A Neural Compositional Paradigm for Image Captioning

Bo Dai
CUHK-SenseTime Joint Lab,
The Chinese University of Hong Kong
bdai@ie.cuhk.edu.h

Sanja Fidler
University of Toronto
fidler@cs.toronto.edu

Dahua Lin
Vector Institute
NVIDIA
dhlin@ie.cuhk.edu.hk

***Abstract*—The general captioning models uses sequential structure to generate captions often creating errors in semantics or lack the detail in the image. The method discussed preserves the semantic content through an explicit factorization of semantics and syntax. Added to all these the method discussed requires less data to train and gives us captions that are more diversified in nature.**

I. Introduction

Image Captioning in the recent years has gained attention and the task we try to accomplish by this is to generate small sentences which state the objects and activities in the image. Basically we describe the image in short sentences. Previous works in this field used the words to be produced in *sequential order* i.e. the probability of each word being selected is proportional to the previous word and the feature of the image.

These issues are dealt with by using a new paradigm in which the extraction of semantics which basically suggest what to say and the construction of captions that are syntactically correct by dividing it into two stages. A *explicit* representation of the semantic content of the image and then through *recursive composition* is applied until a full caption is formed.

This has two main advantages in reference to the previous works.

1. The captions generated by this method are easy to understand to work upon.

2. Through this the hierarchical dependencies among words and phrases are captured.

3. Works when text data is small and generalizes better to new data.

II. Compositional Captioning

The natural language is of hierarchical structure where sentences are parsed in form of trees. The first thing that we do when we get an image is get the noun phrases and then create captions using recursive compositional procedure. This takes into account the non sequential dependencies between words and phrases.

A. Explicit Representation of Semantics

In the framework we use we identity the image by a cluster of noun phrases such as ”a clear sky”, ”a white board” and ”a yellow hat”. These help in acheiving the attribute as well as the object. During research it was found that distinct noun phrases are smaller in number than to the images. So this problem can be considered as a multilabel classification problem.

Fig. 1. Image Captioning

To be specific we get the list of all the distinct noun phrases from the training dataset by parsing the captions and then we treat each of them as a label. From an image we take out the visual features and further give them as an input to two fully connected layers. We take only the noun phrases which have the greatest score among the similar noun phrases.

B. Recursive Composition of Captions

Let us consider the list of all the noun phrases as the phrase pool. Now from this phrase pool each ordered pair are connected using a connecting module such that the two phrases are connected in a plausible way. Now a score is assigned to all the generated large phrase.

Subsequently an Evaluation Module is also used to know if the generated phrase is a complete caption. If the new generated phrase is the complete caption we give it as the output else remove the ordered pair which were required to generate this large caption from the phrase pool and then insert this newly generated phrase in the phrase pool. This continues until a complete caption is generated or the phrase pool contains only one element.

Extension
The framework suggested here can extended to take into account the user preference and other conditions as operations can be controlled and interpreted. Such as one can open can change results by filtering the noun phrases generated in the beginning and then changing the scores.

III. Results

Two different models were trained in each comparison. First for the connecting module and then the evaluating module. This model was compared to other state of the art models such as AdapAtt, Neural Image Captioner(NIC) and TopDown. SPICE metric was used to know about the score of the generated captions. Two different data sets were used to train and test the models namely the Flickr30k and COCO.

Generalization Analysis.
When the models were trained on one and then tested on another it was seen that Compositional Captioning did well in generalization across datasets. It is good at handling out of the domain semantic content and requires much less data then the other three models to learn.

Diversity Analysis.
Two different metrics were used to calculate the diversity of the captions generated. First was to compare diversities in captions generated for different images and second was to compare the captions diverse for a single image. Compositional Captioning did well in all metrics which suggest that the captions generated were diverse.

IV. Conclusion

A novel paradigm has been suggested in this paper for image captioning. It is different form the traditional methods of other models which encode images using feature vectors. This methods uses explicit representation of the input image to generate the noun phrases and this noun phrases are connected to form a caption. The recursive compositional procedure gives it a hierarchical structure similar to that of the human language.

Learning a Deep ConvNet for Multi-label Classification with Partial Labels

***Abstract*—Previous work in the field of image classification has been successful for single label but not for multi-label. This is due to the fact that the input image and output label spaces are more complex than for single label. Annotations are also not known for multi label classification. Thus to handle this aspect of the problem we collect images with partial annotation. Three data sets are used namely MS COCO, NUS-WIDE, and OpenImage.**

I. Introduction

Multi-label classification brings new challenges as previous classifiers used for single label classification cannot be fine-tuned and are not scalable. These challenges are solved using a new loss function and a new method is introduced to fix missing labels.

The first problem addressed is solved by comparing several labeling strategies for multi label datasets. Given experiments show that partially annotating all images is better than annotating small set of images. So given limited budget we would be annotating partially.

The second problem is countered by introducing a new loss function than generalizes the standard binary cross entropy by exploiting label propotion information. Convolutional Neural Networks are highly sensitive to noise and thus a curriculum based learing method is proposed that predicts labels as we progress and then add them to the training set.

Graph based neural network is also introduced as to model the correlation between different labels. As in multi label classification not all the labels are independent hence relation between different labels is important.

II. Learning with Partial Labels

The main goal in this paper is to train convolutional neural network for partial labels. We introduce a new loss function to generalize the binary cross-entropy. Then a graph neural network is also introduced to find the correlations between different labels. Finally this is clubbed together with curriculum based learning to predict labels.

A. Binary cross-entropy for partial labels

Partial binary cross-entropy is used in place of the old binary cross-entropy as normalization causes the back-propagation gradient to be small. Partial binary cross-entropy ignores the categories for unknown labels. The partial binary cross entropy gives the same example to each image just like given in the standard binary cross-entropy. This loss has a behavior of adapting to the proportion of labels.

B. Multi-label classification with GNN

For multi label classification we use a GNN that is a Graphical neural network where each node represents a label and the edge between two nodes show a correlation. A fully-connected graph is used. The message update function used is a multi layered perceptron. The message is first computed by feeding the hidden stated to the perceptron and then taking mean.

C. Prediction of unknown labels

When the model first starts learning it trains on the clean partial labels. Then the algorithm uses this trained model to create new labels which are ease which signifies noise in the labels. Then as we progress we use the combination of these two types of labels to train our model. Different strategies such as Score Threshold strategy and Bayesian Uncertainty strategy is used to add new labels

III. Results

The results are found after training the model on many standard multi label datasets. These datasets are fully labeled and so we drop some labels in random from the datasets to make partial label. In extreme case we also trained to model with only 0.9% of all labels.

The effect of GNN and partial Binary cross-entropy is seen when the data is trained and tested on the MS COCO dataset. Both help the in increasing the performance by a significant amount compared to other standard methods. These experiments show that modeling the correlation between labels is important.

Moreover when an ablation study is done on the MS COCO dataset and with only 10% of known labels it was found that when fine tunning when clubbed with GNN, partial BCE and relabeling gives better results on all 4 metrics namely MAP, 0-1 exact match, Micro-FI and Macro-FI.

IV. Conclusion

A scalable end to end approach is shown in this work for multi label classification of images. The experiments shows that when partial binary cross entropy which is our new loss function significantly improves performance. Bayseian Uncertainity model is the best of all to label the missing labels. The overall recall of our model is high as compared to other models. So this method is a novel method for multi class classification of images.