Enhancing Text Spotting with Visual Context Information

Ahmed Sabir
TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
asabir@cs.upc.edu

***Abstract*—This research proposal addresses the problem of text spotting – automatically detecting and recognizing text in the wild. In this proposal, we approach this problem by drawing inspiration from the successful development of machine learning algorithms with natural language understanding. In particular, deep learning and neural networks with extra information models. Training an end-to-end text recognition system from scratch is a very expensive process. Therefore, our approach focuses on the integration of external prior information. We propose three approaches: first, a trainable language model integration to a pre-trained deep network. Secondly, a visual context bias information model, to exploit the semantic relatedness between word and image context. Thirdly, a combined model of multiple prior knowledge in a fusion based deep model. We also conduct experiments with public datasets, and compare our results to current state-of-the-art benchmarks.**

I. Short research plan

A. Introduction

Research in areas such as text recognition in the wild has not yet reached a mature level in implementation. There are still many challenges due to the many possible variations in textures, backgrounds, fonts, and lighting conditions that are present in such images. Many scene understanding methods successfully recognize objects and regions like roads, trees, sky in the image, but tend to ignore the text on the sign boards. Our goal is to fill this gap in understanding text in the scene. In addition, the automatic detection and recognition of text in natural images, *text spotting*, is an important challenge for visual understanding. There are two stages in text spotting: word detection and word recognition. The detection stage aims to generate the bounding box around a candidate text image. The candidate words in the bounding boxes are recognized in the text recognition stage. In this proposal, we focus on improving the text recognition stage.

B. Related Work

The idea of end-to-end text recognition refers to algorithms able to automatically detect and recognize text in images. A text spotting system should be able to read text images that humans can read. The first text spotting system for image recognition which did not require a list or dictionary was proposed by [1]. The system extracted the character candidates via maximally stable extremal regions (MSER) and eliminated non-text candidates through a trained classifier. The remaining candidates were fed into a character recognition module, which was trained using a large amount of synthetic data. Later, [2] presented a new method for text spotting that combined the advantages of sliding windows and component based methods. In deep learning approach, [3] used a convolutional neural network (CNN) with unsupervised pre-training for text detection and character recognition. PhotoOCR is a text spotting system able to read characters in uncontrolled conditions [4]. It is a deep neural network (DNN) model running on HOG feature instead of image pixels. Another deep learning approach, [5] proposed a new CNN architecture, which allows feature sharing. Convolutional Recurrent Neural Networks (CRNN) presented by [6], a novel neural network architecture that integrates both CNN and RNN for image based sequence recognition. Deep learning based methods have two drawbacks. First, they require a huge dataset to train. Secondly, the computational burden of these methods is extremely high. A hybrid approach between deep learning and classical statistical model opens the possibility to overcome these limitations, and lead to introducing simpler models with better results. In this proposal, we investigate both approaches.

C. Methodology

In this work, we propose a simple approaches for introducing prior knowledge to the text spotting system. Additional knowledge, such as language model and visual context information, are essential to understand text in the scene. Our approaches to tackle these problems will be:

The first approach is an independent N-gram language model. The N-grams compute the probability of the candidate word directly from frequency counts taken from a large corpus. For instance, the word 'on' has higher probability than 'Onn'. The second approach consists of integrating a visual context bias. Contextual information is important to understand text in the scene. For instance, we can use word-embeddings to calculate the similarity distance between a candidate word and its visual context from the image. For example, if the output from a visual context model is 'street', 'car', or 'traffic lights', semantically related words such as 'parking' or 'left turn' are more likely to appear. Finally, we combine both visual context bias and language model information into a single model. The combination of both models should produce more informed predictions, that take into account both word context and linguistic probabilities.

D. Language Model

The language model is based on a N-gram model. N-grams are essential in any task in which we have to identify

words from a noisy and ambiguous input. In speech recognition for instance, the input signal is noisy and most of the words extremely similar.

In this work, we apply a unigram language model over a deep network with pre-defined dictionary [7]. In particular, we extract the most probable words from a *baseline CNN* [7] and pass it into our language model. The outputs from the baseline are the candidate words and the associated probability based on softmax score $P_0(w) = p(w|CNN)$. The unigram language model (ULM) that was trained on a huge English corpus acts as a second independent dictionary. The main purpose of the unigram language model is to increases the probability of most common words in the pre-defined dictionary of the baseline, reranking the most probable words according to this modified probability $P_1(w) = p(w|CNN) \times p(w|ULM)$ from the language model. This hybrid approach opens the possibility of introducing higher-order trainable language models.

We trained two models on different corpora. The first model was trained on *Opensubtitles*[1], a database based on subtitles for movies. The corpus contains around 3 million words (combination of words and digits). Secondly, we trained a bigger model with *google book n-gram*[2] corpus, that contains around 5 million (only words). The overall results show that the language models have improved the baseline accuracy by 2% on both ICDAR03, ICDAR13 and 2.6% on COCO-Text dataset.

E. Visual Context Bias Information

The relation between text and its surrounding environment is very important to understand text in the scene. We propose to integrate visual context bias knowledge to the text spotting pipeline. In particular, we exploit the semantic relatedness between the spotted text and its image context. For example, the word 'parking' is more semantically related to word 'car' than it is to word 'banana' or 'pencil', thus it will be more likely to appear in a visual context where cars are present rather than where fruits or office appliances do.

We use a pre-trained state-of-the-art object classifier [8], [9] to find objects in the image, and we compute the semantic similarity between the candidate word and the detected elements. We use a word embedding approach to estimate the semantic similarity between the spotted word and its visual context information (VCI) and compute $p(w|VCI) = f(similarity(w, context\_objects))$. As in the language model step, this probability can be used to alter the total word probability and re-rank the best words suggested by the baseline CNN: $P_2(w) = p(w|CNN) \times p(w|ULM) \times p(w|VCI)$. We evaluate the performance of the model on the ICDAR 2017 robust reading challenge COCO-Text dataset [10]. The results show that by understanding the semantic relatedness between the spotted text and its visual context, the model improved the baseline by 4%.

F. Fusion Model

A deep model fusion based architectures have been successful in video action recognition [11], which is a joint classification between two deep networks. Our approach uses a similar architecture, to train two streams combined models, multiple prior knowledge and text recognition models. There are different approaches to combine two deep models, early fusion, late fusion and slow fusion, which we plan to explore.

G. Future Work

There are other lines of research that can be approached with this architecture setting:

1) Using higher order language model, such as bi-gram and tri-gram.
2) Re-train from scratch the combined models, text recognition and visual context bias model on the same dataset.
3) Integrate LM with long short term memory (LSTM) models for better word prediction.

[1] *https://www.opensubtitles.org*
[2] *https://books.google.com/ngrams*

ACKNOWLEDGEMENTS

I would like to thanks my supervisors Lluís Padró, Francesc Moreno-Noguer for guidance and Ernest Valveny for fruitful discussions.

REFERENCES

[1] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Asian Conference on Computer Vision*. Springer, 2010, pp. 770–783.

[2] ——, "Scene text localization and recognition with oriented stroke detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 97–104.

[3] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, "Text detection and character recognition in scene images with unsupervised feature learning," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 440–445.

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

[5] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting." in *ECCV (4)*, 2014, pp. 512–528.

[6] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[7] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[10] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.