Cap2TxT: CAPTCHA Image to Text Sequence, An End-to-End Hybrid Neural Network for Captcha Image Text Sequence Recognition

2020 Spring ML Class Final Project Submission

Changwoon Choi, 2014-17733
Electrical and Computer Engineering, Seoul National University
Seoul 08826, South Korea
zzzmaster@snu.ac.kr

Abstract

*Recent developments of deep neural networks including CNN(Convolutional Neural Network) and RNN(Recurrent Neural Network) made object classification and detection process much more easier. However, many real-world sequence learning tasks require the prediction of sequences of labels from noisy, unsegmented input data. In this paper, I propose **Cap2TxT**, a light-weight end-to-end fashion network for captcha image recognition problem. Code has been available at author's github repository.*[1]

1. Introduction

Recently, in the light of the great success in deep learning networks, computer's pattern recognition capabilities has been remarkably improved. Especially, the pattern recognition for image in the computer vision field started to be in the limelight. However, most of the recent deep neural networks are focusing on solving object classification and semantic/instance segmentation problem. This project deals with the problem of image-based text sequence detection, a problem that has been covered a lot in the field of computer vision in the past but rarely has a deep learning-based approach. In particular, we limit the problem domain to the CAPTCHA image consisting only of English alphabets and arabic numbers. I coined my network ***Cap2TxT*** for Captcha Image to Text Sequence.

Unlike image classification in which deep learning and CNN are most actively used, recognizing objects inherently have the properties of sequence(*e.g*. text information in CAPTCHA, the musical score) has some challenges. Those objects that have the feature of sequence can't be handled as image classification problem since the number of labels or classes are infinity. Thus, it is more natural to predict the order or sequence of labels rather than to give a single label. Also, an important difference between objects with ordering properties is that the length of the sequence can vary significantly.(*e.g*. "Bob" versus "Josephine") Therefore, famous existing successful CNN networks such as AlexNet[5], GoogLeNet[7], ResNet[2] and DenseNet[3] cannot be applied directly to this area without modification. So in this work I leveraged RNN to handle those sequence-based objects. The Cap2TxT framework extracts feature sequences through CNN as the first step. Then, the text sequence is predicted for each time step(The time sequence in the RNN is the horizontal sequence of input image.)through the RNN. Finally the loss function between per-sequence-predicted results and ground truth texts(which are unsegmented at sequence level) is computed by the method inspired by [1]. I will describe the details of network architecture later in this paper in section 3.

1.1. Related Works

In this subsection, I would give a brief overview of the research area related to my work including OCR and deep neural networks.

OCR(Optical Character Recognition) is the automated translation of images of typed, printed or handwritten text into coded text, whether from a scanned document, a photo of a document, from subtitle text overlying on an image.[6] About OCR [4]

Deep Neural Networks. About deep neural network

1.2. Main Contribution

The main contribution of this paper is three-fold. In summary, the contributions are as follows:

- in the light of blablablabla this we proposed the network which can recognize any character ... not only

[1] https://github.com/changwoonchoi/mlfinal

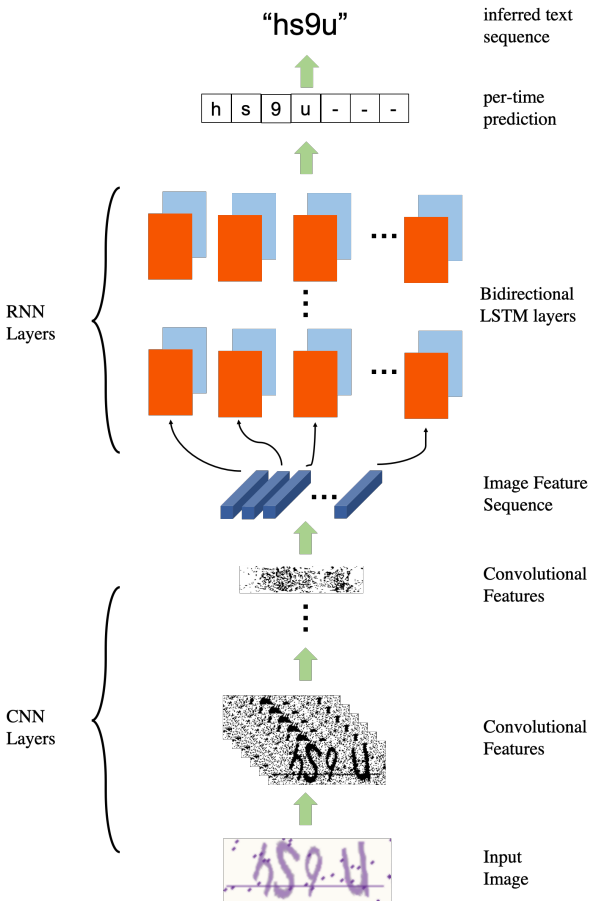

Figure 1: The overview of Cap2Txt architecture. The network is composed of two parts: 1) CNN layers: It takes captcha image as input, extract a feature sequence. 2) RNN Layers:

captcha(English characters and arabic numerals) also any sthsth

- variable length available blabla
- end-to-end blabla
- lexicon-free

The rest of this paper is organized as follows. Section 2 gives a detailed description of the *Cap2TxT* network architecture that I proposed in this final project. In Section 3, I describe the experimental results and methods, and introduce several candidate models for *Cap2TxT* network that have undergone trial and error during the project, and Section 4 concludes.

2. Proposed Network Architecture

The whole *Cap2TxT* network architecture is shown in Figure 1. The network is consisted of three parts.

2.1. Width-Oriented Image Feature Sequence Extraction

2.2. Feature Sequence Predicting

The first part of proposed network is a convolutional layer block. Inspired by [7], as assignment3, this part

2.3. Sequence Squeezing

2.4. Training Methodology

ctcloss [1]

3. Experiments & Discussion

3.1. Experiment Details& Results

3.2. Trials not Adopted as Final Model

I tried some other methods before adopting the final Cap2TxT Network.

4. Conclusion

In this paper, I have proposed an end-to-end deep neural network framework, called *Cap2TxT*, which interprets the captcha image and returns the text sequence in it. Cap2TxT consists of a mixture of

Acknowledgement

Appreciate for all TAs and Professor Yoon. Thanks for interesting, rich class contents.

References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[3] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[4] N. Islam, Z. Islam, and N. Noor. A Survey on Optical Character Recognition System. *arXiv e-prints*, page arXiv:1710.05703, Oct. 2017.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

type	patch size/ stride/ padding	output size (WxHxc)	depth	# 1X1	# 3X3 reduce	# 3X3	# 5X5 reduce	# 5X5	pool proj
convolution	7x7/1/3	160x64x64	1						
ReLU			0						
max pool	3x3/2/1	80x32x64	0						
convolution	1x1/1/0	80x32x64	1						
ReLU			0						
convolution	3x3/1/1	80x32x192	1						
ReLU			0						
max pool	3x3/2/1	40x16x192	0						
inception		40x16x256	2	64	96	128	16	32	32
inception		40x16x480	2	128	128	192	32	96	64
max pool	3x3/2/1	20x8x480	0						
inception		20x8x512	2	192	96	208	16	48	64
inception		20x8x512	2	160	112	224	24	64	64
inception		20x8x512	2	128	128	256	24	64	64
inception		20x8x528	2	112	144	288	32	64	64
inception		20x8x832	2	256	160	320	32	128	128
max pool	3x3/2/1	10x4x832	0						
inception		10x4x832	2	256	160	320	32	128	128
inception		10x4x1024	2	384	192	384	48	128	128
avg pool	4x4/1/0	7x1x1024	0						

Table 1: Network configuration summary of CNN Layer. The first row is input layer, and the last row is final output layer.

Networks	char accuracy	word accuracy
Candidate1	12	13
Candidate2	14	15
Candidate3	14	15
Candidate4	14	15
Candidate5	14	15
Cap2TxT	**88.2%**	**77.9%**

Table 2: Recognition accuracy(%) on test dataset with final Cap2TxT and other candidate network models.

[6] J. Memon, M. Sami, and R. A. Khan. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *arXiv e-prints*, page arXiv:2001.00139, Dec. 2019.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.