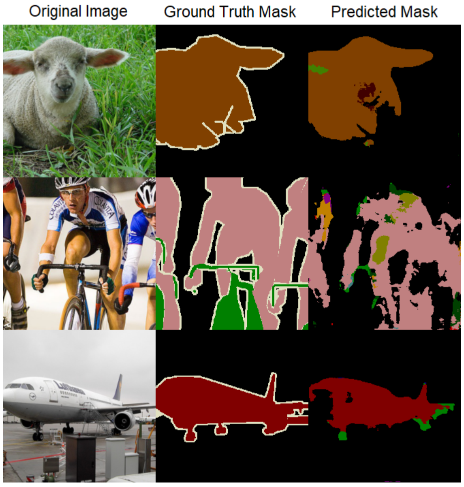

Figure 1: Examples of predicted masks for the PASCAL VOC 2012 using Model 2.

5 Conclusions and Future Work

Results suggest that there are gains in applying a deep mutual learning strategy. Surprisingly, the application of the Region Mutual Information loss does not seem to bring improvements to the predictive performance of the U-Net model in these data sets. Although this can be due to an unknown hyper-parameter that still needs further tuning, some authors suggest that some of the reported results regarding the benefits of using mutual information to optimise deep learning algorithms, may not be attributed to the properties of mutual information alone [6]. Further work should be devoted to the evaluation of different deep image segmentation architectures on the same use-case and to the study of different mutual information properties that could be incorporated in the regularisation of different deep learning algorithms to assess if these methodologies are data set dependent or task-dependent.

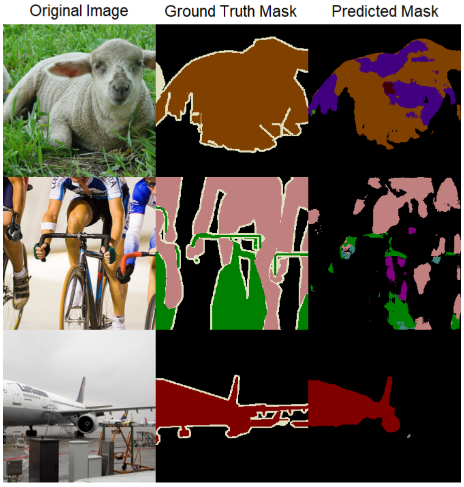

Figure 2: Examples of predicted masks for the PASCAL VOC 2012 using Model 3.

Acknowledgements

This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project NewSpacePortugal, with reference 11, and by FCT – Fundação para a Ciência e a Tecnologia within the PhD grants “2020.06434.BD” and “2021.06945.BD”.

References

[1] Liang-Chieh Chen et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[2] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[3] Tsung-Yi Lin et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[4] Jonathan Long et al. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[5] Adriana Romero et al. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[6] Michael Tschannen et al. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[7] Ying Zhang et al. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[8] Hengshuang Zhao et al. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[9] Shuai Zhao et al. Region mutual information loss for semantic segmentation. *arXiv preprint arXiv:1910.12037*, 2019.