Vinna Gu

Professor Hwang

CS1674: Intro to Computer Vision

27 April 2021

Part I: Paper Introductions

Paper 1: You Only Look Once: Unified, Real-Time Object Detection

In previous object detection approaches, their implementations were much slower in that they either required a fixed sized window that would glide across the entirety of an image or going through multiple processes which can take time. To handle that, they introduced their own approach of object detection called YOLO which uses a single neural network. The main idea is to essentially split an image into a grid of whatever size they choose then generate bounding boxes and a class probability by only looking at an image once. Given that is their goal, they are able to detect objects in real-time; however, this implementation still struggles to identify smaller objects.

Paper 2: Automatic Understanding of Image and Video Advertisements

In these image and video advertisements we see in our daily lives, symbolism and other visual rhetorics are often incorporated to capture a person’s interest. Given that this is the case, the authors’ intention is to implement a way for a system to automatically decipher what these visual advertisements are trying to demonstrate which can be quite complex. To do that, they used a crowdsourcing website to display the various visuals, and obtained opinionated answers from the workers. By taking the responses they have collected, they want to further enhance this automatic understanding by having the ability to display the effectiveness of an ad, how they can target ads based on the user's interest, and how it can benefit those who are impaired visually.

Paper 3: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Given an image, their intention is to be able to manipulate a photo such that it could have similar characteristics like you would see in a painting by Monet or flipping an image depicting a group of horses to depict a group of zebras instead. They referred to this as "image-to-image translation," but their main focus for this is to be able to translate these images without having a pair of trained data. This can be difficult since no information is being given, so in order to tackle this solution, they stated that they wanted to be able to manipulate a photo, and then revert it back to its original state. In that way, there is a mapping that is occuring between the two photos.

Part II: Detailed Paper

Paper 1: VQA: Visual Question Answering

Throughout this paper, they are essentially trying to create a system that can generate natural open-ended answers to images that are being shown. Prior to this, there were other studies working with VQA; however, they had used limited data which in turn resulted in a very limited set of answers. To change that, this group took data from Microsoft Common Objects in Context (MS COCO) and sought out crowd-sourced employees for answers. With their large set of data, they are able to get a diverse response to their questions rather than “yes” or “no”. To get a variety in answers, the questions often consisted of four general topics: location, object, count, and color. They also needed to consider the questions they were asking such that they would receive answers that may consist of nouns, adjectives or verbs.

Based on their observations, they have found that about 90% of the answers they received were single-worded answers; however, given the specificity of the queries, they were not shocked that that was the response they got. Because the respondents were able to give a one word answer, this gave the researchers the conclusion that there was complex thinking for a human to generate such an answer. This can be disadvantageous because while it may be common sense for humans to process how they got their concluding answer, we are working with machines that do not have such capabilities. In that case, they should still describe their thinking and what led them to their conclusion. What they can do is allow for two questions. One question can result in a natural answer and then a follow-up question can request a more fully elaborated answer.

Paper 2: VizWiz Grand Challenge: Answering Visual Questions from Blind People

The goal for VizWiz is to use the data they obtain from blind people and help them on their needs such as figuring out what shirt color they are wearing or what they see. Since computer vision is essentially being used as an artificial set of eyes, the authors are able to obtain more genuine questions that the visually impaired think about in order to assist them in their daily lives. Along with that, their images are more genuine and realistic because it is not often that you will get pictures that are being depicted concisely. Prior to their study, they have already developed an app in 2011 where blind people could take images, ask questions and send them anonymously through the app. One issue they were facing was protecting the user's identity. For them to solve that, they would have trusted people to go through the data to filter out all the information that could potentially expose the user's life such as their location or inappropriate content that may or may not be intentionally shared. This can take a considerable amount of time and can feel quite tedious to do, but this is so they can get crowdsourced answers from the data. Using sites like Amazon Mechanical Turk, they are looking to obtain natural questions and answers that they have about the image as well and describe what they can or cannot see. They found that the majority of the questions starting with the phrase or word "How many" and "Is" often resulted in simple one-worded answers that could have potentially been more descriptive. After getting the crowdsourced answers, they attempted to apply their data using a variety of algorithms. For the most part there were some issues because the training of the images were done on a group of data that did not run into issues like lighting or unfocused images. Given the tools that they have, I think one way they could prevent blurry, unfocused images is send haptic feedback to the user when they are trying to take a photo. That way they are being indicated that the image is not up to a certain standard and request them to take it again.