Predicting Visual Brain Response using Convolutional Neural Networks

Project-I (EC47011) report submitted to
Indian Institute of Technology, Kharagpur in Partial Fulfilment
of the Requirements for the Award of the Degree of

Bachelor of Technology (Hons.)
in
Electronics and Electrical Communication Engineering

by

Awadh Kejriwal
(18EC10078)

Under the supervision of

Dr. Debasis Samanta

Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Autumn Semester, 2021-22

November 2021

Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur
Kharagpur – 721302, India

CERTIFICATE

This is to certify that the project report entitled "**Predicting Visual Brain Response using Convolutional Neural Networks**", submitted by **Awadh Kejriwal** (Roll No. 18EC10078) to Indian Institute Technology, Kharagpur towards partial fulfilment of requirements for the award of the degree of Bachelor of Technology (Hons.) in Electronics and Electrical Communication Engineering is a record of bonafide work carried out by him under my supervision and guidance during the Autumn Semester, 2021-22.

Dr. Debasis Samanta
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Kharagpur – 721302, India

Date - November 20, 2021

DECLARATION

I certify that

a. The work contained in this report is original and has been done by me under the guidance of my supervisors.
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have followed the guidelines provided by the Institute in preparing the report.
d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e. Whenever I have used materials (data, theoretical analysis, figures and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Awadh Kejriwal

Signature of the Student

List of Abbreviations

fMRI	Functional Magnetic Resonance Imaging
EVC	Early Visual Cortex
IT	Inferior Temporal Cortex
CNN	Convolutional Neural Network
RDM	Representational Dissimilarity Matrix
RSA	Representational Similarity Analysis
PCC	Pearson Correlation Coefficient
SNR	Signal to Noise Ratio

Abstract

Understanding the process of visual information extraction by the human brain has been one of the main topics of interest for the researchers in the last few years. The hierarchical structure of the human brain makes it more complex and explorative. Lately, scientists have been trying to develop artificial models of the human brain, which can replicate and reproduce neural activities of the brain in response to visual stimuli. This will help us in understanding the nature of human intelligence, and the knowledge gained can be further used to develop advanced forms of artificial intelligence models which possess the ability to perform complex real-world tasks.

The arrival of deep learning and artificial neural networks have paved the way to predict brain activities. The multi-layered architecture of deep neural networks has been lately used to model the complex hierarchical structure of the human brain. We recreated some widely used convolutional neural networks, namely, AlexNet, VGG and ResNet. Our aim was to identify the model configurations which produce responses to visual stimuli that are most similar to human neural activities, as measured by human fMRI. We found that the layers of AlexNet model best predicted the neural activities of both Early Visual Cortex (EVC) and Inferior Temporal Cortex (IT).

Keywords – Neural Activities, Artificial Intelligence, Deep Learning, Artificial Neural Networks, fMRI, Early Visual Cortex, Inferior Temporal Cortex

Contents

1. Background

1.1. Visual Ventral Stream

Human brain is the most vital and complex organ of a human's body. It is the interpreter of all our senses, movements and controller of our behaviour. Among the various senses, vision is the one which is responsible for making us aware of our surroundings. The first stop for the sense of vision is our eyes, or more specifically, the retina. A train of impulse/signals is generated when light falls on our retina. Neurons act as the carrier of these signals. The signal travels from the optic nerve to the thalamus, followed by the primary visual cortex. The primary visual cortex is subdivided into regions like V1, V2 and V4. After the primary visual cortex, these signals travel to the inferior temporal cortex. This pathway of the impulse is termed as the Ventral Visual Stream.

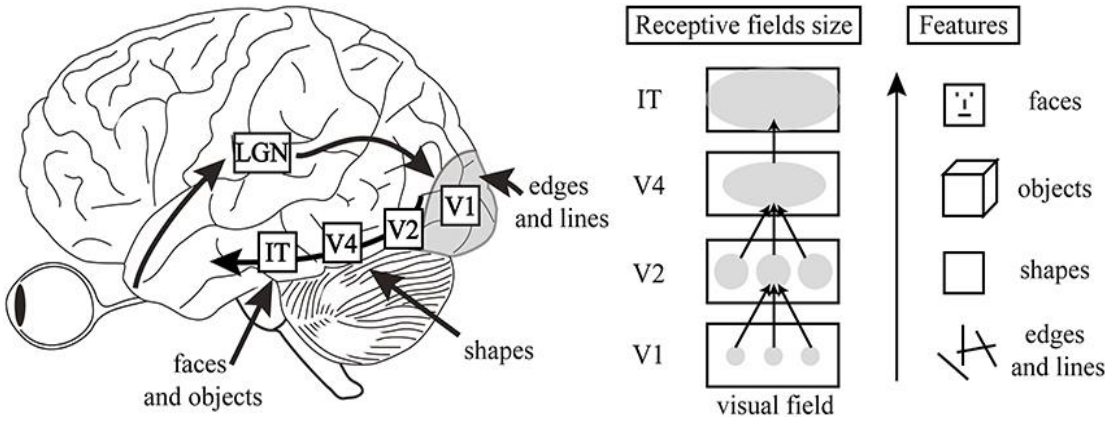

Fig 1.1: *Visual Ventral Stream*

The processing of these signals occurs at almost all the stages of the ventral visual stream. This processing occurs for various purposes, one of them being object recognition. The pathway is built in such a way that the simple features are extracted earlier in the path, and the complex ones happen later as the signal moves deep into the pathway.

Humans can easily recognize different sized objects and put them in the same category. This happens because of the invariances we develop. Whenever we look at any object, our brain extracts the features and in such a way that the size, orientation, illumination, perspective etc don't matter. We remember an object by its shape and inherent features. It doesn't matter how the object is placed, how big or small it is or what side is visible to you. There is a hierarchical build-up of invariances first to position and scale and then to viewpoint and more complex transformations requiring the interpolation between several different object views.

We have cells in our visual cortex that respond to simple shapes like lines and curves. As we move along the ventral stream, we get more complex cells which respond to more complex objects like faces, cars etc. Neurons along the ventral stream show an increase in receptive field size as well as in the complexity of their preferred stimuli. Humans take remarkably little time to recognize and categorize objects. This suggests that there is some form of feedforward processing of information going on. This means that the information processed by the cells in the current level in the ventral stream hierarchy is used by the next level. This helps speed up the process by a huge factor.

1.2. Convolutional Neural Networks

We know the mechanism by which the visual data enters the human visual system and how it's processed. But the problem is that we are still not exactly sure how our brain categorizes and organizes the data. So, we try to extract features from an image and ask our machines to learn from it. There are variations like size, angle, perspective, occlusion, illumination etc. The same object looks very different to a machine when it is presented with a different perspective. Humans, on the other hand, will immediately recognize an object from anywhere. One way to go would be to store all possible sizes, angles, perspectives etc, but this would be

infeasible. It would take an enormous amount of space and time to recognize an object.

Understanding the exact process of object recognition inside a human brain has been an interesting topic for the researchers in the last few decades. Researchers have been trying to artificially model the human brain for this purpose. Among the various techniques, convolutional neural networks (CNNs) have been the most promising when it comes to modelling of the human brain. CNNs not only use the present set of signals, but also learns from the previous instances using backpropagation algorithm. This backpropagation is not possible for the human brain, but it learns through some other mechanisms.

Convolutional neural networks have three main traits that support their use as models of biological vision: (1) they can perform visual tasks at near-human levels, (2) they do this with an architecture that replicates basic features known about the visual system, and (3) they produce activity that is directly relatable to the activity of different areas in the visual system.

1.3. Previous Works

The first major study which established a connection between CNNs and the visual system was Yamins et al. (2014). This study explored many different CNN architectures to determine what leads to a good ability to predict responses of monkey IT cells. For a given network, a subset of the data was used to train linear regression models that mapped activity in the artificial network to individual IT cell activity. The predictive power on held-out data was used to assess the models. A second method, representational similarity analysis, was also used. This method does not involve direct prediction of neural activity, but rather asks if two systems are representing information the same way. This is done by building a matrix for each system, wherein the values represent how similar the response is for two

different inputs. If these matrices look the same for different systems, then they are representing information similarly.

By both measures, CNNs optimized for object recognition outperformed other models. Furthermore, the 3rd layer of the network better predicted V4 cell activity while the 4th (and final) layer better predicted IT. Indicating a correspondence between model layers and brain areas.

Another paper, Khaligh-Razavi and Kriegeskorte (2014), also uses representational similarity analysis to compare 37 different models to human and monkey IT. They too found that models better at object recognition better matched IT representations. Furthermore, the deep CNN trained via supervised learning was the best performing and the best match, with later layers in the network performing better than earlier ones.

2. Work Done

2.1. Introduction

Artificial Intelligence algorithms have been inspired by biological vision since a long time. Simultaneously, progresses in computer vision produced better models to predict neural activities. Recent advancements in convolutional neural network (CNN) models have achieved significant performance on different visual tasks. The layered structure of CNNs have striking resemblance with the stages of biological visual processing. Hence, we can use these CNN models to predict the brain's visual representation space.

CNNs have been widely used for classification tasks for large datasets, having millions of training samples. But the problem is that, it cannot classify objects into new categories without retraining it. So instead of a categorizing model, we can also think about designing a distance model which predicts the similarity between different sets of images, such that images belonging to different categories are far from each other, and the ones belonging to the same category are closer to each other. This can be achieved by using a certain loss function, which maximizes the inter-class distance. The same loss function can also be applied on the corresponding neural data. The two outcomes can be compared and analysed to evaluate the performance of the model.

2.2. Problem Formulation

We were provided with a set of 92 images of different objects, scenes, etc. We were also provided with the dissimilarity matrices of fMRI signals, which were recorded in response to viewing those 92 images, by 15 different subjects. The fMRI signal was collected from two regions of the brain, namely, early visual cortex (EVC) and inferior temporal cortex (IT). Our task was to generate features from the images so that the resulting dissimilarity matrix is closest to the given dissimilarity

matrices. For this, we proposed to use some widely popular convolutional neural networks like AlexNet, VGG and Resnet. We were supposed to push the images though these deep neural networks and use the activations generated at various layers to create dissimilarity matrices, which would then be compared with the dissimilarity matrices generated from actual fMRI data.

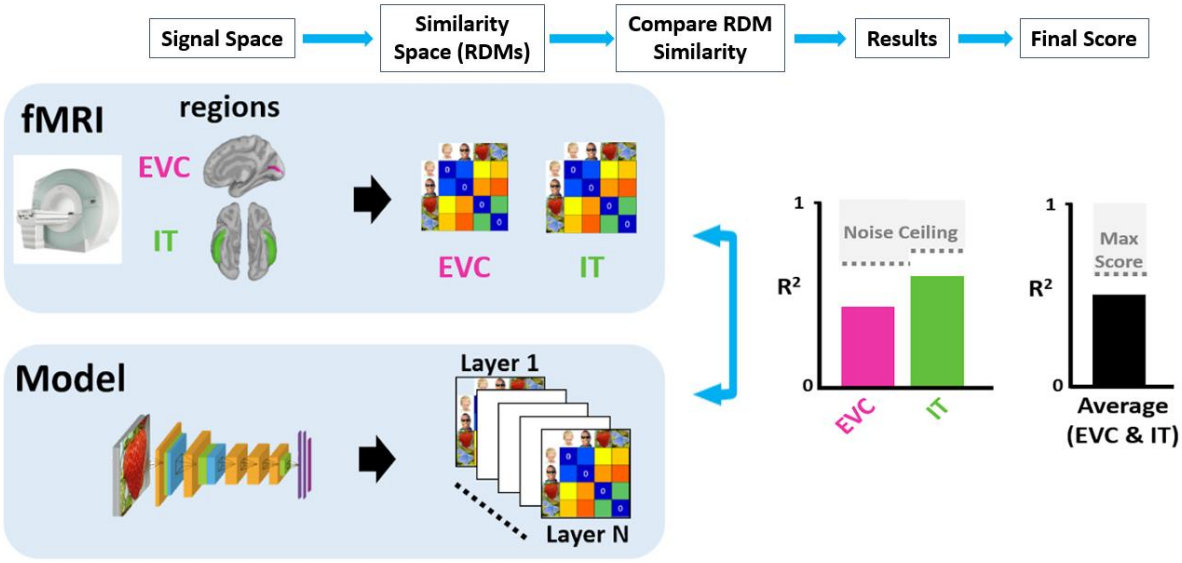

Fig 2.1: Workflow Diagram

2.3. Dataset

The image dataset consisted of 92 silhouette images. The images were of different kinds of objects, animals, scenes, faces, etc. The image set was unlabelled.

The neural activity data was provided in form of dissimilarity matrices computed from fMRI signals. The fMRI signals were recorded from 15 different subjects in response to viewing the 92 images of the image dataset. The signals were recorded from two regions of the ventral visual stream, namely, early visual cortex (EVC) and inferior temporal cortex (IT). Hence, in total 30 dissimilarity matrices were provided - 2 each (EVC, IT), for 15 subjects.

Fig 2.2: *Image Dataset*

2.4. Feature Extraction

Some of the widely used Convolutional Neural Networks (CNNs) were used for the purpose of feature extraction. Deep Neural Networks like AlexNet, VGG and Resnet50 were used. The images were pushed through the layers of these networks and the activations generated at some of the layers were used as features to create dissimilarity matrices. The architecture diagram of the three networks have been shown in the figures below.

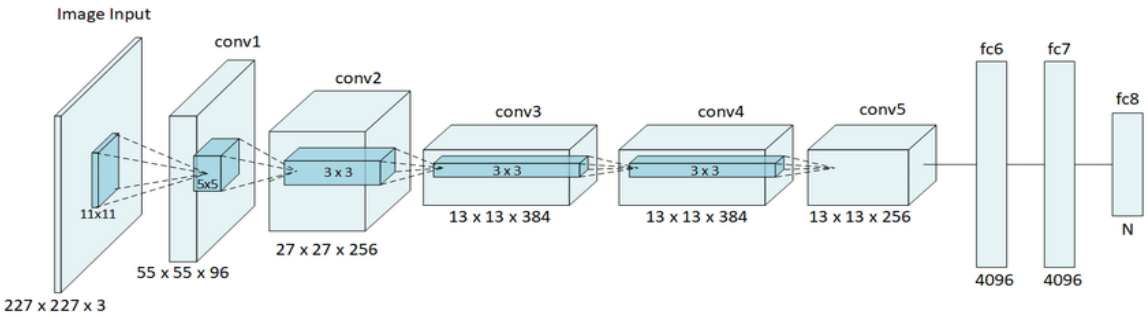

Fig 2.3: *AlexNet Architecture Diagram*

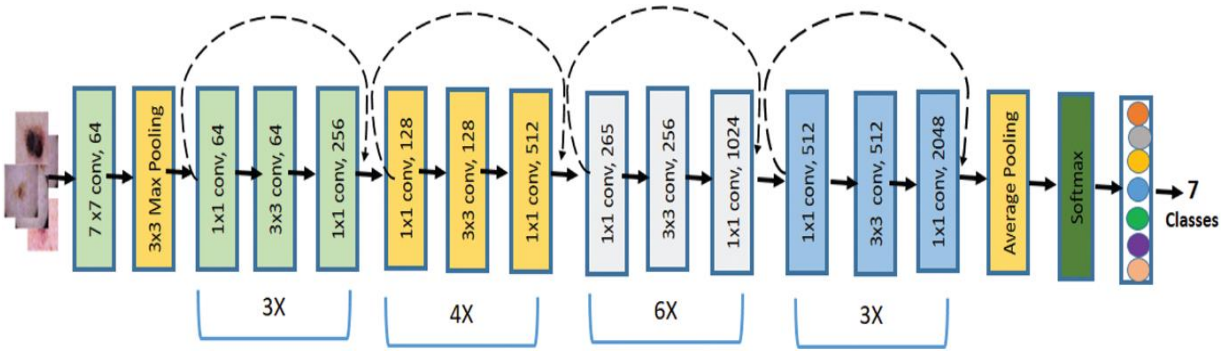

Fig 2.4: *ResNet Architecture Diagram*

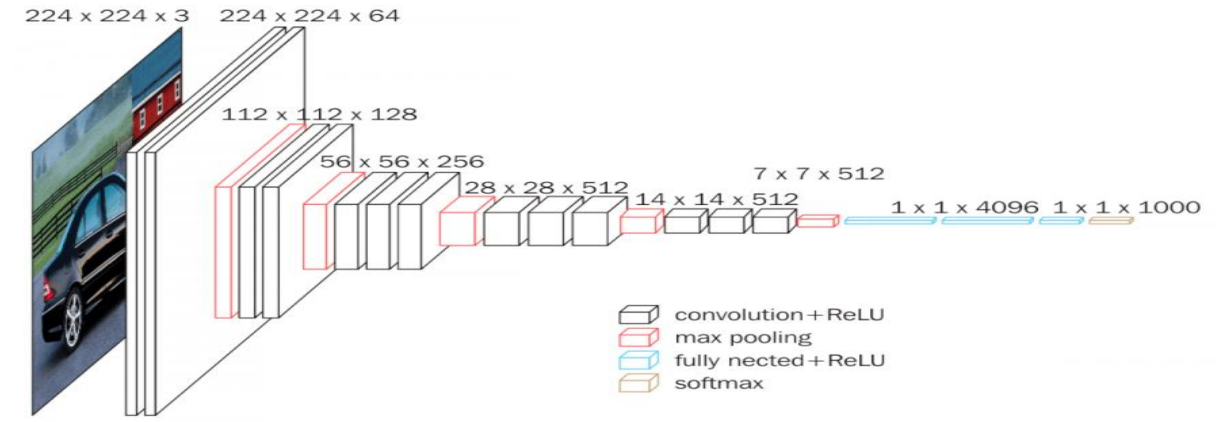

Fig 2.5: *VGG Architecture Diagram*

2.5. RDM Generation

The next step was to create representational dissimilarity matrices (RDMs) for the features obtained at various layers of the CNNs. We used Pearson Correlation Coefficient (PCC) as a metric to compare two different activation vectors. Suppose that the value of PCC for given pair of features is 'r', then the value filled in the RDM was 1-r. This is because the matrix is a measure of the dissimilarity of the elements, whereas, the PCC value indicates the similarity between two features.

2.6. Similarity Analysis

Comparing artificial models and human brains is not an easy task, because of several points of differences between them, one of them being the dimensionality issue. To overcome this, we used a technique called Representational Similarity Analysis (RSA). It relies on the fact that brains and models are equivalent if they treat the same images as similar (or equivalently dissimilar). To perform this analysis, we first obtained RDMs as mentioned in the steps above. The advantage of RDMs is that, for different signal spaces, they have the same dimensions, and hence, are easy to compare. The RDMs can be compared by calculating their similarity using Spearman R. The final score is obtained by squaring the result to R^2, which indicates the amount of variance.

We don't expect the ideal unknown model to have a perfect R^2 score of 1, because of the presence of noise in the data. Hence, to normalize the obtained score, we assume that the subject-averaged RDM is the best model (ideal) which we can obtain. Hence, we calculate the R^2 score value for the subject-averaged RDM, and consider it as the upper bound (noise ceiling) of the final score. Other R^2 values are normalized with respect to this noise ceiling, and hence the final normalized score is termed as Noise Normalized R^2.

3. Results and Discussion

We obtained RDMs for various layers of the Convolutional Neural Networks. These RDMs are compared with the given RDMs for EVC and IT fMRI signals. The fMRI signals were recorded for 15 subjects, hence the obtained RDMs are compared with the RDMs of all the 15 subjects. The average of the 15 correlation values is obtained for both EVC and IT RDMs, which are then normalized with respect to their corresponding noise ceiling values. The obtained normalized scores have been tabulated in the third [EVC] and fourth [IT] column of Table 3.1 given below. To obtain the final average score, the obtained correlation scores (before normalization) for EVC and IT are averaged, and then normalized with respect to the overall average noise ceiling value. The obtained final average scores have been tabulated in the fifth [Avg] column of Table 3.1.

Network	Layer	Noise Normalized R^2		
		EVC	IT	Avg
AlexNet	conv1	6.844	1.253	3.157
AlexNet	conv2	16.793	7.106	10.406
AlexNet	conv3	**17.853**	7.136	10.787
AlexNet	conv4	13.983	**10.662**	**11.793**
AlexNet	conv5	8.029	4.344	5.600
ResNet	block1	8.588	1.533	3.937
ResNet	block2	10.803	1.281	4.525
ResNet	block3	7.448	1.552	3.561
ResNet	block4	12.449	10.018	10.846
VGG	maxpool1	7.235	1.403	3.390
VGG	maxpool2	10.029	1.916	4.680
VGG	maxpool3	14.333	4.758	8.020
VGG	maxpool4	15.746	7.478	10.295
VGG	maxpool5	9.503	5.034	6.556

Table 3.1: *Correlation scores for different models and layers*

The results for the individual CNNs have been plotted in the figures given below, and also the best scores obtained for different networks have been compared.

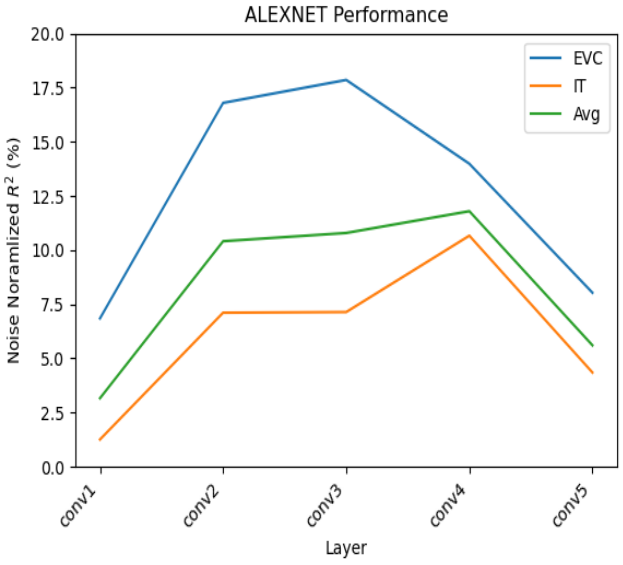

Fig 3.1: *AlexNet Performance*

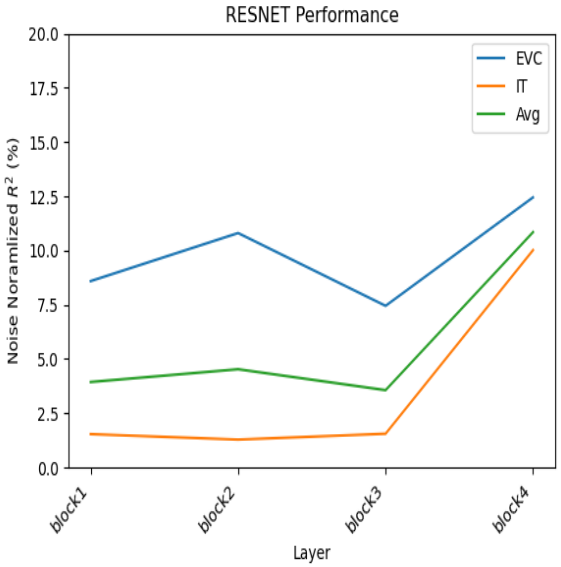

Fig 3.2: *ResNet Performance*

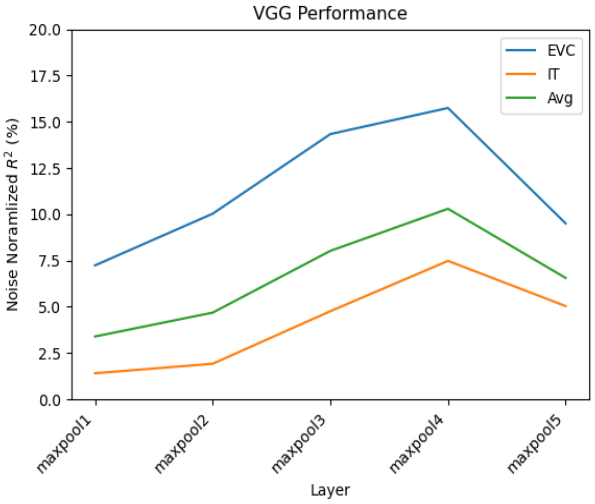

Fig 3.3: *VGG Performance*

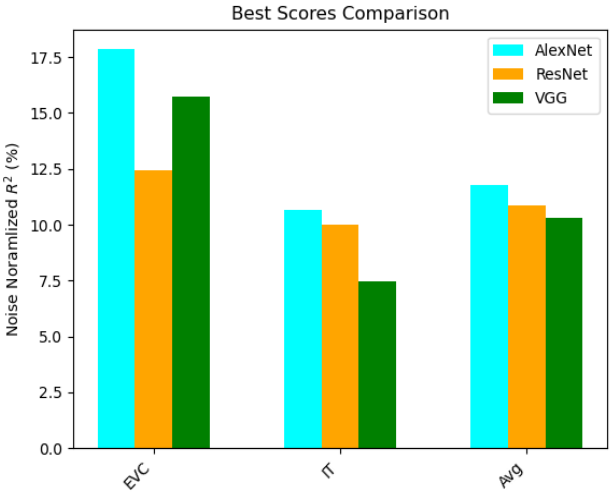

Fig 3.4: *Overall Best Performance Comparison*

AlexNet gave us the best results among the three CNNs. We observe that as we move deeper into the neural networks, the predicted activities get more and more similar to the actual neural activities (except the final layers). To analyse the results further, we plotted the dissimilarity matrix obtained for the 4th convolutional layer of AlexNet (as it gave us the best result). The matrix has been shown in the figure below.

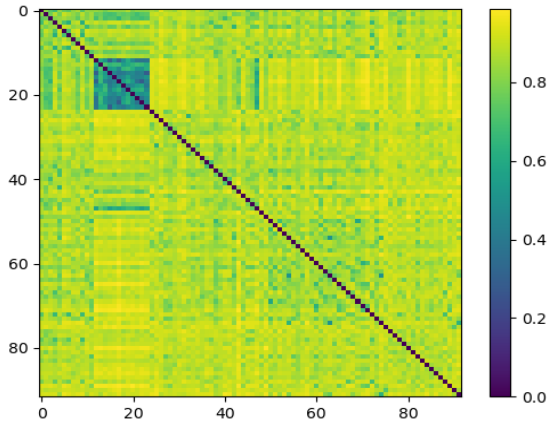

Fig 3.5: *Dissimilarity matrix obtained for the 4th convolutional layer of AlexNet*

We observe that the matrix signifies high similarity among image numbers 12-25 (approx.). As per the dataset, theses images were those of human faces, and hence they had similar shapes. The objects belonging to other classes did not have similar shapes. This explains why the scores obtained for the EVC data is higher as compared to IT. Neurons in early visual cortex (EVC) mostly respond to simple visual features like shapes, edges, etc. whereas neurons in IT respond to more large and complex features like object parts. Hence, we observe that the CNNs were not able to find as much similarity among different objects, as found by the fMRI signals. As compared to other signals like EEG or MEG, fMRI has a low Signal to Noise (SNR) ratio. This could be one of the possible reasons for the low scores.

4. Conclusion

We used some convolutional neural networks to model the hierarchical structure of the human brain visual system. Even though the results were not very promising, but still they were enough to signify that after some fine tuning and further developments in the network architecture, these convolutional neural networks would be able to accurately predict the neural activities of human brain in response to visual stimuli. To model the late visual cortex (or the inferior temporal cortex) we need some more advancements in the architecture of the networks. These advancements would be highly beneficial for the field of computer vision and artificial intelligence, because CNNs would then be able to analyse and classify visual data in a similar way as the human brain does.

5. References

- Abd ElGhany, Sameh et al. “Diagnosis of Various Skin Cancer Lesions Based on Fine-Tuned ResNet50 Deep Network.” Cmc-computers Materials & Continua 68 (2021): 117-135.
- Agrawal, Aakash. “Dissimilarity learning via Siamese network predicts brain imaging data.” arXiv: Neurons and Cognition (2019): n. pag.
- Hemmer, Martin & Khang, Huynh Van & Robbersmyr, Kjell & Waag, Tor & Meyer, Thomas. (2018). Fault Classification of Axial and Radial Roller Bearings Using Transfer Learning through a Pretrained Convolutional Neural Network. Designs. 2. 56. 10.3390/designs2040056.
- Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol 10(11): e1003915.
- Kriegeskorte, N. (2008). Representational similarity analysis connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2, 4.
- Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends in Cognitive Sciences, 17, 401-412.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in neural information processing systems 25 (p. 1097-1105).
- Loukadakis, Manolis & Cano, José & O'Boyle, Michael. (2018). Accelerating Deep Neural Networks on Low Power Heterogeneous Architectures.
- Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. (2019). The Algonauts Project: A Platform for Communication between the Sciences of Biological and Artificial Intelligence. arXiv, arXiv:1905.05675.
- Radoslaw M. Cichy, Dimitrios Pantazis and Aude Oliva. (2016). Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in

Human Cortex During Visual Object Recognition. Cerebral Cortex, 26 (8): 3563-3579.

- Yalda Mohsenzadeh, Caitlin Mullin, Benjamin Lahner, Radoslaw Martin Cichy, and Aude Oliva. (2019). Reliability and Generalizability of Similarity-Based Fusion of MEG and fMRI Data in Human Ventral and Dorsal Visual Streams. Vision, 3(1), 8.
- Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111: 8619–8624.