2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

DOMAIN ADAPTATION FOR BIOMEDICAL IMAGE SEGMENTATION USING ADVERSARIAL TRAINING

Mehran Javanmardi, Tolga Tasdizen

Scientific Computing and Imaging Institute, University of Utah

ABSTRACT

Many biomedical image analysis applications require segmentation. Convolutional neural networks (CNN) have become a promising approach to segment biomedical images; however, the accuracy of these methods is highly dependent on the training data. We focus on biomedical image segmentation in the context where there is variation between source and target datasets and ground truth for the target dataset is very limited or non-existent. We use an adversarial based training approach to train CNNs to achieve good accuracy on the target domain. We use the DRIVE and STARE eye vasculture segmentation datasets and show that our approach can significantly improve results where we only use labels of one domain in training and test on the other domain. We also show improvements on membrane detection between MICCAI 2016 CREMI challenge and ISBI 2013 EM segmentation challenge datasets.

Index Terms— Convolutional Neural Networks, Domain Adaptation, Adversarial Training

1. INTRODUCTION

Recently CNNs have been the method of choice for many visual tasks, including but not limited to image classification [1, 2], image segmentation [3, 4] and object detection [5, 6]. Although CNNs are not new [7], they became popular after their success due to the high computational capability of modern graphical processing units and massive amount of training data provided by datasets such as ImageNet [8] and MSCOCO [9].

However, the success of CNNs is limited to the supervised training using manually annotated datasets. This type of training usually achieves very good results on the testing data if both the training and testing data come from the same distribution. On the other hand, if these distributions differ, accuracy on the test data is lower. The variation between the training data (source domain) and testing data (target domain) is a fundamental and common issue called domain shift. In the context of biomedical applications, domain shift and dataset bias could be due to different reasons such as image acquisition techniques, acquisition device noise, imaging resolution or even more fundamental differences like variations in the tissue structures that are imaged. Researchers are interested in common models that are capable of producing reasonable results regardless of the domain the data is coming from. This usually can be achieved by finetuning the network with the supervised target data, however it comes at the cost of annotating the target domain data.

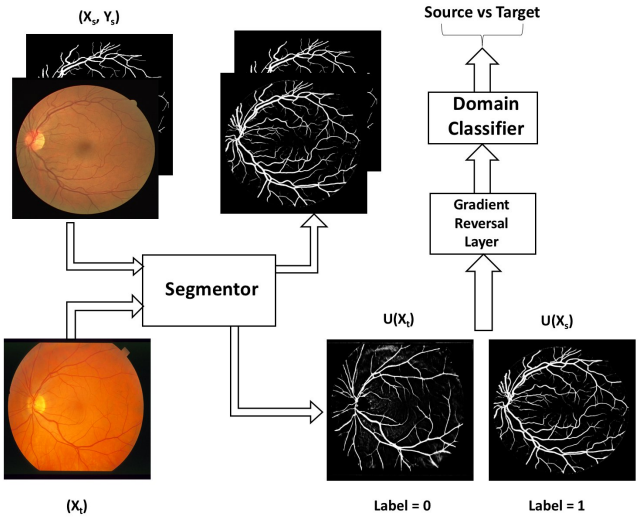

Fig. 1. Our approach: Images from both source and target domain are fed to the segmentor. Supervised pixel-wise loss is calculated for the source images and backpropagated through the segmentor. Outputs of the segmentor are labeled regarding whether they come from source (label = 1) or target (label = 0) domain and fed to the second network which learns to classify the source vs. target domains. This domain loss is backpropagated through the whole network, first the domain classifier and after passing a gradient reversal layer [10] through the segmentor.

Annotating data for the purpose of training is an expensive and time consuming task. This task becomes more costly in the field of biomedical applications if the data should be annotated by experts rather than crowdsourcing platforms like Amazon Mechanical Turk. Furthermore, annotating data for the task of segmentation is much more laborious since fine grained pixel-level annotations are needed compared to other tasks like image classification. Therefore, research to adapt the learning based segmentation approaches to perform ac-

978-1-5386-3636-7/18/$31.00 ©2018 IEEE

curately when the target domain is different than the source domain is imperative. We propose a domain classifier in an adversarial setting on top of a segmentor network (Figure 1) to enforce domain invariance in the feature representations of the segmentor which results in a higher accuracy segmentation when testing data is different from the training data.

2. RELATED WORK

2.1. Image Segmentation

One of the early works that applied CNNs to biomedical images is [11] which uses a sliding window approach and extracts a patch around each pixel and feeds it to a CNN that outputs a class probability. The process is repeated for all pixels in the image to produce the class probability map for the whole image. This approach results in a highly localized output; however, the redundancy due to overlapping patches makes this algorithm inefficient. Another mainstream approach for biomedical image segmentation using CNNs is based on Fully Convolutional Networks (FCN) [3]. FCNs do not have any fully connected layers and the output is an image of class probability vectors that is the same size as the input image. FCNs are highly efficient as they share features extracted by the network for neighboring pixels; however, there is a trade off between localization accuracy and utilization of context in these networks. The deeper the network, more maxpooling layers are used which promote the use of context. On the other hand, these pooling layers reduce the localization capacity of the network. To overcome this drawback, U-net [12], a popular FCN for biomedical applications, uses skip connections between the encoder and decoder part of the network that have the same resolution to preserve localization and granularity of the output. Other works such as [13] extract specialized layer features from different resolutions of a base network to perform eye vasculture segmentation on eye fundus images. Fakhry et al. [14] use residual connections between the encoder and decoder parts of a network similar to U-net to accurately reconstruct neurons in electron microscopy images.

2.2. Domain Adaptation

There has been extensive work on domain adaptation and transfer learning to overcome dataset bias and domain shift problem in learning systems, especially CNNs. As mentioned one approach to mitigate this problem is finetuning the learned networks with few samples of the target domain, however this needs annotations. Domain adaptation techniques usually try to alleviate the shift between the source and target domains by reducing some quantification of it. Minimum Mean Discrepancy (MMD) [15] and correlation distances [16] are common measurements which are optimized to achieve this goal. The network learns to map both source and target domain features into a common feature space. In an alternative approach Ghifary et al. [17] propose to reconstruct the target domain data from the source domain representation which enforces the the network to learn features that are common between both domains.

Domain adaptation techniques based on adversarial training [18] usually consist of two networks, one task specific and one domain specific. The goal is to train the task specific model with reasonable accuracy such that the domain specific model is not able to distinguish if the output of the task specific model is coming from the source or target domain. This idea has been used extensively in the literature to perform adaptation for recognition tasks. Liu et al. [19] propose a coupled generative adversarial network which learns a joint distribution of multi-domain images. Tzeng et al. [20] add a domain classifier to predict domain labels for the input data and use a domain confusion loss to train the domain classifier such that the output labels will have a uniform distribution. Bousmalis et al. [21] propose an unsupervised transformation in the pixel space on the source domain images to be transfered to the target domain and perform classification in the target domain. Ganin et al. [10] introduce a gradient reversal layer that can easily train augmented architectures to learn representations that are discriminative for the main learning task and at the same time are invariant to the source or target domain. While domain adaptation has been used extensively for image classification, we present a novel architecture to learn features that are invariant to the domain and are discriminative for the task of segmentation. In a recent work Hoffman et al. [22] introduce a framework for pixel-wise domain adaptation which minimizes the global domain distribution distances through adversarial training and at the same time optimize a category specific multiple instance loss. This work is different from our approach in the sense that they use small regions corresponding to the natural field of view of the spatial units in the last layer and in the feature space whereas we consider the statistics of the whole output probability map for domain classification.

3. APPROACH

Our proposed architecture consists of two networks. An FCN which performs segmentation on the input images, we call this network segmentor, and a CNN which performs classification on the outputs of the segmentor, we call this network the domain classifier. These two networks are connected through a gradient reversal layer [10] which enables adversarial training. The gradient reversal layer passes its input intact without any modification in the forward pass, however it negates the gradients in the backward pass. This negation of the gradients will update the weights of the segmentor such that it will produce segmentations that are harder for the domain classifier to discriminate. Essentially the segmentor is forced to extract representations that are invariant to the domain while being restricted to accurate segmentations by the supervised pixel-wise loss which is enforced directly to the segmentor.

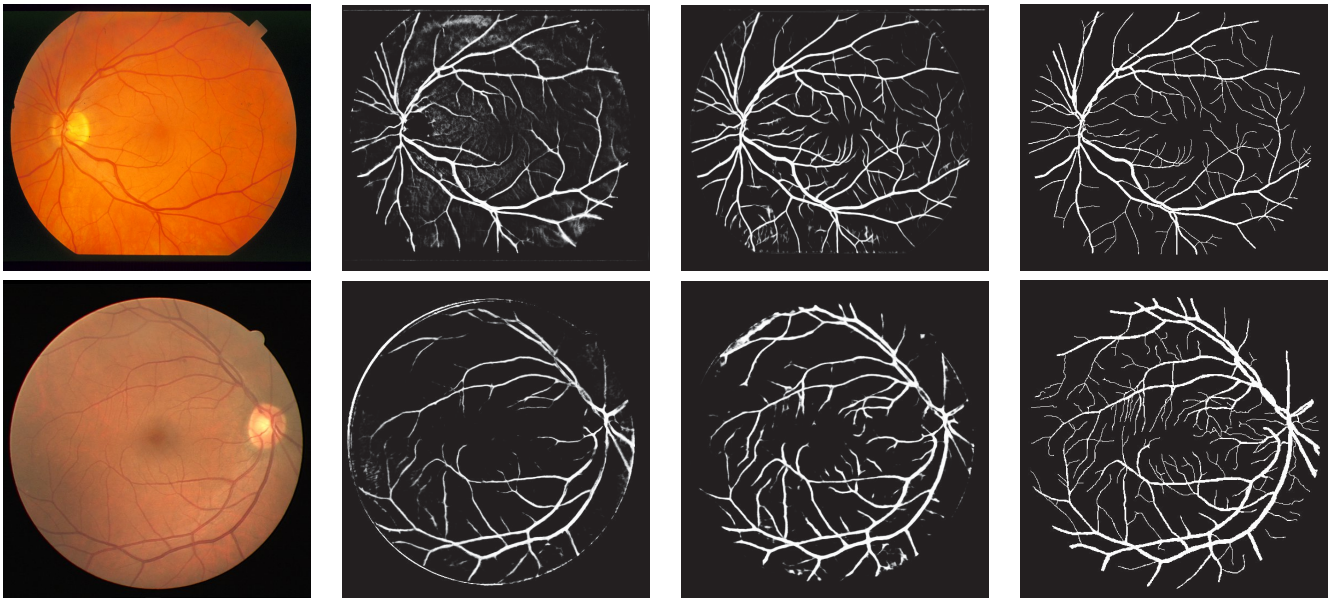

Fig. 2. Results for eye vasculture segmentation in eye fundus image. The first column is the original image, the second column is the results of the segmentor without domain classifier (trained on the source training set and tested on target testing set), the third column is the output of the segmentor when jointly trained with the domain classifier (proposed architecture), the fourth column is the ground truth. The examples in the first row belong to the STARE dataset (source: DRIVE, target: STARE) and images in the second row belong to the DRIVE dataset (source: STARE, target: DRIVE).

Let images from source domain be X_s with pixel-wise labels Y_s and let the target domain images be X_t. We feed X_s and X_t to the segmentor which we call $U(X)$. The outputs from the source domain, $U(X_s)$ are fed into the softmax loss layer $Loss_{sup}(U(X_s), Y_s)$ with their corresponding label ground truth. Note that since we do not have the labels for X_t we exclude them from the supervised loss, therefore no error is backpropagated for the target domain input in this case. The outputs from the segmentor, $U(X_s)$ and $U(X_t)$ are then labeled according to the domain they come from, respectively 1 and 0. Pairs $(U(X_s), 1)$ and $(U(X_t), 0)$ are used to train the domain classifier network which we call $D(X)$. $D(X)$ will be responsible for discriminating between the domains of the segmentations produced by $U(X)$. Let's call the domain classification loss $Loss_{dc}$. The final loss to optimize for the network is:

$$Loss_{sup}(X_s) + \beta\, Loss_{dc}(X_s, X_t)$$

where $Loss_{sup}$ is only calculated for the source domain images and only updates the segmentor network whereas $Loss_{dc}$ is applied to both source and target domain images. The latter backpropagates through both domain classifier and segmentor network with the difference that the gradients applied to the segmentor are the negated gradients.

The intuition behind the proposed architecture comes from the observation that, if there is domain shift, the output segmentation maps for testing images will contain errors that create a visual appearance that is different than the appearance of the segmented structures in training images. More specifically, they can contain background structures detected as false positives and foreground structures that are missed, i.e. false negatives. These will affect the geometry and topology of the output; therefore, they are detectable by the domain classifier. Hence, the segmentor is forced to generate results in the target domain that do not differ visually from the results in the source domain.

4. EXPERIMENTS

4.1. Eye Fundus Images

We use two popular eye vasculture segmentation datasets for validation. The DRIVE dataset [23] consists of 40 images with corresponding pixel labeled ground truth images. We use the standard split of 20 training and 20 testing for this dataset. The STARE dataset [24] contains 20 annotated images, we use 10 for training and 10 for testing. We use U-net as our baseline network in this paper. First the baseline models where we only use the segmentor with $Loss_{sup}(X_s)$ on the source domain and test on the other dataset's testing data. Second, we train the proposed networks jointly using $Loss_{sup}(X_s) + \beta\, Loss_{dc}(X_s, X_t)$ on both the source and target domain data. We use the training data for both source and target (note that the training data for the source is labeled but the target is not) to train and test on the target's testing data. We calculate the f-score value for all the testing results and report them in Table 1.

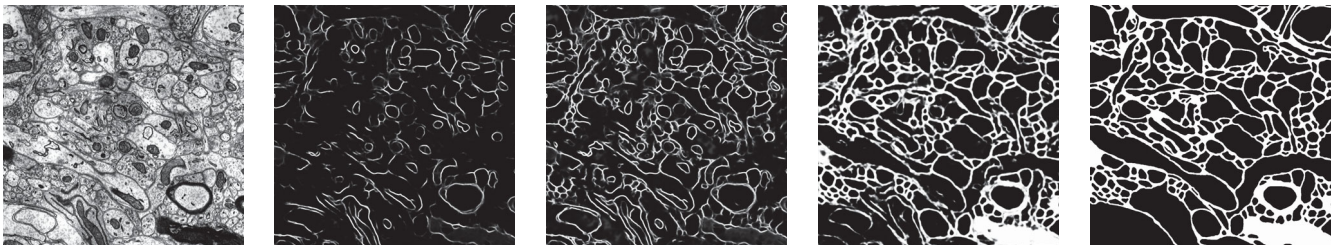

Fig. 3. Results for membrane detection in EM images. The first image is the original image from ISBI challenge, the second image is the results of the segmentor without domain classifier (trained on the MICCAI and tested on ISBI), the third image is the output of the segmentor when jointly trained with the domain classifier (proposed architecture), the fourth image is where we have used 1 labeled images from the target domain (1 out of 100). The last image is the ground truth.

Table 1. The f-score for segmentation results are given for source $\rightarrow$ target. The first column is the baseline and second column is the proposed approach.

	Baseline	Our Approach
STARE $\rightarrow$ DRIVE	62.45	67.09
DRIVE $\rightarrow$ STARE	67.86	76.75

The first column corresponds to the baseline experiments with only the supervised loss. The last column corresponds to the proposed architecture in this paper where we exploit the unlabeled training data from the target domain as well as the labeled training data of the source domain. Using the proposed approach we are able to improve testing accuracy of the DRIVE dataset from 62.45 to 67.09 and the testing accuracy of the STARE dataset from 67.86 to 76.75. Visual results are shown in Figure 2. We should note that when we train our baseline model with the target training data on STARE we achieve a 77.08 testing accuracy and when trained on DRIVE we get a 80.68 testing accuracy. These numbers could be considered as the upper bound that can be achieved.

4.2. Electron Microscopy Images

We also validate our approach on an electron microscopy image segmentation task. In these images the goal is to detect the membranes of each individual neuron so that we are able to fully reconstruct the neuron. We use two datasets for this task. The ISBI 2013 EM segmentation challenge provides 100 images of size 1024x1024 with corresponding connected component ground truth for each individual neuron. The MICCAI 2016 CREMI challenge provides 3 volumes from different parts, we choose volume C to perform our experiments. This volume contains 125 images of size 1250x1250 with corresponding connected component ground truth for each neuron. We dilate each connected component corresponding to neurons on both datasets to obtain the ground truth for membrane detection as the removed pixels. We perform the same set of experiments as we did on eye fundus images. Results are given in Table 2.

Table 2. The f-score for segmentation results are given for source $\rightarrow$ target. The first column is the baseline and second column is the proposed approach. The * indicates the semi-supervised experiments where we include one labeled image from the target domain in the training with source domain.

	Baseline	Our Approach
ISBI $\rightarrow$ MICCAI	35.36	42.60
ISBI $\rightarrow$ MICCAI*	47.50	73.68
MICCAI $\rightarrow$ ISBI	13.16	39.11
MICCAI $\rightarrow$ ISBI*	66.42	77.40

Although we have already shown improvements in Table 2, we note that source and target domain data in EM experiments have a very large domain shift, therefore we include one labeled ground truth image from the target domain in the supervised loss in addition to the source images in a semi-supervised setting and repeat the experiments (we exclude labeled target image used in training from testing). We observe an improved f-score from 47.50, using only the supervised loss (source image + 1 labeled target image), to 73.68 using the proposed approach on the MICCAI data as the target and an improvement from 66.42 to 77.40 on ISBI. We achieve an f-score of 80.63 as an upperbound on MICCAI and 83.91 on ISBI using our baseline trained on target training data.

5. CONCLUSION

We proposed a new architecture to perform segmentation in biomedical images where we have no access to any labeled ground truth for the target domain. We propose a model to adversarially train on a similar source domain dataset against the target domain data. This forces the network to learn feature representations that are invariant to the domain and are at the same time discriminative enough for the segmentation task. We show improvement of target domain testing accuracy using the proposed architecture against the baseline. We also note that the results are to demonstrate proof of concept and results in each experiment could be improved by utilizing application specific baselines.

6. REFERENCES

[1] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 59, pp. 1–35, 2016.

[11] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *NIPS*, 2012, pp. 2843–2851.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[13] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool, "Deep retinal image understanding," in *MICCAI*. Springer, 2016, pp. 140–148.

[14] Ahmed Fakhry, Tao Zeng, and Shuiwang Ji, "Residual deconvolutional networks for brain electron microscopy image segmentation," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 447–456, 2017.

[15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, 2015, pp. 97–105.

[16] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV 2016 Workshops*. Springer, 2016, pp. 443–450.

[17] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *ECCV*. Springer, 2016, pp. 597–613.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[19] Ming-Yu Liu and Oncel Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016, pp. 469–477.

[20] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko, "Simultaneous deep transfer across domains and tasks," in *ICCV*, 2015, pp. 4068–4076.

[21] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," *arXiv preprint arXiv:1612.05424*, 2016.

[22] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016.

[23] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[24] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.