2017 14th IAPR International Conference on Document Analysis and Recognition

Attention Based RNN Model for Document Image Quality Assessment

Pengchao Li*, Liangrui Peng*, Junyang Cai*, Xiaoqing Ding*,Shuangkui Ge†
*Tsinghua National Laboratory for Information Science and Technology
Dept. of Electronic Engineering, Tsinghua University, Beijing, China 100084.
Email: lpc13@mails.tsinghua.edu.cn, {penglr, dingxq}@tsinghua.edu.cn
†Signal Processing and Modern Communication Lab of Wuhan University
Electronic Information School, Wuhan University, Wuhan, China 430072.
Email: xsk@bieta.cn

***Abstract*—Document Image Quality Assessment (DIQA) is an essential step preceding Optical Character Recognition (OCR). In this paper we propose an attention based Recurrent Neural Network (RNN) model for camera based DIQA. Convolutional Neural Network (CNN) and RNN are integrated into our model to capture spatial features for several glimpse regions step by step within an image patch. Reinforcement learning is adopted to train a locator to generate the optimal location of a glimpse region for the next time step so that attention can be payed to the salient part. Given an input document image, patches are generated with a sliding window, and the pure background ones are sifted out. Quality scores are obtained for all the sifted patches by applying the proposed attention based RNN method, and the patch scores are averaged over each input image as the result of DIQA. We conduct experiments on two public datasets and make comparisons with several other reported methods. Experimental results show that our model achieves the state of the art performance.**

I. Introduction

With the advancement of modern smart phones, one can capture document images instead of using traditional scanners. However, it is a known fact that the cameras of the smart phones are not optimized for capturing document images. Taking a photo of a paper document does not ensure that its image quality would be suitable for Optical Character Recognition (OCR) systems. In such circumstances, it is critical to assess document image quality before image enhancement while applying OCR [1].

Document Image Quality Assessment (DIQA) can be roughly separated into two types: reference and no-reference approaches. The reference assessment approaches compare the downgraded document images with the corresponding high quality images, which is not applicable in the circumstances lacking high quality references. Many no-reference DIQA methods have been developed in recent years. Some pioneer methods use hand-craft features[2,3] for DIQA based on black and white connected component analysis. Ye et al. propose an unsupervised feature learning method to learn a visual codebook, and use soft-assignment coding with max pooling for efficient feature representation[4,5]. DIQA is formulated as a regression problem and Support Vector Regression (SVR) with linear kernel is used for regression [5]. Peng et al. propose another unsupervised feature learning based DIQA approach [6]. Raw patches are served as visual words and fed to Latent Dirichlet Allocation (LDA) to learn the quality topics of documents, and then the distributions of those topics are used as features to evaluate the document quality. Peng et al. also apply a semi-supervised feature learning technique to represent document images sparsely, and the target OCR accuracy is integrated into the training phase of sparse representation, which enhances the discriminant capability of the system [7]. Supervised ridge regression is used to predict the quality score [6,7].

Deep learning has attracted attention in computer vision research community for several years. The Convolutional Neural Network (CNN) is one of the most widely used methods for object detection and recognition [8, 9]. Kang et al. implement a CNN based method to evaluate OCR accuracy, where the document images are divided into small patches and then fed into a CNN to produce predicted quality scores [10]. The experimental results on two datasets have shown promising performance. However, when applying CNN based method to real world OCR preprocessing, such method has some potential drawbacks. First, if the main part of some patches are occupied by background, the DIQA system may not work, so the text region of document image must be located accurately before obtaining training samples. Second, the text body must be split into small patches with same size, like 48×48, all of which are fed into CNN. This small patch based approach plus multi-layer convolution operation are time consuming both in training and testing phases.

For human vision, one tends not to process a whole scene in its entirety at once. Instead, one focuses attention on parts of image to acquire information, and combine information from different fixations to build up an internal representation of the scene. Inspired by this idea, attention based RNN model [11] has been developed and achieved good performance on a wide range of machine learning tasks including speech recognition [12], visual object classification [13], image caption generation [14] and handwriting synthesis [15]. The recurrent visual attention model can be implemented by reinforcement learning. We consider the attention process as an agent interacting with the environment. At each time step, the agent observes the environment only in a local region. The agent could control

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.139

how to deploy sensor location, and could also affect the state of the environment by executing actions. Since the environment is only partially observed, the agent needs to maximize reward over time in order to determine how to act and how to deploy its sensor most effectively.

According to this attention based idea, when evaluating the quality of document, one just need to process a sequence of document image patches with different scale to estimate the real OCR accuracy. We propose a novel attention based RNN model for DIQA. Our contributions include three aspects. (1) CNN and RNN are integrated into our model to capture spatial features for several glimpse regions step by step within an image patch. (2) Reinforcement learning is adopted to train a locator to generate the optimal location of a glimpse region for the next time step so that attention can be payed to the salient part. (3) A Mean Square Error (MSE) related reward is designed to train the reinforcement module. This mechanism has two advantages at least. First, we need not to detect the characters accurately when applying preprocessing step. Second, we can reduce the computing cost by reducing the network scale and complexity.

The remainder of this paper is organized as follows. In Section II, we introduce attention based RNN architecture and its training method. Experiments setup and results are described in Section III. We conclude our work in Section IV.

II. DIQA System Framework

The proposed attention based RNN architecture is as Fig. 1 shows, which could be roughly separated into glimpse network, RNN and action network. The glimpse network is a combination of location sensor and glimpse sensor. In glimpse sensor, we add a convolutional layer after a spatial glimpse layer to extract features of document image patches. RNN is used to model the image context information and the attention mechanism. The Action network includes a locator and a linear regression module which controls the reward function.

A. Glimpse Network

Glimpse network aims at extracting spatial glimpse information and location information, while the location sensor in which can extract location information. The inputs of location sensor are x and y coordinates of the glimpse image. We implement fully-connected layer with ReLU [16] as activation function to extract location information. We assume location information as l_{t-1} at time step $t-1$. Then the output of location sensor is $f_{gl}(l_{t-1};\theta_{gl})$, where θ_{gl} represents the parameters of location sensor.

Glimpse sensor extracts a retina-like representation that contains multiple resolution smaller patches given the coordinates of an input image glimpse. Glimpse size and depth are the essential parameters of glimpse sensor. The depth means the number of small patch series extracted from the glimpse location. The successive patch is twice the size of the previous one. We get a high-resolution patch of a small area, and successively lower-resolution patches of larger areas. A spatial convolutional layer and max pooling layer is connected to the spatial glimpse module. Convolutional layer can help generate feature maps to extract essential vision features of images as glimpse information. Both glimpse information and location information are concatenated as a whole to feed into fully-connected layer.

For the glimpse network, at time step t, we assume input image as I_t, then the output of glimpse network is $g_t = f_g(I_t, l_{t-1};\theta_g)$, where θ_g represents the parameters of the glimpse network.

B. Recurrent Neural Network

The network summarize information extracted from several history glimpses to maintain an internal state. The output of glimpse network is fed into a hidden layer to form the input layer of the RNN. This internal state is formed by the hidden units $h_t = f_h(h_{t-1}, g_t;\theta_h)$ of the recurrent layer, where θ_h represents the parameters of the recurrent layer. So as to evade gradient explosion and gradient vanishing problems, some recurrent unit variations have been developed including LSTM[17,18]. Gated Recurrent Unit [19] is a simpler recurrent unit compare to LSTM with no cell and no output gate, which has been proved effective in machine translation, polyphonic music modeling [20] and text line recognition tasks [21]. We implement GRU to form the recurrent network.

C. Action Network

The model performs two actions: deciding location of the next step l_t, and predicting the OCR accuracy. So the action network consists of locator network and linear regression network. The output of locator network is defined as $f_l(h_t;\theta_l)$, where θ_l represents the parameters of the locator network. The output of OCR accuracy prediction network includes an accuracy score and a baseline reward. The output is $f_a(h_t;\theta_a)$, where θ_a represents the parameters of the OCR accuracy prediction network. The location of the next time step is chosen stochastically from a Gaussian distribution parameterized by the locator network. As we can see from Fig. 1, the key module in locator network is the *ReinforceNormal* module implements the REINFORCE [22] algorithm for the Gaussian distribution.

From the perspective of reinforcement learning, the agent gets a reward signal after executing an OCR prediction action. The goal of the agent is to maximize the sum of the reward signal. In DIQA circumstance, the reward should be negative related to the Euclidean distance between the output and the ground truth. Furthermore, a squashing function should be applied to constrain the output between 0 and 1. So the reward function can be defined as follows.

$$R = \frac{1}{2}(sign(\frac{1}{N_b}\sum_{i}^{N_b} exp(-||\mathbf{y}_i - \mathbf{l}_i||_2^2) - th) + 1) \qquad (1)$$

where $\mathbf{y}_i$ is the output of the linear layer in linear regression network, and $\mathbf{l}_i$ is the OCR accuracy, N_b is batch size. Both $\mathbf{y}_i$ and $\mathbf{l}_i$ are vectors since we train the network in batches. Sign function together with *th*, a predefined threshold, are to set the reward to 0 or 1.

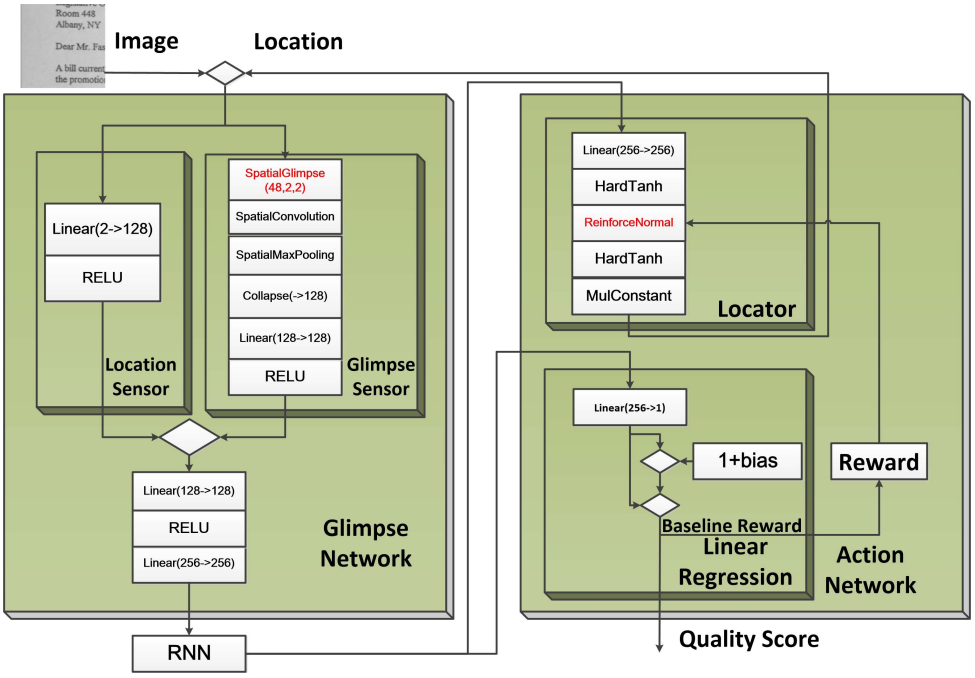


Fig. 1. The architecture of the proposed network.

D. Learning Procedure

The whole network is trained with SGD algorithm. As the picked regions are discrete, the gradients of loss function with regard to the coordinates are not able to be computed. To maximize the sum of reward by standard SGD is impossible. As a result, unlike most Modules, *ReinforceNormal* ignores the input gradient of the next network layer when backpropagation is called. The approximate gradient is given by

$$\nabla_\theta E(R) \approx a \times \sum_i^M \nabla_\theta lnf(x,\mu)(R^i - b) \qquad (2)$$

where a is a scaling factor like learning rate, $\theta = \{\theta_g, \theta_h, \theta_l, \theta_a\}$ is the parameters of the whole network, i represents the ith batch of total M batches, f is the output of the *ReinforcementNormal* module which is a Gaussian stochastic processes, μ is the mean of the Gaussian distribution. Equation (2) is called a variance reduction version of gradient approximation and b is a baseline reward [23], which we learn by reducing the squared error between R^i and b. The REINFORCE algorithm requires that we differentiate the probability density or mass function (PDF/PMF) of the distribution w.r.t. the parameters of *ReinforcementNormal* module, i.e. the mean of the Gaussian distribution μ. So we get $\nabla_\theta lnf(x,\mu)$ in equation (2):

$$\nabla_\theta lnf(x,\mu) = \nabla_\mu lnf(x,\mu) = \frac{x-u}{\sigma^2} \qquad (3)$$

As we use supervised learning with OCR accuracy as ground truth, the MSE is defined as the Euclidean distance between the output of the linear regression network $\mathbf{y}_i$ and the ground truth $\mathbf{l}_i$. A hybrid loss is defined to combine MSE criterion and variance reduction reward criterion for optimization.

III. Experiment Setup and Results

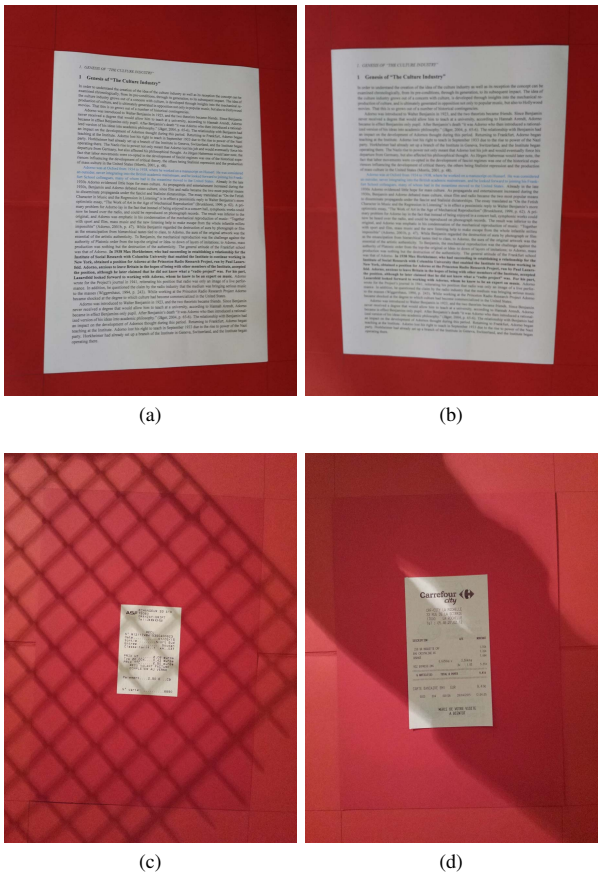
A. Datasets

Two datasets are adopted to train and test our DIQA model.

(1) Sharpness-OCR-Correlation (SOC) dataset [24]: This dataset includes a total of 175 color images with resolution 1840 $\times$ 3264. These images are captured from 25 printed English documents using a cell phone camera. 6-8 photos with varying focal lengths were taken for each document to generate different levels of blur. Fig. 3(a) shows a sample image of this dataset. Three OCR engines (ABBY FineReader, Tesseract and Omnipage) were conducted on each image to get character level accuracies.

(2) Smartdoc-QA dataset [25]: This dataset includes a total of 4260 color images with resolution 3096 $\times$ 4128. These images are captured from 30 documents using Nokia Lumia 920 and Samsung S4. There are three kinds of documents which are contemporary documents, old administrative documents and receipts. 142 different images are captured per document page with varying capture conditions (light, different types of blur and perspective angles) with two smart phones. Two commercial OCR software systems ABBYY Fine Reader and Tesseract were run on each of the images. Fig. 2 shows some samples from Smartdoc-QA with different content, degradation and illumination types.

Every image from both datasets is in high resolution, so it is nontrivial for neural network to process each sample as a whole. As for SOC, because the dataset size is small compare to

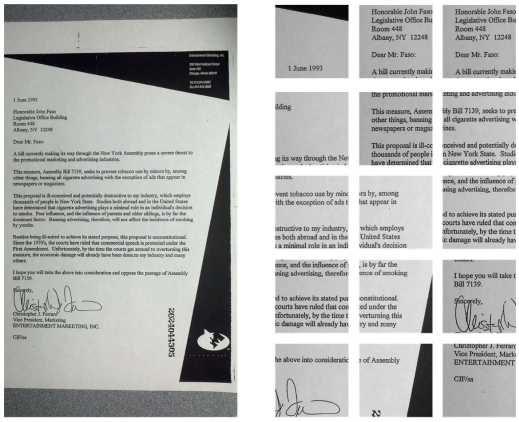


(a) (b) (c) (d)

Fig. 2. Smartdoc-QA dataset samples with different document types and degradations 2a Contemporary document with motion blur 2b Contemporary document with out-of-focus blur 2c Old administrative document with shadow object 2d Receipt of degraded illumination with shadow object.

the weight numbers in the model, we can significantly enhance the generalization ability by enlarge the sample size. Based on these concerns, patch extraction is a prerequisite step before training the model. However, our patch exaction approach is quite different from conventional methods [6,7,10], which is, to performe accurate character detection based on OCR layout analysis, and generate patches with small size. We apply a much bigger sliding window of 256 $\times$ 256 to get patches, and separate the patches into foreground and background ones. Background patches are excluded from the dataset. We can see from Fig. 3 that some patches are not totally occupied by characters, which are also adopted as foreground patches.

In our experiments, we randomly select 80% of the data as training set, and leave the remaining 20% as test set. As a result, for SOC dataset, we get roughly 4 $\times$ 10^3 training patches and 1 $\times$ 10^3 test patches. As for Smartdoc-QA dataset, we get about 3 $\times$ 10^4 training patches and 8 $\times$ 10^3 test patches.



(a) (b)

Fig. 3. (a) Document image from SOC dataset with size of 1860 $\times$ 3264 (b) Cropped image patches of size 256 $\times$ 256.

B. Experiments Setup and Evaluation Protocol

In the training phase, samples are grouped into batches to do optimization. In the testing phase, the predicted accuracy scores of patches that belong to each document image are averaged to obtain a document accuracy score. The initial status of $ReinceforceNormal$ is set randomly for each test sample in the testing phase. As the accuracy scores are generated by different OCR engines for both datasets, in our experiments, we use the average accuracy for the ground truth of training and evaluation process.

Considering the compression and representation ability of input patches, the specific setting of glimpse size and depth is an empirical work. In our experiments, while the size is 1 $\times$ 256 $\times$ 256 (65536 scalars), the output is very small : 2 $\times$ 48 $\times$ 48 (4608 scalars), or about 7% the size of the original document image patch. Fig. 4 shows an intuitional result of glimpse location predication. An example of two test document image patches, which are selected from the test set of SOC dataset. We can get a perceptual idea that the glimpse sensor is able to pay attention to the salient part of a patch in six steps.

In order to evaluate our attention based RNN model for DIQA, we use the Linear Correlation Coefficient(LCC) and the Spearman Rank Order Correlation Coefficient (SROCC). LCC ρ_p is a measure of the degree of linear relationship between two variables, as equation (4) shows.

$$\rho_p = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 \sum_i^n (y_i - \bar{y})^2}} \tag{4}$$

where n is the total number of testing images, x_i and y_i denote the predicted and true OCR accuracy of testing image i, $\bar{x}$ and $\bar{y}$ represent the mean value of predicted and true OCR accuracy on the whole test set. SROCC ρ_s measures how well the relationship between two rank variables can be described

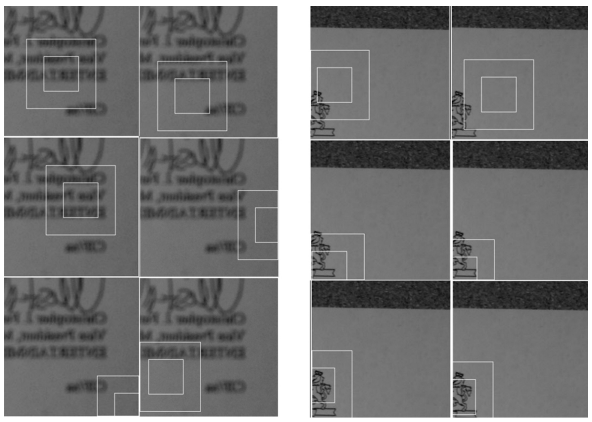


Fig. 4. Glimpse location test result for attention based model on SOC dataset, each patch are given 6 glimpses before a final accuracy score prediction are given.

using a monotonic function, as equation (5) shows.

$$\rho_s = \frac{\sum_i^n (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i^n (u_i - \bar{u})^2 \sum_i^n (v_i - \bar{v})^2}} \tag{5}$$

where u_i and v_i denote the rank order of predicted and true OCR accuracy of th ith testing image i.

C. Experiments on SOC

1) Experiments with and without Convolutional Layer: Fig. 5 demonstrates PLCC and SROCC curve whether network is integrated with convolution layer after spatial glimpse or not. We can conclude that the model with convolutional layer outperforms the one without it in both convergence speed and evaluation protocols, although the model with convolutional layer is larger in size compare to the one without it. The results confirm the effectiveness of combining spatial glimpse and convolutional layer.

2) Comparison with other Methods: In Table I, we show the PLCC and SROCC results of five approaches, including the proposed attention based RNN method and other methods. We can conclude our method provides better PLCC than the one of the other approaches. From the perspective of SROCC, our approach achieves a better result than CNN based method [10] and LDA based features extraction method [6].

For further comparison with CNN based methods, we use a self-built CNN model similar to the one proposed in [10], and train it with patches generated by our patch cropping method. However, the performance is worse than the method proposed in [10], which utilizes accurate character detection and much smaller patch size as Table I shows. This means that the CNN based deep learning method is not appropriate to our preprocessing step.

We also conduct significance test to validate the improvement of proposed method compare with self-build CNN based method as Table II shows. Let μ_0 be the mean PLCC or SROCC of self-build CNN model, μ_1 be the mean PLCC

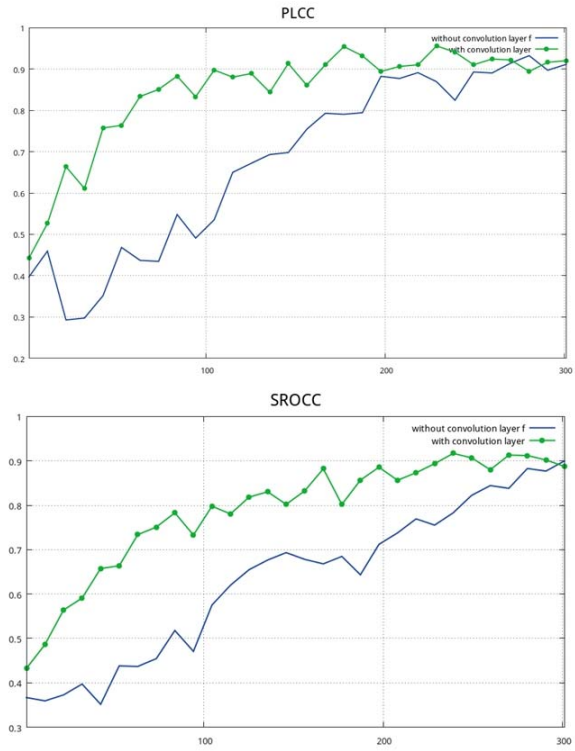


Fig. 5. PLCC and SROCC curve against epochs of training procedure on SOC dataset.

or SROCC of proposed method. We conceive $\mu_0 < \mu 1$ as null hypothesis and $\mu_0 > \mu 1$ as alternative hypothesis. The probability of rejecting the null hypothesis given that it is true (α) is set to 1%. Independent two-sample t-test is adopted to verify the null hypothesis. t-statistic is computed as equation (6):

$$t = \frac{\bar{X_1} - \bar{X_2}}{S_p \sqrt{2/n}} \tag{6}$$

where

$$s_p = \sqrt{\frac{s_{X1}^2 + s_{X2}^2}{2}} \tag{7}$$

Here $\bar{X_1}$ and $\bar{X_1}$ are the statistic averages of the two statistics. s_p is the pooled standard deviation. s_{X1}^2 and s_{X1}^2 are the unbiased estimators of the variances of the two statistics. The degrees of freedom for the student distribution is $2n-2$ where n is the number of participants in each group. In this scenario, we set n to 5. Table II shows us that t-statistic of either PLCC and SROCC is much less than the critical value of Student's t-distribution, which approves the null hypothesis. In other words, we make significance improvement with the attention based RNN method.

D. Experiments on Smartdoc-QA

Table III shows the test result on Smartdoc-QA of self-built CNN method and attention based RNN method. To the

TABLE I
PLCC AND SROCC RESULT ON SOC DATASET

Evaluation Protocol	PLCC	SROCC
CNN Baseline Method (256×256 patch size)	0.926	0.857
CNN[10]	0.950	0.898
LDA[6]	-	0.913
Sparse Representation[7]	0.935	0.928
Proposed method	**0.956**	0.916

TABLE II
PLCC AND SROCC RESULTS OF SELF-BUILT CNN METHOD AND PROPOSED METHOD WITH DIFFERENT DATASET DIVISION

	PLCC		SROCC	
	CNN Baseline Method	Proposed Method	CNN Baseline Method	Proposed Method
1	0.924	0.955	0.855	0.911
2	0.926	0.956	0.852	0.912
3	0.923	0.949	0.853	0.916
4	0.921	0.951	0.857	0.914
5	0.919	0.953	0.849	0.908
t	-16.94		-31.10	
$-t_\alpha(2n-2)$	-2.896		-2.896	

TABLE III
PLCC AND SROCC RESULT ON SMARTDOC-QA DATASET

Evaluation Protocol	PLCC	SROCC
CNN Baseline Method (256×256 patch size)	0.805	0.831
Proposed method	0.814	0.865

best of our knowledge, no published literature has claim the DIQA result on it. Comparing PLCC and SROCC shows in Table III with the ones tested on SOC, the performance is rather worse on Smartdoc-QA. We attribute this mainly to the unbalanced OCR accuracy distribution of the the dataset. Fig. 6 shows the histogram of the OCR accuracy distribution among the images captured by Samsung s4. Adopting the accuracy score by Tesseract OCR engine, we can see that about 40% of all 2160 document images are 0.0%. Another reason for the worse performance on Smartdoc-QA might be the degradation caused by illumination condition. We can see from Fig. 3 that different from SOC dataset, some samples from Smardoc-QA are degraded by shadow object. The patch is not big enough to cover the variation of illumination. So some patches with different illumination will have the same OCR accuracy score, which may mislead the network training procedure.

IV. CONCLUSION

In this paper, we present a DIQA method based on attention based RNN. Unlike the conventional OCR accuracy prediction methods, including unsupervised feature extraction methods and traditional CNN based deep learning methods, our RNN based model is inspired by attention mechanism of human vision. In this framework, spatial glimpse is implemented to select a salient part of document image, and a convolutional layer is used to extract features. We implement reinforcement learning algorithm to train the locator module while stochastic gradient descent is used to train other parts of network. We design a novel reward function and validate the effectiveness of it. The experimental results show that the proposed method is superior to the conventional methods including unsupervised feature extraction based methods and CNN based methods for DIQA.

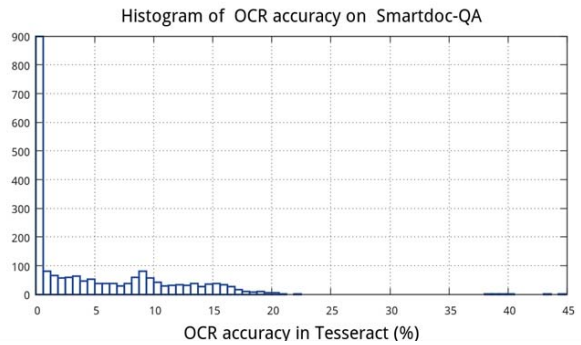


Fig. 6. Histogram of OCR accuracy of Smartdoc-QA dataset, document images are taken by Samsung S4, and their accuracy are give by Tesseract OCR engine.

Our future work includes the optimization of network architecture to enhance the generalization ability. We also plan to explore method combining CNN and RNN to improve the DIQA performance on Smartdoc-QA dataset.

ACKNOWLEDGMENT

We would like to thank Nicholas Lonard for his contribution to the demo code of recurrent visual attention. We would like to appreciate the efforts spent by the reviewers. We are pleased to notice that some of our research work attracts their attentions, and some of the weaknesses they mentioned are very helpful to improve the quality of our work. This work is supported by the National Basic Research Program of China (973 Program) (No. 2014CB340506), and National Natural Science Foundation of China (No. U1636212, U1636124, 61573028).

REFERENCES

[1] P. Ye and D. Doermann, *Document image quality assessment: A brief survey,* in proceedings of International Conference on Document Analysis and Recognition, pp. 723-727, 2013.

[2] M. Cannon, P. Kelly, and J. Hochberg, *Quality assessment and restoration of typewritten document images,* International Journal on Document Analysis and Recognition, vol. 2, no. 2-3, pp. 80-89, 1999.

[3] A. Souza, M. Cheriet, S. Naoi, and C. Y. Suen, *Automatic filter selection using image quality assessment,* in proceedings of Seventh International Conference on Document Analysis and Recognition , pp. 508-512, 2003.

[4] P. Ye, J. Kumar, L. Kang, and D. Doermann, *Unsupervised feature learning framework for no-reference image quality assessment,* in proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1098-1105, 2012.

[5] P. Ye. and D. Doermann, *Learning features for predicting OCR accuracy,* in proceedings of International Conference on Pattern Recognition, pp. 3204-3207, 2012.

[6] X. Peng, H. Cao, and P. Natarajan, *Document image OCR accuracy prediction via Latent Dirichlet Allocation,* in proceedings of International Conference on Document Analysis and Recognition, pp. 771-775, 2015.
[7] X. Peng, H. Cao, and P. Natarajan, *Document Image Quality Assessment Using Discriminative Sparse Representation,* in proceedings of IAPR Workshop on Document Analysis Systems, pp. 227-232, 2016.
[8] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. LeCun, *Learning convolutional feature hierachies for visual recognition,* in proceedings of International Conference on Neural Information Processing Systems, pp. 1090-1098, 2010.
[9] A. Krizhevsky, I. Sutskever, and G. Hinton, *Imagenet classification with deep convolutional neural networks,* in proceedings of International Conference on Neural Information Processing Systems Vol.25, pp.1097-1105, 2012.
[10] L. Kang, P. Ye, Y. Li, D. Doermann, *A deep learning approach to document image quality assessment,* in proceedings of IEEE International Conference on Image Processing, pp. 2570-2574, 2014.
[11] F. Wang, D. Tax, *Survey on the attention based RNN model and its applications in computer vision,* arXiv:1601.06823v1, 2016.
[12] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, *Attention-Based Models for Speech Recognition,* arXiv:1506.07503, 2015.
[13] V. Mnih, N. Heess, A. Graves, et al., *Recurrent models of visual attention,* arXiv:1406.6247., 2014.
[14] K. Xu, J. Ba, R. Kiros et al, *Show, attend and tell: Neural image caption generation with visual attention,* in proceedings of Machine Learning Research, vol. 37 pp. 2048-2057, 2015.
[15] A. Graves, *Generating sequences with recurrent neural networks,* arXiv:1308.0850, 2013.
[16] V. Nair and G. E. Hinton, *Rectified linear units improve restricted boltzmann machines,* in proceedings of International Conference on Machine Learning, pp. 807-814, 2010.
[17] S. Hochreiter and J. Schmidhuber, *Long short-term memory,* Neural computation, Vol. 9, No. 8, pp. 1735-1780, 1997.
[18] F. Gers, *Long Short-Term memory in recurrent neural Networks,* Ph.D.Thesis, University at Hannover, Germany, 2001.
[19] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, *On the Properties of Neural Machine Translation: Encoder-decoder Approaches,* arXiv:1409.1259, 2014.
[20] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,* arXiv:1412.3555, 2014.
[21] P. Li, J. Zhu, L. Peng, Y. Guo, *RNN Based Uyghur Text Line Recognition and Its Training Strategy,* 12th IAPR Workshop on Document Analysis Systems, pp. 19-24, 2016.
[22] R.J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning,* in Machine Learning, vol. 8 no. 3, pp. 229-256, 1992.
[23] R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation,* , in proceedings of International Conference on Neural Information Processing Systems, pp. 1057-1063, 2000.
[24] J. Kumar, P. Ye, and D. Doermann, *A Dataset for Quality Assessment of Camera Captured Document Images,* in proceedings of International Workshop on Camera-Based Document Analysis and Recognition, pp. 39-44, 2013.
[25] N. Nayef, M. Muzzamil Luqman, S. Prum, S. Eskenazi, J. Chazalon and J. Ogier, *SmartDoc-QA: A Dataset for Quality Assessment of Smartphone Captured Document Images Single and Multiple Distortions,* in proceedings of International Workshop on Camera-Based Document Analysis and Recognition, pp. 1231-1235, 2015.