

Sequence-to-sequence Domain Adaptation Network for Robust Text Image Recognition

Anonymous AAAI submission
Supplementary I
Paper ID: 2713

September, 2018

- **Goal:** To reduce the domain shift between the source and target text images
- **Approach:**
 - We develop a sequence-to-sequence domain adaptation network (SDAN) for robust text image recognition, which consists of a sequence encoding unit (mapping input images into a sequence of high-level feature vectors), a gated attention similarity unit (aligning distributions of the source and target domain at character-level) and a sequence decoding unit (converting encoded features into output strings recurrently).
 - The gated attention similarity unit between the encoder and decoder is introduced to adaptively perform domain adaptation on a set of character-level attention context vectors, which focus on the most relevant region towards a specific character instead of global sequence.
 - Through the jointly optimizing of unsupervised similarity loss and source decoding loss, the model is able to learn both domain-invariant and discriminative features that are effective for the shifted target text images.

Contributions:

- We propose a novel Sequence-to-sequence Domain Adaptation Network dubbed SDAN for robust text image recognition, which could be generalized to different scenes, such as natural scene text, handwritten text and mathematical expression recognition.
- We introduce a novel gated attention similarity unit in SDAN to bridge the sequence-like text image recognition and domain adaptation, which could adaptively transfer fine-grained character-level knowledge instead of performing domain adaptation by global features.
- The proposed SDAN is capable of taking advantage of unsupervised sequence data and reducing domain shift effectively.

- Introduction
 - Text image Recognition
 - Related Work
- Proposed Model
 - Overview
 - Sequence Encoding Unit
 - Sequence Decoding Unit
 - Gated Attention Similarity Unit
 - Overall Objective
- Experimental Results
- Ablation Study
- Conclusion

Text image Recognition

- *Text recognition is to read various kinds of text in images, such as natural scene text, printed or handwritten text, and ink words generated from the digital pen.*
- *It is still challenging to build a robust text recognizer that can handle varying data in abundance of scenarios effectively, due to the various types of domain shift.*



Text Image Recognition

Learned to read various kinds of text in images

- *Earlier DNN based methods*: Depending on the segmentation of each character or a non-maximum suppression
 - Wang et al.ICCV'2011; Jaderberg et al. ECCV'2014
- *Sequence learning based methods*: Firstly encoding an entire image text into a sequence of features, and then decoding character sequence recurrently with CTC or attention schemes.
 - CTC based methods
 - He et al.AAAI'2016; Shi et al. TPAMI'2017
 - Attention based methods
 - Shi et al.CVPR'2016; Deng et al ICML'2017; Liu et al. AAAI'2018

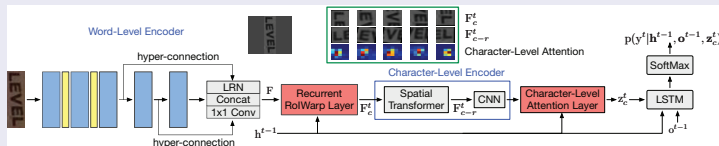
Unsupervised Domain Adaptation

Learned to reduce domain shift across different domains.

- *Feature-level Domain Adaptation:* To learn domain invariant representations by minimizing the distance of feature distribution
 - Maximum Mean Discrepancy (MMD) based methods:
e.g. Baktashmotlagh et al. ICCV'2013; Long et al. ICML'2015; Bousmalis et al. NIPS'2016
 - Correlation Distance based methods:
e.g. Sun et al. AAAI'2016; Sun et al. ECCV'2016
 - Adversarial Loss based methods:
e.g. Ganin et al. JMLR'2016; Tzeng et al. CVPR'2017; Volpi et al. CVPR'2018
- *Pixel-level Domain Adaptation:* To align the raw pixel-level distribution between two domains by adapting source image to appear as if drawn from target image.
 - Bousmalis et al. CVPR'2017; Shrivastava et al. CVPR'2017; Zak et al. CVPR'2018

CharNet: Distorted Scene Text Recognition

- *To introduce character-level spatial transformer to rectify individual characters*



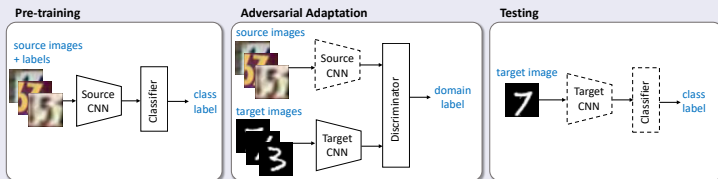
1

- **Advantages:** capable of handling more complicated forms of distortion that cannot be modeled by a single global transformation easily;
- **Disadvantages:**
 - only designed for spatial affine distortion in scene text;
 - hard to generalize to the distortion caused by handwriting style or various structure in mathematical expressions;
 - neglecting the intrinsic domain shift in text image data

¹Figure is from Liu et.al, AAAI'2018

Adversarial Discriminative Domain Adaptation

- To learn domain invariant representations at a global feature level.*



2

- Advantages:** good at finding high-level domain-invariant feature of a global image.
- Disadvantages:** cannot be directly applied on sequential text images with multiple characters, as the domain shift are locally in the characters rather than the global image.

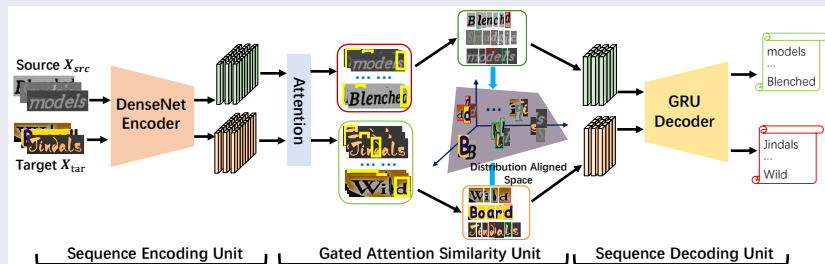
Deficiencies of Previous Methods

- Most text image recognition methods neglected the intrinsic domain shift;
- Some robust text image recognition are for a specific scenario, such as spatial affine distortion in scene text, and cannot be generalized effectively to different task;
- Recent visual domain adaptation works usually focus on non-sequence object recognition with a global feature vector, which is inadequate to transfer a variable-length sequence knowledge.

Proposed Model (I)

Architecture

The general flowchart of the proposed SDAN for unsupervised domain adaptation.



- A sequence encoding unit learns high-level visual representations from an input image.
- A sequence decoding unit generates a sequence of alphanumeric symbols as output, one at every time step.
- A gated attention similarity unit offers the guidance for model to adaptively find character-level common features between the source and target domain.

Proposed Model (II)

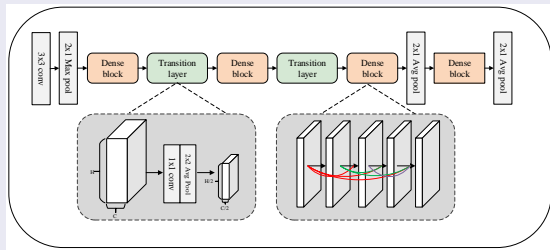
Sequence Encoding Unit

- The sequence encoding unit takes the raw input and produces a feature grid F of size $H' \times W' \times D$

$$F = \{f_1, \dots, f_L\}, f_i \in R^D, \quad (1)$$

where f_i corresponds to i^{th} grid of the encoded image, which preserves specific spatial information of the input image.

- DenseNet Encoder is designed to learn robust features towards different scenes.



Proposed Model (III)

Sequence Decoding Unit

Sequence decoding unit is employed to predict the string of an input text image recurrently.

- A GRU decoder is introduced to leverage the context vector c_t , previous state h_{t-1} and previous predicted character y_{t-1} to generate a new hidden state*

$$h_t = \text{GRU}(h_{t-1}, y_{t-1}, c_t), \quad (2)$$

- the probability of current predicted symbol y_t is computed by :*

$$p(y_t|y_{t-1}, F) = g(W_o \tanh(Ey_{t-1} + W_d h_t + W_c c_t)), \quad (3)$$

- c_t is generated by attention mechanism, which focuses on the most relevant region of current decoding character.*
- g denotes a softmax activation function,*
- W_o , W_d and W_c are the mapping matrices, and E is the embedding matrix.*

Proposed Model (IV)

Gated Attention Similarity Unit i

An attention model is introduced to learn which part in the text image is the most relevant to a decoding character.

- *context vector c_t , which denotes the representation of the most relevant part of encoding feature map F at time-step t*

$$c_t = \sum_{i=0}^L \alpha_{t,i} f_i, \quad (4)$$

- *the attention weights $\alpha_{t,i}$*

$$\alpha_{t,i} = \frac{\exp(s_{t,i})}{\sum_{j=0}^L \exp(s_{t,j})}, \quad (5)$$

$$s_{t,i} = \beta^T \tanh(W_h h_{t-1} + W_f f_i), \quad (6)$$

- *β , W_h and W_f are the parameters to be learnt.*

Proposed Model (V)

Gated Attention Similarity Unit ii

- An attention-driven similarity loss L_{attn} , which takes advantage of fine-grained character-level features, is accordingly introduced to measure similarity on context vector of text images between the source and target domain.

$$L_{attn} = E_{[x_{src} \sim X_{src}, x_{tar} \sim X_{tar}]} \{dist(Attns(x_{src}), Attns(x_{tar}))\}, \quad (7)$$

- $Attns(x)$ denotes the attention context vector set of an input text image x

$$Attns(x) = \{c_1, c_2, \dots, c_T\}, \quad (8)$$

- $Attns(x)$ is adaptively updated by an adaptation gate function $\delta(c_t)$, which is to select effective information to performed adaptation

$$Attns(x) = Attns(x) \otimes \delta(c_t), \quad (9)$$

$$\delta(c_t) = \begin{cases} 1 & \text{if } p(y_t|y_{t-1}, F) > p_{th} \\ 0 & \text{if } p(y_t|y_{t-1}, F) < p_{th} \end{cases}, \quad (10)$$

- \otimes denotes element-wise selecting operator, p_{th} is a confidence threshold.

Proposed Model (VI)

Overall Objective

- *The GAS unit in our model is introduced to offer guidance to learn domain-invariant features between the source and target domain. The learnt robust representations should work effectively on the target domain, where they are also required to be discriminative.*
- *Therefore, the attention similarity loss L_{attn} is combined with the discriminative decoder loss L_{dec} in source domain to learn both domain invariant and discriminative feature.*

$$L_{SDAN} = L_{dec} + \lambda L_{attn}, \quad (11)$$

- *λ is a hyper-parameter to balance two terms.*
- *The model parameters can be directly optimized by minimizing the overall objective through stochastic gradient descent optimization algorithms.*

Experiment

- Scene Text dataset

Synthetic dataset Mjsynth is used as source training data. ICDAR-2003 (IC-03), ICDAR-2013 (IC-13), Street View Text (SVT) and IIIT 5K-words (IIIT5K), which are natural scene text, are used as target test data.



- Handwritten text dataset

IAM is selected, which is partitioned into writer-independent training, validation and test partitions of 6161, 966 and 2915 lines, respectively. That means a total of 55,081, 8,895 and 25,920 words in each partition.

- Handwritten mathematical expression dataset

CROHME 2014 dataset is selected, which contains 8836 training math expressions and 986 test math expressions. The handwritten expressions or LaTeX notations in the test set never appears the train set.

Evaluation

- **Scene text**

The word prediction accuracy is used to evaluate scene text recognition model, following several benchmark(Shi et al, CVPR'2017, Liu et al, AAAI'2018).

- **Handwritten text**

Two metrics are used to evaluate the handwritten text recognition model: the Character Error Rate (CER) and the Word Error Rate (WER) (Bluche et al. NIPS'2016). CER is defined as the Levenstein distance between the predicted and real character sequence of the word. WER denotes the percentage of words improperly recognized.

- **Mathematical expression**

We use a global performance metric expression recognition rate (ExpRate) to denote the percentage of predicted formula sequences matching the real formula sequences (Deng et al. ICML'2017).

Implementation Details

- The architecture of the DenseNet encoder*

Layers	Kenel	Output size
	$[size, stride, channel]$ $[depth, growth_rate]$	
<i>conv_bn_relu</i>	[3, 1, 48]	$H \times W$
<i>max_pool</i>	[(2, 1), (2, 1)]	$H/2 \times W$
<i>dense_block</i>	[6, 24]	$H/2 \times W$
<i>transition_layer</i>	[1, 1, 0.5]	$H/4 \times W/2$
<i>dense_block</i>	[12, 24]	$H/4 \times W/2$
<i>transition_layer</i>	[1, 1, 0.5]	$H/8 \times W/4$
<i>dense_block</i>	[24, 24]	$H/8 \times W/4$
<i>avg_pool</i>	[(2, 1), (2, 1)]	$H/16 \times W/4$
<i>dense_block</i>	[24, 24]	$H/16 \times W/4$
<i>avg_pool</i>	[(2, 1), (2, 1)]	$H/32 \times W/4$

- A bi-directional LSTM is followed DenseNet encoder to capture more context information for attention, and each LSTM has 256 hidden units. The GRU decoder is achieved by a GRU cell with 512 memory blocks.*
- The complete model is initially pre-trained to minimize the decoding loss of the source training data, and then is fine-tuned to minimize the overall domain adaptation objective with unsupervised target data.*

Comparison with Existing Methods (I)

- To validate the performance of our SDAN model, we focus on unconstrained text recognition without any language model or lexicon.
- Three different types of text image data, which are scene text, handwritten text, and mathematical expressions, are evaluated.
- On each type of text image data, we select relative state-of-the-art approaches as counterparts with fair comparison.
- Besides, we also consider a baseline for SDAN as SDAN-base that omits the GAS unit to switch off the domain adaption process. SDAN-base is used to investigate the capability of SDAN for domain adaptation on the text image recognition task.

Comparison with Existing Methods (II)

Results on Scene Text

<i>Model</i>	<i>IIIT5K</i>	<i>SVT</i>	<i>IC-03</i>	<i>IC-13</i>
<i>jaderberg et al 2014a</i>	-	71.7	89.6	81.8
<i>CRNN (Shi, Bai, Yao 2017)</i>	81.2	82.7	91.9	89.6
<i>STAR-Net(Liu et al. 2016)</i>	83.3	83.6	89.9	89.1
<i>R²AM (Shi et al. 2016)</i>	78.4	80.7	88.7	90.0
<i>RARE (Shi et al. 2016)</i>	81.9	81.9	90.1	88.6
<i>Gao et al (Gao et al. 2016)</i>	81.8	82.7	89.2	88.0
<i>Char-Net (Liu, Chen, and Wong 2018)</i>	83.6	84.4	91.5	90.8
<i>SDAN-base</i>	81.1	82.1	91.2	91.0
<i>SDAN</i>	83.8	84.5	92.1	91.8

- *We observe that our model outperforms RARE and STAR-Net, which both employ spatial transformers to rectify the global distortions of text images.*
- *Our model can also achieve comparable results with the best competitor Char-Net, which introduces rectification of the distorted text at character-level.*
- *The two cases indicate that our model is capable of learning domain-invariant features of the distorted scene text between the source and target data.*

Comparison with Existing Methods (III)

Results on Handwritten Text

Method	WER	CER	Average
<i>bluche2015deep</i> (Bluche 2015)	24.7	7.3	16.00
<i>bluche2016joint</i> (Bluche 2016)	24.6	7.9	16.25
<i>sueiras2018offline</i> (Sueiras et al. 2018)	23.8	8.8	16.30
<i>SDAN-base</i>	23.9	9.2	16.55
<i>SDAN</i>	22.2	8.5	15.35

- *Various handwriting styles are primary causes of domain shift in handwritten text recognition. What's more, it may suffer character-touching problem, which is different from scene text.*
- *Our SDAN model can achieve significant improvement on handwritten text, although the performance of our baseline is not better than the best competitor, which demonstrates the effectivity of SDAN in handwritten text recognition.*

Comparison with Existing Methods (IV)

Results on handwritten mathematical expression

<i>Method</i>	<i>ExpRate</i>
<i>I (Mouchere et al. 2014)</i>	<i>37.2</i>
<i>VI(Mouchere et al. 2014)</i>	<i>25.7</i>
<i>VII(Mouchere et al. 2014)</i>	<i>26.1</i>
<i>WYCIWYS(Deng, Kanervisto, and Rush 2016)</i>	<i>28.7</i>
<i>CRNN(Le and Nakagawa 2017)</i>	<i>35.2</i>
<i>IM2TEX(Deng et al. 2017)</i>	<i>38.7</i>
<i>SDAN-base</i>	<i>39.9</i>
<i>SDAN</i>	<i>41.6</i>

- *Handwritten mathematical expression is a more complex problem than traditional scene text or handwriting recognition. In particular, it suffers variant scales of handwritten math symbols with more complicated structure, which results in more difficult domain shift.*
- *Compared to the baseline model SDAN-base and best competitor IM2TEX, SDAN can achieve significant improvement, which shows SDAN is able to capture the complex domain shift in structural images.*

Ablation Study (I)

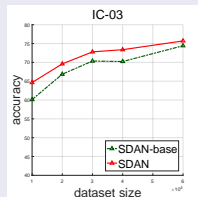
Effectiveness of Different Components in DSAN

Model	V1	V2	V3	V4	V5	V6
VGG	✓	✓				
ResNet			✓	✓		
DenseNet					✓	✓
GAS		✓		✓		✓
WER	32.77	26.88	29.88	27.85	26.38	24.21
CER	15.87	12.60	14.32	13.14	12.40	11.37

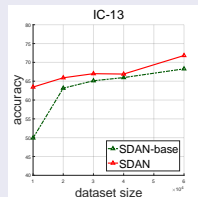
- For the analysis, we choose handwritten text dataset IAM to evaluate model from both character error rate (CER) and word error rate (WER).
- Firstly, we develop three CNN encoders to investigate the effect of encoders in our model. We can observe that DenseNet is a more powerful encoder from the comparisons among the model V1, V3, and V4, where these models are without GAS unit.
- Furthermore, we explore the effectiveness of our GAS unit, the comparison pairs (V1, V2), (V3, V4) and (V5, V6) show that the GAS unit could always improve performance despite of the types of encoders.

Ablation Study (II)

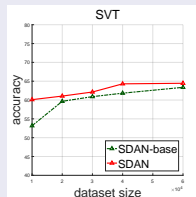
Effectiveness of Unsupervised Data



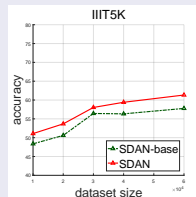
(a) IC-03



(b) IC-13



(c) SVT



(d) IIIT5K

- In order to quantify the effectivity of unsupervised data, we train our model with different size of labeled data and unlabeled data, while keep other hyper-parameters fixed.
- We can observe that using additional unlabeled samples with SDAN can get consistent performance improvement, which implies that SDAN is able to learn the knowledge from unsupervised data.

Ablation Study (III)

From Handwritten Text to Scene text


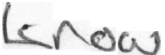
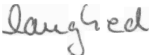
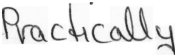
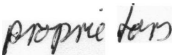
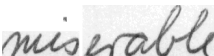
Methods	IIIT5K		SVT		IC-03		IC-13	
	WER	CER	WER	CER	WER	CER	WER	CER
baseline	92.2	87.2	99.4	91.1	90.6	74.4	90.5	77.2
SDAN	91.3	83.4	99.1	87.8	88.8	70.4	89.6	73.4

- *This scenario shows a more complex domain adaptation task, which is from handwritten text to scene text. For the analysis, we evaluate the model from both character error rate (CER) and word error rate(WER).*
- *The baseline model SDAN-base is firstly trained by supervised handwritten text dataset IAM. We can observe that the model trained on IAM has a poor performance on scene text.*
- *SDAN could get some improvement at both character level and word level, where SDAN takes advantage of some unsupervised synthetic scene text from Mjsynth to finetune the model.*

Ablation Study (IV)

Visualization for Recognized Handwritten Text

- Examples showing the recognition result, the left column is the input images, the second column and the last column denote the recognition results with and without domain adaptation, respectively. Each result is shown in the pair of prediction and ground truth.

Input image	SDAN-base	SDAN
	langhed laughed	laughed laughed
	Lenow know	know know
	langhed laughed	laughed laughed
	dRACTically practically	practically practically
	proprietars proprietors	proprietors proprietors
	miscracble miserable	miserable miserable

Conclusion

- In this paper, we present a novel SDAN model for robust text image recognition, which bridges the sequence-like text image recognition and domain adaptation.
- It's capable of taking advantage of unsupervised sequence data to learn more robust feature representations.
- The proposed model could also be generalized to different scenes, which include scene text, handwritten text and mathematical expression recognition.
- Comprehensive experimental results on six datasets and extensive analysis have demonstrated the effectiveness of our algorithm.
- An interesting open issue for future research is to further adjust our proposed SDAN framework to better deal with various sequence domain shift.