Paper: *Deep Visual-Semantic Alignments for Generating Image Descriptions*

Nirajan Koirala

CSC 9010_001

Summary

Authors in this paper use a combination of two techniques: a neural network that is more adaptive for classification of images, CNN and a neural network which is suited for natural language processing, RNN to produce image descriptions. The multimodal RNN architecture uses inferred alignments to learn to generate novel descriptions of image regions. Their method is experimented on datasets like Flickr8K, Flickr30K and MSCOCO. The descriptions produced by their model is able to significantly outperform retrieval baselines on both full images and on a new dataset of region-level annotations. First, they develop a deep neural network model that infers the latent alignment between segments of sentences and regions of the image that they describe. The effectiveness of this approach is validated on image-sentences retrieval experiments. They then introduce a mulitmodal RNN architecture that takes an input image and generates its description in text. The generated sentences are able to significantly outperform the retrieval baselines and produce sensible qualitative predictions. This model is later trained on the inferred correspondences and its performance evaluated on a new dataset of region-level annotations.

In section 3, they put forth a model which first aligns sentence snippets to the visual regions that they describe through a multimodel embedding. Afterwards, these correspondences are treated as training data for a second, multimodal RNN model that learns to generate the snippets. Learning to align visual and language data is described in section 3.1 where they build the model on the approach of Karpathy et al., which learns to ground dependency tree relations to image regions with a ranking objective. Their main contribution is in the use of bidirectional RNN to compute word representations in a sentence, dispensing of the need to compute dependency trees. For representing images, they use h-dimensional vectors where h is the size of the multimodal embedding space. For representing sentences, their bidirectional RNN takes a sequence of N words and transforms them into an h-dimensional vector. Additional details about denoting the position of the word in the sentence is provided later in the sub-sections. In section 3.2, they describe the multimodal RNN for generating descriptions.

For the training of the RNN model, it first learns to combine a word and the previous context to predict the next word. The cost function is to maximize the log probability assigned to the target labels. At test time, image representation and distribution over the first word is computed. A word is sampled from the distribution over the first word, its embedding set as x2 and this process repeated until the END token is encountered. For the optimization, they use SGD with mini-batches of 100 image-sentence pairs and momentum of 0.9 to optimize the alignment model. They also use dropout to reduce overfitting in all layers except the recurrent layers. Section 4 describes the experiments they performed with their models using 3 different datasets. They consider a withheld set of images and sentences and retrieve items in one modality given a query from other by sorting based on the image-sentence score S_{kl}. Their full model is able to outperform previous works and the simpler cost function that they use

also improves the performance. They also found that the bidirectional RNN is able to achieve higher performance than dependency tree relations. The model is able to discover interpretable visual-semantic correspondences even for relatively rare and small objects which are most likely to missed by the models which perform only using images. Their model is also able to modulate the magnitude of the region and word embeddings. It is observed that representations of visually discriminating words like "kayaking, pumpkins" have embedding vectors with higher magnitudes which have a higher influence on image-sentences score. Conversely, pause words like "now, simply, actually, but" are mapped near the origin.

In section 4.2, evaluation of the ability of RNN model to describe images and regions is done using BLEU, METEOR and CIDEr scores. While comparing to the nearest neighbor retrieval baseline, the multimodal RNN is able to achieve significantly higher scores. Also, when compared to other related works, they found that their model prioritizes simplicity and speed at a small cost in performance. Section 4.4 lists the limitations of their model. In particular, their model can only generate description of one input array of pixels at a fixed resolution. The RNN also receive the image information only through bias interactions which are less expressive than multiplicative interactions. Also, the model is not end-to-end and consists of two separate models which adds complexity.

Strengths

- The paper sheds light on the techniques of making two different neural network work together.
- Experiments are thorough and also descriptions of the method they use is very intuitive.

Weaknesses

- I wish the limitations section was a little more detailed and descriptive.

Points of Confusion

- Section 4.3 was a little confusing.

Discussion Questions

- How dependency trees and its relations work?
- How does Markov Random Field helps in creating better alignments of the words to create better descriptions?
- How would an end-to-end model for this task look like?