Error Corrective Boosting for Learning Fully Convolutional Networks with Limited Data

Abhijit Guha Roy[1,2,3], Sailesh Conjeti[2], Debdoot Sheet[3], Amin Katouzian[4], Nassir Navab[2,5], and Christian Wachinger[1]

[1]Artificial Intelligence in Medical Imaging (AI-Med), KJP, LMU München, Germany.
[2]Computer Aided Medical Procedures, Technische Universität München, Germany.
[3]Indian Institute of Technology, Kharagpur, WB India.
[4]IBM Almaden Research Center, Almaden, USA.
[5]Computer Aided Medical Procedures, Johns Hopkins University, USA.

Abstract. Training deep fully convolutional neural networks (F-CNNs) for semantic image segmentation requires access to abundant labeled data. While large datasets of unlabeled image data are available in medical applications, access to manually labeled data is very limited. We propose to automatically create auxiliary labels on initially unlabeled data with existing tools and to use them for pre-training. For the subsequent fine-tuning of the network with manually labeled data, we introduce error corrective boosting (ECB), which emphasizes parameter updates on classes with lower accuracy. Furthermore, we introduce SkipDeconv-Net (SD-Net), a new F-CNN architecture for brain segmentation that combines skip connections with the unpooling strategy for upsampling. The SD-Net addresses challenges of severe class imbalance and errors along boundaries. With application to whole-brain MRI T1 scan segmentation, we generate auxiliary labels on a large dataset with FreeSurfer and fine-tune on two datasets with manual annotations. Our results show that the inclusion of auxiliary labels and ECB yields significant improvements. SD-Net segments a 3D scan in 7 secs in comparison to 30 hours for the closest multi-atlas segmentation method, while reaching similar performance. It also outperforms the latest state-of-the-art F-CNN models.

1 Introduction

Fully convolutional neural networks (F-CNNs) have gained high popularity for image segmentation in computer vision [1–3] and biomedical imaging [4,5]. They directly produce a segmentation for all image pixels in an end-to-end fashion without the need of splitting the image into patches. F-CNNs can therefore fully exploit the image context avoiding artificial partitioning of an image, which also results in an enormous speed-up. Yet, training F-CNNs is challenging because each image serves as a single training sample and consequently much larger datasets with manual labels are required in comparison to patch-based approaches, where each image provides multiple patches. While the amount of unlabeled data rapidly grows, the access to labeled data is still limited due to the labour intense process of manual annotations. At the same time, the

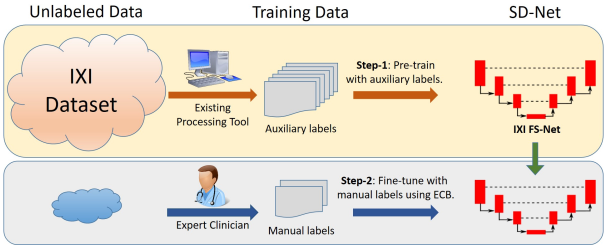

Fig. 1: Illustration of the different steps involved in training of F-CNNs with surplus auxiliary labeled data and limited manually labeled data.

success of deep learning is mainly driven by supervised learning, while unsupervised approaches are still an active field of research. Data augmentation [4] artificially increases the training dataset by simulating different variations of the same data, but it cannot encompass all possible morphological variations. We propose to process unlabeled data with existing automated software tools to create auxiliary labels. These auxiliary labels may not be comparable to manual expert annotations, however, they allow us to efficiently leverage the vast amount of initially unlabeled data for supervised pre-training of the network. We also propose to fine-tune such a pre-trained network using error corrective boosting (ECB), that selectively focuses on classes with erroneous segmentations.

In this work, we focus on whole-brain segmentation of MRI T1 scans. To this end, we introduce a new F-CNN architecture for segmentation, termed SkipDeconv-Net (SD-Net). It combines skip connections from the U-net [4] with the passing of indices for unpooling similar to DeconvNet [2]. This architecture provides rich context information while facilitating the segmentation of small structures. To counter the severe class imbalance problem in whole-brain segmentation, we use median frequency balancing [3] together with placing emphasis on the correct segmentation along anatomical boundaries. For the creation of auxiliary labels, we segment brain scans with FreeSurfer [6], a standard tool for automated labeling in neuroimaging. Fig. 1 shows the steps involved in the training process. First, we train SD-Net on a large amount of data with corresponding auxiliary labels, in effect creating a network that imitates FreeSurfer, referred as FS-Net. Second, we fine-tune FS-Net with limited manually labeled data with ECB, to improve the segmentation incorrectly represented by FS-Net.

Related work: F-CNN models have recently attracted much attention in segmentation. The FCN model [1] up-samples the intermediate pooled feature maps with bilinear interpolation, while the DeconvNet [2] up-samples with indices from the pooling layers, to reach final segmentation. For medical images, U-net was proposed consisting of an encoder-decoder network with skip connections [4]. For MRI T1, eight sub-cortical structures were segmented using an F-CNN model, with slices in [7] and with patches in [8]. Whole-brain segmentation with CNN using 3D patches was presented in [9] and [10]. To the best of our knowledge, this work is the first F-CNN model for whole-brain segmentation. To address the challenge of training a deep network with limited annotations, previous works fine-tune models pre-trained for classification on natural images [11,12]. In fine-tuning, the training data is replaced by data from the target application with additional task specific layers and except for varying the learning rate, the same

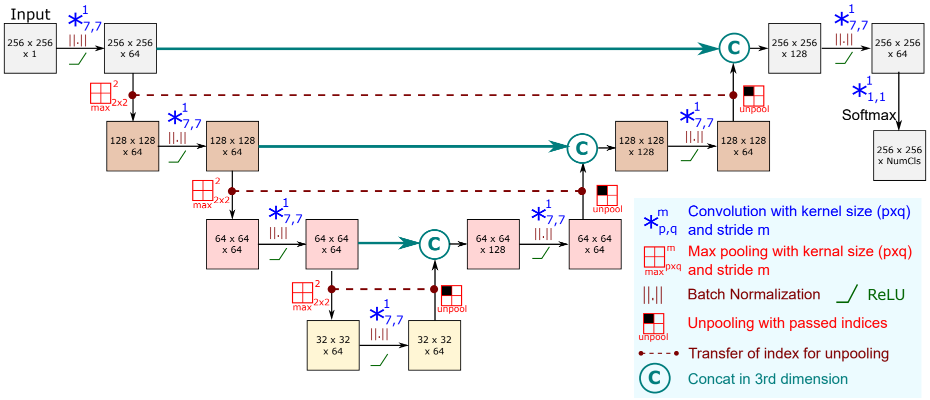

Fig. 2: Illustration of the proposed SkipDeconv-Net (SD-Net) architecture.

training procedure is used. With ECB, we change the class specific penalty in the loss function to focus on regions with high inaccuracies. Furthermore, instead of relying on pre-training on natural images that exhibit substantially different image statistics and are composed of three color channels, we propose using auxiliary labels to directly pre-train an F-CNN, tailored for segmenting T1 scans.

2 Method

2.1 SD-Net for Image Segmentation

We describe the architecture, loss function, and model learning of the proposed SD-Net for image segmentation in the following section:

Architecture: The SD-Net has an encoder-decoder based F-CNN architecture consisting of three encoder and three decoder blocks followed by a classifier with softmax to estimate probability maps for each of the classes. It combines skip connections from U-net [4] and the passing of indices for unpooling from DeconvNet [2], hence the name SkipDeconv-Net (SD-Net). We use skip connections between the encoder and decoder as they provide rich contextual information for segmentation and also a path for the gradients to flow from the shallower decoder to the deeper encoder part during training. In contrast to U-net where upsampling is done by convolution, we use unpooling, which offers advantages for segmenting small structures by placing the activation at the proper spatial location. Fig. 2 illustrates the network architecture for segmenting a 2D image.

Each encoder block consists of a 7×7 convolutional layer, followed by a batch normalization layer and a ReLU (Rectifier Linear Unit) activation function. Appropriate padding is provided before every convolution to ensure similar spatial dimensions of input and output. With the 7×7 kernels, we have an effective receptive field at the lowest encoder block that almost captures the entire brain mask. It therefore presents a good trade-off between model complexity and the capability of learning long-range connections. Each encoder block is followed by a max pooling layer, reducing the spatial dimension of feature maps by half. Each decoder block consists of an unpooling layer, a concatenation by skip connection, a 7 × 7 convolutional layer, batch normalization and ReLU function. The unpooling layer upsamples the spatial dimension of the input feature map by using the saved indices with maximum activation during max pooling of the corresponding encoder block. The remaining locations are filled with zeros. Unpooling does not require to estimate parameters, in contrast to the up-convolution in U-net. The unpooled feature maps are concatenated with the feature maps of the encoder part that have the same spatial dimension. The following convolution layer densifies the sparse unpooled feature maps for smooth prediction. The classifier consists of a 1 × 1 convolutional layer to transfer the 64 dimensional feature map to a dimension corresponding to number of classes (N) followed by a softmax layer.

Loss Function: SD-Net is trained by optimizing two loss functions: (i) weighted multi-class logistic loss and (ii) Dice loss. The logistic loss provides a probabilistic measure of similarity between the prediction and ground truth. The Dice loss is inspired by the Dice overlap ratio and yields a true positive count based estimate of similarity [5]. Given the estimated probability $p_l(\mathbf{x})$ at pixel $\mathbf{x}$ to belong to the class l and the ground truth probability $g_l(\mathbf{x})$, the loss function is

$$\mathcal{L} = \underbrace{-\sum_{\mathbf{x}} \omega(\mathbf{x}) g_l(\mathbf{x}) \log(p_l(\mathbf{x}))}_{\text{LogisticLoss}} \underbrace{- \frac{2\sum_{\mathbf{x}} p_l(\mathbf{x}) g_l(\mathbf{x})}{\sum_{\mathbf{x}} p_l^2(\mathbf{x}) + \sum_{\mathbf{x}} g_l^2(\mathbf{x})}}_{\text{DiceLoss}}. \tag{1}$$

We introduce weights $\omega(\mathbf{x})$ to tailor the loss function to challenges that we have encountered in image segmentation: the class imbalance and the segmentation errors along anatomical boundaries. Given the frequency f_l of class l in the training data, i.e., the class probability, the indicator function I, the training segmentation S, and the 2D gradient operator ∇, the weights are defined as

$$\omega(\mathbf{x}) = \sum_l I(S(\mathbf{x}) == l) \; \frac{median(\mathbf{f})}{f_l} + \omega_0 \cdot I(|\nabla S(\mathbf{x})| > 0) \tag{2}$$

with the vector of all frequencies $\mathbf{f} = [f_1, \ldots, f_N]$. The first term models median frequency balancing [3] and compensates for the class imbalance problem by highlighting classes with low probability. The second term puts higher weight on anatomical boundary regions to emphasize on the correct segmentation of contours. ω_0 balances the two terms.

Model Learning: We learn the SD-Net with stochastic gradient descent. The learning rate is initially set to 0.1 and reduced by one order after every 20 epochs till convergence. The weight decay is set to 0.0001. Mini batches of size 10 images are used, constrained by the 12 GB RAM of the Tesla K40 GPU. A high momentum of 0.9 is set to compensate for this small batch size.

2.2 Fine-Tuning with Error Corrective Boosting

Since the SD-Net directly predicts the segmentation of the entire 2D slice, each 3D scan only provides a limited number of slices for training. Due to this limited availability of manually labeled brain scans and challenges of unsupervised training, we propose to use large scale auxiliary labels for assisting in training the network. The auxiliary labels are created with FreeSurfer [6]. Although these labels cannot replace extensive manual annotations, they can be automatically computed on a large dataset and be used to train FS-Net, which is essentially an F-CNN mimicking FreeSurfer. To the best of our knowledge, this work is the first application of auxiliary, computer-generated labels for training neural networks for image segmentation.

Pre-training provides a strong initialization of the network and we want to use the manually labeled data to improve on brain structures that are poorly represented by the auxiliary labels. To this end, we introduce error corrective boosting (ECB) for fine-tuning, which boosts the learning process for classes with high segmentation inaccuracy. ECB iteratively updates the weights in the logistic loss function in Eq. (1) during fine-tuning. We start the fine-tuning with the standard weights as described in Eq. (2). At epoch $t > 1$, we iteratively evaluate the accuracy a_l^t of class l on the validation set. The weights are updated for each epoch, following an approach that could be considered as median accuracy balancing as shown in Eq. (3).

$$\omega^{(t+1)}(\mathbf{x}) = \sum_l I(S(\mathbf{x}) == l) \; \frac{median(\mathbf{a}^t) - m^t}{{a_l}^t - m^t} \tag{3}$$

with the vector of accuracies $\mathbf{a}^t = [a_1^t, \ldots, a_N^t]$ and the margin $m^t = \min(\mathbf{a}^t) - q$ that normalizes the accuracies with respect to the least performing class. The constant q is set to 0.05, i.e. 5%, to avoid numerical instability. Error corrective boosting sets high weights for classes with low accuracy to selectively correct for errors in the auxiliary labels, which is particularly helpful for whole-brain segmentation with a large number of classes.

3 Results

Datasets: We pre-train the networks with FreeSurfer labels using 581 MRI-T1 volumes from the IXI dataset⋆. These volumes were acquired from 3 different hospitals with different MRI protocols. For the fine-tuning and validation, we use two datasets with manual labels: (i) 30 volumes from the MICCAI Multi-Atlas Labeling challenge [13] and (ii) 20 volumes from the MindBoggle dataset [14]. Both datasets are part of OASIS [15]. In the challenge dataset, 15 volumes were used for training, 5 for validation and 10 for testing. In the MindBoggle dataset, 10 volumes were used for training, 5 for validation and 5 for testing.

We segment the major 26 cortical and sub-cortical structures on the challenge data and 24 on MindBoggle, as left/right white matter are not annotated.

Baselines: We evaluate our two main contributions, the SD-Net architecture for segmentation and the inclusion of auxiliary labels in combination with ECB. We compare SD-Net to two state-of-the-art networks for semantic segmentation, U-net [4] and FCN [1], and also to a variant of SD-Net without Dice loss. For the

⋆ http://brain-development.org/ixi-dataset/

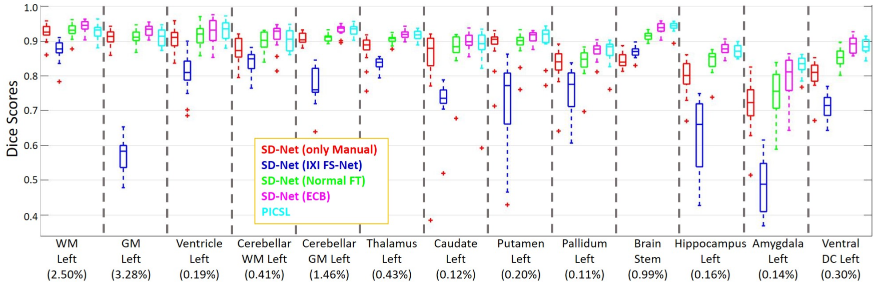

Fig. 3: Boxplot of Dice scores for all structures on the left hemisphere. Comparison of different training strategies of the SD-Net with PICSL. Class probabilities are reported to indicate the severe class imbalance, about 88% are background.

auxiliary labels, we report results for (i) directly deploying the IXI pre-trained network (IXI FS-Net), (ii) training only on the manually annotated data, (iii) normal fine-tuning, and (iv) ECB-based fine-tuning. We use data augmentation with small spatial translations and rotations in all models during training. We also compare to PICSL [16] (winner) and spatial STAPLE [17] (top 5) for the challenge data whose results were available.

Results: Table 1 lists the mean Dice scores on the test data of both datasets for all methods. We first compare the different F-CNN architectures, columns in the table. U-net outperforms FCN on all training scenarios, where the accuracy of FCN is particularly poor on the IXI FS-Net. The SD-Net shows the best performance with an average increase of 2% mean Dice score over U-Net, significant with $p < 0.01$. The SD-Net without the Dice loss in Eq. (1) does not perform as well as the combined loss. We also retrained SD-Net with only limited manual annotated data with $\omega_0 = \{3, 4, 5, 6, 7\}$, resulting in the respective mean dice scores $\{0.85, 0.83, 0.85, 0.84, 0.85\}$. These results show that there is a limited sensitivity to ω_0 and we set it to 5 for the remaining experiments.

Next, we compare the results for the different training setups, presented as rows in the table. Training on the FreeSurfer segmentations of the IXI data yields the worst performance, as it only includes the auxiliary labels. Importantly, fine-tuning the FS-Net with the manually labeled data yields a substantial improvement over directly training from the manual labels. This confirms the advantage of initializing the network with auxiliary labels. Moreover, ECB fine-tuning leads to further improvement of the Dice score in comparison to normal fine-tuning. On

Table 1: Mean and standard deviation of the Dice scores for the different F-CNN models and training procedures on both datasets.

	Multi-Atlas Challenge Dataset				**MindBoggle Dataset**			
Method	IXI FS-Net	Manual Labels	Normal FT	ECB FT	IXI FS-Net	Manual Labels	Normal FT	ECB FT
SD-Net	0.74 ± 0.13	0.85 ± 0.08	0.88 ± 0.06	$\mathbf{0.91} \pm 0.05$	0.71 ± 0.17	0.82 ± 0.06	0.86 ± 0.07	$\mathbf{0.87} \pm 0.06$
SD-Net (No Dice)	0.72 ± 0.14	0.82 ± 0.10	0.84 ± 0.10	0.88 ± 0.06	0.69 ± 0.10	0.80 ± 0.07	0.85 ± 0.10	$\mathbf{0.87} \pm 0.10$
U-Net [4]	0.71 ± 0.15	0.81 ± 0.09	0.82 ± 0.11	0.87 ± 0.06	0.69 ± 0.19	0.76 ± 0.11	0.84 ± 0.07	0.86 ± 0.06
FCN [1]	0.55 ± 0.23	0.70 ± 0.15	0.78 ± 0.12	0.85 ± 0.07	0.45 ± 0.24	0.64 ± 0.23	0.81 ± 0.08	0.83 ± 0.08
Spatial Staple [17]	0.89 ± 0.05				NA			
PICSL [16]	$\mathbf{0.91} \pm 0.04$				NA			

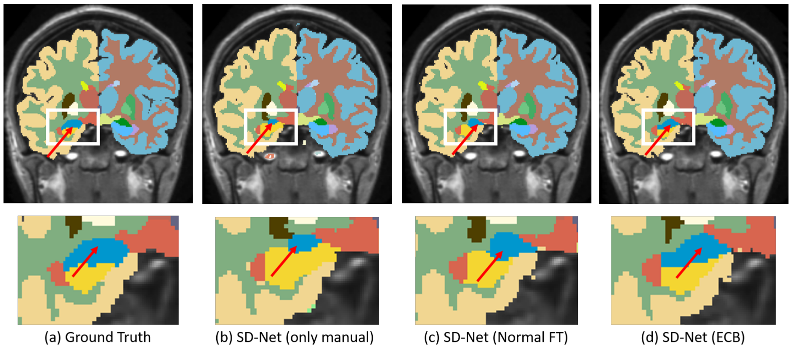

Fig. 4: Comparison of training the SD-Net with only manual labels, normal fine-tuning and ECB together with the ground truth segmentation. A zoomed view of the white box is presented below, where the hippocampus (blue) is indicated by a red arrow.

the challenge dataset, this improvement is statistically significant with $p < 0.01$. Finally, SD-Net with ECB results in significantly higher Dice scores ($p = 0.02$) than spatial STAPLE and the same Dice score as PICSL.

Fig. 3 presents a structure-wise comparison of the different training strategies for the SD-Net together with PICSL. The class probability for each of these structures are also presented to indicate the severe class imbalance problem. There is a consistent increase in Dice scores for all the structures, from training with manually annotated data over normal fine-tuning to ECB. The increase is strongest for structures that are improperly segmented like the hippocampus and amygdala, as they are assigned the highest weights in ECB. Fig. 4 illustrates the ground-truth segmentation together with results from the variations of training the SD-Net. Zoomed in regions are presented for the hippocampus, to highlight the effect of the fine-tuning. The hippocampus with class probability 0.16% is under-segmented when trained with only limited manual data. The segmentation improves after normal fine-tuning, with the best results for ECB.

Segmenting all 2D slices in a 3D volume with SD-Net takes 7 seconds on the GPU. This is orders of magnitude faster than multi-atlas approaches, e.g., PICSL and STAPLE, that require about 30 hours with 2 hours per pair-wise registration. SD-Net is also much faster than the 2-3 minutes reported for the segmentation of eight structures by the patch-based technique in [8].

4 Conclusion

We introduced SD-Net, an F-CNN, encoder-decoder architecture with unpooling that jointly optimizes logistic and Dice loss. We proposed a training strategy with limited labeled data, where we generated auxiliary segmentations from unlabeled data and fine-tuned the pre-trained network with ECB. We demonstrated that (i) SD-Net outperforms U-net and FCN, (ii) using auxiliary labels improves the accuracy and (iii) ECB exploits the manually labeled data better than normal fine-tuning. Our approach achieves state-of-the-art performance for whole-brain segmentation while being orders of magnitude faster.

Acknowledgement: This work was supported in part by the Faculty of Medicine at LMU (FöFoLe), the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), the NVIDIA corporation and DAAD (German Academic Exchange Service). The authors would also like to thank Magdalini Paschali for proof reading and feedback.

References

1. Long J, Shelhamer E, and Darrell T. Fully convolutional networks for semantic segmentation. In CVPR 2015, pp. 3431-40, IEEE.
2. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In ICCV 2015, pp. 1520-28, IEEE.
3. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In arXiv preprint:1511.00561, 2015.
4. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In Proc. MICCAI, Springer 2015, pp. 234-241.
5. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 3DV, 2016, pp. 565-571, IEEE.
6. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, Van Der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. In Neuron, 2002, 33(3), pp. 341-55.
7. Shakeri M, Tsogkas S, Ferrante E, Lippe S, Kadoury S, Paragios N, Kokkinos I. Sub-cortical brain structure segmentation using F-CNNs. In ISBI 2016 pp.269-272.
8. Dolz J, Desrosiers C, Ayed IB. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. In arXiv preprint:1612.03925, 2016.
9. Brebisson A, Montana G, Deep Neural Networks for Anatomical Brain Segmentation, In CVPR Workshops 2015, pp. 20-28.
10. Wachinger C, Reuter M, Klein T, DeepNAT: Deep convolutional neural network for segmenting neuroanatomy, In NeuroImage 2017.
11. Shin H.C, Roth H.R, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. In TMI, 2016, 35(5), pp. 1285-98.
12. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J. Convolutional neural networks for medical image analysis: full training or fine tuning?. In TMI, 2016, 35(5), pp. 1299-312.
13. Landman, B, Warfield, S., Miccai Workshop on Multiatlas Labeling. In MICCAI Grand Challenge, 2012.
14. Klein, A and Tourville, J. 101 labeled brain images and a consistent human cortical labeling protocol. In Front. Neuroscience, 2012, 6, pp. 171.
15. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. In J. cog. Neuroscience. 2010, 12, pp. 2677-84.
16. Wang H, Yushkevich P. Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation. Front. Neuroinform. 2013, 7, pp.27.
17. Asman AJ, Landman BA. Formulating spatially varying performance in the statistical fusion framework. In TMI. 2012, 6, pp. 1326-36.