Tim Tadros

Advisor: Emily Cooper

Thesis Proposal

1. Background and Motivation

Human beings can capture the context of a scene with just a brief glance. In the literature on human scene recognition, it is assumed that local and global properties of a scene are important in allowing humans to perform this scene recognition task. Local properties, such as the objects in an image and the spatial properties of those objects, can be distorted in such a way that the image contains only global features. We can do this by scrambling the image into various pieces of different sizes and rearranging them. Also, global properties, such as color and texture, can be removed from an image by making it grayscale or adding noise. Using distortion techniques and testing human beings on scene recognition can create a direct causal link between the human visual system and the features it detects. For my thesis, I plan on utilizing a state-of-the-art neural network trained for scene recognition to compare human and computer performance on recognizing distorted images.

In order to determine which features are helpful in allowing human beings to recognize scenes, it is important to have a computational model of the task at hand. Deep convolutional neural networks have reached a performance level on object recognition that is comparable to human beings. However, most of the literature on neural networks is focused on object recognition. For human beings, it seems that scene information is captured before object information. It is then assumed that the context and category of the scene in an image can be used to help the human visual system detect objects (Greene, pg. 139). For this reason, it is important to develop a computational model of scene recognition in human beings. If the first step of processing an image in the human visual system is understood, then that may pave the way forward for studying how human beings detect objects.

1. Literature Review

Two previous studies have looked at human and computer performance on scene recognition of jumbled images (Borji et al., Vogel et al.). In Borji et al., they found that models that capture the global features of an image correlate well with human performance on categorizing jumbled images. In Vogel et al., researchers tested many distortion methods and compared them to a semantic model, which used local features, and a computational gist model, which utilized global features. They concluded that global features are more important for scene recognition since the gist model performed better on jumbled images, and both human and computer model performance dropped when images were made grayscale.

While both studies utilize computer models to determine if local or global information is important in recognizing the context of an image, they did not use neural networks for their computer vision algorithms. In the past few years, neural networks have revolutionized the field of computer vision and quickly overtaken other methods to become the current state-of-the-art for many difficult visual tasks. Because neural networks are very successful in image processing, we hope that by using neural networks to process scrambled and distorted images, we will be able to learn more about how humans process information or how the human visual process system may be lacking (if the computer algorithm performs better than human beings). Having a more robust model of the human visual system will hopefully tell us more about the unique failures and successes of visual processing in humans.

1. Plan of Action

In order to test how scene recognition is performed in the human visual system, I will use various distortion methods to determine which features of an image are particularly useful for recognizing a scene. I will compare the accuracy of neural network and other computational models with human beings on the task of identifying if an image is an indoor or outdoor scene. The first stage in this project will be to find and process images. In order to gain the most from the human survey and the computer vision algorithms, it is necessary to pick a representative set of indoor and outdoor images that captures most of the information and variance inherent in the two categories. Once we choose these images, I will process these images using a couple of selected distortion methods.

After the images are chosen and distorted, I will then conduct a survey on those images to collect information about human beings’ performance on the task of categorizing scrambled images as indoor or outdoor scenes. I will be using Amazon Mechanical Turk to run a Qualtrics-made survey. The survey format will ask the participants to select if they think the image depicts an indoor or an outdoor scene. I will measure how scrambling, or otherwise distorting, these images affects people’s ability to perform this task. At the same time, I will be using a state-of-the-art neural net called AlexNet (Zhou et. al) trained to be highly accurate at scene recognition that allows computers to perform the same task (determine if an image is an indoor or outdoor scene) as the survey. By comparing the accuracies of human and computer performance in this manner, we hope to gain knowledge about both how humans process images and how well deep convolutional neural nets can replicate human perception. If computers and humans achieve similar results then that may tell us that the computer algorithm used mirrors how the human visual system processes visual information. The last part of the project involves training a neural network on the same distorted images that we used for AlexNet. It is the goal of this project to both study human visual processing and produce an accurate and robust scene recognition neural network that others can use.

1. Works Cited

Borji, Ali, and Laurent Itti. "Human vs. computer in scene and object recognition." *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

Greene, Michelle R., and Aude Oliva. "Recognition of natural scenes from global properties: Seeing the forest without representing the trees." *Cognitive psychology* 58.2 (2009): 137-176.

Eberhardt, Sven, Jonah Cader, and Thomas Serre. "How Deep is the Feature Analysis underlying Rapid Visual Categorization?." *arXiv preprint arXiv:1606.01167* (2016).

Parikh, Devi. "Recognizing jumbled images: The role of local and global information in image classification." *2011 International Conference on Computer Vision*. IEEE, 2011.

Vogel, Julia, et al. "Categorization of natural scenes: Local versus global information and the role of color." *ACM Transactions on Applied Perception (TAP)* 4.3 (2007): 19.

Zhou, Bolei, et al. "Learning deep features for scene recognition using places database." *Advances in neural information processing systems*. 2014.