CS 766 Computer Vision - Project Proposal
Image Summarization
Hanna Strohm (hstrohm), Nathan White (ntwhite), Joseph Mohr (jvmohr)
February 14th, 2020

Overview
Verbally describing images and their contents is an easy task for humans, but not necessarily so for computers. Both object detection/recognition and the relationships of different objects within a scene can be complex when it comes to image processing. Machine Learning has provided a viable solution to the first portion of this task, but the second portion remains difficult. We want to design an algorithmic approach that will be able to summarize the contents of a given image. This includes recognizing and describing common objects and their relation to other objects within the scene. After this, details regarding the objects of focus will also be added to the description. These two points will be the basis for the scene summarization with the main focus being the objects and how they relate to others within the scene, while extraneous details such as color and size can be added later. One example of this is an image of a human on their cell phone. While it's easy to detect both objects individually, we want to be able to correlate the human with the phone to mean that "the human is using/holding/looking at the phone." Details such as the color of the phone or the person's features can be added later.

Project Impact
The solution to this problem would automate the process of detecting objects, using that information to interpret the relationships between those objects, and finally translate that information into audio. There are a couple of reasons to solve this problem. First, it could provide access to images for people who are visually impaired or blind. The ability for a computer to analyze and describe an image accurately and in a way that is helpful would have the possibility of improving their daily experience navigating the world of images that exists today. Another useful application of image summarization would be to be able to include descriptions of images in an audio book, paper, or other text that is being listened to rather than read. Images can contain information important to a text, and therefore it would be useful to be able to include images in a text-to-speech algorithm.

Current State of the Art
Several studies have been conducted in regards to spatial based object association [1,2]. Falomir et al. [1] explored explaining images through an object's visual characteristics as well as its spatial characteristics. All descriptions are qualitative, and not sentences meant for simply describing the image to the end-user. They also utilized a very small dataset, consisting only of a hallway from their building and the objects within it. Elliot et al. [2] explored the use of a visual dependency representation (VDR) model that looks at recognized object positions in order to determine how sentence structure should be formulated. For their study, they used a pre-trained R-CNN that would detect the objects for them, and utilized the detected objects and their image properties to describe the image's content.

Another approach to this program has utilized a dependency tree recurrent neural network (DT-RNN) in order to associate text with an image [3]. The authors' approach was to be able to map sentences to an image visualization of the text, and vice versa. They did not attempt to generate sentences for the given image, but rather used the sentences provided from their dataset. Lastly, even similarity based search has been explored [4]. Gathering images from Flickr, with their user specified captions, the authors simply search the dataset and attempt to find the image that matches the query the most. Once found, it takes the associated user-specified caption for the dataset image and assigns it to the query image.

Possible Steps

The first step will be detecting what the most prominent objects in the image are, where they are located, and other features about these objects such as color or if two objects are touching. This will likely be done using a CNN, with the ImageNet dataset. We can then calculate distances between objects, which will help us in determining the relationships between certain objects. We can then start to construct simple relationships between objects, based on features that we calculate. Our first goal will be to generate a spatial description of the image. We will need to calculate if an image has an action in it. (For example, an image of a book on a table doesn't require an action.) Natural language processing (NLP) will help with inserting an action into these simple relationships, using our calculated features to help. Finally, also using NLP, we can try to flesh out those relationships into actual sentences using word similarity. As needed, we can go back and grab additional features from the image that will help with this process.

Evaluation & Results

To evaluate our algorithm, we propose multiple methods. The first method would be to test if humans can determine which audio recording or text transcriptions match with which image. This will help us determine if our description is understandable and accurate. The next method would be to test if a computer algorithm can match our description with the correct image using the algorithm created by Socher et al. [3] Finally, we can create a set of descriptions for each image by hand, and present them to people to see if they can pick out the generated description. This will help us to determine how natural our descriptions sound.

We expect to show examples of images and their generated summaries. We can also show what the algorithm produced at different stages of development. We will also outline and analyze the results of our evaluation.

Timeline

Task	Completed By
Project Proposal	February 14th
Multiple Object Detection	February 29nd
Object Boundary Detection	March 13th
Spatial Object Relationships	March 21th
Project Midterm	March 25th
Action Relationship	April 3rd
Object Details	April 24th
Final Project Submission	May 4th

References

1. Falomir, Zoe, et al. "Describing images using qualitative models and description logics." *Spatial Cognition & Computation* 11.1 (2011): 45-74.
2. Elliott, Desmond, and Arjen de Vries. "Describing images using inferred visual dependency representations." *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015.
3. Socher, Richard, et al. "Grounded compositional semantics for finding and describing images with sentences." *Transactions of the Association for Computational Linguistics* 2 (2014): 207-218.
4. Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images using 1 million captioned photographs." *Advances in neural information processing systems*. 2011.