CS 766 Computer Vision - Midterm Report
Image Summarization

Hanna Strohm (hstrohm), Nathan White (ntwhite), Joseph Mohr (jvmohr)

April 1st, 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.

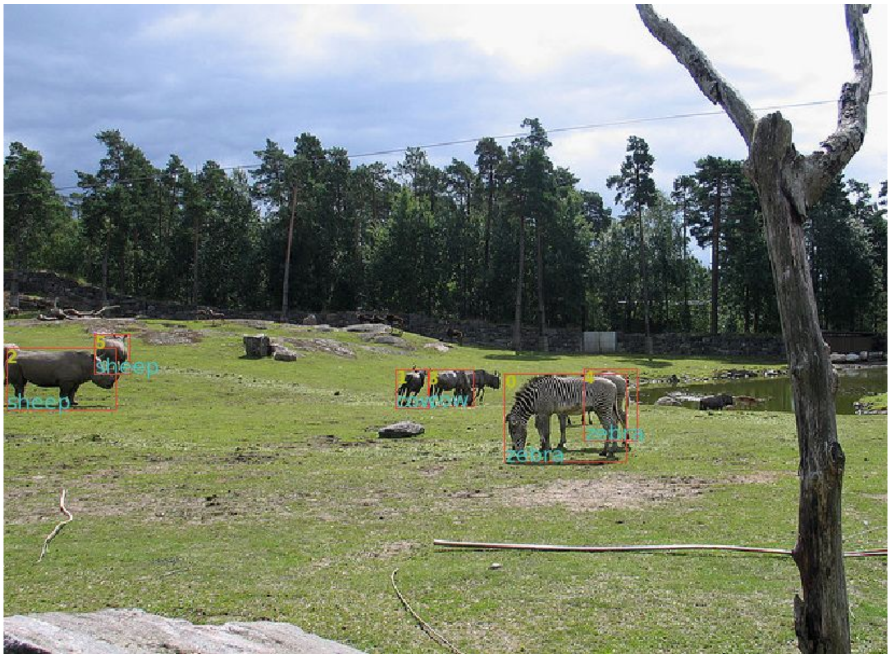
Current State of the Art

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.

Original Plan

We have not deviated too far from the original plan so far. The first step was to detect what the most prominent objects in the image were, where they were located, and other features about these objects such as color or if two objects are touching. We accomplished most of this using a pre-trained Mask R-CNN model that was trained on the COCO dataset, besides the extra features. We are currently working on determining relationships between objects using distance and NLP algorithms, which are a part of our original plan. The final step in our original plan was to use NLP to create actual sentences using word similarity. The following section has more details on what we have accomplished so far, and our updated plans for the future. The section following that details what didn't work for us and what we are currently having issues with.



Current Progress/New Updates

From our original plans, we have completed both object and boundary detections. We're using a pre-trained Mask R-CNN model that was trained on the COCO dataset. This output of the model provides object labels and finds the boundaries of the detected object. From the recognized objects, we have a confidence threshold value of 75% that must be met for us to move forward with that object. This ignores most of the non-focal objects that were “detected.” The image above shows an example of an output image after this process.

At this point, we have a list of recognized objects and their bounding boxes. Our plan is to use the bounding boxes to generate a list of potential prepositions based on how the bounding boxes overlap. From this list of prepositions, combined with the labeled objects, we will use models from NLP in order to find the most likely way the two objects relate. For example, if we have a book above a table, possible prepositions include ‘on’, ‘in’, and ‘above’. The NLP model will inform us which potential phrasing is the most natural. To continue our previous example, our NLP model would tell us that “the book is on the table” is more natural than “the table is on the book.”

We plan to also use a similar NLP-based approach in order to determine the most likely action to be occurring between two objects. While unsure of its potential success, we hope that this will produce moderate results.

Depending on how the above goes, we may have to cut out our plan to add object details. However, we have a plan to get basic information about items from the image. This would likely just add up to two adjectives per sentence, which would help with realism. If rushed, though, it could introduce inaccuracies.

What has failed and how will we fix it?

As a component of our project, we decided to narrow down the results of the object recognition down to the foreground and focused items within an image. This was decided because we can get many images that have lots of things happening in the background, but aren’t the actual focus of the image itself. However, being able to determine which items are the focus (or the main components) of the image has been difficult. We’re using OpenCV’s GrabCut which allows us to remove specific components of an image. We’ve attempted detecting foreground vs background objects by use of the contours within an image.

However, this hasn’t given the desired results, as many non-focused and non-foreground items remain in the resulting image. One potential solution to explore would be to utilize the focus/blur detection techniques discussed in lecture. We could transform our input image using the Fast Fourier Transform and a Laplacian Kernel, then analyze the areas of higher frequency (ignoring the lower frequency parts of the image). This would give us the focused objects, but will require some experimentation with threshold values; as setting the threshold too high might remove some of the image context, while too low will include too much non-focused content.

Evaluation

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. or [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.

Updated Timeline

Task	Completed By
Project Proposal	February 14th
Multiple Object Detection	February 29nd
Object Boundary Detection	March 13th
Project Midterm	March 25th
Focus Object Detection	April 6th
Spatial Object and Action Relationships	April 15th
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.