Margin-aware Unsupervised Domain Adaptation for Cross-lingual Text Labeling

Dejiao Zhang[1] **Ramesh Nallapati**[1] **Henghui Zhu**[1] **Feng Nan**[1]
Cicero Nogueira dos Santos[1] **Kathleen McKeown**[1,2] **Bing Xiang**[1]
[1]AWS AI
[2] Department of Computer Science, Columbia University
dejiaoz,henghui,nanfen,cicnog,rnallapa,mckeownk,bxiang@amazon.com

Abstract

Unsupervised domain adaptation addresses the problem of leveraging labeled data in a source domain to learn a well-performing model in a target domain where labels are unavailable. In this paper, we improve upon a recent theoretical work (Zhang et al., 2019b) and adopt the Margin Disparity Discrepancy (MDD) unsupervised domain adaptation algorithm to solve the cross-lingual text labeling problems. Experiments on cross-lingual document classification and NER demonstrate the proposed domain adaptation approach advances the state-of-the-art results by a large margin. Specifically, we improve MDD by efficiently optimizing the margin loss on the source domain via Virtual Adversarial Training (VAT). This bridges the gap between theory and the loss function used in the original work Zhang et al. (2019b), and thereby significantly boosts the performance. Our numerical results also indicate that VAT can remarkably improve the generalization performance of both domains for various domain adaptation approaches.

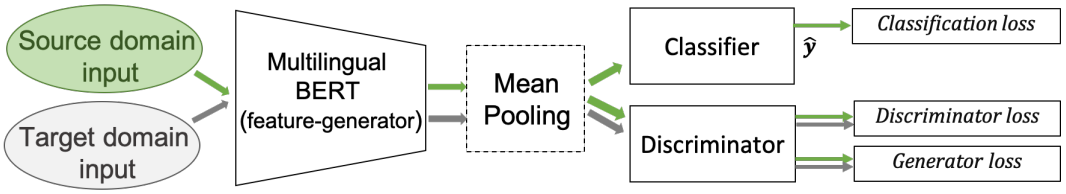
1 Introduction

Unsupervised domain adaptation provides an appealing solution to many applications where direct access to a massive amount of labeled data is prohibitive or very costly (Sun and Saenko, 2014; Vazquez et al., 2013; Stark et al., 2010; Keung et al., 2019). For example, we often have sufficient labeled data for English, while very limited or even no labeled data are available for many other languages. Successfully transferring knowledge learned from the English domain to other languages is of great interest in solving many tasks in natural language processing.

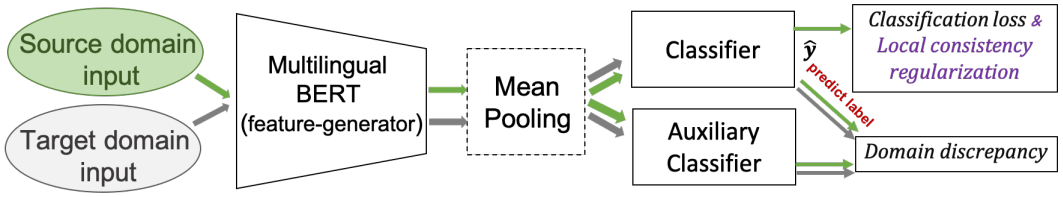
Many recent successes in unsupervised domain adaptation have been achieved by learning domain invariant features that are simultaneously being discriminative to the task in the source domain (Chen et al., 2018; Ganin and Lempitsky, 2014; Ganin et al., 2016; Tzeng et al., 2017). Following this line, Keung et al. (2019) propose a language-adversarial training approach for cross-lingual document classification and NER. They leverage the benefit of contextualized word embeddings by using multilingual BERT (Devlin et al., 2019) as the feature generator, and adopt the GAN framework (Goodfellow et al., 2014) to align the features from the two domains. Keung et al. (2019) show significant improvement over the baseline where the pretrained multilingual BERT is finetuned on the English data alone and testing on the same tasks in other languages. However, Keung et al. (2019), as well as the works mentioned above, are inspired by the pioneering work of Ben-David et al. (2010), which only rigorously studies domain adaptation in the setting of binary classification; there is a lack of theoretical guarantees when it comes to multiclass classification.

In this work, we are instead motivated by a recent work (Zhang et al., 2019b) that focuses on the theoretical analysis of unsupervised domain adaption for multiclass classification and provides explicit guidance for algorithm design. Instead of training a discriminator that predicts if the representations are from the source domain or the target domain (Keung et al., 2019; Ganin and Lempitsky, 2014; Ganin et al., 2016), Zhang et al. (2019b) proposes to optimize an auxiliary classifier which, together with the classifier, minimizes the discrepancy between the two domains via adversarial training. We apply this approach to cross-lingual text labeling tasks, which, as demonstrated in Section 4, outperforms Keung et al. (2019) by a large margin. To the best of our knowledge, we are the first to apply the novel theoretical findings of Zhang et al. (2019b) for unsupervised domain adaptation in NLP.

Another contribution of our work lies in identifying the gap between theory and the actual loss



(a) Keung et al. (2019). The source features and target features are only input to the discriminator to calculate the discriminator loss and the generator loss. For the classification loss, mean pooling is not applied to NER.



(b) **Regularized MDD**. In MDD, the features from the two domains are input to both the classifier and the auxiliary classifier to estimate the domain discrepancy. We improve MDD by effectively optimizing the classification margin loss of the source domain via local consistency regularization, which bridges the gap between theory and the loss function used in the original work Zhang et al. (2019b). Mean pooling is not applied to NER.

Figure 1: Regularized MDD vs. Keung et al. (2019).

function being used in Zhang et al. (2019b). Specifically, Zhang et al. (2019b) use the cross-entropy loss as a proxy to optimize the classification margin loss on the source domain, whereas the cross-entropy loss often leads to poor margins (Liu et al., 2016; Elsayed et al., 2018). To tackle this problem, we augment the cross-entropy loss with Virtual Adversarial Training (VAT) (Miyato et al., 2018). As shown in Zhang et al. (2019a), the local consistency regularization introduced by VAT is capable of promoting large classification margin by optimizing the classification boundary error. This is further demonstrated in Section 4 that the incorporation of VAT leads to remarkable improvement over Zhang et al. (2019b).

Although the pretrained language models (Devlin et al., 2019; Peters et al., 2018; Radford et al., 2019) have provided a good foundation for many downstream tasks, to leverage them for unsupervised domain adaptation, we need to tackle the potential overfitting problem, especially when we only have limited labeled data in the source domain but can require many training iterations to minimize the domain discrepancy. As shown in Section 4, VAT can efficiently prevent overfitting in the source domain, and hence significantly improve the generalization in the target domain. This matches the theoretical insights (Ben-David et al., 2010; Zhang et al., 2019b) that the generalization of the target domain can be boosted as a consequence of the improvement in the source domain.

2 Related Work

Inspired by a pioneering work (Ben-David et al., 2010), there has been a surge of interest in learning domain invariant representations (Ganin and Lempitsky, 2014; Ganin et al., 2016; Keung et al., 2019; Chen et al., 2018; Tzeng et al., 2017) for unsupervised domain adaptation. At a high level, these methods leverage deep neural networks (DNNs) to learn rich representations, and adopt adversarial training (Goodfellow et al., 2014) to promote the emergence of domain invariant representations that are simultaneously being discriminative to the predictor learned in the source domain.

In the mostly related work, Keung et al. (2019) apply such strategy to multilingual document classification and NER, see Figure 1a. Although Keung et al. (2019) have achieved remarkable improvements over the baseline, the underlying theory (Ben-David et al., 2010) is only applicable to binary classification with restrictive 0-1 loss. There is a lack of theoretical understanding of Keung et al. (2019) when it comes to multiclass classification with more general loss functions.

In a recent theory work, Zhang et al. (2019b) extend the previous theories to the multiclass classification setting. Instead of training an additional discriminator, Zhang et al. (2019b) proposes to train an auxiliary classifier that shares the same structure as the classifier. The discrepancy between the two domains is optimized by playing the minimax game between the two classifiers. As illustrated

in Figure 1b, we improve upon this new strategy to address two cross-lingual classification tasks, where the pretrained multilingual BERT is used as the feature generator followed by two identical classifiers with different parameters. Numerical results in Section 4 demonstrate that our proposed approach outperforms both Keung et al. (2019) and Zhang et al. (2019b) by a large margin.

Another line of relevant work is the consistency regularization technique used to force the model output to remain unchanged under input perturbations. As observed in the literature, it can promote large classification margin and significantly improve the performance in semi-supervised learning (Bachman et al., 2014; Miyato et al., 2018; Laine and Aila, 2016; Xie et al., 2019; Berthelot et al., 2019). To effectively optimize the classification margin loss proposed by Zhang et al. (2019b), we augment its original objective with virtual adversarial training (Miyato et al., 2018). By doing so, we bridge the gap between the theory and the loss function used in Zhang et al. (2019b), which in turn yields remarkable improvement on the generalization performance of both domains.

3 Model

We formalize the unsupervised domain adaptation as follows. Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} = \{1, \ldots, K\}$ denote the input and output space of the model, respectively. We consider two domains $\mathcal{S}, \mathcal{T} \in \mathcal{X} \times \mathcal{Y}$, which are referred to as the source domain and the target domain correspondingly. Our ultimate goal is to learn a well-performing classifier on the target domain, while labels are only available for the source domain.

Let $\psi : \mathbb{R}^d \to \mathbb{R}^h$ denote the feature extractor, which we use to transform the minimization of the domain discrepancy from the data space to the representation space. Let $f, f' : \mathbb{R}^h \to \mathbb{R}^K$ denote the scoring functions associated with the classifier and the auxiliary classifier, respectively. Note that, for a scoring function, *e.g.,* f, the outputs of each dimension indicate the prediction confidence. Hence, given an input example x, the prediction is followed as:

$$\widehat{y} = \arg\max_{k \in \mathcal{Y}} f(\psi(x)) \ . \tag{1}$$

Let σ denote the softmax function, *i.e.,*

$$\sigma_j(\mathbf{z}) = \frac{e^{\mathbf{z}_j}}{\sum_{i=1}^{K} e^{\mathbf{z}_i}}, \quad \mathbf{z} \in \mathbb{R}^K \text{ and } j \in \mathcal{Y}. \tag{2}$$

Following Zhang et al. (2019b), we choose the standard cross-entropy loss for the classification task in the source domain,

$$\mathcal{L}_{\mathcal{S}} := \mathbb{E}_{x,y \sim \mathcal{S}} \left[-\log \left[\sigma_y \left(f(\psi(x)) \right) \right] \right] \ . \tag{3}$$

On par with the classification loss, we need to optimize the domain discrepancy between the two domains. Before that, we first introduce the measurements we used to quantify the discrepancy between f' and f on each domain. Let $\widehat{y}_s, \widehat{y}_t$ denote the predictions given by the classifier f (see Equation (1)), then

$$\mathcal{D}_{\mathcal{S}}(f', f) := \mathbb{E}_{x_s \sim \mathcal{S}} \left[-\log \left[\sigma_{\widehat{y}_s} \left(f'(\psi(x_s)) \right) \right] \right] \ ,$$
$$\mathcal{D}_{\mathcal{T}}(f', f) := \mathbb{E}_{x_t \sim \mathcal{T}} \left[\log \left[1 - \sigma_{\widehat{y}_t} \left(f'(\psi(x_t)) \right) \right] \right] \ .$$

As we can see, both $\mathcal{D}_{\mathcal{S}}(f', f)$ and $\mathcal{D}_{\mathcal{T}}(f', f)$ are increasing functions of the difference between the auxiliary classifier f' and the classifier f, *i.e.,* they both increase when the output of f' at the class predicted by f has lower confidence. Following Zhang et al. (2019b), the domain discrepancy is then approximated as,

$$\max_{f'} \left[\mathcal{D}_{\mathcal{T}}(f', f) - \gamma \mathcal{D}_{\mathcal{S}}(f', f) \right], \ \gamma > 1 \tag{4}$$

In other words, given a specific classifier f, the domain discrepancy is induced by f' as the maximal difference between the disparities of f and f' on the two domains. Here γ is proposed by Zhang et al. (2019b) to promote convergence of the optimization of the domain discrepancy. Given $\gamma > 1$ and no restrictions on f', Zhang et al. (2019b) prove that the global minimum of the discrepancy defined in (4) is achieved when $\psi(\mathcal{S}) = \psi(\mathcal{T})$.

Intuitively, solving the inner maximization requires finding a f' that can maximally differ from f on the target domain while staying close to f on the source domain. Minimizing the domain discrepancy naturally induces minimax optimization. The main objective thereby can be formulated as,

$$\min_{f, \psi} \left[\mathcal{L}_{\mathcal{S}} + \max_{f'} \left[\mathcal{D}_{\mathcal{T}}(f', f) - \gamma \mathcal{D}_{\mathcal{S}}(f', f) \right] \right] \ . \tag{5}$$

3.1 Promoting better generalization by VAT

The objective function (5) is identical to the loss function proposed in Zhang et al. (2019b), which, as demonstrated in Section 4, outperforms Keung et al. (2019) by a large margin. However, there are

Algorithm 1 MDD for multilingual document classification.

Input: pretrained BERT model ψ, classifier f, auxiliary classifier f', and the associated learning rates $\eta_\psi, \eta_c, \eta_{f'}$. Let $\theta_\psi, \theta_f, \theta_{f'}$ denote the parameters of different components and γ be the chosen hyperparameter value in (8). We use batch size of 1 for the purpose of illustration.
while not converged **do**
  Solve inner maximization:
    $x_s \leftarrow \text{BatchIterator}(\mathcal{S}), \quad x_t \leftarrow \text{BatchIterator}(\mathcal{T})$
    $h_s = \text{MeanPool}(\psi(x_s)), \quad h_t = \text{MeanPool}(\psi(x_t))$
    $\widehat{y}_s = \arg\max_k f(h_s), \quad \widehat{y}_t = \arg\max_k f(h_t)$
    $\widehat{\mathcal{D}}_\mathcal{S} = -\log[\sigma_{\widehat{y}_\mathcal{S}}(f'(h_s))], \quad \widehat{\mathcal{D}}_\mathcal{T} = \log[1 - \sigma_{\widehat{y}_\mathcal{T}}(f'(h_t))]$
    $\theta_{f'} \mathrel{+}= \eta_{f'} \nabla_{\theta_{f'}} (\widehat{\mathcal{D}}_\mathcal{T} - \gamma \widehat{\mathcal{D}}_\mathcal{S})$
  Solve outer minimization:
    Repeat steps 4 to 7
    freeze f' and update $\theta_\psi \mathrel{-}= \eta_\psi \nabla_{\theta_\psi} (\widehat{\mathcal{D}}_\mathcal{T} - \gamma \widehat{\mathcal{D}}_\mathcal{S})$
  Solve the classification task
    $x_s, \; y_s \leftarrow \text{BatchIterator}(\mathcal{S})$
    $h_s = \text{MeanPool}(\psi(x_s))$
    $\widehat{\mathcal{L}}_\mathcal{S} = -\log[\sigma_{y_s}(f(h_s))]$
    $(\theta_f, \theta_\psi) \mathrel{-}= \eta_c \nabla_{\theta_\psi, \theta_f} \widehat{\mathcal{L}}_\mathcal{S}$
    if use VAT **then**
      $(\theta_f, \theta_\psi) \mathrel{-}= \eta_c \nabla_{\theta_\psi, \theta_f} \widehat{\mathcal{R}}^e_\mathcal{S}$
end while

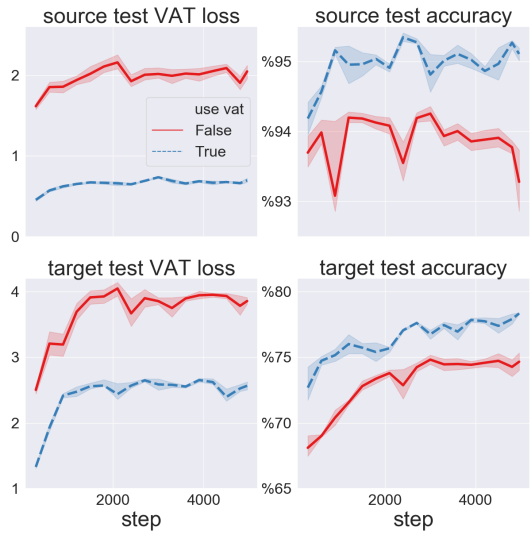


Figure 2: Illustration of the regularization effect of VAT. We apply MDD to solve the cross-lingual document classification problem on MLDoc (Schwenk and Li, 2018). The English corpus is the source domain, and the Italian corpus is the target domain. In each plot, the results are summarized over 4 runs with the solid lines representing the means, and the shaded regions indicating the 75% confidence intervals. The red lines indicate when only MDD is applied, and the blue lines represent the results when VAT is used as well.

still two hurdles we need to cross. Firstly, Zhang et al. (2019b) use the cross-entropy loss, *i.e.,* (3), as a proxy to optimize the classification margin loss on the source domain, which results in a gap between its theoretical results and the loss function being used, especially given that the cross-entropy loss often leads to poor margins (Liu et al., 2016; Elsayed et al., 2018). Secondly, we follow the literature by using a pretrained language model as the feature generator, which provides a good initialization for unsupervised domain adaptation. However, we need to consider the potential overfitting problem, since we usually have limited labels in the source domain, while requiring many training iterations to optimize domain discrepancy via adversarial training. Therefore, the model can overfit to the source domain training data during the training process.

To remedy these two issues, we propose regularizing the source domain classification task via Virtual Adversarial Training (VAT) (Miyato et al., 2018), which is defined as the following,

$$\mathcal{R}_\mathcal{S} := \max_{\delta; \|\delta\| \le \varepsilon} \text{KL}\left[\sigma(f(\psi(x_s))) \| \sigma(f(\psi(x_s + \delta)))\right] . \tag{6}$$

This term regularizes the predictions being consistent within the ε norm ball of each input. As indicated in Zhang et al. (2019a), the local consistency regularization described in (6) can effectively

promote large margin by optimizing the classification boundary error. As demonstrated in Miyato et al. (2018), the maximization in (6) can be well approximated by a pair of forward- and backward-propagations.

Note that, the input is discrete for the language data, hence we apply VAT to the embedding space and consider the following,

$$\mathcal{R}^e_{\mathcal{S}} := \max_{\delta;\|\delta\|\leq\varepsilon} \mathrm{KL}\left[\sigma(f(\psi(e[x_s])))\|\sigma(f(\psi(e[x_s]+\delta)))\right] \tag{7}$$

We use $e[x_s]$ to denote the embedding of the discrete input x_s. In summary, our main objective followed as

$$\min_{f,\psi}\Big(\mathbb{E}_{(x,y)\sim\mathcal{S}}\mathcal{L}_{\mathcal{S}} + \mathbb{E}_{x\sim\mathcal{S}}\mathcal{R}^e_{\mathcal{S}} + \max_{f'}\left[\mathbb{E}_{x\sim\mathcal{T}}\mathcal{D}_{\mathcal{T}} - \gamma\mathbb{E}_{x\sim\mathcal{S}}\mathcal{D}_{\mathcal{S}}\right]\Big) \tag{8}$$

As illustrated in Figure 2, by imposing the local consistency regularization on each data point during training, VAT can remarkably improve the generalization of both domains. This improvement can be explained by the theoretical insights given by Ben-David et al. (2010); Zhang et al. (2019b), which state that the generalization error of the target domain can be upper bounded by the summation of the source error, the domain discrepancy, and a constant value. Therefore, the generalization of the target domain is improved by using VAT to boost the generalization of the source domain.

3.2 Optimization

The pseudo code of our proposed method can be found in Algorithm 1. Note that, in the outer minimization, the domain discrepancy loss is not differentiable with respect to the parameters of the classifier, *i.e.,* f. To address this problem, we follow Zhang et al. (2019b) to instead train the feature exactor ψ to solve the outer minimization of the domain discrepancy, for which the gradients are backpropagated through the auxiliary classifier f', *i.e.,* step 11 in Algorithm 1. However, in Zhang et al. (2019b) the feature extractor ψ is trained through a gradient reversal layer (Ganin and Lempitsky, 2014), which is often not stable and requires extra hyperparameter tuning. In contrast, we optimize f' and ψ alternately, which we find is more stable in practice.

4 Numerical Results

We evaluate the performance of the proposed approach on two different NLP tasks: text classification, where we use the MLDoc corpus (Schwenk and Li, 2018); and named entity recognition, where we use the CoNLL 2002/2003 NER corpus (Tjong Kim Sang, 2002; Sang and De Meulder, 2003). We compare our regularized MDD approach against both Keung et al. (2019) and the baseline. For the baseline, we train the model on the English corpus only, while evaluating on the corpus of the other languages. We also do an ablation study to demonstrate VAT can yield remarkable performance boost for all three approaches evaluated in this section.

We implement all three approaches in PyTorch (Paszke et al., 2017) with the HuggingFace library (Wolf et al., 2019). We use the pretrained cased multilingual BERT (Devlin et al., 2019) as the initialization for the feature extractor, which is followed by a linear classifier of size $768 \times K$ with K indicating the number of classes. We train an additional linear discriminator with size 786×2 for Keung et al. (2019), and an auxiliary classifier with the same size of the primary classifier, *i.e.,* $768 \times K$, for MDD. We use the Adam optimizer (Kingma and Ba, 2015) with batch size of 24 for all approaches. We use a constant learning rate η_c =1e-5 for optimizing the classification loss on the source domain, and use the learning rates η_ψ, $\eta_{f'}$ and η_d to optimize the feature extractor, the auxiliary classifier (MDD), and the discriminator (Keung et al., 2019) correspondingly.

4.1 MLDoc

We first evaluate the performance of our proposed method on the MLDoc corpus (Schwenk and Li, 2018). For each language in MLDoc, it contains four balanced classes extracted from the Reuters News RCV1 and RCV2 datasets. Following the same setting of Keung et al. (2019), we use the labeled *english.train.1000* dataset to optimize the classification loss, while only using the text portion of *english.train.10000* and *target-language.train.10000* to optimize the domain discrepancy measured in MDD and Keung et al. (2019). In this section, we set the perturbation magnitude $\varepsilon = 0.5$ for VAT (see Eq (7)), and use maximal input length of 80. We set $\gamma = 4$, η_ψ =2e-7, $\eta_{f'}$ =2.5e-4 for MDD, and set η_ψ =1e-7, η_d =2.5e-4 for Keung et al. (2019).

VAT improves the generalization of both domains Table 1 shows that VAT can significantly boost the generalization performance of the target domain for all three approaches. As we mentioned

Table 1: Classification accuracy of the MLDoc testing data. English is the source domain. We underline the best results when VAT is not used. As we can see, MDD can outperform Keung et al. (2019) by a large margin on most target domains, no matter whether VAT is used or not. The reported results are averaged over four runs.

	En	**De**	**Es**	**Fr**	**It**	**Ja**	**Zh**	**Ru**
Baseline	94.3	80.7	71.0	67.6	60.6	70.3	70.8	65.9
Baseline + VAT	95.4	86.1	76.7	70.6	68.8	71.1	78.3	67.3
Keung et al. (2019)	-	89.1	76.9	85.2	72.1	75.1	82.9	68.7
Keung et al. (2019) + VAT	-	90.8	78.8	86.6	75.9	**77.4**	86.5	**71.9**
MDD	-	90.6	77.8	87.5	75.6	75.9	85.9	66.7
MDD + VAT	-	**91.9**	**87.2**	**87.9**	**77.9**	77.1	**87.5**	70.1

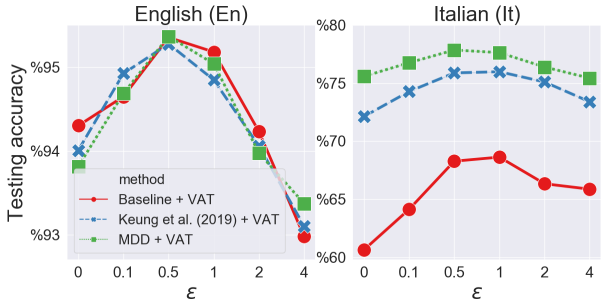


Figure 3: Regularization effect of VAT over different regularization strengths, *i.e.,* different ε values in Equation (7). English (En) is the source domain and Italian (It) is the target domain. The reported testing accuracy are averaged over 4 runs.

before, one possible explanation is indicated by the theoretical insights (Ben-David et al., 2010; Zhang et al., 2019b) that the generalization error of the target domain can be upper bounded by the summation of the source error, the domain discrepancy, and a constant value. Since VAT can effectively improve the generalization of the source domain by imposing the local consistency regularization into the learning objectives, the generalization of the target domain is boosted as a result.

In Figure 3, we evaluate the effectiveness of VAT over different regularization strengths. VAT is capable of enhancing the performance of all three approaches over a wide range of ε values. On the other hand, the improvement is diminishing as we keep increasing the ε values. As shown in Zhang et al. (2019a), the local consistency regularization introduced in Eq (6) can effectively promote large classification margin by optimizing the classification boundary error. Thereby, Figure 3 indicates the trade-off between classification accuracy and classification margin.

MDD outperforms Keung et al. (2019) In Table 1, the comparison between the baseline and the domain adaption approaches demonstrates the effectiveness of optimizing domain discrepancy in successfully transferring knowledge from the source domain to the target domain. On the other hand, Table 1 also shows that MDD can outperform Keung et al. (2019) on most target domains, no matter whether VAT is used or not. We attributed this to the fact that MDD is more theoretically validated, *i.e.,* the underlying theory for MDD directly targets domain adaptation in multiclass classification with more general classification loss function. In contrast, the underlying theoretical support for Keung et al. (2019) only applies to binary classification with the restrict 0-1 loss.

To further compare our regularized MDD approach against Keung et al. (2019), in Figure 4 we report the testing accuracy of all seven target domains over different hyperparameter values. As we can see, the regularized MDD can generally outperform Keung et al. (2019) with VAT over a wide range of hyperparamter values. Moreover, MDD is comparatively more stable than Keung et al. (2019), though they both build upon adversarial training which can cause instability during learning. This again suggests the advantages of MDD over simply training a discriminator to predicts if the representations are from the source domain or the target domain (Keung et al., 2019).

4.2 NER

In this section, we evaluate the proposed approach on the CoNLL 2002/2003 NER corpus (Tjong Kim Sang, 2002; Sang and De Meulder, 2003). We apply VAT to each input. Given that NER requires token level classification, we need to add comparatively large perturbation to guarantee sufficient regularization for each token. Hence, we set $\varepsilon = 4$ for VAT with the maximal input length being 100. We set η_ψ =1e-7 for both MDD and Keung

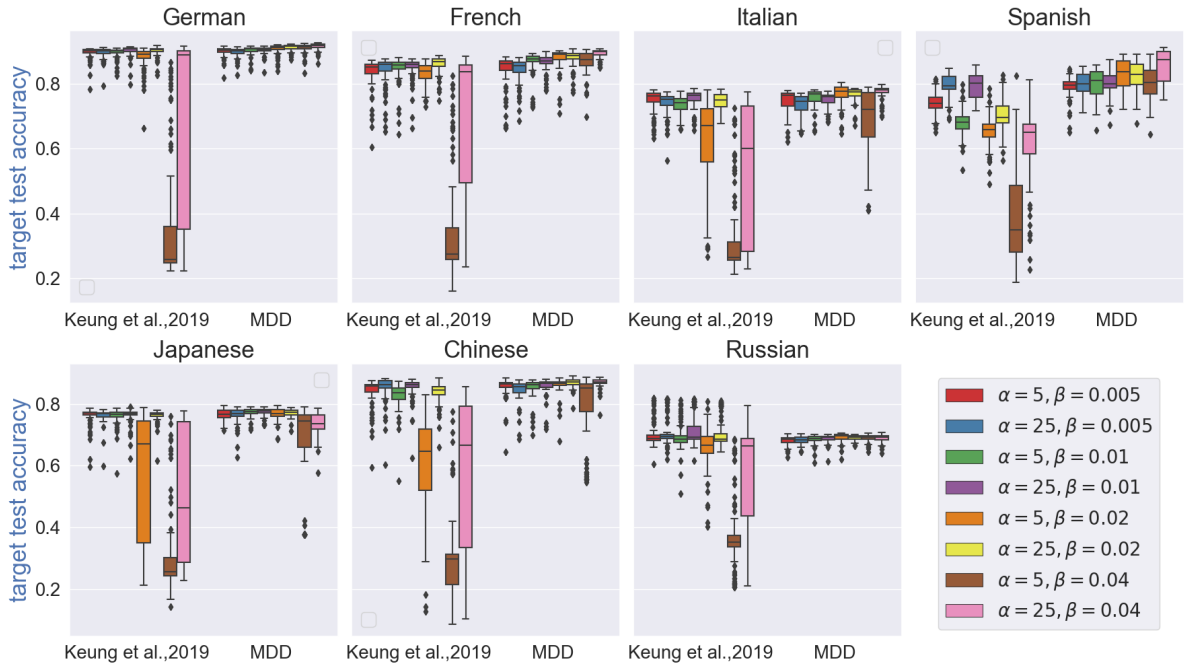


Figure 4: MDD is more stable than Keung et al. (2019) over different learning rates. For both approaches, we use VAT with $\varepsilon = 0.5$. For better visualization, we use α, β to indicate the ratios between the learning rates used for optimizing different components. Specifically, we use $\beta = \eta_\psi / \eta_c$ for both approaches; and use $\alpha = \eta_{f'} / \eta_c$, $\alpha = \eta_d / \eta_c$ for MDD and Keung et al. (2019), respectively. The results are summarized over 4 runs.

et al. (2019), and set $\eta_{f'}$ =5e-5 and η_d =5e-5 for MDD and Keung et al. (2019), respectively.

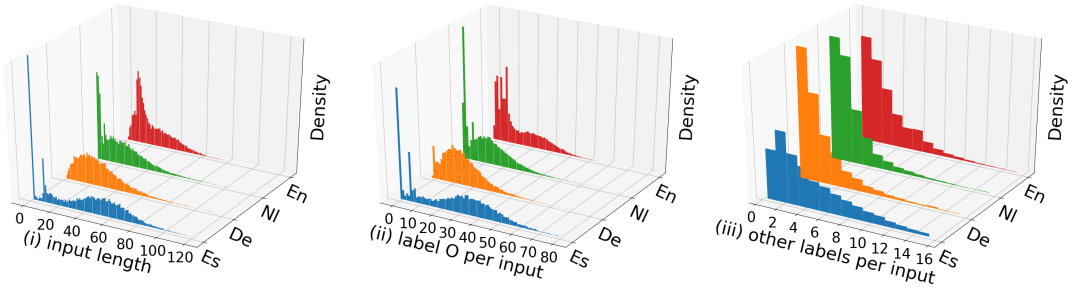
We summarize the data statistics in Figure 5. As indicated by the values of y-axes in Figure 5b, this dataset is highly imbalanced where the "O" label accounts for more than 80% of the labels of each domain. We evaluate all approaches using the F1 score, and the results are summarized in Table 2. Once again, our regularized MDD can generally achieve the best results on most target domains. Moreover, without VAT, MDD constantly outperforms Keung et al. (2019) on all target domains.

To gain more insights into Table 2, we investigate the relationship between domain discrepancy and the generalization on the target domain. In Figure 5a, we plot the statistics of the inputs and the associated labels. As indicated by Figure 5a (i), regarding the input length, Dutch (Nl) is most similar to English (En), while Spanish (Es) shares the least similarity with English. We hypothesize that the comparatively larger similarity shared by Dutch and English explains why all three approaches achieve the best F1 score on Dutch in Table 2. Following this hypothesis, the comparatively smaller similarity between Spanish and English, can also explain why both MDD and Keung et al. (2019) achieve the least improvement over the baseline on the target domain. In other words, the comparatively larger dissimilarity between English and Spanish makes it hard for both Keung et al. (2019) and MDD to effectively optimize the domain discrepancy.

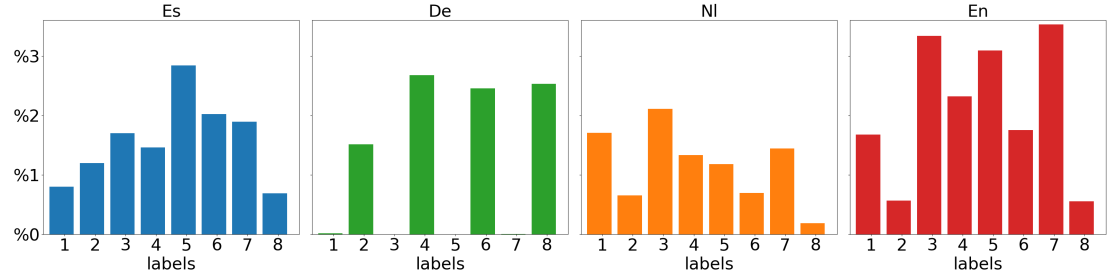
Table 2: F1 scores on the CoNLL2002/2003 testing data. We underline the best results when VAT is not used. Again, MDD achieves the best performance on most target domains. Moreover, without VAT, MDD outperforms Keung et al. (2019) by a large margin on all target domains. The results are averaged over 4 runs.

	En	**De**	**Es**	**Nl**
Baseline	90.0	69.0	73.0	77.3
Baseline + VAT	90.4	69.9	74.3	78.3
Keung et al. (2019)	-	70.7	73.2	78.0
Keung et al. (2019)+VAT	-	**72.3**	74.4	79.3
MDD	-	<u>71.6</u>	<u>73.9</u>	<u>78.4</u>
MDD + VAT	-	72.1	**75.0**	**79.4**

However, Spanish gets a better F1 score than German (De) does for all three approaches, though German shares more similarity with English in terms of the statistics of inputs, as indicated by Figure 5a. We suspect this is caused by the significant difference between German and English in the distribution of labels in the minority group. As



(a) From left to right **(i)** length of each input after BERT tokenization, where the mean and the standard deviation of the input length are En=(21, 13), Nl=(21, 17), De=(28, 16), Es=(44,32) **(ii)** distribution of the number of the label "O" per input; and **(iii)** distribution of the number of the other labels (excluding "O") per input.



(b) Distribution of the labels in the minority group (excluding label "O"). For the purpose of better visualization, we exclude label "O" in each plot, which, as indicated by the values of the y-axes, accounts for more than 80% of the overall labels of each language.

Figure 5: Data statistics of CoNLL 2002/2003.

shown in Figure 5b, German only has four classes besides class "O". In contrast, the other two target domains spread over all the other eight classes. Furthermore, the four classes of German corresponds to four comparatively smaller classes in the English domain.

5 Conclusion

In this paper, we followed the novel theoretical findings of Zhang et al. (2019b), and applied the Margin Disparity Discrepancy (MDD) based unsupervised domain adaptation approach to address the cross-lingual text labeling problems. We demonstrated that MDD can generally outperform the current state-of-the-art model (Keung et al., 2019) by a large margin.

We further improve MDD by identifying the gap between theory and the actual loss function being used in the original work (Zhang et al., 2019b). We resolve the problem by using Virtual Adversarial Training (VAT) (Miyato et al., 2018), which, as demonstrated by our numerical results, leads to remarkable improvement over Zhang et al. (2019b). We attribute this to the fact that VAT is capable of promoting large classification margin by optimizing the classification boundary error Zhang et al. (2019a). This also explains why VAT can generally boost the generalization of the source domain for all three approaches explored in this paper, which in turn leads to the generalization improvement on the target domain.

The remarkable improvement achieved by VAT also motivates us to explore more sophisticated regularization to further improve the performance of various unsupervised domain adaptation approaches. One promising direction is replacing the VAT with adversarial training, which, as proven in Zhang et al. (2019a), yields a reliable classifier that is robust to adversarial attacks in the source domain. To successfully transferring the robustness from the source domain to target domain is of great interest for both theory and practical applications. We leave this as future work.

References

Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles. In *Advances in neural information processing systems*, pages 3365–3373.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. 2018. Large margin deep networks for classification. In *Advances in neural information processing systems*, pages 842–852.

Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1355–1360.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Michael Stark, Michael Goesele, and Bernt Schiele. 2010. Back to the future: Learning shape models from 3d cad data. In *Bmvc*, volume 2, page 5. Citeseer.

Baochen Sun and Kate Saenko. 2014. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, volume 1, page 3.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.

David Vazquez, Antonio M Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. 2013. Virtual and real world adaptation for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):797–809.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's trans-

formers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. 2019a. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*.

Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. 2019b. Bridging theory and algorithm for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7404–7413, Long Beach, California, USA. PMLR.