DOMAIN ADAPTATION VIA MUTUAL INFORMATION MAXIMIZATION FOR HANDWRITING RECOGNITION

Pei Tang[⋆]*, Liangrui Peng*[⋆]*, Ruijie Yan*[⋆]*, Haodong Shi*[⋆]*, Gang Yao*[⋆]*, Changsong Liu*[⋆]*, Jie Li*[†]*, Yuqi Zhang*[†]

[⋆] Beijing National Research Center for Information Science and Technology
[⋆] Department of Electronic Engineering, Tsinghua University, Beijing, China
[†]Shanghai Pudong Development Bank, Shanghai, China

ABSTRACT

Deep learning models for handwriting recognition have been developed in recent years. To improve the model's generalization ability for sequence modeling task, this paper proposes to use domain adaptation with statistical distribution alignment and entropy regularization. For statistical distribution alignment, a domain adaptation loss function is proposed by using both the first and second order statistical information of deep feature representations, which is equivalent to maximizing the mutual information in feature spaces of the source domain and target domain. For entropy regularization, the entropy of the predicted text symbols of unlabeled samples in the target domain is also utilized as an additional loss function, which maximizes the mutual information between the feature space and pattern space in the target domain. Experimental results on the IAM handwriting dataset have demonstrated the effectiveness of the proposed domain adaptation method for sequence modeling task.

Index Terms— Domain Adaptation; Statistical Distribution Alignment; Entropy Regularization; Handwriting Recognition; Deep Learning

1. INTRODUCTION

Handwriting recognition is a challenging sequence modeling task due to the huge variations in handwriting styles. In recent years, deep learning based handwriting recognition models perform well under the assumption of the training and testing data from the same distribution, which is hardly satisfied in real applications. Therefore, it is crucial to improve the model's generalization ability with limited labeled training samples.

Domain adaptation [1, 2] is a promising way to alleviate the domain shift problem between the source domain and target domain. There are two kinds of major domain adaptation approaches: the statistical feature distribution alignment and adversarial training. The former one, including maximum mean discrepancy (MMD) [3, 4], multi-kernel MMD (MK-MMD) [5], and correlation alignment distance (CORAL) [6, 7], etc., computes and minimizes the discrepancy of statistical distributions between the source domain and target domain explicitly. The latter one [8, 9] trains a domain classifier to distinguish the features' domain, and a feature extractor to confuse the classifier. In this way, the feature distributions in the source domain and target domain are aligned implicitly.

The above methods have been mainly used for image classification tasks [5] in stead of sequence modeling tasks, such as isolated character recognition [10]. For domain adaptation of handwritten word recognition, Zhang et al. [11] compare the effectiveness of MMD, CORAL and adversarial training, and Kang et al. [12] use adversarial training to transfer the learned model parameters from synthetic images to real images. However, these two methods have not conducted experiments at text line level.

For deep learning based handwriting recognition model, images are usually processed by a convolutional neural network (CNN) to extract feature maps, then converted into feature representations by a recurrent neural network (RNN) based encoder. For the outputs of encoder, there are two main decoding techniques to convert them into recognized results: connectionist temporal classification (CTC) [13] and attentional decoder which is implemented by another RNN [14, 15, 16]. Recently, attempts have also been made to use Transformer [17, 18] with self-attention mechanism for both encoding and decoding.

In this paper, we propose a novel domain adaptation method for handwriting recognition based on statistical distribution alignment and entropy regularization. For statistical distribution alignment, we proposed a distribution discrepancy estimator called MECOV by using both the mean vectors and covariance matrices of samples in the source domain and target domain, which combines the advantages of MMD and CORAL. For entropy regularization, we enlarge margins of features across different character classes by minimizing the entropy of predicted probabilities [19] without supervision in the target domain. Both the proposed MECOV and entropy regularization can be interpreted from the perspective

This research is supported by a grant from the Institute for Guo Qiang, Tsinghua University. The second author is also supported by National Key R&D Program of China.

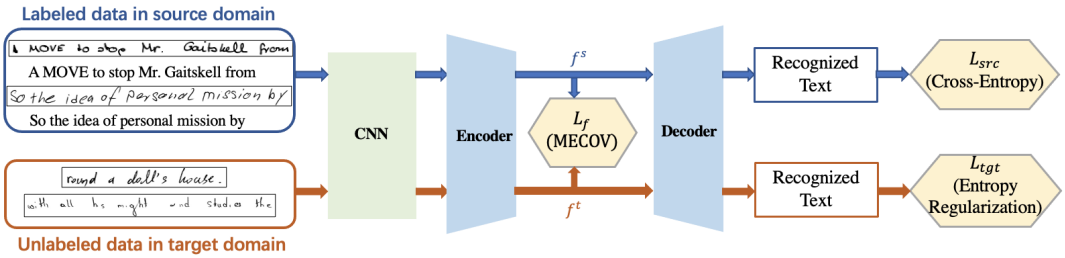


Fig. 1. System framework.

of mutual information maximization.

We construct a handwriting recognition model that obtains feature representations from text line or word images by using a CNN and an RNN encoder, and decodes the features into recognized text by a Transformer decoder. Experiments are carried out on the IAM handwriting dataset [20] at both word level and text line level.

The main contributions of this paper are as following:

1. A novel statistical distribution alignment method which uses both the first and second order statistical information is proposed for domain adaptation in sequence modeling task.
2. An entropy regularization method is proposed by adding an objective function of the Shannon entropy of the predicted text symbols in the target domain.
3. The proposed methods are incorporated into a model consisting of a CNN, an RNN encoder and a Transformer decoder. Experimental results on the IAM handwriting dataset [20] at both text line level and word level demonstrate the effectiveness of the proposed methods.

2. METHODOLOGY

2.1. System framework

The system framework is shown in Fig. 1, which is composed of a CNN, an RNN encoder and a Transformer[17] decoder. The CNN is a modified version [21] of ResNet [22]. The RNN encoder is a two-layer LSTM network, and the Transformer contains two decoder layers. The objective function includes the cross-entropy loss L_{src} for labeled data in the source domain with supervised learning, and both the losses of statistical distribution alignment L_f and entropy regularization L_{tgt} for domain adaptation.

2.2. Statistical distribution alignment

The statistical distribution alignment aims at minimizing the discrepancy of feature representations between the source domain and target domain. Different methods of statistical distribution alignment focus on utilizing different aspects of statistical information. Among these methods, MMD [3, 4] and MK-MMD [5] use the mean of features which is the first order statistics, and CORAL [6, 7] uses the covariance matrix of features which contains the second order statistics. The proposed statistical distribution alignment is called MECOV, which uses both the mean vectors and covariance matrices of feature representations.

Let $\mathbf{X}_s$ denotes the random variable of features for samples in source domain and $\mathbf{X}_t$ for that in target domain, MMD [4] for classification task is defined as

$$\mathrm{MMD}\left(\mathbf{X}_s, \mathbf{X}_t\right) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi\left(X_i^s\right) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi\left(X_j^t\right) \right\|, \tag{1}$$

where n_s and n_t are the numbers of samples in the source domain and target domain respectively. $\phi(\cdot)$ is a mapping to kernel space. When $\phi(\cdot)$ is a linear composition of several kernel functions instead of one kernel function, MMD turns to be MK-MMD[5].

CORAL[7] is defined as

$$\mathrm{CORAL}\left(\mathbf{X}_s, \mathbf{X}_t\right) = \frac{1}{4d^2} \left\| C_s - C_t \right\|_F^2, \tag{2}$$

where C_s and C_t are covariance matrices, $\|\cdot\|_F$ denotes the Frobenius norm, d is the dimension of a feature vector.

By utilizing both the mean vectors and covariance matrices, we propose MECOV to measure the discrepancy between the source domain and target domain.

$$\mathrm{MECOV}\left(\mathbf{X}_s, \mathbf{X}_t\right) = \frac{\lambda}{d} \left\| \overline{X_s} - \overline{X_t} \right\|^2 + \frac{1}{d^2} \left\| C_s - C_t \right\|_F^2, \tag{3}$$

where λ is a hyper parameter set to a constant such as 1.0, $\overline{X_s}$, $\overline{X_t}$ and C_s, C_t are the mean vectors and covariance matrices for deep representations of samples in the source domain and target domain respectively. For sequence modeling tasks, each feature at different time frame is regarded as an independent feature while calculating the mean vectors and covariance matrices in our method.

We can interpret the statistical distribution alignment from the perspective of mutual information. The mutual informa-

tion between $\mathbf{X}_s$ and $\mathbf{X}_t$ is defined as

$$I\left(\mathbf{X}_s, \mathbf{X}_t\right) = H\left(\mathbf{X}_s\right) + H\left(\mathbf{X}_t\right) - H\left(\mathbf{X}_s, \mathbf{X}_t\right), \quad (4)$$

where $H(\cdot)$ denotes Shannon entropy of a random variable, and $H\left(\mathbf{X}_s, \mathbf{X}_t\right)$ is the joint entropy of $\mathbf{X}_s$ and $\mathbf{X}_t$. MECOV minimizes the discrepancy between the source domain and target domain with respect to mean and covariance of feature representations, thus minimizes the joint entropy $H\left(\mathbf{X}_s, \mathbf{X}_t\right)$, and maximizes the mutual information $I\left(\mathbf{X}_s, \mathbf{X}_t\right)$.

2.3. Entropy regularization (ER)

For unlabeled samples in the target domain, we propose to use entropy regularization (ER) as an additional objective function for sequence modeling task. For the deep feature representation $\mathbf{x}_i^t$ of the i^{th} unlabeled samples in the target domain, the decoder output predicted character codes with softmax values at each decoding step $k, 1 \leq k \leq K_i$, where K_i is the sequence length of the decoded text. The softmax values at the decoding step k can be regarded as the posterior probability $p(y_j^k|x_i^t), 1 \leq j \leq C$, where C is the number of pattern classes. For all the unlabeled samples in the target domain, the conditional entropy $H(\mathbf{Y}|\mathbf{X}_t)$ can be used as an entropy regularization term in the objective function.

$$\begin{aligned} ER &= H(\mathbf{Y}|\mathbf{X}_t) \\ &= -\frac{1}{n_t}\Sigma_{i=1}^{n_t}\frac{1}{K_i}\Sigma_{k=1}^{K_i}\Sigma_{j=1}^{C} p(y_j^k|x_i^t) log\ p(y_j^k|x_i^t), \end{aligned} \quad (5)$$

where $H(\mathbf{Y}|\mathbf{X}_t)$ is the entropy of $\mathbf{Y}$ conditioned on $\mathbf{X}_t$, i.e. the entropy of the predicted text symbols with the features $\mathbf{X}_t$. For the calculation of entropy, the logarithm to the base e is used in our method, and the unit of entropy is nat.

The mutual information between the feature space $\mathbf{X}_t = \{x_i^t\}_{i=1}^{n_t}$ and pattern space $Y = \{y_j\}_{j=1}^{C}$ in the target domain is defined as:

$$\begin{aligned} &I(\mathbf{X}_t, \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}_t) \\ &= -\Sigma_{j=1}^{C} p(y_j) log\ p(y_j) \\ &\quad + \frac{1}{n_t}\Sigma_{i=1}^{n_t}\frac{1}{K_i}\Sigma_{k=1}^{K_i}\Sigma_{j=1}^{C} p(y_j^k|x_i^t) log\ p(y_j^k|x_i^t). \end{aligned} \quad (6)$$

As $H(\mathbf{Y}) = -\Sigma_{j=1}^{C} p(y_j) log\ p(y_j)$ is the entropy in the pattern space, $H(\mathbf{Y})$ is constant for a given task. If we use the conditional entropy $H(\mathbf{Y}|\mathbf{X}_t)$ as the entropy regularization term, minimizing $H(\mathbf{Y}|\mathbf{X}_t)$ is equivalent to maximizing the mutual information $I(\mathbf{X}_t, \mathbf{Y})$ between the feature space and pattern space in the target domain.

2.4. Objective function

The objective function used in the training stage is defined as

$$L = L_{src} + k_1 L_{tgt} + k_2 L_f, \quad (7)$$

where L_{src} is the cross-entropy loss between labels and predicted results of samples in the source domain, L_{tgt} is the entropy regularization of predicted results in the target domain, and L_f is MECOV loss between feature distributions of the source domain and target domain. In the following experiments, the hyper parameters k_1 and k_2 are empirically set to 0.01 and 0.1, respectively.

3. EXPERIMENTS

We first select our baseline model by comparing different combinations of RNN and Transformer in the encoder-decoder framework. Then we apply the proposed MECOV and entropy regularization into the baseline model, and compare different re-implemented domain adaptation methods with our method on the handwritten text line recognition task. At last, we compare our method with other reported methods on the handwritten word recognition task.

Our experiments are conducted on the IAM handwriting dataset [20], which contains 10042 English text lines written by 657 writers. The labeled data are provided at both text line and word levels. The following experiments are carried out mainly on the IAM handwritten text line images unless otherwise indicated. In our experiments, we take the training set as the source domain and the test set as the target domain for comparison with other work. In our method, neither additional synthetic training data nor external language models are used, so as to focus on the domain adaptation technique.

The metrics used to evaluate the recognition performance are Character Error Rate (CER) and Word Error Rate (WER). CER is defined as the Levenshtein distance [23] between ground-truth and prediction. WER is the percentage of improperly recognized words.

All models are trained for 200 epochs and optimized with ADADELTA. The learning rate is set to 0.5 for the first 120 epochs, 0.1 for the next 60 epochs and 0.01 for the last 20 epochs. In the domain adaptation stage, all models are trained for another 100 epochs. The learning rate is initialized to 0.01 and decays by a factor of 0.5 after every 30 epochs.

3.1. Selection of baseline model

For the selection of baseline model, we compare encoders and decoders with different combinations of RNN [14] and Transformer [17] as shown in Tab. 1, and the combination of RNN based encoder and Transformer based decoder is selected as our baseline model. In the decoding process, a linear layer converts the output data of the decoder into logits by a linear layer at each decoding step, and a softmax function turns the logits into probabilities.

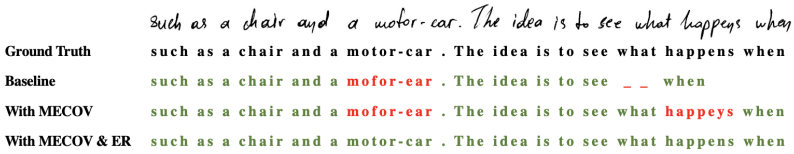
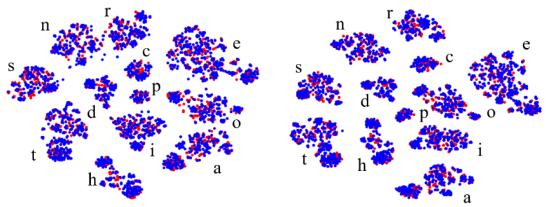


Fig. 2. Examples of recognized results using different methods.



(a) Before domain adaptation (b) After domain adaptation

Fig. 3. Data distributions of the source domain (red) and target domain (blue).

Table 1. Comparison to select baseline model.

Encoder	Decoder	CER(%)	WER(%)
RNN	RNN	6.31	19.02
Transformer	Transformer	6.34	18.91
Transformer	RNN	6.60	20.14
RNN	Transformer	**6.01**	**17.57**

3.2. Comparisons of domain adaptation methods

The comparisons of different domain adaptation techniques are shown in Tab. 2. We re-implemented MK-MMD [5] with five Gaussian kernels, CORAL [6], and the adversarial learning [8] with a three-linear-layers domain classifier. The re-implemented methods are incorporated into our baseline model respectively. Compared with the baseline model, our method with both MECOV and entropy regularization can reduce CER by 1% and WER by 2%, which is better than other re-implemented methods. By comparing different versions of our method, it shows that entropy regularization contributes most. Examples of recognized results using different methods are shown in Fig. 2.

To visualize the data distributions from the source domain and target domain before and after applying our method, we select 100 images from each domain, and plot the data distributions of 12 most frequent characters by using t-SNE. The data for visualization are the logits in the decoding process, thus the corresponding character labels can be marked. As shown in Fig. 3, some similar characters are easily distinguished from each other after the utilization of MECOV and entropy regularization, e.g. n/r, o/a.

According to Equ. (5) and Equ. (6), the amount of the increased mutual information $I(\mathbf{X}_t, \mathbf{Y})$ of the feature space and pattern space in the target domain is the same as the amount of the decreased $H(Y|X)$ with the utilization of MECOV and entropy regularization, which is 0.028 (nat) in our experiments.

Table 2. Comparison of domain adaptation methods.

Methods		CER(%)	WER(%)
Baseline		6.01	17.57
Re-implemented	MK-MMD	9.62	23.95
	CORAL	5.94	17.54
	Adversarial	6.64	19.31
Ours	MECOV	5.93	17.48
	ME+ER	5.11	15.73
	COV+ER	5.08	15.56
	MECOV+ER	**5.05**	**15.53**

3.3. Comparison with other reported methods

Our method can be applied to handwriting recognition at both word and text line levels. The comparison results with other reported domain adaptation methods on the IAM handwritten word images are shown in Tab. 3. Our method with both statistical distribution alignment and entropy regularization has achieved competitive results, and the WER is lower than previously reported results.

Table 3. Comparison with other reported methods.

Method	CER(%)	WER(%)
Zhang et al.[11]	8.50	22.20
Kang et al.[12]	**6.75**	17.26
Ours	7.22	**15.53**

4. CONCLUSION

We address the issue of domain adaptation for sequence modeling task to improve the generalization ability of deep learning model, and propose to incorporate statistical distribution alignment and entropy regularization from the perspective of mutual information maximization. Experimental results on the IAM handwriting dataset have demonstrated the effectiveness of the proposed methods.

5. REFERENCES

[1] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Trans. on Knowledge Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *ICML*, 2011, p. 513–520.

[3] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[4] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[5] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015.

[6] Baochen Sun, Jiashi Feng, and Kate Saenko, "Return of frustratingly easy domain adaptation," *arXiv preprint arXiv:1511.05547*, 2015.

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

[8] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[9] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 7167–7176.

[10] Yejun Tang, Bing Wu, Liangrui Peng, and Changsong Liu, "Semi-supervised transfer learning for convolutional neural network based Chinese character recognition," in *ICDAR*, 2017, pp. 441–447.

[11] Yaping Zhang, Shuai Nie, Wenju Liu, Xing Xu, Dongxiang Zhang, and Heng Tao Shen, "Sequence-to-sequence domain adaptation network for robust text image recognition," in *CVPR*, 2019, pp. 2740–2749.

[12] Lei Kang, Marçal Rusiñol, Alicia Fornés, Pau Riba, and Mauricio Villegas, "Unsupervised writer adaptation for synthetic-to-real handwritten word recognition," in *WACV*, 2020, pp. 3502–3511.

[13] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[15] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *NIPS*, 2015, pp. 577–585.

[16] Zbigniew Wojna, Alexander N. Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz, "Attention-based extraction of structured information from street view imagery," in *ICDAR*, 2017, pp. 844–850.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, p. 6000–6010.

[18] Lu Yang, Peng Wang, Hui Li, Zhen Li, and Yanning Zhang, "A holistic representation guided attention network for scene text recognition," *Neurocomputing*, vol. 414, no. 1, pp. 67–75, 2020.

[19] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," *arXiv preprint arXiv:1711.10288*, 2017.

[20] U-V Marti and Horst Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *IJDAR*, vol. 5, no. 1, pp. 39–46, 2002.

[21] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *ICCV*, 2017, pp. 5086–5094.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[23] Vladimir I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.