Intramodality Domain Adaptation using Self Ensembling and Adversarial Training

Zahil Shanis[1,2], Samuel Gerber[1], Mingchen Gao[2], and Andinet Enquobahrie[1]

[1] Kitware Inc, Carrboro NC 27510, USA
[2] SUNY at Buffalo, Buffalo, NY 14260, USA

Abstract. Advances in deep learning techniques have led to compelling achievements in medical image analysis. However, performance of neural network models degrades drastically if the test data is from a domain different from training data. In this paper, we present and evaluate a novel unsupervised domain adaptation(DA) framework for semantic segmentation which uses self ensembling and adversarial training methods to effectively tackle domain shift between MR images. We evaluate our method on two publicly available MRI dataset to address two different types of domain shifts: On the BraTS dataset[11] to mitigate domain shift between high grade and low grade gliomas and on the SCGM dataset[13] to tackle cross institutional domain shift. Through extensive evaluation, we show that our method achieves favorable results on both datasets.

1 Introduction

Existence of domain shift between related datasets pose a serious challenge for CNN based tasks like segmentation which require a large amount of annotated data for training. Unlike in the natural images, the problem of domain shift is ubiquitous in biomedical image analysis as images acquired by various institutions belong to different domains due to difference in image acquisition parameters used for capturing data. In addition, tumors and cancers of different grades and severity may belong to different distributions, limiting the ability of single segmentation model in labeling cancerous tumors of varying severity and growth (Figure 1). To tackle this issue, unsupervised domain adaptation has been extensively studied to enable CNN to achieve competitive performance in a domain different than the training domain [19].

In this paper, we study intramodality domain adaptation where both source and target domains belong to same modality, but have different distributions due to difference in image acquisition parameters or tumor severity. Intramodality domain shift is often neglected in biomedical image analysis as most of the deep learning based networks are trained and tested on a mixture of data collected from different institutions and devices, disregarding the associated domain shift. This often results in unpredictable performance if test set is from a data source different than training.

Numerous unsupervised domain adaptation methods have been proposed in the literature, with a growing emphasis on learning domain invariant representation

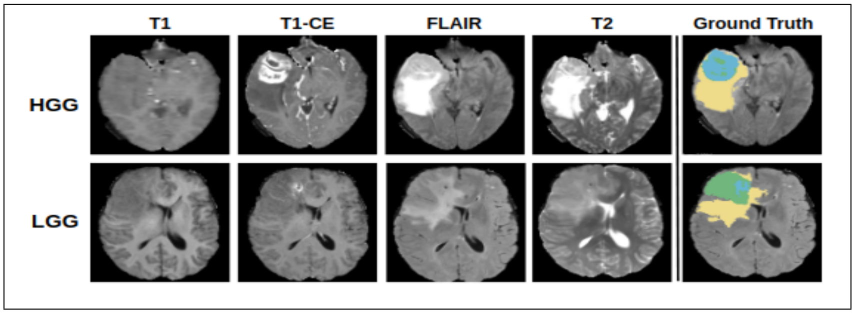

Fig. 1: Tumor size variability in BraTS dataset. Top row: Axial slices of high grade(HGG) tumors, bottom row: low grade(LGG) tumors. In Ground Truth(GT), union of all colors=whole tumor, green=enhanced tumor and blue=core tumor. HGG and LGG have different size and distributions for tumor regions.

to implicitly learn the feature mapping between domains [19]. These methods can be broadly classified as divergence minimising methods [10, 3, 17] which propose to minimise the distribution statistics between domains and adversarial methods [20, 5, 16] which use discriminators for aligning feature spaces. In contrast, French et.al [4] employed self-ensembling for domain adaptation and achieved state-of-the-art results on VisDA-2017 domain adaptation challenge. This technique is based on the Mean-Teacher Network [18] introduced for semi-supervised learning and requires extensive task-specific data augmentation. Additionally, pixel space translation [2] and modulating batchnorm statistics [9] are also explored in detail for domain adaptation and achieved promising results [19].

In biomedical imaging, Kamnitsas et.al's [7] work on brain lesion MRI domain adaptation using adversarial training demonstrated the effectiveness of adversarial loss for unsupervised domain adaptation on medical datasets. The latest study on medical data that is closely related to our work is [12], which performed unsupervised domain adaptation using self ensembling techniques for spinal cord grey matter segmentation and achieved promising results.

Current research trends in domain adaptation are directed towards combining multiple techniques to achieve superior performance in various computer vision tasks [6, 15]. Following this direction, we propose a combined network which uses domain invariant feature training with self ensembling technique for MRI domain adaptation in the context of semantic segmentation. We demonstrate the performance of our method on two publicly available MRI datasets: 1) On BraTS [11, 1] dataset for multiclass tumor segmentation using high grade to low grade glioma domain adaptation, 2) On SCGM [13] Segmentation dataset for grey matter segmentation using cross institutional DA. To the best of our knowledge, our work here is the first to perform high grade to low grade glioma domain adaptation and the first one to use a combination of self-ensembling and adversarial training for medical image domain adaptation.

2 Methodology

2.1 Overview of the Proposed Model

Our domain adaptation network consists of three modules as shown in Figure 2: A student segmentation network G, a teacher segmentation network $\overline{G}$ and a discriminator D. First, we forward source images with labels through segmentation network G and update its weights. Then we pass unlabeled target images through G and obtain its pre-softmax layer predictions. Predictions from both the domains are passed through discriminator D to distinguish whether the input belongs to source or target domain. Adversarial loss from D is then back-propagated through G to update network weights to learn domain invariant feature representation. Teacher network $\overline{G}$ weights are then updated as the exponential moving average (EMA) of student network(G) weights. Finally, we compute consistency loss between student and teacher networks predictions for target images and back-propagate through student network(G). Figure 2 illustrates the proposed algorithm.

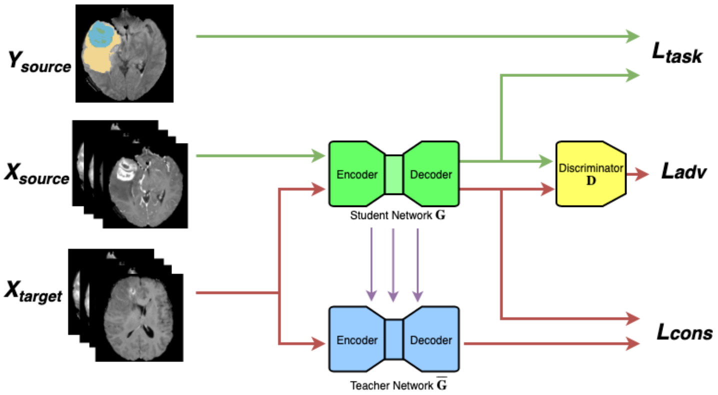

Fig. 2: Our proposed architecture. Green arrows correspond to source data and red arrows correspond to target data. Teacher Network weights are updated via EMA.

2.2 Adversarial Training

The objective behind adversarial training is to adapt the segmentation network invariant to variations between source and target. This is achieved by using a fully convolutional discriminator network(D) to distinguish the domain of input data. D is trained with a cross entropy loss using source and target domain predictions. For target images predictions, we compute an adversarial loss($\mathcal{L}_{adv}$) and back-propagate it to segmentation network(G) to fool the discriminator by pushing the feature representation to a domain invariant space.

2.3 Self Ensembling and Mean Teacher

We combine adversarial training with self ensembling using Mean-Teacher in our network. Although initial self ensembling papers [8, 18] were specifically designed for semi-supervised learning, French et.al extended mean-teacher algorithm for UDA in his seminal paper [4]. Their proposed architecture consists of a student network and a teacher network where the student network is trained with back-propagation while the teacher network weights are an exponential moving average of student network weights. We use self ensembling as a regularizer to smoothen the weights of our feature space domain adaptation network. Student network weights are updated by task loss and adversarial loss which is then exponentially averaged over time to update teacher network weights. We finally use teacher network for making predictions. For our mean teacher self ensembling model, we use the same architecture proposed by [4].

C.Perone et.al [12] has adapted and implemented this network for domain adaptation for medical imaging segmentation and achieved favorable results. A key difference between their work and ours is that their model uses only self ensembling for domain adaptation while we combine it with adversarial training as a regularizer for feature-space domain adaptation.

2.4 Objective Function

With the proposed network, we formulate the final loss function for domain adaptation as follows:

$$\mathcal{L} = \mathcal{L}_{task}(I_s) + \lambda_{adv}\mathcal{L}_{adv}(I_t) + \lambda_{cons}\mathcal{L}_{cons}(I_t) \tag{1}$$

where I_s, and I_t are inputs from source and target domains respectively. $\mathcal{L}_{task}(I_s)$ is the segmentation task loss computed on the paired input data. We use dice loss for segmentation which is commonly employed in biomedical image segmentation due to its low sensitivity to class imbalance. Adversarial loss $\mathcal{L}_{adv}(I_t)$ is computed as a cross entropy loss on target images to adversarially align feature representation of both domains. Consistency loss $\mathcal{L}_{cons}(I_t)$ measures the difference between predictions from teacher and student networks for distilling the knowledge on the student model for self ensembling. We use mean squared error(MSE) for $\mathcal{L}_{cons}(I_t)$ as suggested by [12]. Additionally, discriminator network is trained using source and target feature representations using a standard cross-entropy discriminator loss ($\mathcal{L}_{disc}(I_s, I_t)$).

2.5 Model Architecture

Discriminator Network : For Discriminator, we use a fully convolutional neural network consisting of four convolutional layers with 4 × 4 kernels and stride of 2. Except for the last layer, each convolution layer is followed by a leaky ReLU parameterized by 0.2. Discriminator is trained with Adam as optimizer with default set of parameters and a polynomial decay function for learning rate.

Segmentation Network :We use UNet [14] as our segmentation network with 15 layers, batch normalization and dropout. Network is trained using Adam as optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Both student and teacher networks have identical UNet architecture and only student network weights are updated by back-propagation. Performance of the model is validated using teacher network on validation data from both domains.

3 Datasets

We used two publicly available MRI datasets to evaluate our methodology. We performed HGG to LGG domain adaptation on BraTS dataset [11, 1] and cross institutional domain adaptation on SCGM segmentation challenge dataset [13].

BraTS 2018 [11, 1] dataset consists of 285 MRI samples (210 HGG and 75 LGG) each with T1, T1-contrast enhanced, T2-weighted and FLAIR volumes with ground truth voxel-wise labels for enhancing tumor, peritumoral edema and necrotic and non-enhancing tumor core. Both HGG and LGG volumes are splitted into train and test and we use train HGG as source and train LGG as target for domain adaptation experiments. Since we are using 2D-Unet for segmentation, we slice 3D voxels into 2D axial slices of 128×128 and concatenated all four MRI modalities to get a 4-channel input. More information about dataset can be found at [11].

Spinal Cord Gray Matter Challenge(SCGM) [13] dataset contains single channel Spinal Cord MRI data with grey matter labels from 4 different centers. Data is collected from four centers (UCL, Montreal, Zurich, Vanderbilt) using three different MRI systems (Philips Acheiva, Siemens Trio, Siemens Skyra) with institution specific acquisition parameters. From each center, 10 MRI volumes are publicly available which we center cropped 2D axial slices of 200×200 for our experiment. We use our network to perform cross institutional domain adaptation on this dataset with centers 3 and 1 as source and center 2 as target and validate the performance on all four centers.

4 Experiments and Results

In this section, we present experimental results to validate the proposed domain adaptation method for semantic segmentation on both datasets. First we evaluate model performance on SCGM dataset for cross institutional domain adaptation. Second, we carry out experiments for HGG to LGG domain adaptation on BraTS dataset. We also conduct extensive experiments and ablation studies on both dataset to substantiate the efficacy of our proposed architecture. For a fair comparison and analysis, all experiments are run for the same number of epochs with the same set of parameters for optimizers and learning rate decay. Model performance is evaluated using the dice coefficient. For each dataset we conduct the following experiments:

1. Training the segmenter network (with no DA) on combined source and target data and test separately on heldout sets (*super-all*).
2. Training the segmenter network (with no DA) on source data alone and test separately on source and target (*super-source*).
3. Domain adaptation using only adversarial training (*da-adv*).
4. Domain adaptation using only self ensembling (*da-ensemble*).
5. Proposed domain adaptation algorithm with both adversarial training and self-ensembling(*da-combined*).

4.1 Spinal Cord Cross Institutional Domain Adaptation

All networks for cross institutional DA are trained for 350 epochs with centers 3 and 1 as source and center 2 as target. Weights for adversarial and consistency losses($\lambda_{adv}, \lambda_{cons}$) are optimized separately using *da-adv* and *da-ensemble* models. We found $\lambda_{adv} = 0.001$ and $\lambda_{cons} = 2$ to have best performance on individual domain adaptation models and used them for the combined DA model as well.

Experiment	**Center1**	**Center2**	**Center3**	**Center4**
super-all	87.5	87.9	87.8	87.96
super-source	87.48	77.11	87.19	85.25
da-adv	87.27	79.43	87.49	87.2
da-ensemble	87.7	84.76	**87.59**	87.33
da-combined	**87.93**	**85.75**	87.56	**87.43**

Table 1: Dice score for cross institutional domain adaptation.

We present experimental results for cross-institutional domain adaptation in Table 1. Combined supervised model achieved similar dice scores on all held-out sets while source-only supervised model produced poor results for center 2. This substantiates the existence of intramodality domain shift among multi institutional MRI data and validates the importance of medical image domain adaptation. In contrast, all domain adaptation networks achieved improved results on center2, showing the effectiveness of DA techniques in mitigating domain shift. Our proposed model achieved highest dice score on 3 out of 4 centers and produced results on par with supervised training using combined data. Figure 3 presents some example results for adapted segmentation using combined model. Although domain adaptation models are adversarially trained against center2, model performance has improved for all centers. This suggests that DA with the proposed architecture can be used for domain generalisation as well.

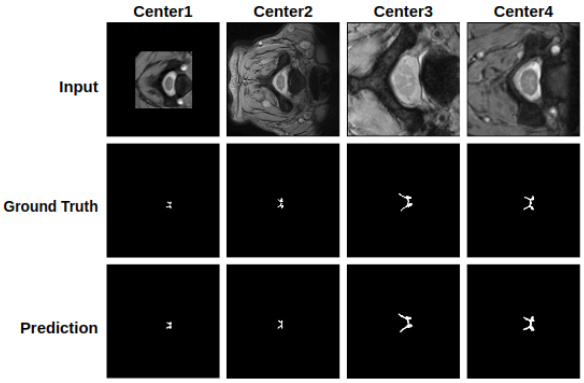

Fig. 3: Example results of adapted segmentation for SCGM Dataset. Model is trained using combined adversarial and self ensembling domain adaptation.

4.2 Brain Tumor Segmentation using Domain Adaptation

We trained all experiments for 150 epochs with HGG as source and LGG as target. Networks are trained with 4-channel sliced 2D axial MRI images to perform 4-class segmentation (background, enhanced tumor, whole tumor and core tumor). Performance scores for all experiments with class wise dice scores are presented in 2. Supervised model results clearly show the domain shift between high grade and low grade gliomas in BraTS dataset. LGG heldout set produced inferior results when the network is trained only using HGG volumes. Our proposed domain adaptation method mitigated this domain shift to an extent and achieved noticeable improvement in segmenting whole and core tumor regions in LGG dataset.

Experiment	**HGG**				**LGG**			
	Whole	**Enh**	**Core**	**Overall**	**Whole**	**Enh**	**Core**	**Overall**
super-all	85.51	67.84	67.13	78.47	85.23	38.22	55.14	64.34
super-source	85.66	66.84	66.59	77.34	79.29	33.09	44.11	58.44
da-adv	85.47	59.01	64.63	73.44	80.09	30.35	44.90	60.07
da-ensemble	85.90	66.84	66.59	77.61	82.97	**33.84**	46.87	60.97
da-combined	85.80	66.43	67.11	78.23	**84.11**	32.67	**47.11**	**62.17**

Table 2: Dice scores for BraTS domain adaptation.

5 Conclusion

In this paper, we presented a novel approach to intra-modality domain adaptation using adversarial training and self ensembling. We evaluated our model on two publicly available MRI datasets to address cross institutional domain shift and tumor severity domain shift. The results showed improved segmentation performance on both datasets. Superior performance on two different datasets validates the generalisability of our proposed model which can be extended to other intra-modality DA applications for biomedical image segmentation. Future work includes extensive hyperparameter tuning for improved segmentation for unsupervised domain adaptation.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4** (9 2017)
2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 95–104 (2017)
3. Eric Tzeng, Judy Hoffman, N.Z., Darrell., T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)

4. French, G., Mackiewicz, M., Fisher, M.H.: Self-ensembling for domain adaptation. CoRR **abs/1706.05208** (2017)
5. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2016)
6. Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. CoRR **abs/1711.03213** (2017)
7. Kamnitsas, K., Baumgartner, C.F., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. CoRR **abs/1612.08894** (2016), http://arxiv.org/abs/1612.08894
8. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. CoRR **abs/1610.02242** (2016), http://arxiv.org/abs/1610.02242
9. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. CoRR **abs/1603.04779** (2016)
10. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: On ICML - Volume 37. pp. 97–105. ICML'15 (2015)
11. Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahaniy, K., Kirby, J., Burren, Y., Porz, N., Slotboomy, J., Wiest, R., Lanczyi, L., Gerstnery, E., Webery, M.A., Arbel, T., B. Avants, B., Ayache, N., Buendia, P., Collins, L., Cordier, N., Van Leemput, K.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **99** (12 2014)
12. Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J.: Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. CoRR **abs/1811.06042** (2018), http://arxiv.org/abs/1811.06042
13. Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M.J., Conrad, B.N., Datta, E., Dvid, G., Leener, B.D., Dupont, S.M., Freund, P., Wheeler-Kingshott, C.A.G., Grussu, F., Henry, R., Landman, B.A., Ljungberg, E., Lyttle, B., Ourselin, S., Papinutto, N., Saporito, S., Schlaeger, R., Smith, S.A., Summers, P., Tam, R., Yiannakas, M.C., Zhu, A., Cohen-Adad, J.: Spinal cord grey matter segmentation challenge. NeuroImage **152**, 312 – 329 (2017)
14. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. LNCS, vol. 9351, pp. 234–241. Springer (2015)
15. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. arXiv preprint arXiv:1712.02560 (2017)
16. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. CoRR **1704.01705** (2017)
17. Sun, B., Saenko, K.: Deep CORAL: correlation alignment for deep domain adaptation. CoRR **abs/1607.01719** (2016), http://arxiv.org/abs/1607.01719
18. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems 30, pp. 1195–1204 (2017)
19. Wilson, G., Cook, D.J.: Adversarial transfer learning. arXiv **1812.02849** (2018)
20. Yaroslav Ganin, Evgeniya Ustinova, H.A.P.G.H.L.F.L.M.M., Lempitsky, V.: Domain-adversarial training of neural networks. In: Journal of Machine Learning Research. p. 17(59):135. IEEE (2016)