Research Review of the Paper:

Andrej Karpathy *Li Fei-Fei*
Department of Computer Science, Stanford University
{karpathy,feifeili}@cs.stanford.edu

Deep Visual-Semantic Alignments for Generating Image Descriptions

Submitted by: Vadym Serpak, v.serpak@gmail.com

Course: Artificial Intelligence Nanodegree, Udacity

Date: 06/22/2017

1. Introduction

The paper introduces a model that generates natural language descriptions of images and their regions. Described approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data.

1.1. Alignment Model

The Alignment Model was presented first. It is based on a novel combination of **Convolutional Neural Networks** over image regions, bidirectional **Recurrent Neural Networks** over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. Then these correspondences had been treated as training data for a second, multimodal **RNN** model that learns to generate the snippets. The way is the following:

- Learning to align visual and language data
  - Representing images
  - Representing sentences
  - Alignment
  - Decoding text segment alignments to images

1.2. Generation Model

The **Multimodal RNN architecture** that uses the inferred alignments to learn to generate novel descriptions of image regions had been described. This model is going to learn from that dataset in order to generate descriptions given an image. The model takes in an image and feeds it through a **CNN**. The **softmax** layer is disregarded as the outputs of the fully connected layer become the inputs to another **RNN**. A typical size of hidden layer of the **RNN** is 512 neurons. The **RNN** is trained to combine a word and the previous context to predict the next word.

1.3. Optimization

It was used:

- **SGD** with mini-batches of 100 image-sentence pairs and momentum of 0.9 to optimize the alignment model
- **Cross-validation** for the learning rate and the weight decay
- **Dropout** regularization in all layers except in the recurrent layers and clip gradients **elementwise** at 5 (important)

1.4. Experiments

Datasets: It was used Flickr8K , Flickr30K and MSCOCO datasets with 8,000, 31,000 and 123,000 images respectively and each is annotated with 5 sentences using Amazon Mechanical Turk. For Flickr8K and Flickr30K, it was used 1,000 images for validation and for testing, the rest for training. For MSCOCO it was used 5,000 images for both validation and testing.

Data Preprocessing: All sentences had been converted to lowercase, discard non-alphanumeric characters. The words had been filtered to those that occur at least 5 times in the training set. Next steps:

- Image-Sentence Alignment Evaluation
- Generated Descriptions: Fulframe evaluation
- Generated Descriptions: Region evaluation
- Limitation

2. Conclusions

- Introduced a model that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences.
- The approach features a novel ranking model that aligned parts of visual and language modalities through a common, multimodal embedding.
- Described a **Multimodal RNN architecture** that generates descriptions of visual data.
- Evaluated its performance on both **fullframe** and **region-level** experiments - in both cases it outperforms retrieval baselines.