CS6101-15

CNN-RNN for Image Annotation with Label Correlations

Xiao Junbin, Hu Zikun, Lim Joo Gek, Tan Kay Pong

Abstract

Convolutional Neural Networks (CNNs) have shown great success in image recognition where one image belongs to only one category (label), though in multi-label prediction, their performances are suboptimal mainly for their neglect of the label correlations. Recurrent Neural Networks (RNNs) are superior in capturing label relationships, such as label dependency and semantic redundancy. Hence, in this project we implemented a CNN-RNN framework for multi-label image annotation by exploiting CNN's capability for image-to-label recognition and RNN's complement in label-to-label inference. We experimented on the popular benchmark IAPRTC12 to show that CNN-RNN can help improve the performance on the CNN baseline.

1. Introduction

Multi-label image annotation is to predict the co-presence of certain objects of interest in images[1]. It is of great importance in facilitating image management. Yet, it is also challenging because of the numerous images of diverse content shared on the Internet. In this project, we applied CNN to perform multi-label image annotation. Meanwhile, we found some objects often appear together, for instance, "sky" and "cloud", "waterfall" and "cliff", "car" and "street"/"road" (Fig.1). This motivated us that some challenging objects (e.g., small size, heavy deformation and occlusion) or rare objects, though cannot be predicted from the image directly, can still possibly be recalled, due to their co-occurrence relationship with other objects which are much easier to be predicted. Thus, we further employed RNN to complement CNN, in explicitly learning the relationship between labels.

Figure 1. Multi-label image annotation examples.

2. Methodology

As is shown in Fig.2, our method comprises mainly two components:

- The CNN part for image-to-label prediction. We designed a light CNN similar to [3] , and trained it from scratch by using binary cross-entropy loss.
- The RNN part for label-to-label inference. Specifically, we adopted the Bi-directional LSTM to model relationships between labels. The label order was arranged according to [1], i.e., rare label comes first.

For testing, we fed the image to CNN to obtain our first prediction result. Meanwhile, we extracted image feature from the penultimate layer, and fed it to RNN together with the predicted word at each time step ("start" for the initial step) to get the RNN predictions. The final result is obtained by max-pooling the predicted distribution in CNN and the Softmax representation at each step in RNN. We used beam size 2 in searching the most likely words at each time step.

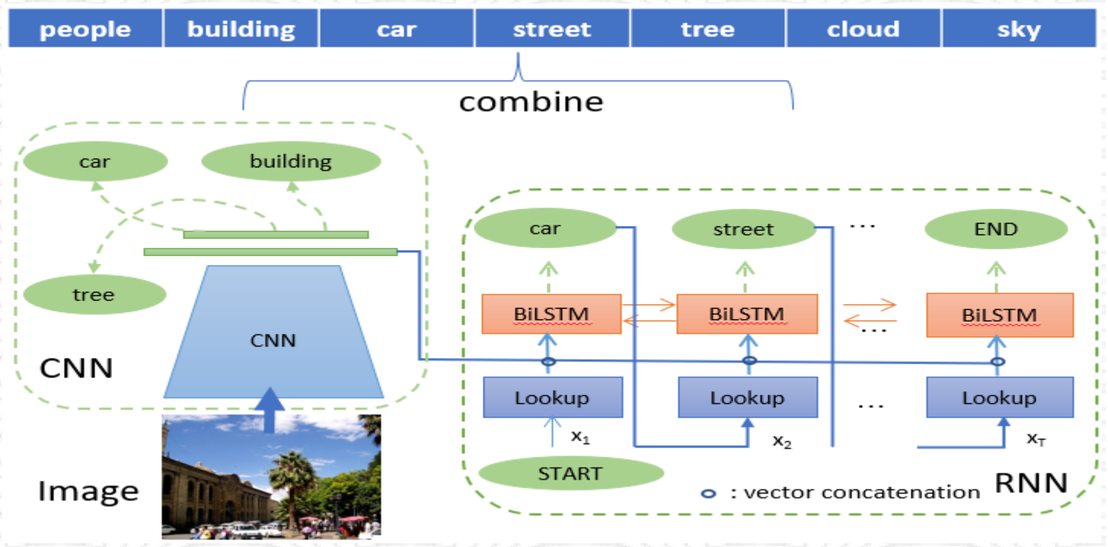

Figure 2. CNN-RNN framework for multi-label image annotation.

3. Experiment

3.1 Dataset

We evaluated our method on the popular multi-label annotation dataset IAPRTC12[4]. We summarized the dataset information in Table 1.

Table 1. IAPRTC12 dataset

Total labels	Labels per image	Images per label	Training images	Testing images
291	5/23 (avg/max)	363/4553	17664	1960

3.2 Result

We followed established works[1,2] to report macro-precision (AP), recall (AR), and F1 score (AF1) and also their micro versions (IP, IR, IF1) to evaluate multi-label performance.

Table 2. Experimental results

Method	AP	AR	AF1	IP	IR	IF1
CNN	0.2506	0.2689	0.2322	0.3222	**0.4888**	0.3884
RNN	**0.2761**	0.2518	0.2516	**0.3438**	0.3395	0.3416
CNN -RNN	0.2246	**0.315**	**0.2534**	0.3283	0.4868	**0.3921**

The result has proven that CNN-RNN framework is effective for multi-label image annotation. The explicitly learned relation information does improve the final prediction results according to F1 score. Besides, We also found that RNN achieved a relatively higher precision than CNN. This could be due to the capability of RNN in self-determining the annotation length. Our result is not as comparable to that of [1], because our CNN is much smaller, and we did not use pre-trained model. Generally, the lower evaluation performance may somewhat attribute to the subjective annotations, e.g., the model predicted the labels "sky" and "cloud" correctly, but they may not appear in ground-truth annotations.

References

1. Wang J et al. (2016). CNN-RNN: A unified framework for multi-label image classification. CVPR, 2285-2294.
2. Jin J., & Nakayama H. (2016). Annotation order matters: Recurrent image annotator for arbitrary length image tagging. ICPR, 2452-2457.
3. Yang F. et al. (2018). Learning to compare: Relation network for few-shot learning. CVPR.
4. Grubinger et al. (2006). The IAPR Benchmark: A New Evaluation Resource for Visual Information Systems. ICLRE.

NUS National University of Singapore · INFOCOMM MEDIA DEVELOPMENT AUTHORITY · sea connecting the dots · indeed · SGINNOVATE · WING Web Information Retrieval / Natural Language Processing Group