Comparison between Human Vision and Computer Vision to perceive Kanizsa Illusion.

Adheena Reji Sagar B Dollin Akshat Govekar
MSc Cognitive Science, IIT Delhi.

Project code is available at this link: https://github.com/SagarDollin/Computer_Vision_to_perceive_illusion

Abstract

In this project, we start with a fundamental question of how human vision is different from computer vision. We explore the problem of comparing the standard Computer Vision (CNN) model with human vision in perceiving the Kanizsa illusion. Then we use the model defined by Zhaoyang Pang et al. in their paper *Predictive coding feedback results in perceived illusory contours in a recurrent neural network* based on Rao and Ballard's theory, to perceive the illusion and discuss why a predictive model is vital to perceive illusions and why such a model is more realistic for various applications of computer vision than the standard CNN models.

Introduction to Inferior Temporal Cortex

The process of visual information processing begins at the retina with the transduction of photons into electrical signals. In primates, visual processing happens through two pathways emerging from V1: the primary visual cortex; - the dorsal pathway and the ventral pathway. The dorsal pathway is commonly known as the 'where pathway' and is considered the primitive one that evolved earlier. The ventral pathway is called the 'what pathway' and is involved in object discrimination and recognition. The foremost importance of visual perception and object recognition has been identified in the primate cortex, including humans. These major areas of the ventral visual pathway in the primate brain include area V1, V2, V4, and Inferior Temporal Cortex (ITC), the highest ventral cortical area. The inferior temporal cortex is the cerebral cortex portion located on the inferior complexity of the temporal lobe that plays a predominant role in visual object recognition, classification, and perceptual learning. This pathway is seen to be exceedingly hierarchical in terms of information processing. Top levels of the hierarchy show a progressive increase in receptive field size, object transformations tolerance degree, degree of selectivity to "complex" shapes, etc.

The anatomical organization of ITC can be explained with the help of Brodmann areas. The ITC corresponds to Brodmann areas 20 and 21. It extends from the anterior to the inferior occipital sulcus to a few millimeters posterior to the temporal pole and also from the fundus of the superior temporal sulcus to the occipitotemporal sulcus lateral wall. ITC forms extensive connections with regions of the peristriate cortex, including visual areas V2, V3, and V4, with multimodal areas like the superior temporal sulcus (STS) and the prefrontal cortex, limbic system, and temporopolar pro isocortex.

The first stage of processing the retinal image arises in the V1 layer called the primary visual cortex. The region V1 is characterized by two types of cells: simple cells and complex cells whose functionalities seem to provide the commencing steps for satisfying the selectivity and invariance of visual recognition. These are the core constraints of the visual recognition problem. The simple cell extracts selective edge information in the first step of processing which is highly relevant for image recognition. The complex cells as the second step ensure tolerance in the edge extraction concerning small position shifts. However, the neuronal population in the V1 region does not show robust tolerance to complex image transformations. The V1 region then sends feedforward signals to higher visual areas like V2, V3, and V4. The mid-level ventral areas like V4, which sends the dominant cortical input to the inferior temporal cortex, exhibit object selectivity and variation tolerance at an intermediate level. The later stages of ventral pathway visual processing occur at the inferior temporal cortex. The processing here shows selectivity for global shape and high-level integration. The synthesis of the signals from early visual areas might take place here at this later processing stage. This top-level processing seems to support invariant object categorization over a broad range of real-time tasks.

The inferior temporal cortex can be broadly divided into two areas: the anterior Inferior temporal area called the TE area, which borders with the Superior temporal sulcus (STS). The perirhinal cortex ventral and rostral to the TE area. The final stage of visual processing is represented in the TE area. The IT cortex has a columnar organization with patterns of horizontal activation in the TE region neurons. It is observed that TE has columnar modules with neurons responding to similar features and selective cell responses to complex object features. The column neurons of the TE region are not identical but are correlated with respect to their stimulus selectivity. IT cortex also shows partial overlapping of columns responding to different but related features showing continuous maps of associated features in the anterior IT.

Computer Vision and CNN's.

As an interdisciplinary field of science, computer Vision seeks to attain the processing and analysis of visual data by imitating the functionality of the biological human visual system and brain components. The field of computer vision and its attempt to procure its objectives had originated back in the 1950s. With the growth and advancement of neural networks and deep learning, the field has gained great empowerment. Among various approaches to deep learning, the one that has made considerable contributions to computer vision is Convolutional Neural Networks (CNN's). They form an essential form of computational models in computer vision research. These models are particularly suited for modeling vision since they were explicitly designed for inputs in vision.

Deep Convolutional Neural Networks function by modeling small basic information consecutively and then combining them in deeper networks. E.g., the first layer serves by forming templates for edge detection and executes edge detection. The subsequent layers combine it into simpler shapes and eventually convert them into templates of various

illumination, object positions, scales, etc. The deeper layers match the input image with all the templates and make a final prediction using a weighted sum. Using these ideas, CNN's have attained the capacity to model complex behaviors and variations by providing a high degree of accuracy in their predictions.

Studies have shown that deep learning provides the most comprehensive computational models that encode and extract hierarchically organized features from natural arbitrary pictures and videos [1]. Among the deep learning methods, specifically, the deep CNN's are constructed and trained with organizational and coding principles similar to that of the feedforward visual cortical network [2]. Studies show evidence for CNN's capability to partially explain the brain's responses to natural picture stimuli [3]. However, there is no clear idea about CNN's role in defining and decoding brain responses to natural video stimuli. The main difference between dynamic natural vision and CNN's for computer vision is that dynamic natural vision has associated feedforward, recurrent, and feedback connections in them [4]. But the CNN involves only feedforward processing, and it operates on instantaneous input without considering the recurrent or feedback network interactions [5].

Many CNN-based architectures are offering high performance in image classification. Some of the most famous architecture includes Alex Net (2012), GoogLeNet / Inception (2014), VGGNet (2014), ResNet (2015).

Even though computer vision has shown significant progress in recent years with the advancement of CNN, there is still a considerable gap between computer vision and biological vision in humans. Even though they can perform specific computer vision tasks at superhuman levels, mimicking human visual perception is still limited. Visual object recognition is an important cognitive task that seems to be automatic and instantaneous to humans. However, this task offers daunting challenges to computer vision research. An ideal visual object recognition theory, implemented by a computational model, is obligated to explain certain visual phenomena characteristics like transformation tolerance, Generic shape recognition, Selectivity, Speed, etc.

The human visual system is highly adaptable and flexible with its capability to recognize objects and images from various angles, distances, different object pose, under distortion, under occlusion, etc. Under many other sources of variations and noise. An essential functional difference between human vision and a deep feedforward convolutional network of computer vision is that the computer vision model fails in perceiving illusory contours and, therefore, fails to detect illusions like the illusion of Kanizsa squares.

Problem statement:

We have discussed the popular method used for Computer Vision, i.e., Convolutional Neural Networks, our problem statement is relatively straightforward. Can Computers see the way humans see? If it is different, How is it different? There could be several computer vision models when I say computer vision, but we will stick to the standard CNN model used for classification. Before diving deep into comparing the Computer Vision model and Human vision,

let's define our problem statement in a more precise manner. We intend to compare a standard CNN classifier model to test whether it sees Kanizsa square illusion with humans vision.

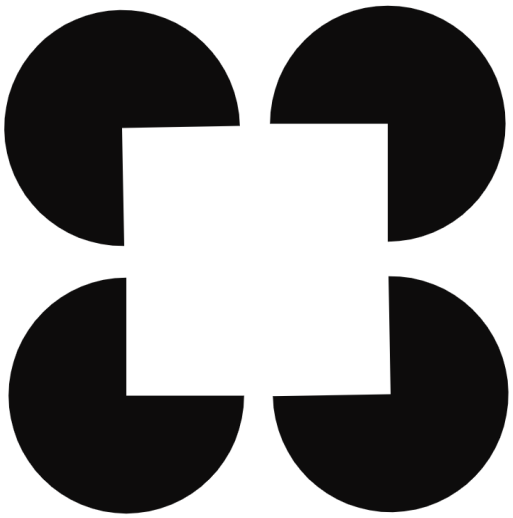


Figure 1: Kanizsa Square

Image credits: Adheena Reji, MSc Cognitive Science, IIT Delhi

This is a Kanizsa square; there is no actual square here, just an illusion created by four Pacman-like objects organized in a particular fashion. To humans, this appears to be a square, and our vision tends to fill in information like edges that are not even present there. I mean to say, between the gap of each Pacman, we tend to see an edge of the square which isn't present. This is because our visual system is tuned to fill in information whenever there is partial information presented.[12] As Daniel Dennette would say, this is because of the expectations we have about seeing particular objects. Also, if you observe, the square might appear brighter than the white background, but in reality, they are of the same brightness; this is mainly once our brain has considered that there is a square, it expects that there should be a figure-ground separation. This also could be the reason for the non-existent edges we see.

We use a CNN model to see whether such a standard and basic model can be used to detect such illusions. Next, we use a predictive coding model from the paper; *Predictive coding feedback results in perceived illusory contours in a recurrent neural network*.[13] In the mentioned paper, we see a predictive coding network capable of forming the expectations discussed earlier. The idea of predictive coding is from Rao and Ballard's paper [14] "*in the hierarchical network. The higher layers try to predict the activity of the lower layers, and the errors made in this prediction are then used to update their activity."* This form of prediction leads to expectation. Here in our project, we try to implement a predictive coding model and compare it with the standard CNN-based architecture in recognizing the Kanizsa illusion as a square. The Predictive coding model is expected to perform better in recognizing the illusion as square since it is designed to form expectations similar to neurons in the human brain.

Predictive Coding feedback:

The model consists of 3 feedforward encoding layers, e1,e2,e3, and three corresponding decoding generative layers d0,d1,d2, whose errors are used to update the activity of the encoding layer below them at each timestamp. The error of a layer n at any timestamp t can be given as,

$$\epsilon_n = e_n - d_n.$$

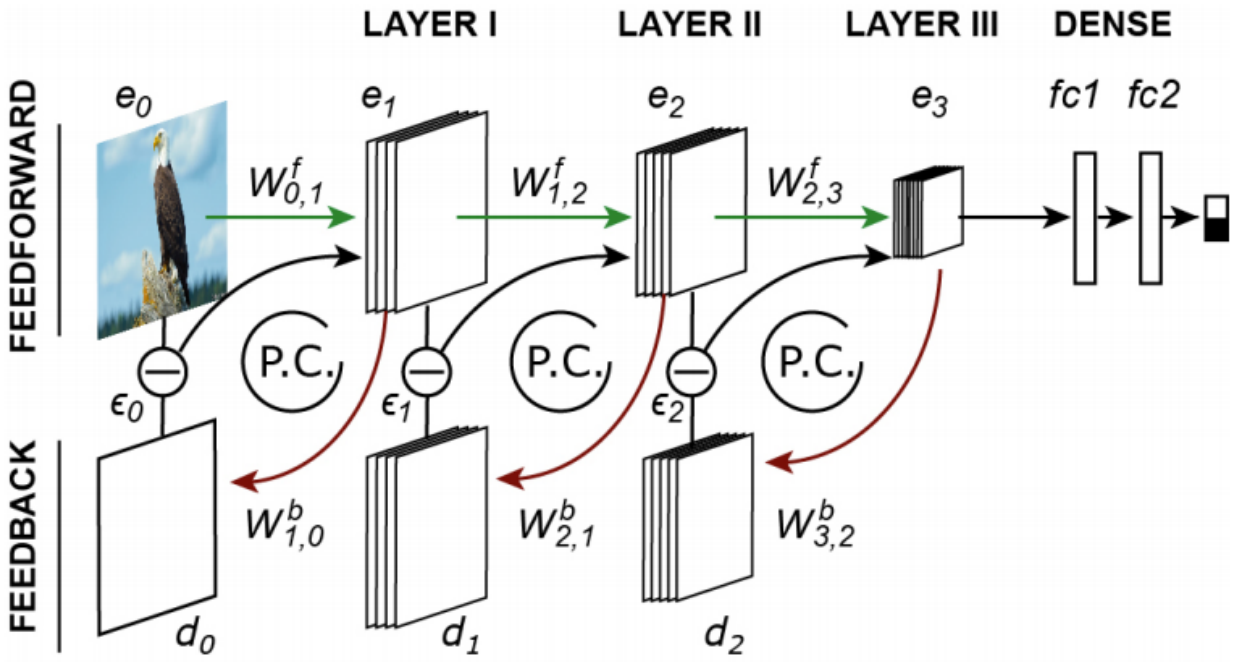


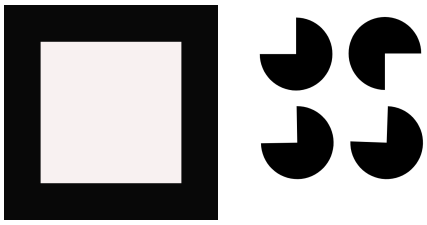
Image source: *Predictive coding feedback results in perceived illusory contours in a recurrent neural network*

Image 2: e1,e2,e3 are the feedforward encoding layers whose activities are updated(P.C loops) using the errors($\in_n$) produced by the corresponding d0,d1,d2 feedback generative layers that try to predict the activity of the lower e_{n-1} layers.

To the hierarchical encoding layer, we attach a binary classifier. That is used to classify images as either square or Pacman.

The encoder model with the predictive coding network is initially trained on the Cifar100 dataset of natural images(tuning). It has been found that it is essential for neural networks to be

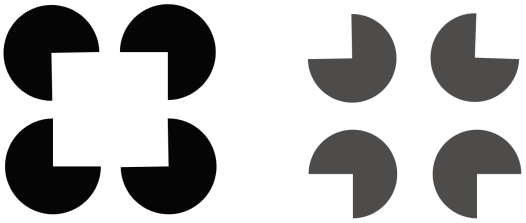
trained on natural images to develop expectations for recognizing patterns. Then we add the classification head to the encoding model and train the complete model on square images and random Pacman orientation images, as shown below.



(a) (b)

Image 3: (a) shows the square image and (b) shows Pacman in random orientations.

Once the model is trained on square and random Pacman orientation, we test our model on the Kanizsa square and Pacman orientation so that all are facing up, as shown below.



(a) (b)

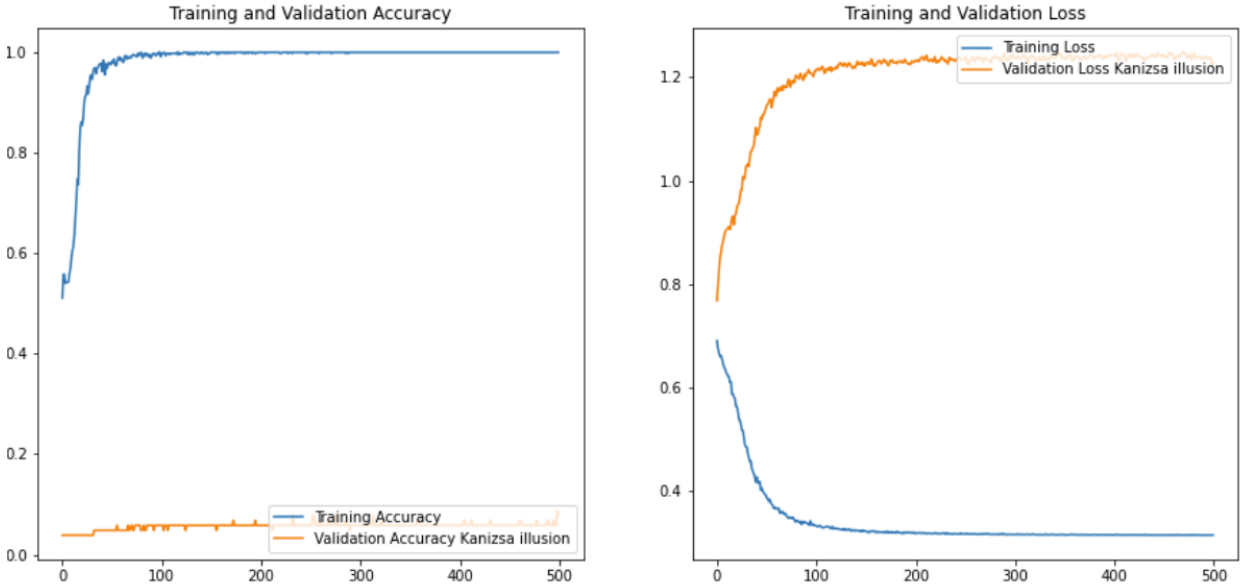
Image 4:(a) shows the kanizsa square image and (b) shows Pacman in outward orientations.

We can intuitively see how this works; the model is initially trained on real square images and Pacman in random orientations. It is then tested whether it identifies the Kanizsa illusion as a square or as Pacman. And for all outward Pacman orientations, it is expected to identify it as Pacman. Hence our classifier model had a binary classifying head; the image is classified as either square or Pacman.

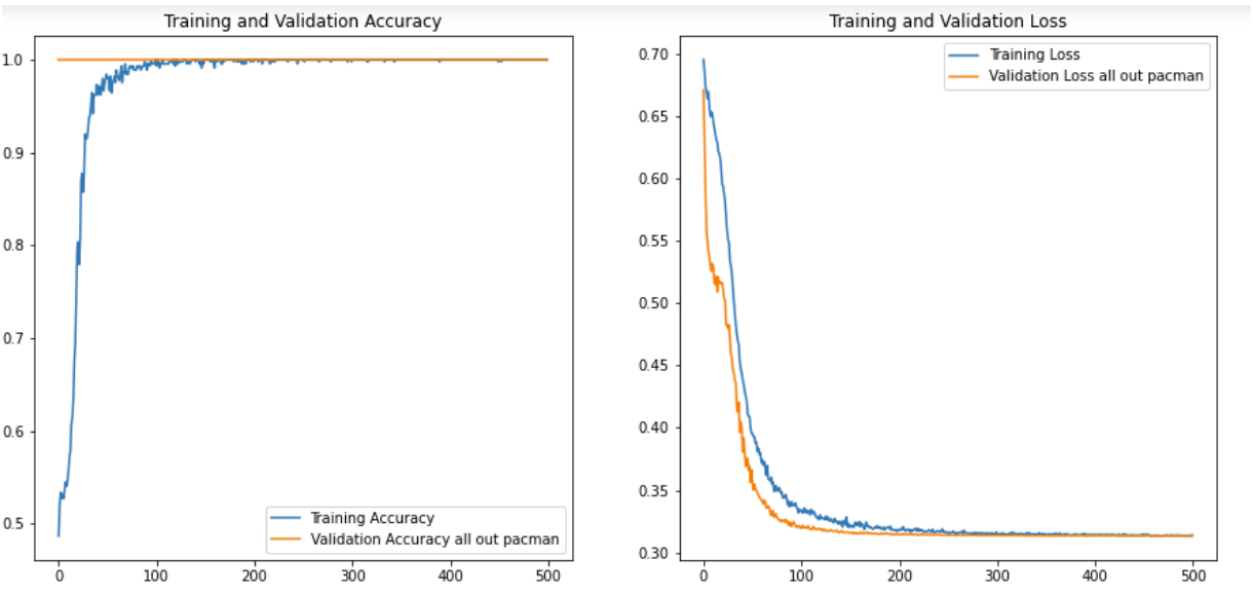
Results and discussion

The results of the model are astonishing. Even though the authors who devised this model to perceive the Kanizsa illusion claim that. Our model identifies the Kanizsa illusion as a square with almost 50% certainty. However, the results I got by implementing the model contradict the author's claims. The predictive coding network implemented using the method author suggests does not perceive the illusion as the square at all. Now we can think that this

might be because there must be some issues in the parameter tuning. However, I tried different hyperparameters for the model, and all yield the same results. Look at the training and validation graph given below trained for 500 epochs .



As you can see the model does not perceive the test set of Kanizsa illusion as square, however, the model performs quite well when tested on Pacman images. It recognizes Pacman images very well.



Now does this mean this model is not good enough, I think not. I think the intuition for such a model is very valid and with a few improvements, we should achieve such a model very soon.

At the beginning of the paper we nowhere claimed that such a model must work, we were only paraphrasing the author's views. Also, note that at the time we did this project, The

author hasn't published their code in the Github repository they gave in. Hence The code is completely developed from scratch using some libraries like Keras, Opencv, etc.

Our sole intention of this project was to implement a computational model to evaluate the claims of the paper we referred to.

Why is such a model important?

A model that can see illusions? Now, why would that be of any importance? Let me take you back to the general question we started with, i.e., Can computers see how we do? And then, we went on to explore a more specific problem of Kanizsa illusion. In doing so, we explored the kind of model that would make computers see the illusion that we see. This gave us a model that resembles closer to how neurons in hierarchical layers interact in the brain. This model can also be extended to use for more complicated tasks where one requires to model expectation. Humans are good at filling in the missing details when partial data is provided, and this step we are taking in predictive networks will make the AI systems better at this task than they are today. This kind of model is promising for self-driving vehicles where the vehicle's vision does not have access to all the visual information needed to make decisions. Still, a predictive feedback neural network model can help create expectations and fill. Our in the missing details.

REFERENCES:-

[1] LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature. 521: 436–444.

[2] Dicarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? Neuron. 73:415–434

Yamins DL, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci. 19: 356–365.

[3] Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci.111(23):8619–8624.

Eickenberg M, Gramfort A, Varoquaux G, Thirion B. 2016. Seeing it all: convolutional network layers map the function of the human visual system. NeuroImage. doi:10.1016/j.neuroimage. 2016.10.001.

[4] Callaway EM. 2004. Feedforward, feedback and inhibitory connections in primate visual cortex.Neural Netw. 17(5): 625–632.

[5] Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. 2012. Canonical microcircuits for predictive coding. Neuron. 76(4):695–711.

Polack PO, Contreras D. 2012. Long-range parallel processing and local recurrent activity in the visual cortex of the mouse. J Neurosci. 32(32):11120–11131.

[6] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, Zhongming Liu, Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision, *Cerebral Cortex*, Volume 28, Issue 12, December 2018, Pages 4136–4160, https://doi.org/10.1093/cercor/bhx268

[7] Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). *Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences, 111(23), 8619–8624.* doi:10.1073/pnas.1403112111.

[8] Efremova, N.A., Inui, T. An inferior temporal cortex model for object recognition and classification.*Sci.Tech.Inf.Proc.***41,**362–369(2014). https://doi.org/10.3103/S0147688214060045

[9] Conway, B. R. (2018). *The Organization and Operation of Inferior Temporal Cortex. Annual Review of Vision Science, 4(1).* doi:10.1146/annurev-vision-091517-034202

[10] Sigala N. Visual categorization and the inferior temporal cortex. Behav Brain Res. 2004 Feb 4;149(1):1-7. doi: 10.1016/s0166-4328(03)00224-9. PMID: 14739004.

[11] http://www.scholarpedia.org/article/Inferior_temporal_cortex

[12] T. E. Parks, Rock's cognitive theory of illusory figures: a commentary, Perception 30 (5) (2001) 627–631.

[13]Predictive coding feedback results in perceived illusory contours in a recurrent neural network: Zhaoyang Pang, Callum Biggs O'May, Bhavin Choksi, Rufin VanRullen

https://arxiv.org/abs/2102.01955

[14] Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects, Rajesh P. N. Rao and Dana H. Ballard