Visual Attention based OCR

Yuntian Deng and **Qi Guo** [*]
School of Computer Science, Carnegie Mellon Univeristy
Pittsburgh, PA 15213, USA
yuntiand@cs.cmu.edu, qiguo@andrew.cmu.edu

Abstract

Optical Character Recognition (OCR) has been a long standing language-vision joint research topic. Recently, deep learning based methods have become more and more popular since they are end-to-end trainable and have achieved the state-of-the-art results. The structure of such methods typically include an encoder and a decoder, where the encoder encodes the image to a fixed-length vector from which the decoder generates the character sequence. However, a fixed length representation is unsuitable for images of various aspect ratios. Inspired by the success of attention-based models in neural machine translation, we propose to treat a text image as a sequence of vertical image slices and formulate OCR as a sequence-to-sequence problem, where we employ a visual attention mechanism in the decoder. Experiments show that our model can do character decoding and image-text alignment simultaneously.

Introduction

Optical character recognition (OCR) in natural scenes (Chen and Yuille 2004) goes beyond processing document photos, tries to perform OCR in natural scenes, like city streets. This can scaffold autonomous driving by recognizing street signs, refine maps from street view images, enable mobile phones to search directly with photos and assist blind people. While being more generally applicable, scene text recognition introduces more challenges. The difficulty of OCR is in general two-fold: character-level recognition and image-word alignment.

Traditionally, character recognition is traditionally done based on hand-engineered features (Graves et al. 2009; Wang and Belongie 2010; Su and Lu 2014) or mid-level features learned on top them (Gordo 2015; Lee et al. 2014; Yao et al. 2014). Recently convolutional neural network (CNN) is explored again for OCR (Jaderberg et al. 2015; Jaderberg et al. 2016; Wang et al. 2012) and proven to work well.

Alignment for OCR is challenging since the input is unsegmented. The font size and the spacing is unknown. Geometry distortions and different illumination conditions like specular highlight add more to the difficulty. Previous models used sliding windows (Mishra, Alahari, and Jawahar 2012b), over-segmentation (Bissacco et al. 2013), pictorial structures (Wang, Babenko, and Belongie 2011) and recurrent neural network (RNN) with connectionist temporal classification (CTC) (Graves et al. 2006; Graves et al. 2009; Shi, Bai, and Yao 2015). While CTC seems to be an appealing choice for solving the alignment problem for unsegmented sequence data, it assumes the network outputs are conditionally independent given the network's hidden states. Character-level recognition and image-word alignment are tackled at different stages in the model. Other probabilistic graphical models like HMM (Rabiner and Juang 1986) and CRF (Lafferty, McCallum, and Pereira 2001) also have some dependence assumptions for tractable inference. So do their hybrids with RNN.

Most previous approaches resize input images to the same size, ignoring the sequential nature

[*]The two authors contributed equally.

of the input. While not explicitly stating it, Shi et al. (2015) approached OCR as a sequence-to-sequence problem. If we treat the word image as a sequence of vertical image slices, and the word as a sequence of characters, then the scene text reading problem is formulated as a sequence-to-sequence problem. It is a general class of problem including machine translation, speech recognition and OCR etc. Recent sequence-to-sequence neural network models are usually proposed under the encoder-decoder framework, allowing variable input/ output lengths. The encoder generates a fixed-length vector to represent the whole input sequence. The vector is feed into the decoder, which generates the output sequence from it.

Recently, attention model has been proven to work well for sequence-to-sequence learning tasks like machine translation (Bahdanau, Cho, and Bengio 2014; Luong, Pham, and Manning 2015), syntax parsing (Vinyals et al. 2015) and speech recognition (Chan et al. 2016). The idea of *visual attention* has been applied to image-to-sequence tasks, like image captioning (Xu et al. 2015; Karpathy and Fei-Fei 2015). Lee and Osindero (Lee and Osindero 2016) tried to integrate a weighting scheme for image features. It used a CNN with fully connected layers, so the weights cannot be matched to certain receptive fields in the image. In contrast, each character in the text corresponds to a specific image region in reality, thus the idea of visual attention is quite intuitive to apply.

In this paper, we propose to use visual attention based encoder-decoder framework for OCR, coupling character recognition and alignment together in the model. Our model better exploits the sequence-to-sequence nature of the task, which enables us to take variable-length inputs and outputs.

Visual Attention based Model

Problem Formulation

Our model takes as input a gray-scale image $I \in \mathbb{R}^{h^I \times d^I}$, where h^I and d^I are image height and width. These images are word images cropped from natural scene images. Our model outputs a sequence of characters $\mathbf{y} = \{y_1, y_2, \cdots, y_C\}$, where y_i is a character, C is the length of the output word.

Note that, many previous approaches are lexicon-dependent, where there is a lexicon of small size (usually 50) for each word image, or a global lexicon containing the set of possible words. Recent end-to-end trainable systems starts to show promising results in the lexicon-free setting (Shi, Bai, and Yao 2015; Lee and Osindero 2016). Following them, we are concerned with the lexicon-free setting.

Model Structure

Our model first extracts image feature sequences using CNN. Then encodes the sequence with an RNN. The encoded sequence is feed into the decoder with a visual attention mechanism. The decoder RNN outputs a character sequence as predicted word text. The structure is illustrated in Figure 1.

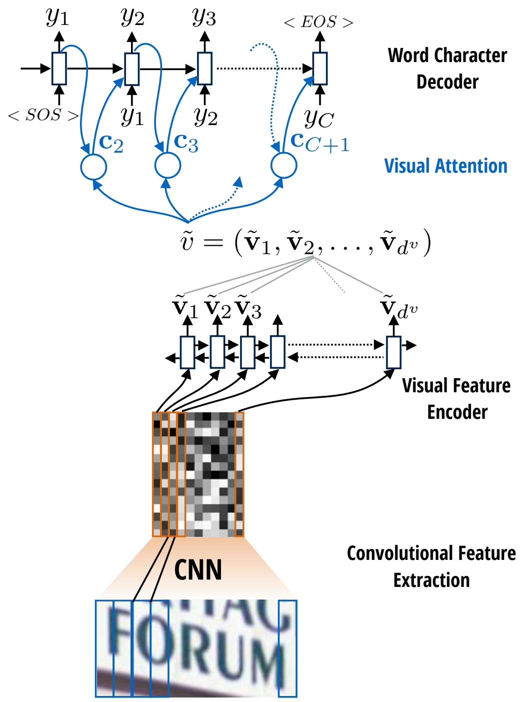
y_1
y_2
y_3
$< EOS >$
Word Character Decoder
$< SOS >$
y_1
y_2
y_C
$\mathbf{c}_2$
$\mathbf{c}_3$
$\mathbf{c}_{C+1}$
Visual Attention
$\tilde{v} = (\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_{d^v})$
$\tilde{\mathbf{v}}_1$ $\tilde{\mathbf{v}}_2$ $\tilde{\mathbf{v}}_3$
$\tilde{\mathbf{v}}_{d^v}$
Visual Feature Encoder
CNN
Convolutional Feature Extraction
FORUM

Figure 1: Visual attention based encoder-decoder network structure.

Convolutional Feature Extraction

At the start of the model, we extract the visual features of a image using a CNN. Unlike some recent work (Jaderberg et al. 2015; Lee and Osindero

2016), we do not use fully-connected layers at the end of the neural network. It only contains convolution layers and max-pooling layers. These layers preserve the locality of image features, which offers the possibility of incorporating the concept of visual attention into the model. The CNN produces feature maps of size $c^v \times h^v \times d^v$, where c^v denotes the number of channels. We split and reshape them into d^v vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_{d^v})$, each of length $c^v \times h^v$. Each vector corresponds to a vertical slice/frame of the image, forming a sequence. Note the receptive fields overlap, which is not illustrated in Figure 1 for simplicity. The specification of the CNN for our implementation is described later .

RNN Encoder

We feed the CNN features into a RNN, encoding sequential dependencies into the visual features. In image captioning (Xu et al. 2015), image feature vectors are directly fed into the decoder, without additional encoding. Because in image captioning, object location matters much less than their semantic meanings. Wherever the bird appears will only slightly affect where we will put the word "bird" in the sentence. The word order of generated caption mainly depends on the language model. So even if we shuffle the image feature vectors, i.e., reorganizing locations of objects in the image, we would expect similar image captions. Convolutional features represent well local semantic information. However, for OCR, we expect the visual features fed in to the decoder contains the sequential order information in addition. In general terms, an RNN has a hidden state $\mathbf{h}_t$. At time t, it is updated with input $\mathbf{x}_t$ in the following manner:

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t)$$

. The output is $f_o(V\mathbf{h}_t)$, where f_o can be linear and softmax etc. We will elaborate on our choices of RNN later . Here, the encoder RNN takes $\mathbf{x}_t^{\text{enc}} = \mathbf{v}_t, 1 \leq t \leq d^v$ as input. The output is an *annotation vector* $\tilde{\mathbf{v}}_t$ for each frame t.

Visual Attention based RNN Decoder

The decoder RNN generates sequence of characters $\{y_i\}$, based on a sequence of annotation vector $\{\tilde{\mathbf{v}}_t\}$. At each time steps i, for predicting a character, the RNN takes into account all annotation vectors in the sequence. However, not all annotation vectors are equal. Intuitively, a character in the word has a corresponding region in the image. So annotation vectors representing local vertical image slices should be attended differently. We use an *attention model* $a(\cdot)$ to model the alignment:

$$\begin{aligned}
\mathbf{e}_i &= a(\mathbf{h}_{i-1}, \{\tilde{\mathbf{v}}_t\}) \\
\alpha_i &= \text{softmax}(\mathbf{e}_i) \\
\mathbf{c}_i &= \phi(\{\tilde{\mathbf{v}}_t\}, \alpha_i)
\end{aligned}$$

α_i are the weights calculated based on $\mathbf{e}$. The weight vector α_i and all annotation vectors $\{\tilde{\mathbf{v}}_t\}$ are combined to form a *context vector* $\mathbf{c}_i$. Note there are different choices for a and ϕ, we use $e_{it} = \beta^T \tanh(W_h \mathbf{h}_{i-1} + W_v \tilde{\mathbf{v}}_t)$ and $\mathbf{c}_i = \sum_t \alpha_{it} \mathbf{v}_t$. Finally $\mathbf{c}_i$ and y_{i-1} are together fed in to the decoder RNN as input at step i, for predicting y_i.

Experiments

Dataset

We use Synth (Jaderberg et al. 2014), a large synthetic data set for training the model. The testing is done on the following real image datasets without ad hoc modifications.

ICDAR03 (Lucas et al. 2005) consists of 860 test word images cropped from 230 scene images. Following Wang et al. (2011) and others for comparison, images with nonalphanumeric chracters or word length less than three are excluded.

ICDAR13 (Karatzas et al. 2013) is extended from ICDAR 03, with 1015 cropped test word iamges.

SVT (Wang, Babenko, and Belongie 2011), i.e., Street View Text, consists of 647 word images cropped from 249 Google Stree View images.

IIIT5k (Mishra, Alahari, and Jawahar 2012a) contains 3000 cropped test word images.

Some of the above datasets has a preselected lexicon for each image, which are ignored as we are concerned with the lexicon-free recognition task.

Implementation Details

We resize the input grayscale image to be of fixed height 32 while preserving the aspect ratio. Different for resizing all images to the same size, it preserves some visual features, like character

Layer	Specification
Conv	c:512, k:(2,2), s:(1,1), p:(0,0), bn
MP	po:(2,1), s:(2,1)
Conv	c:512, k:(3,3), s:(1,1), p:(1,1)
Conv	c:512, k:(3,3), s:(1,1), p:(1,1), bn
MP	po:(2,1), s:(2,1)
Conv	c:256, k:(3,3), s:(1,1), p:(1,1)
Conv	c:256, k:(3,3), s:(1,1), p:(1,1), bn
MP	po:(2,2), s:(2,2)
Conv	c:128, k:(3,3), s:(1,1), p:(1,1)
MP	po:(2,2), s:(2,2)
Conv	c:64, k:(3,3), s:(1,1), p:(1,1)
Input	$32 \times d^v$ grayscale image

Table 1: CNN specification. ‘Conv‘: convolution layer, ‘MP’: Max-pooling layer. ‘c’:channel, ‘k’: kernel size, ‘s’: stride, ‘p’: padding size, ‘po’: , ‘bn’: with batch normalization

shape and spacing. We only consider alphanumeric characters without discriminating upper and lower cases, resulting in a target vocabulary $\{y_i\}$ consisting of A-Z, 0-9, and two special tokens <SOS>/<EOS> marking the start and end of a character sequence.

The structure of the CNN (Table 1) is adapted from Shi et al. (2015). The activation function we use is ReLU. We do batch normalization along the channel axis before activation to alleviate internal covariate shift in the model for convolution layers. We have $c^v = 512, h^v = 1, d^v = \frac{1}{4}d^I$ using this configuration. We use a bidirectional LSTM for the encoder RNN. A two-layer LSTM is used for the decoder RNN. Our model has about 8M parameters in total. The encoder bidirectional LSTM has 256 hidden units for each LSTM. The decoder 2-layer LSTM has 128 hidden units each. We train our model with SGD, using the AdaDelta algorithm to automatically tune the learning rate. The minibatch size is set to be 128.

We developed a pretraining scheme for our model. The pretraining is based on masking the target output. We first train our model to predict the first character of the word, while masking other characters. Whatever the model predict for other characters does not affect the gradient. Then we extend to predict first two, four, eight and finally all characters. Intuitively, when we only train for the first character, we are basically just training the CNN, as the output is not sequential yet. Later on, we are gradually training the RNNs. At each stage, we first fix the parameters for CNN from last stage, only training the RNN parameters. Then we make the CNN trainable, training it jointly with RNN. This pretraining strategy worked well experimentally for our model, and it may be able to generalize to other sequential models.

Results

Qualitative Results In order to visualize the inferred attention weights, we up-sample the attention map to the size of the original image, and apply a Gaussian filter for smoothing. Then we multiply it with the original image to show the attention results (Figure 2). As illustrated, our model have successfully aligned the characters between the text and the original image.

Quantitative Results We compare our method to other end-to-end recognition methods by word recognition accuracy based on the edit distance, following the convention (Table 2). The listed methods are Jaderberg15 (Jaderberg et al. 2015), Jaderberg16 (Jaderberg et al. 2016), PhotoOCR (Bissacco et al. 2013), Shi et al. (2015) and R^{2}AM (Lee and Osindero 2016).

Method	IIIT	SVT	IC03	IC13
Jaderberg15	-	71.7	89.6	81.8
Jaderberg16*	92.7	80.7	93.1	90.8
PhotoOCR	-	78.0	-	87.6
Shi et al.	78.2	80.8	89.4	86.7
R^{2}AM	78.4	80.7	88.7	90.0
Ours	67.7	67.6	77.1	74.2

Table 2: Word recognition accuracy (%). *: not strictly lexicon-free, as with a global dictionary.

Conclusion and Future Work

We have proposed a visual attention based sequence-to-sequence model for OCR. Unfortunately, our model is not achieving the state-of-art performance. The performance might be improved by hyperparameter selection via cross-validation etc. On a more technical perspective, one drawback of our model might be the unconstrained attention mechanism. Our current attention model allows the decoder attending to any position at

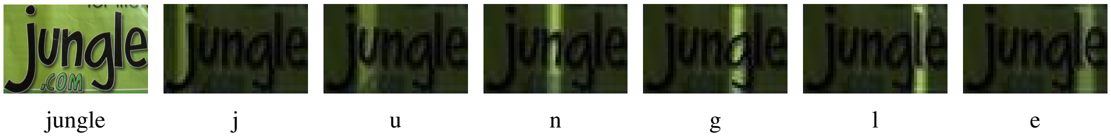

Figure 2: Attention Visualization

any time step. For OCR, however, the attention should strictly follow a left-to-right order. For future work, we will incorporate this knowledge into the attention mechanism. For example, we can develop a strategy like the local attention scheme (Luong, Pham, and Manning 2015), while forcing the attention to move forward.

References

[Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Bissacco et al. 2013] Bissacco, A.; Cummins, M.; Netzer, Y.; and Neven, H. 2013. Photoocr: Reading text in uncontrolled conditions. In *Proceedings of the IEEE International Conference on Computer Vision*, 785–792.

[Chan et al. 2016] Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4960–4964. IEEE.

[Chen and Yuille 2004] Chen, X., and Yuille, A. L. 2004. Detecting and reading text in natural scenes. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, II–366. IEEE.

[Gordo 2015] Gordo, A. 2015. Supervised mid-level features for word image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2956–2964.

[Graves et al. 2006] Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376. ACM.

[Graves et al. 2009] Graves, A.; Liwicki, M.; Fernández, S.; Bertolami, R.; Bunke, H.; and Schmidhuber, J. 2009. A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(5):855–868.

[Jaderberg et al. 2014] Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

[Jaderberg et al. 2015] Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2015. Deep structured output learning for unconstrained text recognition. *ICLR*.

[Jaderberg et al. 2016] Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2016. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision* 116(1):1–20.

[Karatzas et al. 2013] Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Gomez i Bigorda, L.; Robles Mestre, S.; Mas, J.; Fernandez Mota, D.; Almazan Almazan, J.; and de las Heras, L.-P. 2013. Icdar 2013 robust reading competition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, 1484–1493. IEEE.

[Karpathy and Fei-Fei 2015] Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.

[Lafferty, McCallum, and Pereira 2001] Lafferty, J.; McCallum, A.; and Pereira, F. C. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, volume 951, 282–289.

[Lee and Osindero 2016] Lee, C.-Y., and Osindero, S. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. *arXiv preprint arXiv:1603.03101*.

[Lee et al. 2014] Lee, C.-Y.; Bhardwaj, A.; Di, W.; Jagadeesh, V.; and Piramuthu, R. 2014. Region-based discriminative feature pooling for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4050–4057.

[Lucas et al. 2005] Lucas, S. M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; Young, R.; Ashida, K.; Nagai, H.; Okamoto, M.; Yamamoto, H.; et al. 2005. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)* 7(2-3):105–122.

[Luong, Pham, and Manning 2015] Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *EMNLP*.

[Mishra, Alahari, and Jawahar 2012a] Mishra, A.; Alahari, K.; and Jawahar, C. 2012a. Scene text recognition using higher order language priors. In *BMVC 2012-23rd British Machine Vision Conference*. BMVA.

[Mishra, Alahari, and Jawahar 2012b] Mishra, A.; Alahari, K.; and Jawahar, C. 2012b. Top-down and bottom-up cues for scene text recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2687–2694. IEEE.

[Rabiner and Juang 1986] Rabiner, L. R., and Juang, B.-H. 1986. An introduction to hidden markov models. *ASSP Magazine, IEEE* 3(1):4–16.

[Shi, Bai, and Yao 2015] Shi, B.; Bai, X.; and Yao, C. 2015. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *arXiv preprint arXiv:1507.05717*.

[Su and Lu 2014] Su, B., and Lu, S. 2014. Accurate scene text recognition based on recurrent neural network. In *Computer Vision–ACCV 2014*. Springer. 35–48.

[Vinyals et al. 2015] Vinyals, O.; Kaiser, Ł.; Koo, T.; Petrov, S.; Sutskever, I.; and Hinton, G. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, 2755–2763.

[Wang and Belongie 2010] Wang, K., and Belongie, S. 2010. Word spotting in the wild. *Computer Vision–ECCV 2010* 591–604.

[Wang, Babenko, and Belongie 2011] Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 1457–1464. IEEE.

[Wang et al. 2012] Wang, T.; Wu, D. J.; Coates, A.; and Ng, A. Y. 2012. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, 3304–3308. IEEE.

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of The 32nd International Conference on Machine Learning*, 2048–2057.

[Yao et al. 2014] Yao, C.; Bai, X.; Shi, B.; and Liu, W. 2014. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4042–4049.