NLP Project Proposal - Visual Relationship Detection

Makkunda Sharma 2015CS50459
Madhur Singhal 2015CS10235

March 20, 2018

1 Hypothesis and Goal

We hypothesize that images contain a huge deal of semantic information about the world and that much of this information can be expressed in terms of "relationships" or "interactions" between objects. We also observe that images often co-occur with natural language in the wild and an analysis (like a dependency parse) of sentences co-occuring with an image gives a lot of semantic information about the contents of the image. Thus our goal will be to train a model, which given an input image produces various 'relationships' present in the image. We view these relationships as 3-tuples containing two object names (or synsets) and one relationship identifier along with bounding boxes corresponding to the two objects in the image. Some examples are (man,hugs,woman), (child,rides,elephant) and (branch, on,tree). We will also explore zero-shot learning of never before seen relationships using zero shot learning and relationship embeddings.

2 Datasets

We identified two datasets suitable for our purposes. First is the Flickr30k dataset which contains captions 158k captions for 30k images along with bounding boxes for the entities mentioned in the sentences. Second is the Visual Genome dataset which contains 100k images with 2.3 million relationships which are of the form we require. This dataset also contains 5.4 million captions of various regions of images which can be useful.

3 Literature/Code Search

We were inspired by the paper "Visual Relationship Detection with Language Priors" which explores this idea. A more recent paper from last year "Phrase Localization and Visual Relationship Detection with comprehensive Image-Language Cues" uses cues like sentence part of speech to better learn relationships. Some code is available for both of these papers though we have not fully explored their capabilites.

4 Evaluation

We can evaluate our model based upon the number of relationships present in test images that it recognizes correctly. Mean Average Precision and Recall are metrics commonly utilized. Further we can use sentence creation from relationships and use captioning metrics for evaluation too.

5 Demo

We can make a demo app which takes an image and gives the various relationships by highlighting the objects with bounding boxes and showing the most probable relationship keywords.