A study of the impact of CNN architecture variation on predicting brain activity using feature-weighted receptive fields

Vlad Murgoci

Supervisor: Xucong Zhang

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Vlad Murgoci
Final project course: CSE3000 Research Project
Thesis committee: Xucong Zhang, Nergis Tömen

Abstract

This study investigates the relationship between deep learning models and the human brain, specifically focusing on the prediction of brain activity in response to static visual stimuli using functional magnetic resonance imaging (fMRI). By leveraging intermediate outputs of pre-trained convolutional neural networks (CNNs) with feature-weighted receptive fields, it becomes possible to estimate brain activity in the visual cortex. The primary objective of this research is to analyze how different CNN architectures affect the accuracy of predicting brain activity. To accomplish this, we utilize the novel fMRI Natural Scenes Dataset, which provides a large-scale data set for comprehensive analysis. Through this investigation, we aim to gain insights into the impact of CNN architectures on the prediction accuracy of brain activity in the context of visual stimuli.

Keywords

Neural Activity, Brain Modeling, Convolutional Neural Networks, Visual Stimulus, Deep Learning, fMRI.

1 Introduction

Brain activity prediction based on external stimuli has made significant advances, primarily due to improvements in machine learning and functional magnetic resonance imaging (fMRI) technology, as presented in the surveys of Cao et al. [1] and Chen et al. [2]. There have been multiple encoding techniques proposed, such as Bidirectional Deep Generative Networks (Changde et al. [3]), Gabor wavelet filter-based networks (Yibo et al. [4], [5]), and semantic models (Naselaris et al. [6]).

Accurate prediction of brain activity from visual stimuli enables scientists to test and validate new theories about brain function, investigate the role of specific brain areas in visual processing, study the development of visual perception throughout one's lifespan, and understand the contribution of different neural pathways to various visual functions (e.g. Henderson et al. [7]).

The recent emergence of large-scale datasets, such as the Natural Scenes Dataset [1] (NSD) of Kay et al. in [8], presents a new opportunity for more extensive research in predicting brain activity in response to visual stimuli. This research paper aims to train and evaluate brain activity encoding models based on convolutional neural networks on the NSD, using feature-weighted receptive fields, proposed by St-Yves and Naselaris in [9], a task that has not yet been attempted due to the novelty of the dataset.

Kriegeskorte and Nikolaus [10] have shown that there are similar internal representations between convolutional neural networks (CNNs) and the human brain. St-Yves and Naselaris [9] have proposed a brain activity prediction technique that exploits the feature maps of intermediate CNN layers, the feature-weighted receptive field (fwRF). fwRF demonstrates high accuracy in predicting brain activity in a voxel (volumetric pixel, cubic section of the brain) based manner and has been used in the encoding part of the state-of-the-art image synthesis NeuroGen model (Gu et al. [11]), which is capable of generating images that achieve predetermined brain activations when presented to a subject.

There is an opportunity to investigate how model architecture variation affects performance in predicting brain activity in visual regions of interest within the visual cortex. Studying how the variation of CNN model architecture affects the performance of the fwRF framework could help us gain further insight into the links between CNNs and the human brain.

In this research paper, we aim to address these gaps by training and comparing fwRFs using the large-scale NSD dataset which has not been thoroughly investigated in previous studies that used more compact datasets such as vim-1 (Kay et al. [12]), BOLD5000 (Chang et al. [13]), and GOD (Horikawa and Kamitani [14]). The expected outcome is a clear view of the performance of various models trained on NSD, under various experimental settings.

The rest of the paper is structured as follows: Section 2 offers background information regarding the core topics. Section 3 presents the used methodology, which involves leveraging convolutional neural networks (CNNs) with feature-weighted receptive fields. Section 4 provides an overview of the experimental setup and details the utilization of the large-scale fMRI NSD dataset. The results of our analysis, including the performance of five chosen CNN architectures in predicting brain activity, are presented in Section 5. The results are discussed in Section 6. Throughout the study, we prioritize responsible research practices, as highlighted in Section 7, and finally, in Section 8, we conclude the paper by summarizing our findings and outlining potential avenues for future work.

2 Background

2.1 Functional Magnetic Resonance Imaging (fMRI) and BOLD Signals

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive imaging technique that allows for the observation and measurement of brain activity. It works by detecting changes in blood flow to different parts of the brain, which is indicative of neural activity in those regions. The primary measure used in fMRI is the blood oxygenation level-dependent (BOLD) signal. This signal is based on the fact that oxygenated and deoxygenated blood have different magnetic properties. When a particular brain region is active, there is an increased demand for oxygen, leading to a change in the BOLD signal. This change can be detected and used to infer brain activity.

2.2 Natural Scenes Dataset (NSD)

The Natural Scenes Dataset (NSD) [8] is a large-scale dataset that contains fMRI data from 8 subjects viewing a variety of natural scenes. The dataset is unique in its size and scope,

[1] https://naturalscenesdataset.org/

providing a rich resource for investigating the relationship between visual stimuli and brain activity. The NSD includes both the stimuli (images of natural scenes) and the corresponding brain activity (measured using fMRI), allowing for a comprehensive analysis of the relationship between these two elements. The use of natural scenes as stimuli is particularly relevant for studying visual perception, as these stimuli closely resemble the types of visual input that the human visual system is designed to process.

In this study, we leverage the NSD to train and evaluate our models, providing a robust test of their ability to predict brain activity in response to visual stimuli.

2.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep learning models that have proven to be highly effective in tasks related to image processing and recognition. They are designed to automatically and adaptively learn spatial hierarchies of features from images. CNNs are generally composed of one or more convolutional layers, followed by one or more fully connected layers.

The architecture of the CNN is inspired by the structure and behavior of the visual cortex of the human brain. This similarity can be used to uncover hidden functionalities of the brain, in a non-invasive way. The CNNs can be dissected, layer by layer, as they tend to capture visual features in a hierarchical manner, early layers capture low-level features such as shapes and edges with certain orientations, while latter layers can extract faces, objects, and other complex features.

The CNNs used in our study are pre-trained on the large scale ImageNet [2] dataset (Deng et al. [15]) and have achieved high classification accuracies.

2.4 Feature-Weighted Receptive Field (fwRF)

The feature-Weighted receptive field (fwRF) [9] is an encoding model used to find the relationship between the features extracted by a CNN and brain activity (stimulus-to-voxel model). The fwRF assumes that the response of a neuron (or voxel in fMRI) can be predicted by a weighted sum of the features in a particular region of the input space (the receptive field). The weights are learned from the data, allowing the model to determine which features are most relevant for predicting the response of each neuron or voxel, as it can be seen in Figures 1 and 2.

The feature pooling field is modeled as an isotropic 2D Gaussian blob as in Equation 1, where $\mu = (\mu_x, \mu_y)$ is the feature pooling field center and σ_g is the radius of the feature pooling field.

$$g(x, y; \mu_x, \mu_y, \sigma_g) = \frac{1}{\sqrt{2\pi\sigma_g}} \exp\left[-\frac{(x-\mu_x)^2 + (y-\mu_y)^2}{2\sigma_g^2}\right] \tag{1}$$

The response $\hat{r_t}$ of a voxel to an image S_t is modeled as follows:

$$\hat{r_t} = \sum_k^K w_k \int_{-D/2}^{D/2} \int_{-D/2}^{D/2} g(x, y, \mu_x, \mu_y, \sigma_g) \phi^k_{i(x)j(y)}(S_t) dx dy \tag{2}$$

where D is the perceived visual angle - a measure of the pooling field's size compared to the whole image, and $i(x), j(y)$ represent the discretization of spatial coordinates into pixel indices.

2.5 Visual Cortex

The visual cortex of the brain is a part of the cerebral cortex that processes visual information. It is located in the occipital lobe of the primary cerebral cortex, at the back of the brain. Visual information from the eyes is sent to the primary visual cortex via the optic nerve and the lateral geniculate nucleus in the thalamus. The visual cortex is hierarchically organized into several areas which play different roles in processing aspects of vision. These areas include the primary visual cortex (V1), secondary visual cortex (V2), tertiary visual cortex (V3), and the fourth human visual field map (hV4).

V1

V1, the primary visual cortex, is the first area that processes incoming visual input in the brain. V1 is one of the best-understood brain areas, and it's known to execute basic operations, such as simple filtering to enhance contours and edges.

V2

V2, the secondary visual cortex, receives strong feedforward connections from V1. It continues the processing of visual information initiated in V1, handling features such as the orientation of illusory contours, more complex shapes, and color constancy.

V3

The tertiary visual cortex is part of the dorsal stream of the visual system. Much is not known about the function of V3, but it is suspected that it is involved in handling spatial location, motion, and colors.

hV4

The exact role of area V4 is still under debate, but it is probably involved in recognizing shapes, and it appears to be essential for perceiving selective colors. hV4 sends outputs to areas involved in face and scene recognition.

3 Methodology

3.1 Experimental Methodology

In order to investigate how the variation of architecture affects prediction performance, 5 different CNN architectures have been chosen. The choice of the architectures was limited by training time, and the tight schedule of the research plan, but should provide a general perspective for the research question.

The CNNs are used to extract feature maps from the NSD visual stimulus. The images used in the NSD experiment

[2] https://image-net.org/

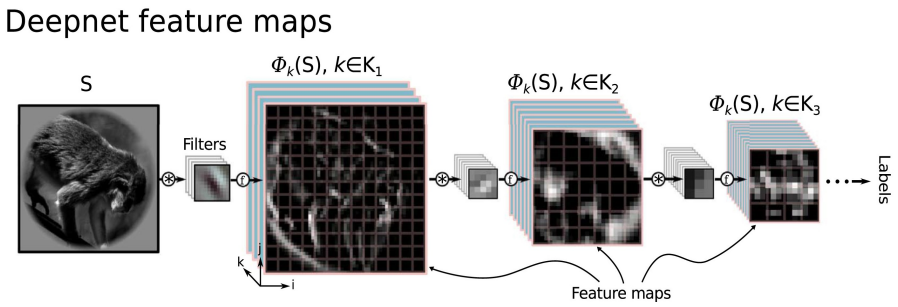


Figure 1: After input S is fed into the neural network, each stacked feature map $\phi_k(S), k \in K_i$ is extracted from layer K_i. Adapted from [9] in NeuroImage, 2018, Volume 180, Part A. Available at https://www.sciencedirect.com/journal/neuroimage/vol/180/part/PA. Licensed under Elsevier.

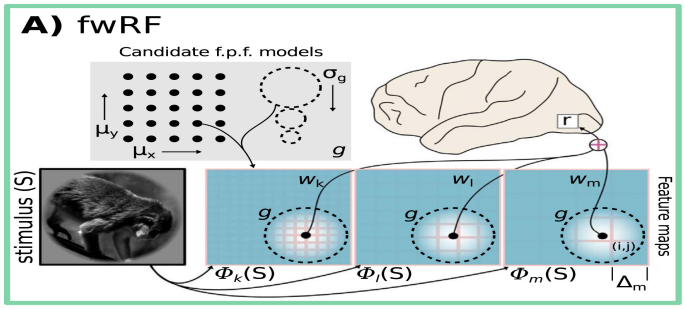


Figure 2: A representation of the fwRF, predicting the activation of a single voxel r in response to the stimulus S. After the feature maps $\phi_{K\,i}$ are extracted from the model inference using stimulus S, they are filtered by a 2D Gaussian feature pooling field, g, that is selected by conducting grid search based on the position of the center - μ_x, μ_y - and the radius of the pooling field σ_g. The hyperparameters are the same for each feature map. The output of each pooling operation, for each feature map ϕ_i is multiplied by a learned weight w_i and summed up to produce a prediction of the activity in voxel r. Adapted from [9] in NeuroImage, 2018, Volume 180, Part A. Available at https://www.sciencedirect.com/journal/neuroimage/vol/180/part/PA. Licensed under Elsevier.

are the 2017 train/val subset of the COCO dataset (Lin et al. [16]). The difference between training data and inference data used during the experiment (ImageNet and COCO respectively) should not be a topic of concern because CNNs are able to generalize features and adapt to unseen data. Also, the datasets are similar, both ImageNet and COCO contain images of everyday scenes, common objects, and persons.

These features are then used to train the Feature-Weighted Receptive Field (fwRF) encoding model [9]. The fwRF model is a powerful tool for predicting voxel-wise brain activity, particularly within the visual Region of Interest (ROI) of the human brain.

The training process involves fitting the fwRF model to the features extracted by the CNNs during inference using NSD samples. The performance of the model is then evaluated based on its ability to accurately predict brain activity in the visual ROIs.

This methodology allows us to systematically assess the impact of different model architectures on the performance of brain activity prediction. By comparing the performance of different pre-trained models, we can gain insights into how the choice of model architecture influences prediction accuracy.

In [9] the fwRF model was fitted to feature maps of a standard CNN architecture, with basic convolutional layers. This research aims to investigate both standard and more complex architectures and their impact on the accuracy of the fwRF.

3.2 Model selection

AlexNet (Krizhevsky et al. [17]) and **VGG-13** (Simonyan and Zisserman [18]) possess classical CNN architectures and provide decent accuracy on the ImageNet dataset. Their straightforward structure serves as a baseline for the experiment.

GoogleNet (Szegedy et al. [19]) was chosen due to its Inception architecture that mimics the brain's sparse connectivity between neurons by using 1x1 convolutional filters, connected to a small subset of the inputs. Another similarity with the brain is the multi-scale processing of the input in the inception blocks. Convolutions with filters of size 1x1, 3x3, and 5x5 allow the network to learn and combine features of different scales.

ResNet-18 (He et al. [20]) was selected because it is known for its deep structure and "shortcut" or "skip connections" that pass on output from early layers to distant, later layers, a technique used to address the problem of vanishing/exploding gradients. One could draw a loose parallel between the brain's ability to rewire itself after suffering partial lesions as demonstrated by Johansen-Berg in [21] and the skip connections that bypass regions of the network. The con-

	Top-5 Accuracy	Top-1 Accuracy
AlexNet	79.066%	56.522%
EfficientNetV2_S	96.878%	84.228%
GoogleNet	89.53%	69.778%
ResNet18	89.078%	69.758%
VGG-13	89.246%	69.928%

Table 1: Top-5 and Top-1 accuracy percentages of models on the ImageNet-1K dataset.

	PCC
GoogleNet	0.3
VGG-13	0.293
ResNet18	0.291
AlexNet	0.288
EfficientNetV2_S	0.279

Table 2: Averaged Pearson correlation scores for each model.

nection of one layer to several other layers, through skip and normal connections, resembles the multiple feedforward connection of early visual ROIs with other layers (V2 has feedforward connections to both V3 and hV4).

Finally, the experiment will make use of a more performant model, namely the small version of **EfficientNetV2** proposed by Tan and Le in [22]. The model manages to keep low computational complexity while delivering high classification performance, as can be seen in Table 1, by leveraging progressive training and layers of specialized blocks called fused-MBConv blocks.

Table 1 presents the Top-5 and Top-1 accuracy of each network on the widely known ImageNet-1K dataset [15]. EfficientNet achieves the greatest accuracy, followed by Resnet, GoogleNet, VGG and finally AlexNet.

4 Experimental setup

4.1 Data Collection and Preparation

The fMRI and stimulus data are extracted from the NSD dataset. The dataset offers 40 stimulus trials for each subject. During each trial, a subject is presented with 750 stimulus images, and fMRI data is collected based on BOLD signals (Blood-oxygen-level-dependent).

The used BOLD signals are prepared as 1.8mm volume betas (difference in percentage between the resting state of the voxel) and denoised using a GLM-denoise technique proposed by Kay et al. in [23].

For each subject, brain ROI masks were prepared, which select only the voxels in the V1, V2, V3, and hV4 and exclude the rest. Masks are prepared individually since the brain structures of the subjects differ. On average, each mask includes 4000 voxels. Figure 3 illustrates both the complete brain mask of a subject, representing the overall shape of the brain (left), and the region of interest (ROI) mask specifically highlighting the voxels within the visual cortex. Each fMRI scan is partitioned into an average of 80 vertical slices (varies depending on the subject), with each slice corresponding to a specific height within the brain.

4.2 Model Preparation and Evaluation

Our comparison will be done for AlexNet [17], VGG-13 [18], ResNet-18 [20], GoogleNet [19] and EfficientNetV2-S [22].

Each selected CNN is pre-trained on the ImageNet dataset. In order to extract both high and low-level features, we chose to extract feature maps from early, middle, and late model layers. To maintain consistency between models, 6 or 7 layers were selected for feature extraction, with similar distributions in terms of position in the network: 2 early layers, 2 or 3 middle layers, and 2 late layers. [9] does not offer a detailed heuristic for layer selection, but indicates that there should be a uniform distribution in terms of hierarchical location. For each layer, we also filtered out feature maps that showed lower variance, with a maximum number of feature maps per layer of 512.

After conducting feature map filtering and selection, we generated a grid of candidate parameters for the fwRF, by varying center locations and radii. Grid search is then done for 393 candidate models. We execute grid search separately for each subject, with varying training samples per subject. We select the best model from the pool by selecting the ones that achieve the lowest root mean squared error (RMSE). Each model is evaluated on a validation set comprised of 2000 held-out samples.

The Pearson Correlation Coefficient (PCC) is a preferred metric for evaluating model performance due to its ability to capture the linear relationship between variables and because it is used as an accuracy metric in multiple papers that discuss encoding models [9], [24], [25], [26]. The PCC provides a measure of the strength and direction of the linear association between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), while a PCC of 0 indicates no correlation. The PCC's ability to capture linear relationships, interpretability, its computational efficiency and common adoption among the similar scientific writings makes it a preferred metric for evaluating model performance in the current setting.

5 Results

5.1 General Performance

The obtained results exhibit a clear and concise pattern, which allows for a succinct presentation in the results section. One particular model demonstrates superior performance across all evaluated areas, while the remaining models can be also ranked in a straight-forward hierarchical fashion.

Using the Pearson Correlation Coefficient (PCC) as a measure of performance for our models, the average performance of the fwRF model increases by at most 7.5% (worst model compared to best). Table 1 presents mean correlation coefficients for each model, averaged over each subject. Each model was trained on 4000 samples per subject.

Googlenet presents the best performance across all subjects and all testing setups. VGG and Resnet follow with a difference of 2-3% accuracy, while AlexNet and EfficientNet (especially) present the worst performance. The Googlenet-based encoding model seems to be able to converge faster than the other models to a point of diminishing returns.

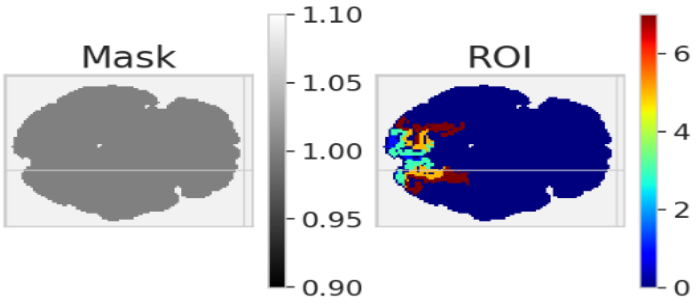


Figure 3: Brain (left) and ROI (right) masks for one slice. The ROI mask portrays voxels in different regions of the visual cortex with different colors.

5.2 Performance across different subjects

The multisubject experiment was conducted with a training set of 4000 samples for each subject, with a grid search of 393 candidate models for each voxel. A bigger training size was not chosen due to the previously mentioned time constraints.

The performance across subjects varies in terms of absolute value from subject to subject, but it is consistent with our previous results. Googlenet outperforms all the other models, for each subject. The best accuracies are achieved when fitting subjects 1, 2 and 3, while subjects 6 and particularly 8 exhibit low accuracy.

Figure 5 portrays the scores of a resampling with replacement validation test. The validation input data is ran through the model, and mean correlation coefficients are computed for 64 resampling runs. This is a procedure that analyzes the stability of correlation coefficients to data variability. The distribution of mean PCC for the resampling runs follows a normal distribution, with a mean standard deviation of approximately 0.00928 for each model, which verifies the validity of our results.

5.3 Model performance variation based on training size

In order to assess the performance evolution under different training sizes. Models were trained with varying training sizes from 100 to 8900 samples, for subject 1. Figure 5 depicts the performance of each model (left), and the percentual advantage of GoogleNet over each model (right). The models seem to converge when using 4000 or more training samples, while GoogleNet constantly outperforms the other models. AlexNet and EfficientNet demonstrate similar performance for bigger training sizes, but AlexNet has an edge on training sizes lower than 4000. VGG and ResNet are within 2-3% of GoogleNet for higher training sizes.

We can observe that GoogleNet has better performance for intermediate training sizes, 2000-4000 samples, where it outperforms every network by 3-8%, while still achieving a good performance-to-training-set-size ratio.

5.4 ROI performance

Figure 6 depicts the performance of each model for different regions of interest (ROI). GoogleNet is the best-performing model in every ROI. The general model performance suggests that models capture early layers features better than the latter layers, with the mean PCC decreasing from 0.399 for V1 to 0.266 for hV4.

EfficientNet struggles to learn voxel activations for the latter layers when the training sizes are low to medium, and the performance for larger training sizes suggests that the model is catching up in performance.

All models experience a drop in accuracy for the V3 ROI, when trained with 5000 training samples. This is a curious observation since the other ROIs do not present a drop in PCC. This may be caused by isolated overfitting, high data variability for V3, or a skewed sample distribution for a training size of 5000 samples.

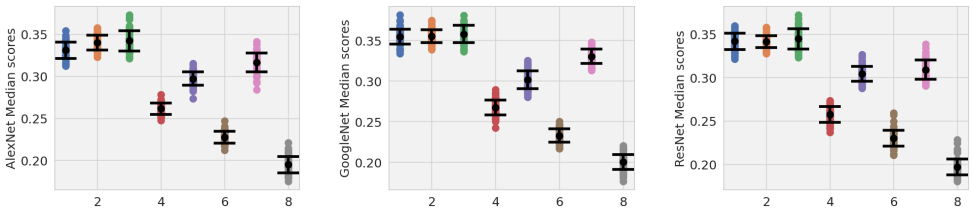
5.5 Bigger Grid Search

In order to test the potential performance of the models we trained the models on a bigger grid of 1414 candidate models. The selection of models was enlarged by halving the minimum distance between the possible centers of the Gaussian pooling function. Besides VGG, each network experiences an increase in performance of approximately 1.7%. This suggests that there is performance to be found by performing more detailed grid searches, that fine-tune the position of the pooling field's center.

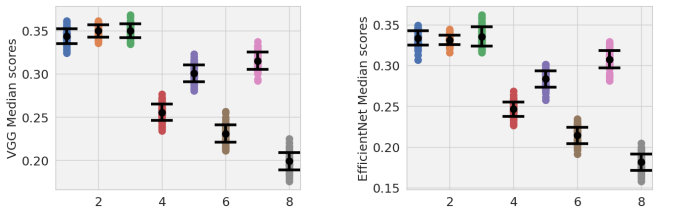
6 Discussion

[9] does not offer exact prediction accuracies for the DeepNet-fwRF, therefore we cannot directly compare the results of our experiments with the original paper's results. [9] compares the accuracy of the deepnet-based model against a Gabor-wavelet-based fwRF and against a layer-wise Deepnet regression model. The main observations are that the model outperforms the other two, especially in latter visual ROIs, and that it achieves state-of-the-art performance. [9] uses a smaller training and validation set, more specifically, 1750 training samples and 120 validation samples.

The performance for each ROI is in line with the community consensus about the limitations of CNN-based encoding models. CNNs are known to struggle with the prediction of high-level visual regions. As a solution to this problem, Xiao et al. [26] proposed fitting a DNN on data from early visual cortex layers, in order to predict activations in the latter layers. The stimulus-to-voxel encoding model that uses the CNN's feature maps is leveraged for its high accuracy on early layers, and the voxel-to-voxel DNN model is fed



(a) AlexNet CC scores (b) GoogleNet CC scores (c) ResNet CC scores



(d) VGG-13 CC scores (e) EfficientNetV2_S CC scores

Figure 4: PCC scores for each model, based on subject. The black dots indicate the mean PCC achieved over 64 resampling runs.

with the predictions of the fwRF in order to predict activity in high-level layers.

GoogleNet's performance edge could be justified by its unique inception architecture, which has a biological resemblance to early visual cortex areas. Firstly, the Inception architecture adopts a multi-scale processing approach, by using filters of different sizes ,on the same layer, 1x1, 3x3, and 5x5, similar to how the early visual cortex processes visual information across different visual scales. Inception modules in the Googlenet model use 1x1 convolutions to reduce the dimensionality of feature maps before applying larger convolutions. This sparse connectivity mimics the observed neural connectivity in the early visual cortex, where neurons have local receptive fields and selectively connect to nearby neurons. The Inception architecture also incorporates parallel convolutional pathways with different receptive field sizes. This parallel processing strategy is reminiscent of the parallel pathways observed in the early visual cortex, such as the magnocellular and parvocellular pathways. These parallel pathways specialize in processing different visual features, such as motion and color, respectively. Similarly, the parallel convolutional pathways in Googlenet can capture different types of visual information and facilitate the model's ability to extract diverse and meaningful features.

Even though EfficientNet has a similar architecture to VGG, it still cannot reach the same level of performance. The PCC of EfficientNet, the model with the highest accuracy on ImageNet, represents a solid argument for the use of brain-optimized neural networks to the detriment of goal-optimized neural networks, as proposed by St-Yves et al. [25], where the underlying neural network takes part in joint training with the feature pooling field. Since the classification accuracy of the network appears to be of no direct use to the encoding models, this allows us to tweak the parameters in intermediate layers toward a voxel-focused prediction.

The research could be improved by testing the performance of more CNN architectures, especially of more performant models because they can help identify what is the bottleneck of the ensemble, the model, or the fwRF technique.

7 Responsible Research

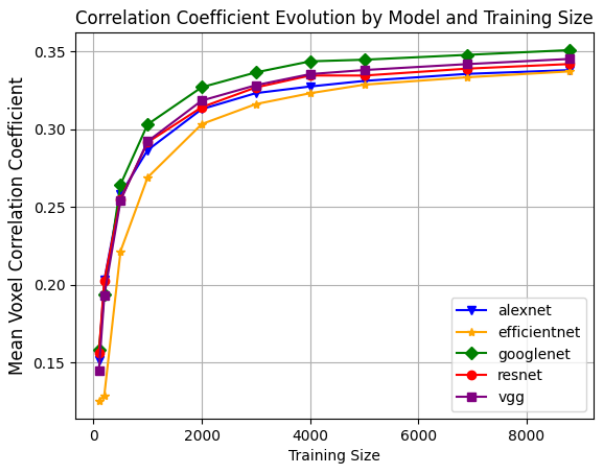
This segment is dedicated to discussing the ethical principles adhered to in this project, with a particular emphasis on the ability to reproduce the methods and outcomes. The aspect of reproducibility gains heightened significance in studies involving comparisons; a simplified replication process enables researchers to assess a variety of detectors and generate a larger volume of results with more ease.

7.1 Ethical implications

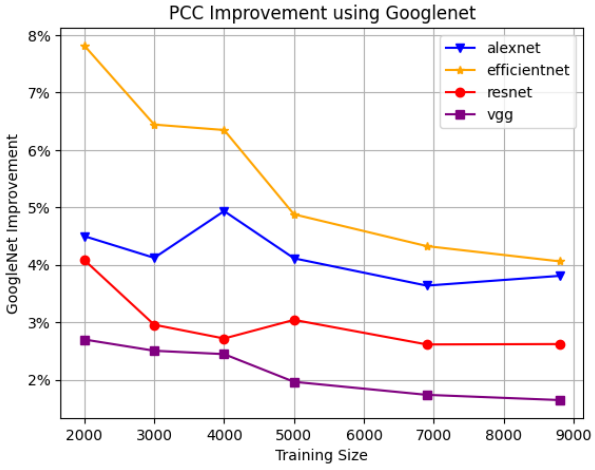
In this research, we have taken great care to ensure that our work is conducted in an ethically responsible manner. The use of machine learning models to predict brain activity has profound implications, not only for the field of neuroscience but also for society at large.

The potential to predict and understand brain activity could lead to significant advancements in medical diagnostics and treatments, particularly for neurological disorders. However, it also raises important ethical questions about privacy and consent. As we move forward with this research, we must ensure that any applications of our work respect individuals' rights to privacy and are used in a manner that benefits society. Furthermore, the use of machine learning models in neuroscience research also raises questions about algorithmic bias and fairness. It is crucial to ensure that our models do not perpetuate existing biases in the data or in the pre-trained models.

Therefore, the anonymity of the subjects that consent to share their biometric data should be kept throughout experimental incursions. Models trained to predict the activity of a subject's brain response to visual stimuli must not be used for commercial use or monetary gain. Fortunately, since every person has different brain reactions to the same visual stimuli, our encoding model is not able to accurately generalize



(a) PCC change based on training size



(b) GoogleNet advantage compared to other models.

Figure 5: PCC scores for each model trained on subject 1 fMRI data, with varying training sizes (Left) and the difference in accuracy (in percentage) between AlexNet, EfficientNet, ResNet, and VGG compared to GoogleNet, where the Y axis represents the increase in performance by using GoogleNet compared to a specific model (Right).

predictions to unknown persons.

7.2 Reproducibility of the results

In order to guarantee the reproducibility of the results, a GitHub repository [3] has been made available with the notebooks used for data preparation, model training, and evaluation. The repository is based on another GitHub repository [4] created by the author of the fwRF paper [9] and the CNN models are adapted from the Pytorch torchvision repository [5]. The data used during the experiment is available on an Amazon S3 bucket, but access must be requested through the NSD website [6].

8 Conclusion and Future Work

This study has provided an investigation into the impact of Convolutional Neural Network (CNN) architecture variation on predicting brain activity using feature-weighted receptive fields. Our research leveraged intermediate outputs of five pre-trained CNNs - AlexNet [17], EfficientNetV2_S [22], GoogleNet [19], ResNet-18 [20], and VGG13 [18] - with feature-weighted receptive fields to estimate brain activity in the ventral stream in response to static visual stimuli. The primary objective was to analyze how different CNN architectures affect the accuracy of predicting brain activity.

Our findings suggest that the performance of the encoding models can be influenced by the architecture of the CNNs used. The Inception architecture used by GoogleNet provided the most promising results, outperforming the other models in every ROI. This opens up a new avenue for improving the performance of such models by testing more state-of-the-art CNN architectures, and improved versions of the ones we tested (e.g. Inception v2, Inception v3, ResNet-50, VGG16). It also raises important questions about the nature of the bottlenecks in the ensemble, the model, or the fwRF method, which warrant further investigation.

Another significant discovery is the low prediction accuracy of the EfficientNet model, the one with the highest classification accuracy on ImageNet-1K [15], indicating that the high image classification accuracy of an architecture does not necessarily translate to additional performance for the corresponding brain encoding model.

The coupling of deep learning feature extraction, the fwRF model, and other mathematical, medical, and behavioral techniques could present novel methods of non-invasive brain activity analysis. This could have far-reaching implications, benefiting both the neurological and technological worlds.

Looking ahead, future research should focus on how the performance varies based on the subject and region of interest. It would be particularly interesting to further explore whether the fwRF is able to predict voxel activity better in the latter parts of the visual cortex that process high-level features and whether the answer is mainly influenced by the quality of the feature maps from corresponding layers in the network, or by the method itself.

In terms of reproducibility, we have made every effort to ensure that our methods and results can be replicated by other researchers. To this end, we have provided a GitHub repository with the Python notebooks used for data preparation, model training, and evaluation. The data used during the experiment is available on an Amazon S3 bucket, with access available upon request.

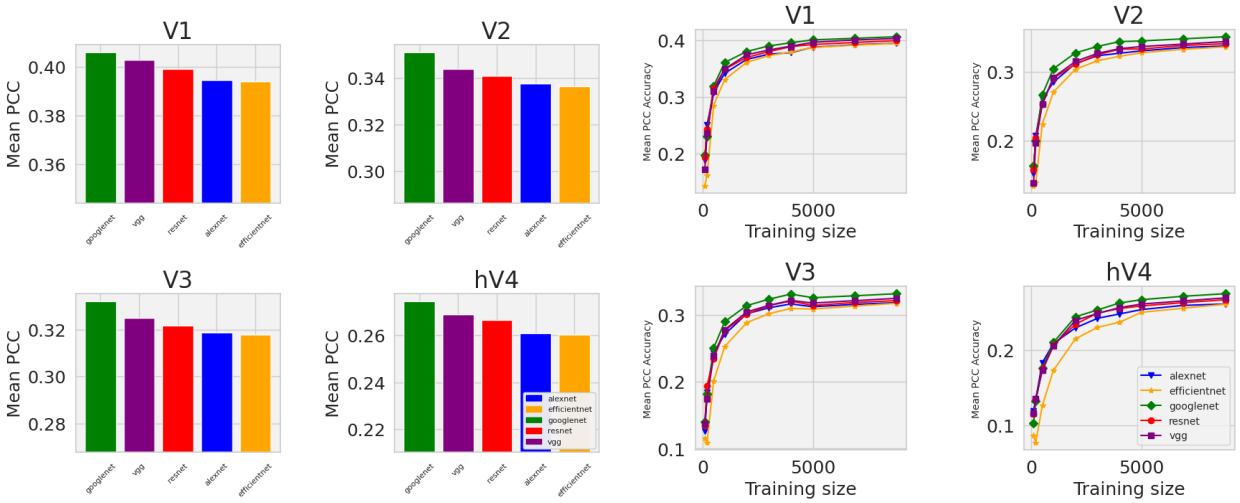
In conclusion, this study has shed light on the potential of CNN architectures to influence the accuracy of predicting brain activity. The findings pave the way for future research

[3] https://github.com/VladMurgoci/BrainEncoding

[4] https://github.com/styvesg/nsd/tree/master

[5] https://github.com/pytorch/vision/tree/main/torchvision/models

[6] https://naturalscenesdataset.org/



(a) Mean PCC by ROI and model

(b) PCC evolution by ROI and model

Figure 6: ROI PCC scores for each model, based on subject.

to further optimize these models and expand their application to other areas of the brain.

References

[1] L. Cao, D. Huang, and Y. Zhang, "When computational representation meets neuroscience: A survey on brain encoding and decoding," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 4339–4347, International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track.

[2] M. Chen, J. Han, X. Hu, X. Jiang, L. Guo, and T. Liu, "Survey of encoding and decoding of visual stimulus via fmri: an image analysis perspective," *Brain Imaging and Behavior*, vol. 8, pp. 7–23, Mar 2014.

[3] C. Du, J. Li, L. Huang, and H. He, "Brain encoding and decoding in fmri with bidirectional deep generative models," *Engineering*, vol. 5, no. 5, p. 948 – 953, 2019. All Open Access, Gold Open Access.

[4] Y. Cui, K. Qiao, C. Zhang, L. Wang, B. Yan, and L. Tong, "Gabornet visual encoding: A lightweight region-based visual encoding model with good expressiveness and biological interpretability," *Frontiers in Neuroscience*, vol. 15, 2021. All Open Access, Gold Open Access, Green Open Access.

[5] Y. Cui, C. Zhang, L. Wang, B. Yan, and L. Tong, "Dense-gwp: An improved primary visual encoding model based on dense gabor features," *Journal of Mechanics in Medicine and Biology*, vol. 21, no. 5, 2021. All Open Access, Hybrid Gold Open Access.

[6] T. Naselaris, D. E. Stansbury, and J. L. Gallant, "Cortical representation of animate and inanimate objects in complex natural scenes," *Journal of Physiology-Paris*, vol. 106, no. 5, pp. 239–249, 2012. New trends in neurogeometrical approaches to the brain and mind problem.

[7] M. M. Henderson, M. J. Tarr, and L. Wehbe, "A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex," *Journal of Neuroscience*, vol. 43, no. 22, pp. 4144–4161, 2023.

[8] E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. Kay, "A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence," *Nature Neuroscience*, vol. 25, pp. 116–126, Jan 2022.

[9] G. St-Yves and T. Naselaris, "The feature-weighted receptive field: an interpretable encoding model for complex feature spaces," *NeuroImage*, vol. 180, pp. 188–202, 2018. New advances in encoding and decoding of brain signals.

[10] N. Kriegeskorte, "Deep neural networks: A new framework for modeling biological vision and brain information processing," *Annual Review of Vision Science*, vol. 1, no. 1, pp. 417–446, 2015. PMID: 28532370.

[11] Z. Gu, K. W. Jamison, M. Khosla, E. J. Allen, Y. Wu, G. St-Yves, T. Naselaris, K. Kay, M. R. Sabuncu, and A. Kuceyeski, "Neurogen: Activation optimized image synthesis for discovery neuroscience," *NeuroImage*, vol. 247, p. 118812, 2022.

[12] M. Lescroart, K. Kay, T. Naselaris, R. Prenger, M. Oliver, and J. Gallant, "fmri of human visual areas in response to natural images," 2011.

[13] N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, and E. M. Aminoff, "BOLD5000, a public fMRI dataset while viewing 5000 visual images," *Sci Data*, vol. 6, p. 49, May 2019.

[14] T. Horikawa and Y. Kamitani, ""generic object decoding (fmri on imagenet)"," 2019.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[16] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[21] H. Johansen-Berg, "Structural plasticity: Rewiring the brain," *Current Biology*, vol. 17, no. 4, pp. R141–R144, 2007.

[22] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," *CoRR*, vol. abs/2104.00298, 2021.

[23] K. Kay, A. Rokem, J. Winawer, R. Dougherty, and B. Wandell, "Glmdenoise: a fast, automated technique for denoising task-based fmri data," *Frontiers in Neuroscience*, vol. 7, 2013.

[24] H. Wang, L. Huang, C. Du, D. Li, B. Wang, and H. He, "Neural encoding for human visual cortex with deep neural networks learning "what" and "where"," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 4, pp. 827–840, 2021.

[25] G. St-Yves, E. J. Allen, Y. Wu, K. Kay, and T. Naselaris, "Brain-optimized neural networks learn non-hierarchical models of representation in human visual cortex," *bioRxiv*, 2022.

[26] W. Xiao, J. Li, C. Zhang, L. Wang, P. Chen, Z. Yu, L. Tong, and B. Yan, "High-level visual encoding model framework with hierarchical ventral stream-optimized neural networks," *Brain Sciences*, vol. 12, no. 8, 2022. All Open Access, Gold Open Access, Green Open Access.