1. Introduction

In this project, we will design a image retrieval model that would identify an image from a set of given images by a natural language specification. The natural language specification is the ground truth caption labeled by human for a target image. Some examples are shown in figure 1 (Vinyals et al., 2016).

It is worth to notice that the caption is not specifically tagged to distinguish the target image with other images. The image identification task is illustrated in the Figure 3.

2. Our Approach

In our model, the major approach to identify the target image is by learning and measuring a similarity between the given natural language specification and the representation of each image.

2.1. The Baseline Model

One possible way is that a similarity can be obtained by taking cosine similarity between the average of word embeddings (Pennington et al., 2014) of the input query (the natural language specification) and a generated caption for each image from a caption model. We consider this approach as our baseline. The specific image captioning we would use is from (Vinyals et al., 2016).

2.2. The Proposed Model

In this section, we would propose a new model which compute the similarity directly from visual features from an image and language features extracted from an input query. In the final step, we would generate a saliency map (visual attention map) for the target image and the natural language specification by using a single back-propagation. This saliency map could provide an interpretable explanations of the model.

The visual representation of images are computed by Convolutional Neural Networks which are widely used.
for image tasks and is currently state-of-the-art for object recognition and detection. (Vinyals et al., 2016; Simonyan & Zisserman, 2014; He et al., 2015)

The similarity model would also consist of Recurrent Neural Networks, which is widely used in natural language understanding tasks like machine translation and captioning systems (Vinyals et al., 2016; Devlin et al., 2015; Wu et al., 2016). The specific RNN architecture we would use for image-caption similarity task is the Long Short Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997).

In this work, we would use the output from the deep CNN as the initial state for RNN, the language model.

2.3. Finding Saliency Map

We try to find visual grounding of the classifier with a saliency map. Given a natural language specification input, our goal is to find the most salient regions in an image. Zhang et al. (Zhang et al., 2016) proposed a method to find a saliency map with an excitation back-prop. We extends their works for our purposes.

Figure 3. An illustration of saliency maps given a word. (Zhang et al., 2016)

3. Experiment

3.1. Dataset

The data set we would use is the MSCOCO (Lin et al., 2014) dataset with caption annotation.

We would also use the Flickr30K dataset (Plummer et al., 2015). This dataset contains 31,783 images on people and animals, and 158,915 English captions (5 captions for each image). It also has bounding boxes for objects in an image, as shown in Figure 4.

Figure 4. An example of an image in Flickr30K. (Plummer et al., 2015)

4. Evaluation Metrics

We use the test split of Flickr30K dataset to evaluate our performances. We generate positive pairs with ground-truth captions and the corresponding image and negative pairs with an image and irrelevant captions from other images. We measure Rank-1,3, and 5 accuracy for our image identification task.

References


Lin, Tsung-Yi, Maire, Michael, Belongie, Serge J., Bourdev, Lubomir D., Girshick, Ross B., Hays,


