ROBUST WHITE MATTER HYPERINTENSITY SEGMENTATION ON UNSEEN DOMAIN

Xingchen Zhao\textsuperscript{1} Anthony Sicilia\textsuperscript{2} Davneet S. Minhas\textsuperscript{3} Erin E. O’Connor\textsuperscript{5}
Howard J. Aizenstein\textsuperscript{4} William E. Klunk\textsuperscript{4} Dana L. Tudorascu\textsuperscript{4} Seong Jae Hwang\textsuperscript{1,2}

\textsuperscript{1}Intelligent Systems Program - University of Pittsburgh
\textsuperscript{2}Computer Science, \textsuperscript{3}Radiology, \textsuperscript{4}Psychiatry - University of Pittsburgh
\textsuperscript{5}Department of Diagnostic Radiology & Nuclear Medicine - University of Maryland, Baltimore

ABSTRACT

Typical machine learning frameworks heavily rely on an underlying assumption that training and test data follow the same distribution. In medical imaging which increasingly begun acquiring datasets from multiple sites or scanners, this identical distribution assumption often fails to hold due to systematic variability induced by site or scanner dependent factors. Therefore, we cannot simply expect a model trained on a given dataset to consistently work well, or generalize, on a dataset from another distribution. In this work, we address this problem, investigating the application of machine learning models to unseen medical imaging data. Specifically, we consider the challenging case of Domain Generalization (DG) where we train a model without any knowledge about the testing distribution. That is, we train on samples from a set of distributions (sources) and test on samples from a new, unseen distribution (target). We focus on the task of white matter hyperintensity (WMH) prediction using the multi-site WMH Segmentation Challenge dataset and our local in-house dataset. We identify how two mechanically distinct DG approaches, namely domain adversarial learning and mix-up, have theoretical synergy. Then, we show drastic improvements of WMH prediction on an unseen target domain.

Index Terms— Domain Generalization, Image Segmentation, Deep Learning, White Matter Hyperintensity

1. INTRODUCTION

In traditional machine learning, an underlying assumption is that the model is trained on training data that is well representative of the testing data. That is, both the training and testing samples come from an identical distribution. However, this assumption has become difficult to satisfy in the modern medical imaging analysis community as it rapidly grows with multiple sites or scanners [1,2]. Notably, datasets often exhibit heterogeneity (e.g., differing distributions of intensity) due to systematic variability induced by various site/scanner dependent factors, and commonly, imaging protocols.

In this work, we view sites or scanners as distinct distributions or domains and consider learning in the presence of multiple domains where the identical distribution assumption no longer holds. Under this new training regime, a model is trained on data arising from some sites/scanners and tested on samples from a new site/scanner unseen during the training process. Formally, at train-time, we observe \( k \) domains (sites/scanners) refer to as sources which have distributions \( P_1, P_2, \ldots, P_k \) over some space \( \mathcal{X} \). At test-time, we test the model on a distinct target domain with distribution \( Q \) over \( \mathcal{X} \).

This multi-domain construct appears in the literature in several medical image segmentation problems. When a model pretrained on sources is further trained on additional target samples with labels, this is often referred to as Transfer Learning [3]. Adding constraints to this setup, Domain Adaptation (DA) assumes access to samples from \( Q \) but without their labels [4,5]. This is a prevalent situation for tasks requiring costly data annotations (e.g., manual tracing of brain lesions). Without labels, DA techniques may still utilize the target information to learn domain-agnostic features [4,6,7], so that task performance is promising irrespective of the input domain. Importantly, these solutions still assume unlabeled data from \( Q \), and this assumption of pre-existing knowledge may be too strong in some practical scenarios.

We therefore consider the problem of Domain Generalization (DG) where we make no assumptions about the target distribution, i.e., no access to the samples from \( Q \) during training. This recently developing construct has been deemed more challenging, but also, more useful to real world problems where we want to generalize without any data from our target [8,9]. Yet, while a few DG approaches exist [10,11], DG is still a nascent concept in medical imaging especially on segmentation problems. Several reasons contribute to this hindered progress, including the lack of theoretical understanding of the existing DG methods and their non-trivial adaptation to segmentation models such as U-Net. Nonetheless, a domain generalizable model is expected to bring practical benefits: retrospectively, we may better leverage existing siloed multi-domain datasets, and prospectively, we can reliably use these models on unseen datasets.

Contribution. In this work, we ask the following question: Can we devise a segmentation model that generalizes
well to unseen data? We answer this as follows: (1) We investigate two mechanically different DG methods, namely domain adversarial neural network and mixup, and identify their theoretical commonality. (2) We use these DG methods to build upon a U-Net segmentation model, tackling the WMH segmentation problem on a multi-site WMH Challenge Dataset and drastically improving performance of a traditional U-Net on the DG task. (3) We further test our model on our local data which is not a part of the aforementioned multi-site dataset. We make our code publicly available.

2. METHODS

Intuition tells us that if our features are invariant to the domain, then the main task should not be affected by the domain of the input. In fact, recent theoretical argument formally suggests such domain invariance in the feature space as a solution for DG. This motivates our proposed approach. We first employ the common domain adversarial training algorithm DANN which learns domain invariant features that fool a domain discriminator. We further show the data-augmentation algorithm mixup may also be viewed as promoting domain invariance. We then propose to use both DANN and mixup after identifying their theoretical connection in DG.

2.1. Domain Adversarial Neural Network for DG

Based on the seminal theoretical work of Ben-David et al. [13, 14], Ganin and Lempitsky [5] proposed the commonly used algorithm DANN which learns domain invariant features that fool a domain discriminator. We further show the data-augmentation algorithm mixup may also be viewed as promoting domain invariance. We then propose to use both DANN and mixup after identifying their theoretical connection in DG.

2.2. Mixup

Besides application of DANN, we also propose to use the common data-augmentation algorithm Mixup [12]. At first glance, this extension is fairly simple. Still, in the presence of multiple domains \( P_1, \ldots, P_k \), this algorithm complements the optimization objective in Eq. (1) because it also aims to produce domain invariance in our learning algorithm (discussed in the next subsection). The algorithm is defined as follows. Suppose we have a batch of data-points \( \{x_i\}_i \) with \( x_p \) and \( x_q \) two distinct data-points from this sample. Further, let \( \lambda \sim \text{Beta}(\alpha, \alpha) \). Then, we define the new mixed data-point:

\[
\tilde{x}_{p,q} = \lambda x_p + (1 - \lambda) x_q. 
\]

With a certain probability (e.g., 0.5 in all our experiments), we can then substitute every data-point in the batch \( \{x_i\}_i \) with a mixed counterpart by randomly pairing the elements of \( \{x_i\}_i \) and using Eq. (3) to combine them. We do remark that the original proposal for mixup also mixes across the label space. Our segmentation setting is slightly different due to the common use of a Dice score loss. Thus, we adopt a loss balancing strategy similar to [12]: \( \mathcal{L}(\tilde{x}_{p,q}) = \lambda \mathcal{L}^s(x_{p,q}) + (1 - \lambda) \mathcal{L}^t(\tilde{x}_{p,q}) \) where \( \mathcal{L}^t \) is the loss with the label \( y_i \) of \( x_i \).

2.3. Theoretical Motivations

DANN and the \( \mathcal{H} \)-divergence. While many works have used the motivation of domain invariant features for DG [8, 16], we note that the original theoretical motivation of DANN was based on the domain adaptation theory proposed by Ben-David et al. [14]. Further, recent work [9] has extended this theory to demonstrate that the application of DANN in DG is justified. In particular, it is shown that the \( \mathcal{H} \)-divergence – a measure of the difference between two domains – appears in an upper bound of the error on an unseen target. Therefore, to proxy minimization of the unseen target error, we can minimize the \( \mathcal{H} \)-divergence in the upper-bound (amongst other quantities, such as the error on the sources). The objective described in Eq. (1) can be interpreted as minimizing the \( \mathcal{H} \)-divergence in a fairly formal sense because the \( \mathcal{H} \)-divergence

https://github.com/xingchenzhao/MixDANN
measures a classifier’s ability to distinguish between domains. Motivated by this, we learn invariant features in Eq. (1) by maximizing the errors of the domain classifier \(d_\mu\).

A Formal Discussion of Mixup. Mixup [12] is a special case of Vicinal Risk Minimization [17]. Usually, in machine learning, we use Empirical Risk Minimization which suggests estimating the true data-distribution by the empirical measure \(P(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i, (x_i, y_i)}\), with the Dirac measure \(\delta_{x, y}\) estimating density at each data-point. In the more general case of Vicinal Risk Minimization, we allow freedom to use a density estimate in the vicinity of the data-point \(x_i\) using the vicinal measure \(u_\lambda\), as \(P_{\lambda}(x, y) = \frac{1}{n} \sum_{i=1}^{n} v_{\lambda}(x_i) \delta_{y_i}(y)\).

In Mixup, the modifications to our data described in Eq. (3) equate to sampling from a certain vicinal distribution \(u_\lambda\): \(u_\lambda(x) = \frac{1}{\lambda} \sum_{j=1}^{\lambda} E_{\lambda} \delta_{x, \lambda x_j + (1 - \lambda)x_j}\), where \(\lambda \sim \text{Beta}\).

Connecting Mixup to the \(\mathcal{H}\)-divergence. From the perspective of the mentioned theory motivating DANN for DG (above), this form of density estimation has interestingly been linked to invariant learning by data-augmentation [17]. In particular, the vicinal measure \(v_{\lambda}\) may be defined to promote learning which is invariant to features of our data (e.g., augmentation by noise seeks to make a neural network’s prediction invariant to noise in the input). By applying Mixup as defined in Eq. (3), we coincidently mix features across the domains \(P_1, \ldots, P_k\) because \(x_p\) and \(x_q\) may be drawn from differing domains. In this sense, the proposed vicinal distribution \(u_\lambda\) estimates the density of a data point \(x_i\) to promote invariance to domain features in our learning algorithm. Hence, Mixup in the presence of multiple domains may be viewed as a technique complementary to DANN. It too is aimed at training \(r_\theta\) and \(c_{\theta}\) to be invariant to domain features so that the errors of the domain discriminator \(d_\mu\) are maximized, and subsequently, the \(\mathcal{H}\)-divergence is minimized. Alg. 1 shows MixDANN, our algorithmic and theoretical combination of Mixup and DANN for DG.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. We evaluate on two WMH datasets consisting of FLAIR (Fluid Attenuated Inverse Recovery), T1, and manual WMH segmentation for each subject: (1) a multi-site public MICCAI WMH Challenge Dataset [1] from three sites (Amsterdam (A), Singapore (S), Utrecht (U)) and (2) our local in-house dataset [18] from a single site (Pittsburgh (P)). See Fig. 2 for distinct scanners/protocols and [1] [18] for details.

Metrics. We use five evaluation metrics to assess the WMH prediction mask \(Y\) against the ground truth segmentation mask \(Y\) (TP: true positive, FP: false positive, FN: false negative). (1) Dice Similarity Coefficient (DSC): \(2(|Y \cap \hat{Y}|) / (|Y| + |\hat{Y}|)\), (2) Hausdorff Distance (H95): \(H_{95} = \max\{\sup_{x \in Y} \inf_{y \in \hat{Y}} d(x, y), \sup_{y \in \hat{Y}} \inf_{x \in Y} d(x, y)\}\) using the 95-th percentile distance, (3) Absolute Volume Difference (AVD) between the predicted and true WMH volume, (4) Level Recall: Computes the \# of correctly detected WMH over the \# of true WMH, (5) Level F1: TP / (TP + 0.5(FP+FN)).

Baseline Models. These baselines build on U-Net [19], and we report their results by a recent DA/DG work [4]: (1) DeepAll is the baseline U-Net with no DA or DG mechanisms which the following models build on. (2) UDA [4] is an unsupervised DA method using all target scans but not their labels. Note this DA method requires target thus has an advantage over DG methods. (3) BigAug [10] is a state-of-the-art DG medical imaging segmentation method with heavy data augmentations. See [4] [10] for full details.

Our Model. Our proposed approaches also build on U-Net. Standard data augmentations (rotation, scale, shear) are applied. (1) We implement our own DeepAll comparable to the DeepAll by [4] for our best effort for fair comparison. (2) DANN: We introduce the domain discriminator (Conv-Conv-FC-FC-FC) \(d_\mu\) to U-Net to the output of the second downsampling layer (Fig. 1). For the purpose of DANN (Eq. (1)), we can treat the U-Net as \(c_{\theta} \circ r_\theta\) where \(r_\theta\) is before and \(c_{\theta}\) is after the second downsampling. We slowly introduce \(\mathcal{L}_D\) to \(r_\theta\) by setting \(\gamma = (2 - \xi) / (1 + e^{-\gamma (\kappa - p)}) - 1\) in Eq. (1) with \(p = \text{epoch} / \max\text{epoch}, \kappa = 3,\) and \(\xi = 0.1\). (3) Mixup: We randomly mix the samples following Eq. (3) to induce domain invariance. (4) MixDANN: Our final model combines DANN and MixDANN. The initial learning rate is 2e-4 for all models. We trained on 80% of the training data against the comparing methods using 2 x NVIDIA RTX2080 Ti.

3.2. Results and Analysis

Each model tests on one target domain after training on the remaining sources (Sources→Target). Generally, the com-
comparison in DG is subjective across different experimental setups. For instance, despite the near identical architectures, our DeepAll slightly under-performs on some targets. As such, we pay special attention to the relative performance gain over the respective DeepAll.

Exp 1: DG within WMH Challenge Dataset. We test on each target by training on the remaining two sources. Table 1 shows the results of all targets and the average across them. We see that despite our weaker DeepAll, MixDANN shows the best absolute avg on three metrics (DSC, AVD, and Lesion F1) and the best relative gain on all metrics. We ablate and see improvements in the order of DANN, Mixup, and MixDANN, also visualized in Fig. 3. We note that AVD assessing the accuracy of the predicted WMH volume, which is often considered as a biomarker of vascular pathology, is most accurate by MixDANN. We also pay special attention to the improvement in the hardest case of A+S→U: This exactly exemplifies a possible scenario where a DeepAll may fail on an unseen dataset but MixDANN can assure robustness.

Exp 2: A+S+U→Pitt. We do a “cross-dataset” DG: train on A+S and test on Pitt. We could not include Pitt as a source since it only has 5 consecutive slices of manual segmentation available. Nonetheless, when we test it as a target as shown in Table 2, we again see consistent improvements over DeepAll. Interestingly, DANN best performs, implying that Mixup and DANN may also individually bring benefits. Our MixDANN with a single U-Net is now ranked 6th on the leader board of the WMH Challenge [1], competitive against other top ensemble U-Nets. We observe poor performance by Mixup on the WMH counting metrics (Recall and F1), suspecting this to be from the different manual annotation standards between the two datasets. We consider this as our future work.

Table 1: WMH Challenge Results. Sources→Target: Trained on Sources and tested on Target among Amsterdam, Singapore, and Utrecht. avg: Average of the target results. gain: avg gain over the respective DeepAll.

<table>
<thead>
<tr>
<th>Model</th>
<th>A→S→U</th>
<th>U→S→A</th>
<th>A→U→S</th>
<th>avg</th>
<th>gain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSC</td>
<td>H95</td>
<td>AVD</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>DeepAll (Setup)</td>
<td>0.430</td>
<td>0.624</td>
<td>0.682</td>
<td>0.595</td>
<td>-</td>
</tr>
<tr>
<td>BigAug [10]</td>
<td>0.534</td>
<td>0.691</td>
<td>0.711</td>
<td>0.645</td>
<td>0.050</td>
</tr>
<tr>
<td>UDA [1]</td>
<td>0.529</td>
<td>0.737</td>
<td>0.782</td>
<td>0.683</td>
<td>0.087</td>
</tr>
<tr>
<td>DeepAll (for Ours)</td>
<td>0.183</td>
<td>0.619</td>
<td>0.781</td>
<td>0.528</td>
<td>-</td>
</tr>
<tr>
<td>DANN</td>
<td>0.315</td>
<td>0.674</td>
<td>0.773</td>
<td>0.587</td>
<td>0.060</td>
</tr>
<tr>
<td>Mixup</td>
<td>0.619</td>
<td>0.691</td>
<td>0.815</td>
<td>0.715</td>
<td>0.187</td>
</tr>
<tr>
<td>MixDANN</td>
<td>0.694</td>
<td>0.700</td>
<td>0.839</td>
<td>0.744</td>
<td>0.217</td>
</tr>
</tbody>
</table>

Table 2: WMH Segmentation Results of A+S+U→Pitt

<table>
<thead>
<tr>
<th>Model</th>
<th>DSC</th>
<th>H95</th>
<th>AVD</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepAll</td>
<td>0.384</td>
<td>0.760</td>
<td>0.895</td>
<td>0.463</td>
<td>-</td>
</tr>
<tr>
<td>BigAug</td>
<td>0.476</td>
<td>0.764</td>
<td>0.851</td>
<td>0.374</td>
<td>-</td>
</tr>
<tr>
<td>UDA (Full Set)</td>
<td>0.545</td>
<td>0.709</td>
<td>0.822</td>
<td>0.360</td>
<td>-</td>
</tr>
<tr>
<td>DeepAll</td>
<td>384.19</td>
<td>43.28</td>
<td>12.81</td>
<td>150.24</td>
<td>-</td>
</tr>
<tr>
<td>DANN</td>
<td>134.03</td>
<td>26.65</td>
<td>24.09</td>
<td>23.40</td>
<td>26.48</td>
</tr>
<tr>
<td>Mixup</td>
<td>42.69</td>
<td>18.05</td>
<td>4.56</td>
<td>21.77</td>
<td>-</td>
</tr>
<tr>
<td>MixDANN</td>
<td>20.57</td>
<td>12.75</td>
<td>3.10</td>
<td>12.14</td>
<td>-7.63</td>
</tr>
<tr>
<td>Lesion Recall</td>
<td>0.634</td>
<td>0.692</td>
<td>0.641</td>
<td>0.656</td>
<td>-</td>
</tr>
<tr>
<td>A→S→U</td>
<td>0.643</td>
<td>0.709</td>
<td>0.691</td>
<td>0.681</td>
<td>0.025</td>
</tr>
<tr>
<td>UDA (Full Set)</td>
<td>0.652</td>
<td>0.841</td>
<td>0.754</td>
<td>0.749</td>
<td>0.093</td>
</tr>
<tr>
<td>DeepAll</td>
<td>0.309</td>
<td>0.623</td>
<td>0.703</td>
<td>0.546</td>
<td>-</td>
</tr>
<tr>
<td>BigAug</td>
<td>0.349</td>
<td>0.630</td>
<td>0.740</td>
<td>0.573</td>
<td>0.028</td>
</tr>
<tr>
<td>UDA (Full Set)</td>
<td>0.556</td>
<td>0.700</td>
<td>0.790</td>
<td>0.682</td>
<td>0.136</td>
</tr>
<tr>
<td>Mixup</td>
<td>0.604</td>
<td>0.685</td>
<td>0.797</td>
<td>0.695</td>
<td>0.150</td>
</tr>
<tr>
<td>MixDANN</td>
<td>0.561</td>
<td>0.673</td>
<td>0.592</td>
<td>0.609</td>
<td>-</td>
</tr>
<tr>
<td>UDA (Full Set)</td>
<td>0.546</td>
<td>0.739</td>
<td>0.649</td>
<td>0.645</td>
<td>0.035</td>
</tr>
<tr>
<td>DeepAll</td>
<td>0.288</td>
<td>0.554</td>
<td>0.697</td>
<td>0.531</td>
<td>-</td>
</tr>
<tr>
<td>DANN</td>
<td>0.309</td>
<td>0.610</td>
<td>0.708</td>
<td>0.542</td>
<td>0.029</td>
</tr>
<tr>
<td>Mixup</td>
<td>0.515</td>
<td>0.642</td>
<td>0.724</td>
<td>0.627</td>
<td>0.114</td>
</tr>
<tr>
<td>MixDANN</td>
<td>0.602</td>
<td>0.651</td>
<td>0.728</td>
<td>0.660</td>
<td>0.147</td>
</tr>
</tbody>
</table>

Do we learn domain-invariance? Fig. 4 shows the t-SNE plots of the second downsampling layer (rθ) output. We see that the features by DeepAll can easily identify certain domains while those by MixDANN blur the boundary among domains as intended. In the context of WMH prediction, MixDANN explicitly suppresses the site/scanner dependent information, thus is more robust when test on unseen data.

Fig. 3: Visualization of WMH predictions (A+S→U). True Positive: Green, False Positive: Yellow, False Negative: Red.

Fig. 4: t-SNE plots of U-Net features trained on Amsterdam (pink), Singapore (gold) and Utrecht (blue) as sources.

4. CONCLUSION

We investigate the domain generalizability of a WMH segmentation deep model to be trained on sources and operate well on an unseen target. We identify a theoretical connection between two DG approaches, namely DANN and Mixup, and jointly incorporate them into U-Net. Using a multi-site WMH dataset and our local dataset, we show our domain invariant learning frameworks bring drastic improvements over other DA/DG methods in both relative and absolute performances.
5. ACKNOWLEDGMENTS
This work was supported by the NIH/NIA (R01 AG063752, RF1 AG025516, P01 AG025204, K23 MH118070), and SCI UR Scholars Award. We report no conflicts of interests.

6. COMPLIANCE WITH ETHICAL STANDARDS
The WMH Challenge study was conducted retrospectively using open access data by WMH Segmentation Challenge (https://wmh.isi.uu.nl/). Ethical approval was not required as confirmed by its license. The Pitt study performed in line with the principles of the Declaration of Helsinki was approved by the Ethics Committee of the University of Pittsburgh.

7. REFERENCES


