Understanding the Effects of Biological Predispositions in Human Vision in Thwarting Adversarial Attacks

Varshini Reddy
Project Report for
MIT 6.819 - Advances in Computer Vision
varshinibogolu@fas.harvard.edu

Abstract

The structure of deep nets aspire loosely to emulate the human brain, beginning from the introduction of nodes to imitate neurons in the brains to adjusted strengths of connections between neurons to reflect their associations similar to that in the visual cortex. In short, there is a close association of cognitive and neuroscience with computer vision.

One of the grand challenges of computer vision lies in understanding how the brain recognizes objects in the visual world to improve machine vision i.e. understand the underlying representations. To this extent, this work aims to incorporate certain human vision traits such as acuity and initial blurred vision into known state of the art object recognition architectures, which is VGG-16 in this case. Additionally, given that humans are immune to adversarial attacks, we also try to understand whether human motivated computer vision architectures can help in mitigating the effects of adversarial attacks.

1. Introduction

Object detection and recognition has always been at the forefront of computer vision, even before era of neural networks. There have been many perspectives of how object recognition takes place. Humphreys and Bruce, in 1989, proposed a model of object recognition that fits a wider context of cognition. According to them, the recognition of objects occurs in a series of stages. First, a sensory image is generation, following a perceptual classification, were the information is compared with previously stored descriptions of objects. Another approach to this problem was proposed by Marr and Nishihara, where we store an object as a 3D model which can be used to make predictions, thus making them transformations invariant. This was followed by a proposal by Biedermann. Regardless of the technique, one thing remained constant among all proposals. The approach used to achieve object recognition, which was to try emulate the human interpretation of vision detection and recognition.

Nature has caused humans to be born with multiple predispositions which might seem to be sensory function limitations initially, however they have been proven to be beneficial in the long run. An example of this is babies preferring sweet as compared to bitter food. This has been studied to help them eat energy packed and while discouraging the consumption of toxins. The motivation behind this work are 2 such biological predisposition. One babies starting at an acuity of 20/600 (which causes blurred and squinted vision) [6]. The second is the limited perception of colour in infants [1]. While the former helps focus on interpreting local patterns, the latter helps focus on detecting edges and other shapes rather than learning object recognition using only colours.

Regardless of the size of the image corpus' models, today, are trained on, such as ImageNet [2], they are remain susceptible to simple Adversarial attacks and failure in identifying Out-of-Distribution images. However, humans have the capacity to classifying objects or images they have never seen before successfully, though it may be into a broad category. The motivation behind this paper is to understand whether there are any traits of human vision which enable us to perform so well on unseen and adversarial images.

2. Related Works

There have been immense efforts in understanding the workings of human brains and discovering the visual features and representations used by the brain to recognize objects [3]. Further, efforts have been focused on applying this knowledge to improve computer vision. Recently, neural network models of visual object recognition, including biological and deep network models, have shown remarkable progress and have begun to rival human performance in some challenging tasks. These models are trained on image examples and learn to extract features and rep-

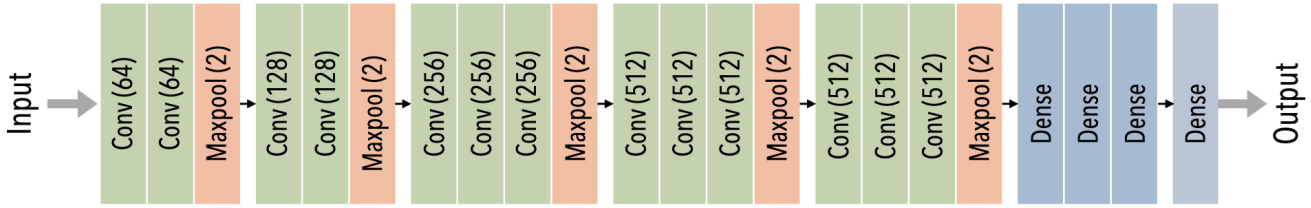

Figure 1. Baseline architecture VGG-16

resentations and to use them for categorization. It remains unclear, however, whether the representations and learning processes discovered by current models are similar to those used by the human visual system [10].

In a more recent work [4], the authors propose to improve machine vision by incorporating context representation of scenes as perceived by humans without objects (in most cases). With a systematic paradigm developed to capture human contextual expectations they were able to improve model performance as compared to generic image augmentation techniques. In yet another work [7], the authors talk of the differences in the perceptual distance of objects between between human and computer vision. To this extent, they propose fixing the biases that cause machines to underestimate the distances between symmetric objects compared to human perception.

3. Proposed Methodology

In this work, we attempt to cover the effects of three human traits in general object detection and classification and in mitigating adversarial attacks.

3.1. Dataset and Pre-Processing

For the purpose of this work, we use the a subset of the ImageNet dataset. In particular, we use the ImageNet 2012 Validation set for both model training and evaluation [8]. This dataset contains 50,000 images belonging to 1000 classes, all of which are balanced. However, owing to the limited computational capacity, we use only 15,000 images for training and 5,000 images for evaluation. Each image is resized to dimensions 224x224x3. The test set is further augmented to create a new set of 5,000 images and made an adversarial version of the same images described in the later sections.

3.2. Proposed Model Architecture

For incorporating all these, we will be using the VGG-16 architecture [9] as the baseline. The weights of this model trained on the original ImageNet dataset is what is going to be used to reduce the required training epochs. The baseline architecture is shown in Fig. 1.

3.2.1 Approach 1 (Reduced Visual Acuity)

As discussed in the previous sections, infants are born with lower acuity than normal adults. As the child grows this acuity naturally improves. This initial disadvantage allows us to learn local pattern relationships by forcing a smaller receptive field and blurred vision. To incorporate this into machine vision, we input the same image at various stages of the network as input. The image at each stage is preprocessed to have a varying level of blur, starting with highly blurred images. The blurring is done using a Gaussian kernel. The proposed architecture to achieve this is shown in Fig. 2. The Gaussian blurring parameters for each input position is given in the following table.

	Input 1	Input 2	Input 3	Input 4
Blur Level	15	10	8	5
σ	200	100	50	10

3.2.2 Approach 2 (Depth Perception)

Another advantages humans have is the ability to perceive depth when they look at a scene. Though recent studies have shown that humans do not directly encode information in 3D, but rather as a sequence of 2D views, the essence of depth is still encoded in the information stored. This allows for an altered context and object representation as compared to how the learned representations are in computer models. In an effort to incorporate this we process the input image to compute its depth map. The input to the baseline architecture is a stack of the original input image with its corresponding depth map. This is depicted in Fig. 3.

3.2.3 Approach 3 (Edge Detection)

As seen in both, humans and machines, the kernels (Gabor filter in humans) learn small edges and other contrasting features initially followed by larger image features. To reduce the training requirement and find if reducing noise which makes the model sensitive we can improve its adversarial robustness. This is also to take into account the presence of rods and cones in the eyes, which enables both during day and night. For this find the edges of the image

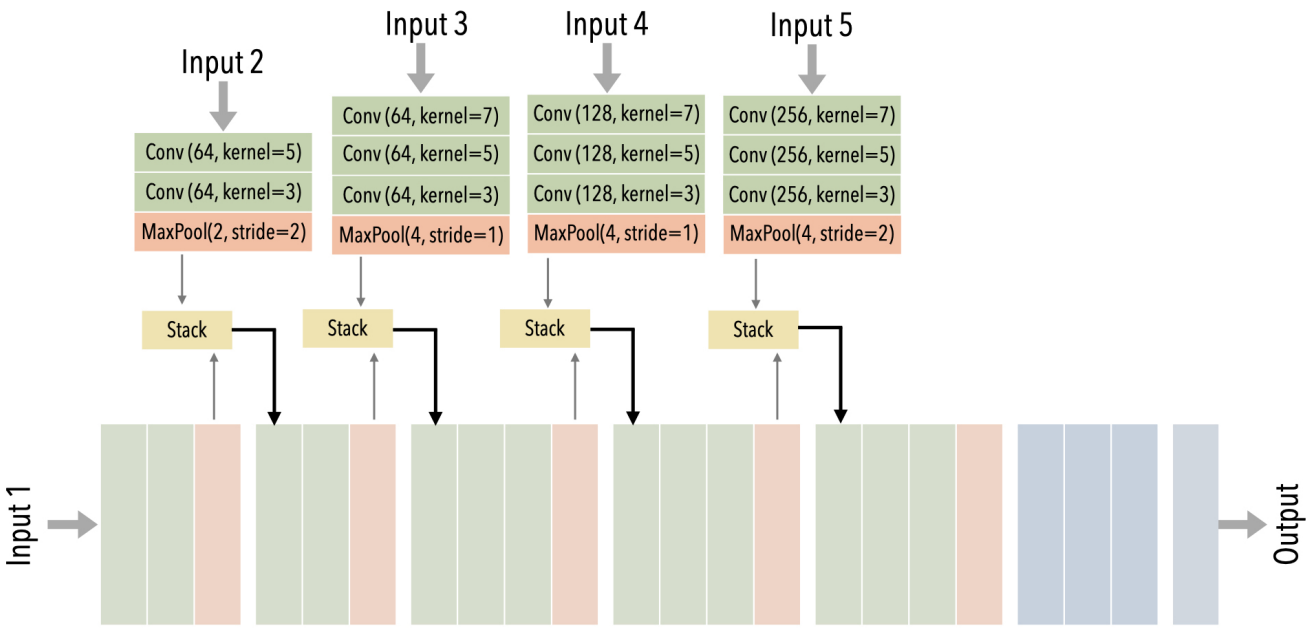

Figure 2. Proposed architecture to incorporate blurred images

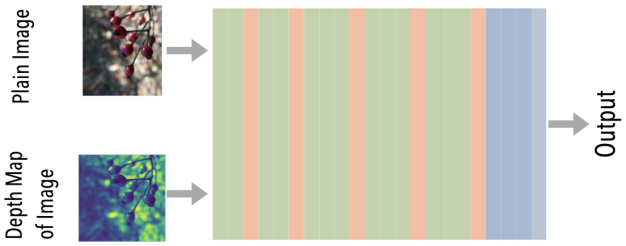

Figure 3. Proposed architecture to incorporate depth information

and input it to the baseline architecture. The structure to achieve this is depicted in Fig. 4.

3.3. Training and Evaluation

For each of these models, we use 70% of the 15,00 images as train and the remaining 30% for validation.

For the architecture that takes as input the blurred image, we freeze the weights of the baseline architecture. We train the weights for the added layers and the full connected section for the first set of epochs and finally train all the layers for the second set of epochs. Similarly, for the depth and edge information incorporation experiments, we trained the model only for the weights that were not part of the original architecture along with the fully-connected layer. For the later part of the training, we briefly train all the layers.

To evaluate these models, we compute the performance as a function of accuracy (since it is a balanced dataset) on the original ImageNet test set. Additionally, to compute the adversarial robustness, we create adversarial samples from the same test set and compare the model accuracy. For adversarial image generation we used an existing white-box attack PGD algorithm described in [5].

Thus, we mainly compute the difference between the accuracy of the baseline architecture, proposed architecture where the images are not modified (all inputs are the plain image) and finally the proposed architectures with the processed images.

4. Results and Conclusion

4.1. Approach 1

The accuracy for both the test and augmented data seems to have dropped significantly for the proposed architecture with appropriately processed blurred images. However, there is a 15% improvement in accuracy as compared to Vanilla VGG16. The improvement in robustness against adversarial images seems to be due to the fact that that blurred images have some noise added to them. There is an intuition that the lower accuracy on the test and augmented data is maybe because of the high noise added. This could seem like a trade-off, however it should be improved with more training.

4.2. Approach 2

This architecture does not differ much from the original VGG16. It shows similar results to that of Vanilla VGG16 on the test data. The results for the augmented test data has dropped. This could be attributed to the fact that the depth map information for most of the augmented images

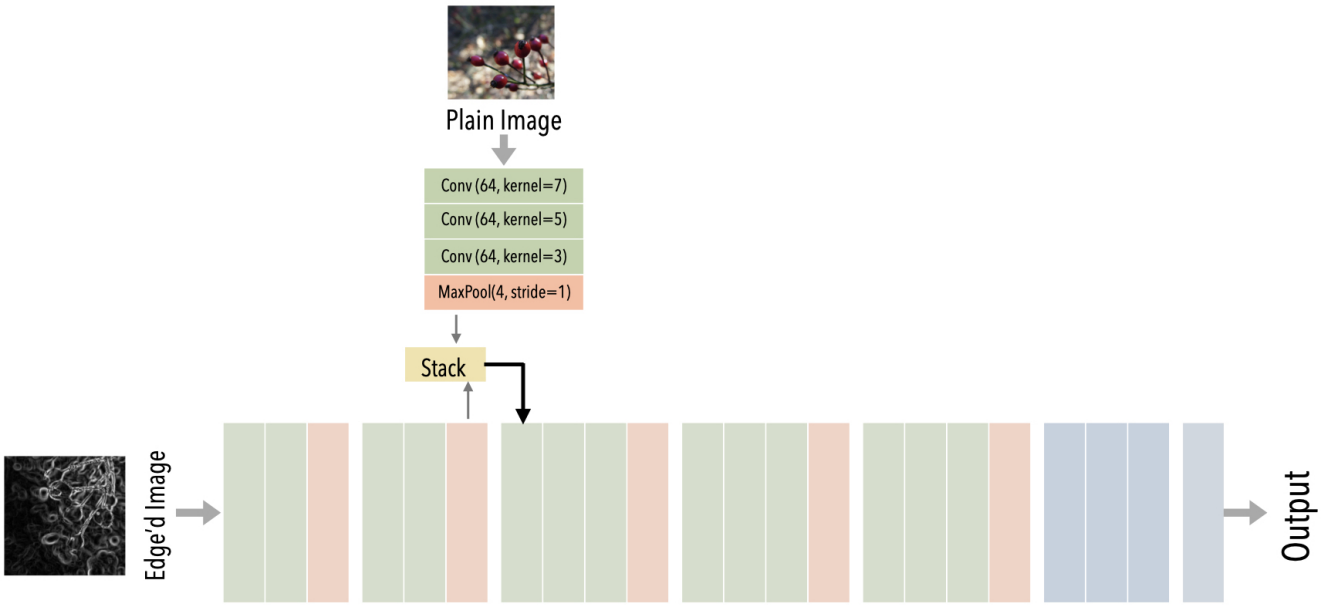

Figure 4. Proposed architecture to edge information

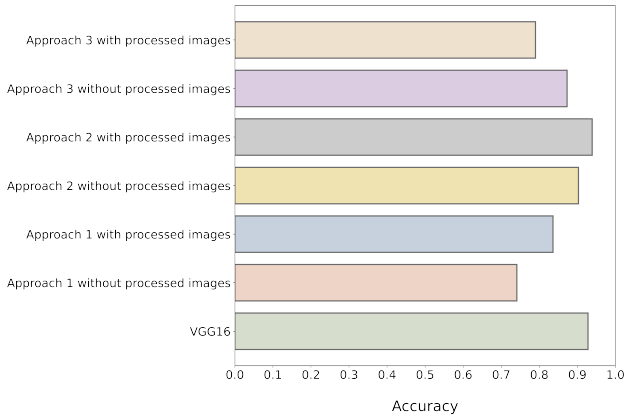

Figure 5. Results of all the model on the test dataset

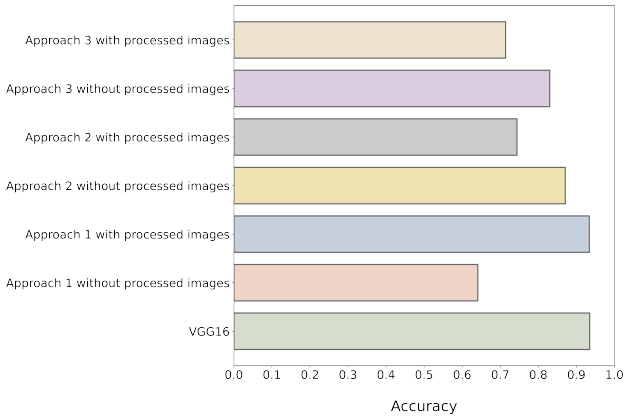

Figure 6. Results of all the model on the augmented version of the test dataset

is similar to random noise. There seems to be a slight improvement in adversarial predictions, but that could just be a result particular to this test set.

4.3. Approach 3

As for the architecture that incorporates edge information, we can see for that the accuracy drops significantly for both the test and augmented sets. There is a drop in the performance for adversarial images as well. The reason for this could be that VGG-16's pre-trained weights are tuned to look for color images as the input, which seems to govern the weights of the atleast a few convolutions initially. However, similar results (there was a slight improvement) were observed when the position of the true ImageNet image and the edge detected image's position were swapped. Regardless, training a network with such edge information does not provide any useful information to help the model classify. In fact, it seems to goes against the trained weights, thus ruining their existing representations.

4.4. Conclusion

In conclusion, there seems to be a positive indication when using architectures inspired by human vision. This is true especially for the model with blurred images which shows significant improvement for adversarial image classification over the vanilla network. The edge information incorporated did no help to the existing model, in fact it

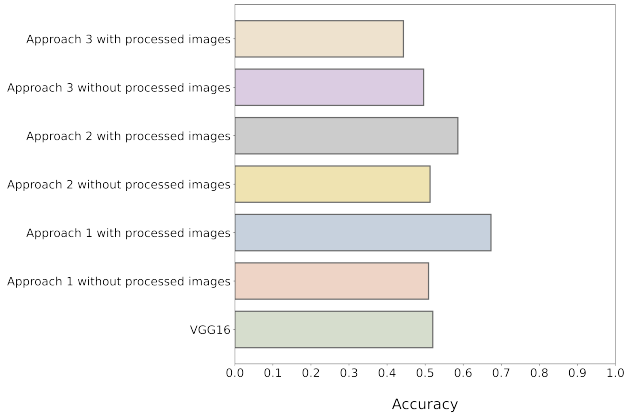

Figure 7. Results of all the model on the adversarial version of the test dataset

seemed to have caused model performance to deplete significantly. The depth information of an image does seem to help in general image classification, however further experiments have to be performed to find out how they can be used for augmented images. Regardless, there seems to be sufficient evidence to say that incorporating human vision traits can help with general classification, as well as adversarial image classification. The results for all models are depicted by Fig. 5, Fig. 6 and Fig. 7.

5. Future Work

The results of the first proposed architecture is much better than expected. Hence, it trying this model on other images and computing other metrics seems like a natural flow. Additionally, it seems to be worthwhile to try and combine the blurred images with depth information to get a better accuracy on the test set as well. As for the augmented images, maybe a better look at the saliency maps to understand why the results on augmented are very bad. There is also a possibility that with more training epochs the results would be improved. Training with slightly augmented data would aid in improving the model performance.

References

[1] Human infant color vision and color perception. *Infant Behavior and Development*, 2:241–273, 1979. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[3] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network, 2021. 1

[4] Harish Katti, Marius V. Peelen, and S. P. Arun. Machine vision benefits from human contextual expectations. *Scientific Reports*, 9(1):2112, Feb 2019. 2

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017. 3

[6] Daphne Maurer and Terri L. Lewis. Visual acuity: the role of visual input in inducing postnatal change. *Clinical Neuroscience Research*, 1(4):239–247, 2001. 1

[7] R. T. Pramod and S. P. Arun. Improving machine vision using human perceptual representations: The case of planar reflection symmetry for object classification. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):228–241, Jan 2022. 32750809[pmid]. 2

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 2

[10] Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences of the United States of America*, 113(10):2744–2749, Mar 2016. 26884200[pmid]. 2