Jenish Maharjan **Week 11**

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

Summary:

Karpathy and Fei-Fei introduce a model that generates descriptions of images and their regions. Their approach involves learning about the inter-modal correspondences between the words and the image data based on datasets of images and sentences describing them. They combine Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences and a structured objective to align the two modalities through a multimodal embedding. In their paper, they describe their deep neural network model that learns the latent alignment between regions in the images and segments of sentences. They also introduce a multimodal Recurrent Neural Network which takes as input an image and generates textual descriptions based on the sentence segments and image region correspondence learned by the previous model.

In order to achieve their objective of generating textual descriptions from images, they start with a alignment model that is responsible for learning alignment between image regions and sentence segments. They use a RCNN (Region Convolutional Neural Network) pre-trained on ImageNet and Finetune on the 200 classes of ImageNet Detection Challenge in order to detect objects in every image. The top 19 detected locations in addition to the whole image are used to compute the representation based on the pixels inside each bounding box – every image is represented as a set of h-dimensional vectors. In order to represent sentences, they use a bidirectional RNN which takes a sequence of N-words and transforms each one into an h-dimensional vector while representation of each word is enriched by a variably-sized context around that word. The alignment between the image and the sentence segments is learned by formulating image-sentence score as a function of the individual region-word scores. A sentence-image pair should have a high matching score if the words have a confident support in the image.

In the second part of the study, they design a model that can predict a variable-sized sequence of outputs given an image. They implement a simple extension to RNNs used for text generation that focuses the generative process on the input image content. This multimodal RNN takes the image pixels and a sequence of input vectors during training and computes a sequence of hidden states and a sequence of outputs. The RNN is trained to combine a word to the previous context to predict the next word. During testing, they found that beam search could improved results. To optimize their approach, they used mini-batch Stochastic Gradient Descent and cross-validated learning rate and the weight decay. The generative RNN achieved the best results using RMSprop as the loss function.

They used the Flickr8K, Flickr30K and MSCOCO datasets in their experiments to train and evaluate their models. Each of these datasets contains images each of which is annotated using Amazon Mechanical Turk. All sentences were converted to lowercase and all non-alphanumeric characters were discarded. Redundant words were also filtered out from the dataset. The authors evaluated the quality of the inferred text and image alignments with ranking experiments using an image-sentence score. The evaluation of their full model shows that it outperforms pervious work which used a single image representation and an RNN over the sentence. They found that their simpler cost function improved the performance. Their full model also outperformed dependency tree relations. Since the dependency relations were previously shown to work better than singe words and bigrams, they note that their model takes advantage of contexts longer than two words. For the MSCOCO data, since other published ranking results were not available, they reported results on a subset of 1000 images and the full set of 5000 test images for future comparisons. Qualitatively, their model discovers visual-semantics correspondences that can be interpreted even for small or relatively rare objects. The authors also evaluate the ability of their RNN model to describe images and regions using BLEU, METEOR and CIDEr scores computed with coco-caption code. They note that the region model outperforms full frame model and the baseline. The authors note that although the results are "encouraging", their model is subject to limitations such as the ability of the model to only generate a description of one input array of pixels at a fixed resolution. The use of two models in their approach is also noted as a limitation.

Strengths:

- The content of the paper is well organized. Important points or the gist of many sections have been presented in bold style which is a great help.
- Each of the components of the model are described really well in detail.
- They reported results on a subset of 1000 images and the full set of 5000 test images so that they could be used for future comparisons, which is a great contribution.

Weaknesses:

- They have a limitations section which could have been elaborated.

Confusions:

- What does "learning to modulate the magnitude of the region and word embedding" mean?
- When using the region level model, are they looking for descriptions for each region only or do they combine the descriptions from each region for an image?

Discussion Questions:

- What are dependency tree relations?
- What could be other sources of data for the same task?
- What do they mean when they say "A more sensible approach might be to use multiple saccades around the image to identify all entities, their mutual interactions and wider context before generating a description"?