End-to-End Chinese Image Text Recognition with Attention Model

Fenfen Sheng[1,2(✉)], Chuanlei Zhai[1,2], Zhineng Chen[1], and Bo Xu[1]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{shengfenfen2015,zhaichuanlei2014,zhineng.chen,xubo}@ia.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100190, China

Abstract. This paper presents an attention-based model for end-to-end Chinese image text recognition. The proposed model includes an encoder and a decoder. For each input text image, the encoder part firstly combines deep convolutional layers with bidirectional Recurrent Neural Network to generate an ordered, high-level feature sequence, which could avoid the complicated text segmentation pre-processing. Then in the decoder, a recurrent network with attention mechanism is developed to generate text line output, enabling the model to selectively exploit image features from the encoder correspondingly. The whole segmentation-free model allows end-to-end training within a standard backpropagation algorithm. Extensive experiments demonstrate significant performance improvements comparing to baseline systems. Furthermore, qualitative analysis reveals that the proposed model could learn the alignment between input and output in accordance with the intuition.

Keywords: Chinese images text recognition · End-to-end · Attention · Segmentation-free

1 Introduction

Chinese Image Text Recognition (ChnITR), which aims to read Chinese text in natural images, is an essential step for various commercial applications, such as geolocation, caption reading, and image-based machine translation. Despite the maturity of researches on Optical Character Recognition (OCR) [1], text recognition in natural images instead of scanned documents still remains challenging. Difficulties mainly come from the diversity of text patterns (e.g. low resolution, low contrast, and blurring) and cluttered background.

There are many ChnITR systems [2–6] focus on developing powerful word-level classifiers to improve performance. Most of them follow a three-step pipeline: first segmenting text into words, then recognizing each individual word, and finally employing post-processing to combine words results back into text. There is a trend to adopt Deep Neural Networks (DNN) for word representation learning in step 2. However, ChnITR accuracy is still confined by the word-level recognition, which lacks meaningful context-dependent information of

D. Liu et al. (Eds.): ICONIP 2017, Part III, LNCS 10636, pp. 180–189, 2017.
https://doi.org/10.1007/978-3-319-70090-8_19

whole text and further influences the following text combination step. Besides, the pre-segmentation step is quite tricky, which could bring in unavoidable troubles such as getting error text cutting points.

To deal with these problems, some methods regard ChnITR as a sequence labelling task. They are mainly based on Recurrent Neural Networks (RNN) [7]. Zhai et al. [8] combine bidirectional Long Short-Term Memory Recurrent Neural Networks (BiLSTM) with Connectionist Temporal Classification (CTC) for ChnITR without segmentation pre-processing. However, this system can not be trained end-to-end because it uses manually extracted HOG features as input, which also limit the learning of high-level features.

Inspired by recent advances on machine translation [9], this paper adopts the encoder-decoder model with attention mechanism, and proposes several meaningful modifications to accommodate ChnITR. For each input image, the encoder, i.e., deep convolutional layers (CNN) with BiRNN, is employed to generate a discriminative feature sequence. And the decoder, an attention-based network, recurrently outputs recognition result by decoding relevant contents from the feature sequence, which is determined by its attention mechanism at each step. Because of these, the whole attention-based architecture can be jointly trained end-to-end to maximize correction of the recognized result.

The contributions of this paper are as below: First, an attention-based architecture is adopted for ChnITR, which could automatically learn the alignment between input and output, and therefore dispense with the segmentation preprocessing. Second, a CNN-BiRNN structure is applied in encoder to extract both high-level and context-dependent features from the input. The whole system can be trained end-to-end without any pre or post processing thus could better explores information contained in the original text image. Experiments show that a significantly performance improvement has been achieved comparing to baseline systems. There are also a detailed ablation study to examine the effectiveness of each proposed components. Furthermore, qualitative analysis reveals that the proposed model finds the alignment between the source image and destination text as mentioned above.

2 Related Work

The ChnITR problem has been discussed for a long time but still remains unperfectly solved. By using segmentation methods, most of ChnITR systems capture individual words from text as the first step in the recognition pipeline. Mishra et al. [10] and Bai et al. [11] apply different binarization algorithms separately to distinguish text pixels from non-text's, then use existing OCR system to segment text into words and do word-level recognition work. A method from Bai et al. [4] is based on over-segmentation process. It firstly finds all potential cutting points from input text, then combines beam search with a language model to get the proper cutting points dynamically. After these steps, a word-level classifier is exploited to recognize each word. We use method [4] as a baseline to compare performances between our system and traditional ChnITR approach.

The advances of DNN for image representation encourage the development of more powerful word classifiers. A multi-layer DNN system is proposed for ChnITR in [12]. SHL-CNN system [13] employs a shared-hidden-layer deep convolutional neural network to extract word-level features. Similarly, Zhong et al. [14] put forward a multi-pooling layer on top of final convolutional layer; which is robust to spatial layout variations and deformations in multi-font printed Chinese words. And a 7-layer CNN based ChnITR architecture with synthetic data is presented by X Ren et al. [15].

Approaches above do not utilize whole-text context information because isolated word classification and subsequent words combination are treated separately. Usually, they have to design complicated optimization algorithm to train their systems and post-processing steps to refine output results. In order to leverage context-dependent information of the whole text, some researches regard text reading as a sequence labelling task avoiding explicitly segmentation. Ronaldo et al. [16] come up with a Chinese handwritten text recognition system based on Multi-Dimensional Long Short-Term Memory Recurrent Neural Network (MDLSTM-RNNs) trained with CTC. In [8], a novel BLSTM-CTC method is proposed to solve ChnITR problem. However, the use of HOG feature in [8] restricts its performance and end-to-end optimization. We use this system as another baseline to validate the effectiveness of our end-to-end system.

The proposed system is motivated by the recent advances of "attention-based" model, which has been proven successful for machine translation [9], image caption generation [17,18], and speech recognition [19]. Attention mechanism has been widely applied to the recognition of symbols sequences [20–23]. Baoguang et al. [22] extend the STN framework [24] with an attention-based sequence recognizer to train the whole model end-to-end. In [21], the authors come up with a new lexicon-free photo OCR framework that incorporates recursive CNNs for image encoding, RNNs for language modeling, and attention-based mechanism for better image feature usage. At the same time, Theodore et al. [20] present an attention-based model to transcribe complete paragraphs of text without an explicit line segmentation with a multi-dimensional LSTM network. Similar to English text recognition task, this paper attempts to solve ChnITR problem end-to-end by extending attention-based model without segmentation pre-processing.

3 An Attention-Based Model for End-to-End ChnITR

This section describes the proposed method for ChnITR, including an encoder and a decoder. The whole architecture is depicted in Fig. 1.

3.1 Encoder: Combination of Deep CNN and BiRNN

As illustrated in Fig. 1, there are several convolutional layers at the bottom of the encoder. It should be noted that, each image I has been resized into a fixed height before being fed to the encoder, while keeping its original aspect

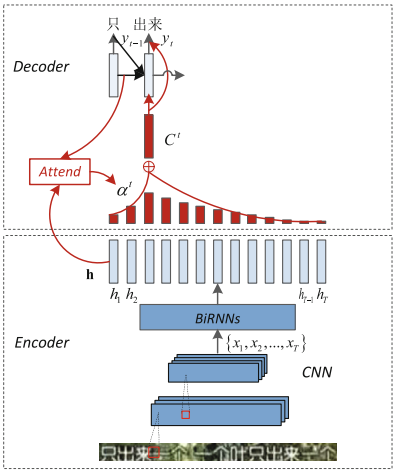

Fig. 1. The end-to-end architecture for ChnITR. The encoder combines deep CNN with BiRNN to extract feature sequence (**h**) from each input image. The decoder then generates the text result based on **h**.

ratio unchanged. After applying deep CNN to the resized image, we get a CNN sequence $\{x_1, x_2, \ldots, x_T\}$, each component of which contains high-level semantic information due to powerful feature extraction ability of deep CNN. What's more, the translation invariance property of deep CNN enables each x_i have a corresponding local image region, i.e. receptive fields.

Restricted by the size of receptive fields, the generated CNN sequence leverages limited context-dependent information. To get long-distance dependency, BiRNN is adopted on top of deep CNN. The BiRNN has also shown strong capability in learning meaningful structure from an ordered sequence. The BiRNN consists of a forward and a backward RNN. The forward RNN $\overrightarrow{f}$ reads the input sequence $\{x_1, x_2, \ldots, x_T\}$ as it is ordered (from x_1 to x_T) and calculates a sequence of forward hidden states $\left\{\overrightarrow{h_1}, \ldots, \overrightarrow{h_T}\right\}$. The backward RNN $\overleftarrow{f}$ reads the sequence in reverse order (from x_T to x_1), resulting in a sequence of backward hidden states $\left\{\overrightarrow{h_T}, \ldots, \overrightarrow{h_1}\right\}$. We obtain a feature sequence $\mathbf{h} = \{h_1, \ldots, h_T\}$ from $\{x_1, x_2, \ldots, x_T\}$ as the encoder output, each of which h_j concatenates the forward hidden state $\overrightarrow{h_j}$ and the backward one $\overleftarrow{h_j}$, i.e. $h_j = \left[\overrightarrow{h_j}; \overleftarrow{h_j}\right]$. In this way, the encoder vector h_j contains the summary of both the preceding words and the following words. As the BiRNN tends to focus on recent input x_j, the encoder vector h_j performs a better representation for the word around x_j.

3.2 Decoder: Recurrent Word Generator

The decoder generates text result, based on feature sequence $\mathbf{h} = \{h_1, \ldots, h_T\}$ from the encoder. Specifically, the decoder defines a probability over the recognition text $\mathbf{y}$ by decomposing the joint probability into ordered conditionals:

$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t \,|\{y_1, \ldots, y_{t-1}\}, c_t). \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_T)$, and c_t is a context vector at time t generated from the encoder sequence $\mathbf{h}$. With one RNN layer, each conditional probability is modeled as

$$p(y_t \,|y_1, \ldots, y_{t-1}, c_t) = g(y_{t-1}, s_t, c_t). \tag{2}$$

where g is a nonlinear function that outputs the probability of y_t, and s_t is the hidden state of the RNN, computed by

$$s_t = f(s_{t-1}, y_{t-1}, c_t). \tag{3}$$

It should be noted that, the probability for each target word y_t is conditioned on a distinct context vector c_t. The context vector c_t depends on the feature sequence $\mathbf{h}$:

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j. \tag{4}$$

The weight α_{tj} of each feature vector h_j is computed by:

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})}, \tag{5}$$

where

$$e_{tj} = a(s_{t-1}, h_j). \tag{6}$$

is an alignment model which scores how well the inputs around position j matches the output at position t. The score is based on the RNN hidden state s_{t-1} and the j-th feature vector h_j of the input sentence. The attention model a is parameterized as a feedforward neural network which is trained jointly with all the other components of the proposed system. With this approach, the information can be spread throughout the encoder feature sequence, which can be selectively retrieved by the decoder accordingly.

4 Experiments

4.1 Dataset

This paper carries out experiments on the dataset described in [4]. The dataset consists of 13 TV channels with various text patterns and backgrounds.

It includes 6633 real labeled text images (i.e., frames extracted from TV channels) and 1617636 artificial text images generated automatically by using the method in [4], since the number of labeled text images is far from enough for model training. Similarly to [4], we use all artificial text images for training, 4846 labeled text images for fine-tuning, and the rest 1787 for testing. For more details about dataset, please refer to [4].

4.2 Implementation Details

The proposed architecture is shown in Fig. 1. For the encoder, the CNN part has 7 convolutional layers, whose filter size, number of filters, stride and padding size are respectively $\{3, 64, 1, 1\}$, $\{3, 128, 1, 1\}$, $\{3, 256, 1, 1\}$, $\{3, 256, 1, 1\}$, $\{3, 512, 1, 1\}$, $\{3, 512, 1, 1\}$, $\{2, 512, 1, 0\}$. 2×2 max pooling follows the first, second, fourth, and sixth convolutional layers. The third, fifth and seventh convolutional layers are followed by a batch normalization operation separately. The input of the CNN is a resized $W \times 32$ gray scale image, and the output is a T length CNN sequence (different text image has different T because of different W). A BiRNN is on the top of the CNN, each of which has 1024 hidden units. For the decoder, the system uses a one-layer RNN network with 1024 cells and a softmax layer with 40000 output units. The number of the softmax function is equal to the class number of Chinese words in the dataset. The whole network is trained with the adadelta algorithm [25]. After that, beam search is employed to determine the recognition result that approximately maximizes the conditional probability.

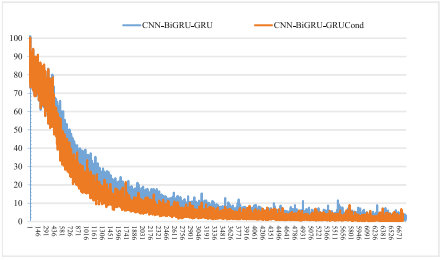

Fig. 2. Convergence process of models trained with different RNN units.

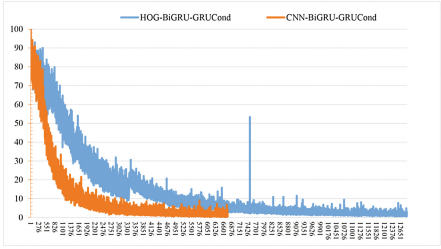

Fig. 3. Convergence process of models trained with different feature extractors.

4.3 Ablation Study

In this section, we empirically investigate the contributions made by different components in the proposed model, namely: different RNN units for model performance and different feature extractors for the input text image. After that, we cite methods [4,8] as baselines, and use character error rate (CER) as evaluation metric to compare their recognition performances.

Table 1. *Performance of models trained with different RNN units.*

Models	CNN−BiGRU−GRU	CNN−BiGRU−GRUCond
CER(%)	10.17	**9.06**

Different RNN Units for Model Performance. Gated Recurrent Unit (GRU) [26], one of RNN units, could overcome the vanishing gradient problem in standard RNN and transmit the gradient information consistently over long time. Compared to LSTM, GRU is simpler because it only has two gates but no memory cell or peephole connection. Recently, a novel conditional GRU with attention (GRUCond) is presented in [27]. One GRUCond layer with attention mechanism consists of three components: two GRU state transition blocks and an attention mechanism in between. The way of combining RNN blocks is different from normal GRU. Readers are referred to [27] for more details.

In this paper, we build two encoder-decoder models: CNN-BiGRU-GRU and CNN-BiGRU-GRUCond to validate the influence of different RNN units on the system's performance. Table 1 shows that CNN-BiGRU-GRUCond has lower character error than CNN-BiGRU-GRUCond, and converges faster as Fig. 2 indicated. As mentioned above, one GRUCond layer can be regarded as two GRU layers, which implicitly deepen and widen the network and thus enable our model to learn more abstract information in a certain extent.

Table 2. *Performance of models trained with different feature extractors.*

Feature extractors	Time (hour/epoch)	Epochs	CER(%)
HOG-BiGRU	**18.4**	6	10.78
CNN-BiGRU	29.3	**2**	**9.06**

Different Feature Extractors for the Input. HOG is the most commonly used feature before the rise of deep learning [4,8], while CNN becomes popular now because of its more powerful feature extraction capability than traditional features.

In this section, we combine HOG and CNN respectively with BiGRU as differnent feature extractors in the encoder: CNN-BiGRU and HOG-BiGRU. Results are presented in Table 2. Compared to HOG-BiGRU, CNN-BiGRU achieves better recognition performance, which agree well with the our intuition. It should also be noted that, though with more training time for each epoch, CNN-BiGRU conveges after 2 epoches, faster than HOG-BiGRU which needs 6 epoches, indicated in Fig. 3. Based on what we described above, we apply CNN-BiGRU as the feature extractor in the following experiments.

Table 3. *Performance of different methods on the 13 TV channels.*

Channels	Images	Bai's	Zhai's	Proposed	Channels	Images	Bai's	Zhai's	Proposed
AHTV	109	18.14	10.47	**10.20**	HuNtv	218	24.53	16.26	**13.10**
BJTV	159	35.69	15.56	**7.99**	JXTV	54	34.86	18.35	**9.54**
CCTV1	134	13.73	10.26	**7.31**	SDTV	109	18.63	12.08	**5.92**
CCTV4	133	14.44	**7.47**	8.74	SZTV	54	24.38	15.30	**8.72**
CQTV	155	20.13	10.07	**5.5**	XJYV	128	29.33	20.56	**13.89**
DFTV	306	14.74	12.23	**9.92**	YNTV	87	25.70	18.09	**9.21**
HeNTV	141	10.66	5.84	**3.58**	ALL	1787	21.99	13.27	**9.06**

Compared to the Baselines. Two typical systems are established as baselines in our experiments. The system proposed by [4] performs ChnITR following the three-step pipeline, i.e., over-segmentation, word classification and beam search, while the system proposed by [4] is a BiLSTM-CTC based segmentation-free method. Compared to [8] whose system can not be optimized at once because of the use of HOG features, our system achieves end-to-end in the real sense.

To quantitatively analyze our proposed pipeline, we test the text images one-by-one using our method and baseline systems respectively. The CERs(%) are listed in Table 3, which are grouped channel-by-channel. Our method gets the best results for ChnITR, where 58.80% and 31.73% relative CER reductions are observed when compared with [4,8].

4.4 Attention Weight Visualization

Attention-based model can automatically learn the alignment information between input and output. For CNN-BiGRU-GRU and CNN-BiGRU-GRUCond, Figs. 4 and 5 provide an intuitive way to inspect this alignment by visualizing the attention weights α_{tj} in Eq. 5 respectively. Each row of a matrix in each plot indicates the weights associated with the feature sequences. From this we could see the source image and destination text align well.

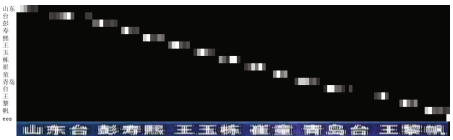

Fig. 4. One alignment exemplar obtained from CNN-BiGRU-GRU.

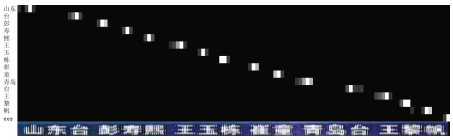

Fig. 5. One alignment exemplar obtained from CNN-BiGRU-GRUCond.

5 Conclusion

Most conventional approaches on ChnITR are based on costly pre-segmentation process, where unsatisfactory segmentation may harm the whole recognition.

This paper proposes an encoder-decoder model with attention mechanism to address this problem. Given an text image, the whole system can be trained end-to-end to maximize the probability of correct recognition. Extensive experiments show that the proposed ChnITR system significantly improve the recognition performance. Furthermore, from the qualitative analysis where we investigate the alignment generated by the proposed system, we conclude that the model can correctly align each target word with the relevant source word in the text image. In the future, we plan to investigate the end-to-end Chinese Image Text Spotting problem by combining text recognition with text detection.

Acknowledgments. This work was supported by National Key Technology R&D Program of China under No. 2015BAH53F02.

References

1. Nagy, G.: Twenty years of document image analysis in PAMI. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 38–62 (2000)
2. Xu, L., Yin, F., Wang, Q.F., Liu, C.L.: An over-segmentation method for single-touching Chinese handwriting with learning-based filtering. Int. J. Doc. Anal. Recogn. (IJDAR) **17**(1), 91–104 (2014)
3. Saidane, Z., Garcia, C., Dugelay, J.L.: The image text recognition graph (iTRG). In: IEEE International Conference on Multimedia and Expo, ICME 2009, pp. 266–269. IEEE (2009)
4. Bai, J., Chen, Z., Feng, B., Xu, B.: Chinese image text recognition on grayscale pixels. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1380–1384. IEEE (2014)
5. Song, Y., Chen, J., Xie, H., Chen, Z., Gao, X., Chen, X.: Robust and parallel Uyghur text localization in complex background images. Machine Vision and Applications, vol. 28, pp. 755–769. Springer, Heidelberg (2017). doi:10.1007/s00138-017-0837-3
6. Fang, S., Xie, H., Chen, Z., Zhu, S., Gu, X., Gao, X.: Detecting Uyghur text in complex background images with convolutional neural network. Multimedia Tools Appl. **76**(13), 15083–15103 (2017)
7. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649. IEEE (2013)
8. Zhai, C., Chen, Z., Li, J., Xu, B.: Chinese image text recognition with BLSTM-CTC: a segmentation-free method. In: Tan, T., Li, X., Chen, X., Zhou, J., Yang, J., Cheng, H. (eds.) Chinese Conference on Pattern Recognition, pp. 525–536. Springer, Singapore (2016). doi:10.1007/978-981-10-3005-5_43
9. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
10. Mishra, A., Alahari, K., Jawahar, C.: An MRF model for binarization of natural scene text. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 11–16. IEEE (2011)
11. Bai, J., Feng, B., Xu, B.: Binarization of natural scene text based on L1-Norm PCA. In: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–4. IEEE (2013)

12. Bai, J., Chen, Z., Feng, B., Xu, B.: Chinese image character recognition using DNN and machine simulated training samples. In: Wermter, S., Weber, C., Duch, W., Honkela, T., Koprinkova-Hristova, P., Magg, S., Palm, G., Villa, A.E.P. (eds.) ICANN 2014. LNCS, vol. 8681, pp. 209–216. Springer, Cham (2014). doi:10.1007/978-3-319-11179-7_27
13. Bai, J., Chen, Z., Feng, B., Xu, B.: Image character recognition using deep convolutional neural network learned from different languages. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 2560–2564. IEEE (2014)
14. Zhong, Z., Jin, L., Feng, Z.: Multi-font printed Chinese character recognition using multi-pooling convolutional neural network. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 96–100. IEEE (2015)
15. Ren, X., Chen, K., Sun, J.: A CNN based scene Chinese text recognition algorithm with synthetic data engine. arXiv preprint arXiv:1604.01891 (2016)
16. Messina, R., Louradour, J.: Segmentation-free handwritten Chinese text recognition with LSTM-RNN. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 171–175. IEEE (2015)
17. Cho, K., Courville, A., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks. IEEE Trans. Multimedia **17**(11), 1875–1886 (2015)
18. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
19. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems, pp. 577–585 (2015)
20. Bluche, T., Louradour, J., Messina, R.: Scan, attend and read: end-to-end handwritten paragraph recognition with MDLSTM attention. arXiv preprint arXiv:1604.03286 (2016)
21. Lee, C.Y., Osindero, S.: Recursive recurrent nets with attention modeling for OCR in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2231–2239 (2016)
22. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4168–4176 (2016)
23. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755 (2014)
24. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
25. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)
26. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
27. Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A.V.M., Mokry, J., et al.: Nematus: a toolkit for neural machine translation. arXiv preprint arXiv:1703.04357 (2017)