Conditional Entropy as a Supervised Primitive Segmentation Loss Function

Sundaresh Ram$^{1,2}$ and Mert R. Sabuncu$^{1,2}$

$^1$ School of Electrical & Computer Engineering, Cornell University, Ithaca, NY, USA
sr2555@cornell.edu

$^2$ Nancy E. & Peter C. Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA

Abstract. Supervised image segmentation assigns image voxels to a set of labels, as defined by a specific labeling protocol. In this paper, we decompose segmentation into two steps. The first step is what we call “primitive segmentation”, where voxels that form sub-parts (primitives) of the various segmentation labels available in the training data, are grouped together. The second step involves computing a protocol-specific label map based on the primitive segmentation. Our core contribution is a novel loss function for the first step, where a primitive segmentation model is trained. The proposed loss function is the entropy of the (protocol-specific) “ground truth” label map conditioned on the primitive segmentation. The conditional entropy loss enables combining training datasets that have been manually labeled with different protocols. Furthermore, as we show empirically, it facilitates an efficient strategy for transfer learning via a lightweight protocol adaptation model that can be trained with little manually labeled data. We apply the proposed approach to the volumetric segmentation of brain MRI scans, where we achieve promising results.

Keywords: Image segmentation, deep learning, conditional entropy

1 Introduction

Supervised image segmentation often starts with a labeling protocol, which describes how the structure or region boundaries should be delineated by an expert. Furthermore, each label is endowed with an anatomical or biological meaning and can be used to establish correspondence across images. Yet, the labeling protocol can be arbitrary and harmonization across studies can be challenging, even for well-defined tasks such as the segmentation of hippocampal sub-fields [12].

Once a labeling protocol is established, the typical approach is to have human experts manually trace the label boundaries on a set of images. This is a time consuming, error prone, and expensive task. Thus manually segmented datasets, especially those that have been carefully assessed by multiple experts, are invaluable. Most state-of-the-art segmentation algorithms rely on such datasets to train a model that learns to “mimic” the manual segmentation [3, 8–11]. Many of these models can only handle one protocol, and cannot flexibly adapt to different
labeling rules. This can constrain the potential use of the derived segmentation tools, while restricting the accuracy due to limited training datasets.

In practice, for a given application domain such as structural brain MRI scans, there is usually several segmentation protocols - a number bounded by the number of manually labeled datasets. These protocols collectively determine a common set of primitives, which can be defined as sub-parts that make up the individual labels in the protocols. For example, one protocol might simply delineate the cerebral cortex as a whole, while another might divide it up into major sulci and gyri (E.g., see Fig. 5). Given these two protocols, the individual gyri and sulci can be considered as the primitives, as their combinations will define the cortical lobes. Once an image is segmented into primitives, obtaining a label map for any of these protocols should be straightforward.

Thus a primitive segmentation model should ideally be trained on multiple training datasets with different labeling protocols. The method should be able to elegantly handle the variable numbers of labels in the protocols, while respecting the non-trivial relationship between the different labels. For example, both protocols might have a label for the “hippocampus”, while they might differ in where exactly they place the boundaries. While it might be possible to hand-code some of these relationships [6], this approach will not scale well and might miss subtle differences.

In this paper, we propose a novel loss function to train a primitive segmentation model and flexibly handle the aforementioned challenges. The proposed loss function is the entropy of the protocol-specific “ground truth” label map conditioned on the estimated primitive segmentation. The conditional entropy is invariant to the actual values of the labels and thus does not require correspondence of labels between the training images. We apply this loss function to the problem of brain structural MRI segmentation, a popular problem in biomedical image analysis. Our results demonstrate that the proposed approach can yield state-of-the-art accuracy. Furthermore, we show how the primitive segmentation model can be coupled with a protocol adaptation model to generate segmentations for a new set of labeling rules. Our experiments suggest that the protocol adaptation model can be lightweight, which can in turn be trained on a relatively small number of manually labeled datasets.

2 Methods

Let us consider a 3D image \( I(x) \) defined \( \forall x \in \Omega \subset \mathbb{R}^3 \). Let \( P(x) : \Omega \rightarrow \{1, \cdots, p\} \) be a partition of the image into \( p \) primitives, and let \( S(x) : \Omega \rightarrow \{1, \cdots, l\} \) be the protocol specific label map, with a total of \( l \leq p \) regions in it. In this paper, we consider a supervised learning problem of segmenting the image \( I(x) \) to obtain a primitive segmentation \( P(x) \), from which we can then obtain any protocol specific label map \( S(x) \) using a protocol adaptation model. We propose to train the primitive and protocol-specific segmentation models separately.

2.1 Conditional Entropy Loss Function

Let’s first describe our proposed loss function that we use for training a primitive segmentation model. The entropy of the protocol specific “ground truth” label
map conditioned on the estimated primitive segmentation, \( H(S|P) \) is given by:

\[
H(S|P) \equiv \sum_{i=1}^{p} \Pr_S(i) H(S|P = i) = \sum_{i=1}^{p} \sum_{j=1}^{l} \Pr_{Sp}(i,j) \log \left\{ \frac{\Pr_S(j)}{\Pr_{Sp}(i,j)} \right\},
\]

where \( 0 \log 0 = 0 \), \( \Pr_{Sp} \) is the joint normalized histogram of \( P \) and \( S \), and \( \Pr_S \) is the marginal normalized histogram of \( S \). I.e., \( \Pr_S(j) \) is the fraction of voxels in \( S \) that have label \( j \), and \( \Pr_{Sp}(i,j) \) is the fraction of voxels with protocol label \( i \) in \( S \) and primitive label \( j \) in \( P \). This loss function is always non-negative and achieves zero if and only if the protocol labels are determined by the primitive labels. In other words, the conditional entropy will be minimal if and only if each primitive label is associated with only one protocol label in a given ground truth segmentation. In practice, the primitive segmentation network will therefore aim to reduce the uncertainty in the association from the primitive labels to the protocol labels.

For a single protocol, the primitive segmentation that is a 1-1 mapping to the protocol labels will minimize the conditional entropy. In the more interesting case of multiple protocols, however, the learned primitives will represent the sub-parts of the labels across the different protocols. In other words, the primitive labels will capture intersections of protocol-specific labels.

**Remark:** The proposed conditional entropy loss function is fundamentally different from cross-entropy (aka negative log-likelihood), which is widely used in the machine learning literature. Cross-entropy is a general classification loss function (not specific to images) and is defined as an expectation with respect to ground truth label probabilities, assuming that labels across samples adhere to shared semantics. Furthermore, cross-entropy relies on a probabilistic prediction of the label. The proposed conditional entropy loss, on the other hand, leverages the joint statistics of the predicted and ground truth labels over voxels and does not require a probabilistic prediction. Moreover, conditional entropy only assumes label correspondence within an image (or image patch) and not across images. I.e., label \( X \), for example, is supposed to carry a common semantic meaning within the image but not necessarily between images or image datasets. Due to the piecewise smoothness of the input images, the proposed conditional entropy loss function results in a spatially consistent segmentation.

### 2.2 Protocol Adaptation

Once we have learnt a primitive segmentation model, we can couple it with a protocol adaptation model that uses the primitive segmentation \( P(x) \) to compute a protocol-specific label map \( S(x) \). As we present below, this can be a lean model that is trained independently on little data from a specific protocol.

### 2.3 Implementation

**Primitive Segmentation Network:** We adopt a 3D U-Net architecture [9] as the basis of our primitive segmentation network with the conditional entropy...
loss function. We employ three up/down sampling steps and rectified linear units (ReLU) as the activation function. Each layer of the contraction and expansion paths consists of two $3 \times 3 \times 3$ convolutions and a $2 \times 2 \times 2$ max pooling for contraction, and a $2 \times 2 \times 2$ up-sampling for expansion paths. The last layer uses a $1 \times 1 \times 1$ convolution to obtain a single output channel containing the primitive segmentation map $P$, which is a 3D scalar volume with integer-valued (quantized) intensities that correspond to primitive labels. In total the segmentation network has 10 convolutional layers (see Fig. 1-left). The input to our segmentation network are 128 $\times$ 128 $\times$ 128 3D image patches, along with three channels corresponding to x-, y-, and z-coordinate grids that encode the location of the patch in the original image space. The output patches are of size 72 $\times$ 72 $\times$ 72.

**Protocol Adaptation Network:** We use a lightweight 3D fully convolutional network that takes as input the primitive segmentation and the original image, and produces a protocol specific segmentation label map $S(x)$. Our network has three layers of $1 \times 1 \times 1$ convolutions followed by a ReLU (Fig. 1-right). These layers essentially find the mapping that converts primitive labels into segmentation labels. The output layer has multiple channels, each corresponding to a probabilistic estimate of the protocol-specific segmentation label. Thus, the number of output channels is equal to the number of labels in the protocol.

**Other Details:** The proposed primitive segmentation network is trained with stochastic gradient descent on conditional entropy loss. To make the conditional entropy differentiable, in our implementation, the primitive segmentation voxels have real values, which make a soft contribution to the histogram. So, for example a voxel with an intensity value of 2.3 makes a 0.7 contribution to the label bin at 2 and a 0.3 contribution to the label bin at 3. This way, we can compute analytic gradients that are used in the Adam [7] algorithm. We use a mini-batch size of 4 and batch normalization (BN) before each ReLU. The learning rate is initialized to $\alpha = 0.01$ and decreased by a factor of 10 every 10 epochs. For all other parameters we use the default parameters as suggested in [7]. We apply standard data augmentation, where each image is randomly rotated between $-10$ to $+10$. 

---

**Fig. 1:** Proposed segmentation framework. Gray boxes represent multi-channel feature maps, and white boxes are copied features. The number of channels is above each box.
degrees, and randomly scaled between 0.9 and 1.1 to achieve 10× data augmentation. In our experiments, we tile the 3D volume by the output patches in a non-overlapping fashion (except for the boundaries where overlaps are unavoidable). Our code is freely available at https://github.com/sundareshram/cenet.

3 Experiments and Results

3.1 Data

We consider the segmentation of neuroanatomical structures in 3D brain MRI scans with 1mm³ isotropic voxels. We preprocess all the brain MRI scans by orienting them along the same direction. We experiment with three datasets, each manually segmented by experts using different labeling protocols. The first dataset (Buckner) [2] has 39 subjects. For this dataset, we consider following labels in our evaluation experiments: white matter (WM), cerebral cortex (CT), hippocampus (HP), thalamus (TH), amygdala (AM), pallidum (PA), lateral ventricle (LV), putamen (PU), caudate (CA), and cerebellum (CB).

The second dataset (Hammers) [5] has 30 subjects, and we consider nine anatomical regions of interest (ROIs) for performance evaluation, eight of which have the same semantic label as the Buckner structures (HP, TH, AM, PA, LV, PU, CA, CB). These labels differ in their exact placement of structure boundaries. The ninth label, the cerebral white and gray matter (CWM), was created by merging all cerebral regions in original labeling protocol.

The third dataset [4] contains 33 pediatric brain MRIs and we consider 16 structures: superior temporal gyrus (STG), inferior temporal gyrus (ITG), anterior cingulate gyrus (ACG), posterior cingulate gyrus (PCG), middle frontal gyurs (MFG), precentral gyrus (PrG), anterior orbital gyrus (AOG), inferior frontal gyrus (IFG), superior frontal gyrus (SFG), medial orbital gyrus (MOG), lateral orbital gyrus (LOG), posterior orbital gyrus (POG), lingual gyrus (LG), straight gyrus (SG), postcentral gyrus (PsG), and superior parietal gyrus (SPG).

Note that the original protocols label the left and right structures individually (which we combined) and contain more structures than we use in our evaluations. We reserve 29 Buckner subjects (25 training plus 4 validation), 22 Hammers subjects (18 training plus 4 validation), and 22 pediatric subjects (18 training plus 4 validation) for learning the models. Remaining samples (10 Buckner, 8 Hammers, and 11 pediatric) were set aside as independent test subjects.

3.2 Benchmarks and Proposed Method

We implemented three benchmark models, all based on a 3D version of U-Net [9], with output channels equal to the number of protocol-specific evaluation labels, and Dice overlap combined with cross entropy as the loss function. As in the primitive segmentation network, the input consisted of four channels: the 3D image patch, and three axis grids corresponding to the location of the patch in the original image. First model was trained on the 25 training subjects from the Buckner data (Buckner U-Net), whereas the second model was trained on the 18 training subjects of the Hammers data (Hammers U-Net). The third model
was trained on the Buckner and Hammers training subjects together (Combined U-Net), where the Buckner and Hammers labels were treated separately. During training, the soft-max output was computed for each subject based on labels of its protocol. Validation subjects were used to set hyper-parameters, such as the relative weight of the cross-entropy loss, and stop the optimization.

For proposed approach, we trained the primitive segmentation model on the Buckner and Hammers training subjects \((N = 43)\), using the original manual segmentations. The validation subjects were then used to train the protocol adaptation network for each dataset separately, using the evaluation labels.

### 3.3 Performance Evaluation

We evaluated the benchmark and proposed models using the mutual overlap measure between automatic and manual segmentations (Dice) for the aforementioned label sets, computed on the test subjects (see Figure 2). We observe that the proposed method achieves the best Dice score for every single structure. In Fig. 3 we show the average Dice scores (mean over all ROIs) for each test subject. We observe that the proposed method outperforms the benchmarks for all test subjects.

#### 3.4 Protocol Adaptation on a New Dataset

The primitive segmentation model is capable of learning rich primitives, which when coupled with a lightweight protocol adaptation model can generate seg-

---

**Fig. 2:** Dice scores for compared models. In each box the central mark is the median, the edges of the box are the 25th and 75th percentiles. The whiskers extend to 2.7 standard deviations around the mean.

**Fig. 3:** Average Dice score for each test subject (connected with a line). Left: Buckner. Right: Hammers.
Fig. 4: Dice scores on pediatric dataset. Dashed horizontal lines are for the baseline. Curves are for the proposed model trained on different number of subjects. Black corresponds to average over ROIs. Individual ROIs are indicated with different colors.

Fig. 5: Example slice from pediatric dataset. Protocol adaptation models convert top row to bottom row. If zoomed in, it can be appreciated that each region in the primitive segmentation corresponds to a label in the protocol segmentation.

4 Conclusion

In this paper, we divide supervised image segmentation into two parts: primitive segmentation and protocol-specific labeling. For the first part, we propose to train a segmentation model, such as a U-Net, using a novel loss function, namely the entropy of the ground truth labels conditioned on the primitive segmentation. This loss function is flexible and can elegantly handle heterogeneous train-
ing datasets of images manually labeled with different protocols. The loss function is minimized when the mapping from primitive segmentations to protocol-specific segmentations carries little uncertainty. Hence, the proposed strategy discovers primitive labels that correspond to intersections of the different labels that exist in the training data. This supervised approach is therefore, substantially different from prior unsupervised primitive segmentation techniques, such as super-pixels [1]. Next, we demonstrated that a pre-trained primitive segmentation model can be coupled with a lightweight protocol adaptation model to produce protocol-specific segmentations. Our experiments show that the protocol adaptation model can be trained on little data and still produce highly accurate results. Our approach, we believe, promises a new direction for transfer learning between protocols in image segmentation.

5 Acknowledgements

This work was supported by NIH grants R01LM012719 and R01AG053949 and the NSF NeuroNex Neurotechnology Hub grant 1707312.

References

1. Achanta, R., et al.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern. Anal. Mach. Intell. 34(11), 2274–2282 (May 2012)
2. Fischl, B., et al.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33(3), 341–355 (Jan 2002)
3. Fischl, B., et al.: Sequence-independent segmentation of magnetic resonance images. NeuroImage 23, S69–S84 (Jan 2004)
4. Gousias, I.S., et al.: Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. NeuroImage 40(2), 672–684 (Apr 2008)
5. Hammers, A., et al.: Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19(4), 224–247 (Aug 2003), http://brain-development.org/
6. Iglesias, J.E., et al.: An algorithm for optimal fusion of atlases with different labeling protocols. NeuroImage 106, 451–463 (Feb 2015)
7. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: Proc. 3rd Intl. Conf. Learning Representations (2015), https://arxiv.org/pdf/1412.6980.pdf
8. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imag. 34(10), 1993–2024 (Dec 2015)
9. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. Med. Image Comput. Comput. Assist. Intervention (MICCAI). pp. 234–241. Springer (2015)
10. Sabuncu, M.R., et al.: A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imag. 29(10), 1714–1729 (Oct 2010)
11. Soares, J.V., et al.: Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. IEEE Trans. Med. Imag. 25(9), 1214–1222 (Aug 2006)
12. Yushkevich, P.A., et al.: Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRL towards a harmonized segmentation protocol. Neuroimage 111, 526–541 (May 2015)