Simulating IT Cortex Face Responsivity Through Deep Learning Models

Lisa Gavronskiy
Mathew Maradin
University of Waterloo
April 23rd, 2025

Abstract

The inferior temporal cortex (IT) an important role in object and face recognition, with neurons exhibiting selective responses to specific facial features and body parts [1]. Motivated by neurophysiological experiments showing that IT neurons respond strongly to intact faces but weakly to objects and partial or scrambled faces. Specifically, macaque monkey experiments that examined neuronal responses in the IT to facial versus non-facial object stimuli and manipulation of facial stimuli by selectively removing or blurring features, such as the eyes, leads to assess the impact on neuronal responses [2,3]. This study investigates how convolution neural networks (CNNs) replicate these responses. The CNNs of increasing complexity were implemented in TensorFlow: a shallow network for edge detection, an intermediate model for mid-level feature learning, and a deep residual network for high-level abstraction. Each model was trained on a dataset containing human faces and other object categories (e.g., bikes, horses, flowers), and then evaluated on both clean and feature-distorted versions of the facial images. Accuracy levels and internal activation maps were analyzed, interpreting those as neuronal activity and facial recognition. Results demonstrated that the shallow CNN failed to distinguish faces from non-faces, with accuracy hovering around 50% and high variability. The intermediate CNN outperformed both others, showing strong edge-based recognition and robustness to blur, while the deeper residual CNN excelled under partial facial feature removal but was more prone to overfitting. Activation maps and probabilistic outputs revealed processing patterns that align with hierarchical visual processing in the brain, particularly the importance of eye and nose regions, consistent with IT findings from previously done experimental work. These findings support the hypothesis that hierarchical CNNs mirror biological vision mechanisms and provide a useful framework for bridging machine learning with computational neuroscience. Future work for this project can involve biologically-informed CNN designs that map onto specific vision processing areas (e.g., V1, V2, V4) using receptive field properties, and test using more precise facial feature manipulations (e.g., isolating or removing only eyes, nose, or mouth) to gain a deeper understanding into facial stimuli.

1 Introduction

Understanding mechanisms behind facial recognition has been an increasing focus in computational neuroscience. The IT contains neurons that are selective to visual stimuli, including faces and facial parts, and damage to this region can impair the ability of facial recognition [4]. Previous neuroscience

studies have shown that removing facial features in experimental images (e.g., eyes or mouth) leads to a reduced firing rate in IT neurons, demonstrating sensitivity to facial configurations [4]. Modeling these neural responses helps in understanding how changes in brain function can affect human perception and cognition.

The IT plays a crucial role in the hierarchical processing of visual information. In the initial stages, the lateral geniculate nucleus (LGN) receives input from ganglion cells in the retina, which preserves the spatial organization of the visual field [4]. This information is then transmitted to the primary visual cortex (V1), where simple cells detect basic features such as edges and orientations [4]. As visual signals progress, from V1 to areas like V2 and V4, neurons integrate features over larger spatial areas. This integration enables the encoding of more complex visual information, including textures and object parts, for more advanced object recognition [4]. Neural networks, particularly CNNs, are models that are extensively used computational neuroscience, which mirrors biological visual processing pathways. Convolution neural networks can detect features in images, such as lines and edges, analogous to simple cells in the V1. As data is passed through multiple layers of convolution filters, models are able to detect more complex features, similar to the hierarchical process of the vision system. Additionally, each filter in a convolution layer acts on a certain region of the input image, for example, one filter might detect a horizontal line while another detects vertical lines in the center of an image. Likewise, the visual system's receptive fields are sensitive to different regions in the field of view.

In previous studies on facial stimuli in macaque monkeys using fMRI to monitor blood flow, it was found that 97% of visually responsive neurons in a specific region of the IT were face-selective [2]. This conclusion was drawn by presenting the monkeys with images of faces, non-face objects (such as clocks and fruits), and scrambled images, and analyzing the neural responses., revealing that IT neurons responded exclusively to faces [2]. In another fMRI study, monkeys were shown faces followed by the same faces with the eye region selectively removed. The removal or distortion of the eyes led to a significant reduction in neuronal activity in the IT cortex, indicating that specific facial features play a critical role in generating responses from face-selective neurons [3]. This project aims to evaluate how CNNS of increasing complexity replicate or differ from the feature sensitivity observed in IT neurons, and how alterations to facial features or face images more broadly affect the responsiveness of these neurons. The output of the models, including classification accuracy, activation maps, and convolution filters, are analyzed to represent neural firing patterns and compare them to studies on facial feature sensitivity in monkeys, as well as to build upon those findings.

2 Methods

2.1 Model Architecture

Three CNN models were constructed in TensorFlow, each designed to reflect increasing levels of visual abstraction. In developing the three models for this project, the primary strategy involved progressively increasing model complexity. The process began with the implementation of a baseline model, followed by a linear increase in complexity while maintaining the same architecture. Finally, the architecture was transitioned to a more advanced residual network. At each stage, the goal was to enhance feature learning and gradient stability.

The simplest network (later referred to as CNN1) represents the baseline for testing, it's a shallow model consisting of two convolution layers with ReLU activations and max pooling. This network simulates early visual processing, focusing on edge and texture detection. Following this, the intermediate network (later referred to as CNN2) expands upon the first network by scaling each parameter of the network, as it includes more convolution layers and dropout layers after pooling to mitigate overfitting. Additionally, batch normalization was applied after each convolution to stabilize training and help prevent vanishing or exploding gradients [5]. This network aims to target mid-level feature representations. Finally, the most complicated network (later referred to as CNN3) alters the model architecture to a residual network. This model was chosen to avoid the exploding/vanishing gradients that are commonly observed when linearly scaling the layers of a network. This architecture

employs residual blocks to enable deeper learning without degradation [6]. It reflects the functional complexity observed in deeper layers in the vision processing system and maintains gradient flow during backpropagation [6]. This structure aligns with the brain's method of processing visual information, in which initial stages focus on simple features and subsequent stages focus on features with increasing complexity.

2.2 Data Preprocessing

The dataset used in this project includes two classes: human faces and non-facial objects such as bicycles, flowers, and horses, as shown in Figure 1. To reduce computational requirements and standardize the input format, all images were converted to grayscale and resized to 60×60 pixels, since the original images were significantly larger and would require substantial processing power. The original dataset was imbalanced, containing many more non-facial images than human faces. To address this, an equal number of samples from each class were selected, ensuring balanced representation during training and testing to improve model performance. Additionally, only 50% of the full dataset was used to limit resource usage. The final sample distribution consisted of 523 human images and 536 other images in the training set, and 138 human images and 126 other images in the testing set. From this, training and testing data loaders were created with a batch size of 100.

Figure 1: Fifteen randomly selected examples images from the preprocessed dataset, showcasing images from both the human face and non-face classes.

2.3 Experimental Design

Each network was trained for 10 iterations, with a training accuracy recorded after each iteration and a testing accuracy computed at the end of the 10 iterations. The primary goal of this project is to investigate how various modifications to facial images affect visual recognition performance. To carry out this analysis, two types of image alterations were introduced. First, a 16-pixel-wide white stripe was applied across the images both horizontally and vertically, shifting in 10-pixel steps for a total of 13 positions, as seen in Figures 2 and 3. This approach allows for analysis of which regions of the image are most critical for face recognition. Second, a Gaussian blur was applied to the images using five progressively increasing blur levels, as seen in Figure 4. This helps assess whether the models rely on fine-grained facial features or more general shapes and structures for accurate classification.

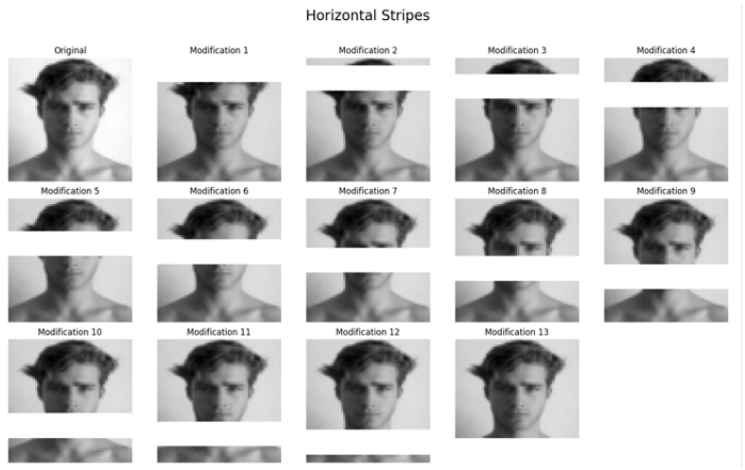

Figure 2: Original unmodified image, followed by 13 modified versions with a 16-pixel-wide horizontal stripe progressively removed moving down across the face with a 10-pixel-wide step.

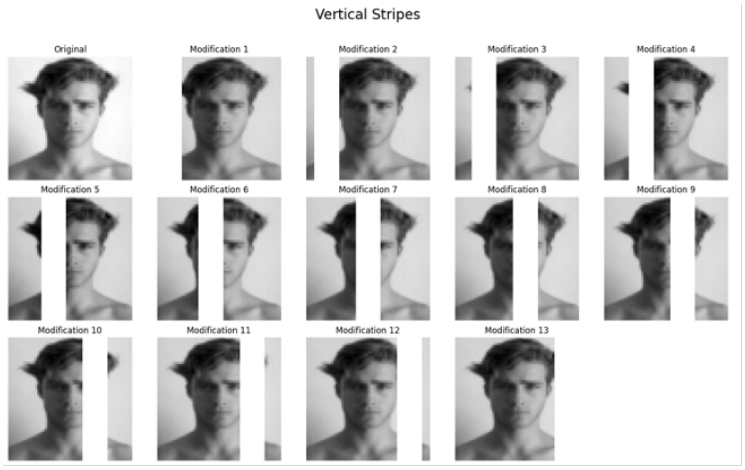

Figure 3: Original unmodified image, followed by 13 modified versions with a 16-pixel-wide vertical stripe progressively removed moving left to right across the face with a 10-pixel-wide step.

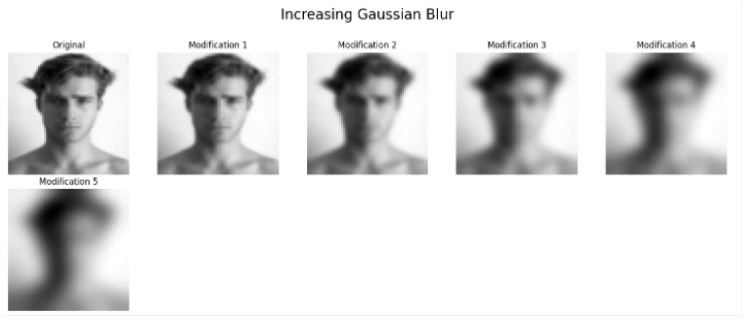

Figure 4: Original unmodified image, followed by 5 modified versions with an increasing applied Gaussian blur.

2.4 Model Evaluation

To evaluate overall model performance, classification accuracy was calculated by measuring the proportion of correctly predicted images relative to the total number of samples. A 5-fold cross-validation was conducted to ensure statistical reliability in the models. In this setup, the dataset is split into five equally sized subsets; in each fold, four subsets (80%) are used for training while the remaining one (20%) is used for testing. Each fold uses a different subset for testing, ensuring that all data points are used exactly once for evaluation and never repeated, which leads to a more reliable and unbiased estimate of the model's performance. This method provides a more comprehensive evaluation by mitigating variance due to a particular train-test split. Running the model across five folds allows for the computation of statistical measures such as the average accuracy and standard deviation across both training and testing tests [7]. During testing, the gradient tracking was disabled to ensure that the model's parameters were not updated while making predictions on the data.

The same k-fold approach was also applied to the modified images. In each fold, the subset reserved for testing was modified using the stripe and blur techniques to evaluate how these alterations impact the model's performance. This performance analysis provides insight into the models, simulating the biological visual system, for example, understanding which image regions or levels of detail are essential for recognition can reflect how human perception prioritizes visual information. Additionally, comparing accuracies between the models helps to identify which ones are more biologically accurate and allows for an investigation into the specific features that contribute to this accuracy.

In addition to classification accuracy, probabilistic output values were tracked for a single sample image across all modifications. This deeper analysis goes beyond binary correctness by observing the confidence levels of the model's predictions, which may be biologically analogous to varying degrees of neuronal activation in the brain's visual pathways. To gain further insight into feature processing, the final activation maps from each convolution filter were analyzed for the selected image. These activation maps visualize how different filters respond to various features in the input, helping to identify what the network has learned to detect, such as edges, textures, or certain facial features. Furthermore, the convolution filters themselves were also examined to determine whether they were capturing meaningful visual patterns.

3 Results

When analyzing the training accuracy over 10 iterations for the three networks, CNN1 shows the lowest accuracy, which remains around 50% throughout training, with minimal improvement. In contrast, both CNN2 and CNN3 demonstrate significant increases in accuracy over the iterations, with CNN2 slightly outperforming CNN3 at each step, as shown in Figure 5. Additionally, CNN1 exhibits the highest variability in performance when evaluated using the k-fold testing method. In terms of testing performance, a similar pattern is observed (as seen in Figure 6): CNN1 exhibits the greatest variability and the lowest performance, while CNN2 and CNN3 perform much better. Notably, CNN2 achieves the highest testing accuracy, reaching the high 80s.

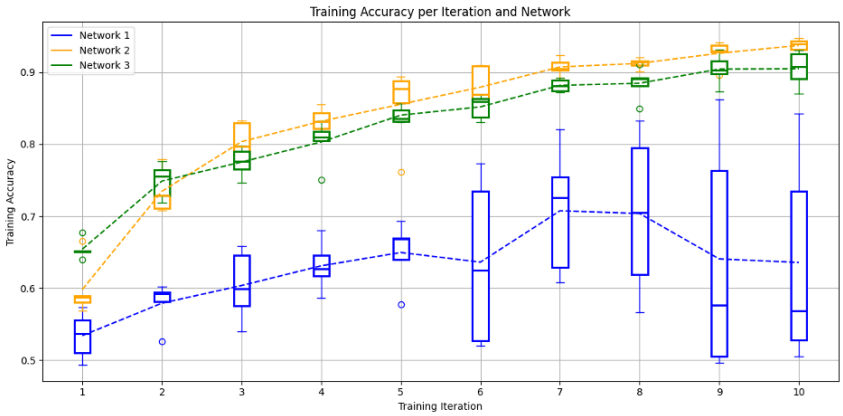

Figure 5: Training accuracy over 10 iterations for CNN1, CNN2, and CNN3 (labeled as Network 1, 2, and 3), illustrating the performance progression of each model during training.

When analyzing the training accuracy over 10 iterations for the three networks, CNN1 shows the lowest accuracy, remaining around 50% throughout training with minimal improvement. This can be attributed to the simplicity of the model architecture; CNN1 is weak and unable to effectively differentiate between faces and non-faces, causing it to produce a 50/50 output and struggle with classification. In contrast, both CNN2 and CNN3 show significant increases in accuracy over the iterations, with CNN2 slightly outperforming CNN3 at each step, as shown in Figure 5. The improved performance of CNN2 can be attributed to its increased complexity and additional layers, allowing it to better capture features in the data. Additionally, CNN1 exhibits the highest variability in performance when evaluated using the k-fold testing method. In terms of testing performance, a similar pattern is observed (as seen in Figure 6): CNN1 shows the greatest variability and poorest performance, while CNN2 and CNN3 perform much better. Notably, CNN2 achieves the highest testing accuracy, reaching the high 80s. While CNN3 is significantly more complex and theoretically should perform better, the overfitting to irrelevant patterns or noise in the data undermines its performance, demonstrating the trade-offs associated with deeper architectures.

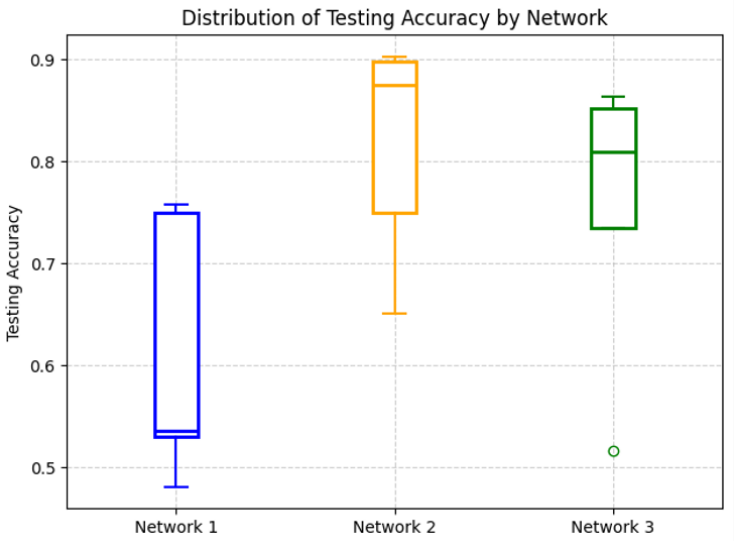

Figure 6: Testing accuracy for CNN1, CNN2, and CNN3 (labeled as Network 1, 2, and 3), illustrating the difference in model performance.

To further analyze how image modifications affect network performance, a single representative image was selected for a detailed examination of the activation maps corresponding to each type of modification, as shown in Figures 2, 3, and 4. In addition to the activation maps, the output probabilities were computed for each modification to simulate neuronal firing rates. Figure 7 visualizes the model's output probability across all modifications, plotted in the order shown in Figures 2-4, and includes the baseline average testing accuracy for unmodified images. Consistently, CNN1 hovers around 50% confidence across all modifications, further supporting the fact that it is unable to effectively distinguish between faces and non-faces.

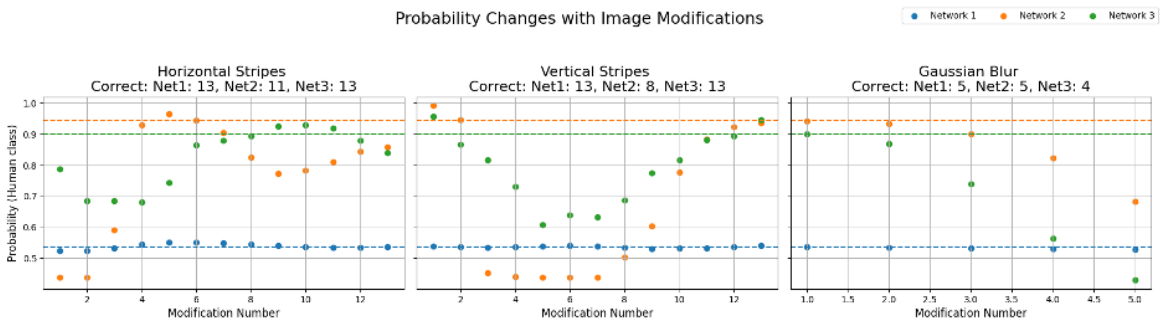

Figure 7: Model output probabilities for a single representative image across all modifications (horizontal stripes, vertical stripes, and Gaussian blur), shown in the same order as Figures 2–4. The baseline average testing accuracy for unmodified images is represented by the dashed line for each model.

3.1 Horizontal Stripes

In the case of horizontal stripe modifications, CNN2 and CNN3 show a drop in performance around modifications 2–4, which correspond to the eye region, a critical area for face recognition. Interestingly, in the subsequent modifications (4–6), targeting the eye-nose region, CNN3 begins to outperform CNN2. This suggests that CNN3, with its deeper and more complex architecture, is able to extract higher-level facial features and can partially compensate when key regions are removed, unlike CNN2 which relies more heavily on mid-level features.

3.2 Vertical Stripes

A similar pattern is observed with vertical stripe modifications. Performance begins to degrade significantly once the stripe intersects the left eye region. This drop in accuracy persists as the occlusion moves across the facial midline. Despite this degradation, CNN3 consistently performs better than CNN1 and CNN2, indicating that its deeper convolution layers and hierarchical feature extraction allow it to recognize facial structures even when presented in distorted or partially occluded forms. This robustness is likely due to its ability to integrate complex features from across the entire face rather than relying on specific regions.

3.3 Gaussian Blur

When Gaussian blur is applied, both CNN2 and CNN3 exhibit decreasing accuracy as the blur intensity increases. However, CNN2 consistently outperforms CNN3 under these conditions. This may be due to CNN2's emphasis on lower- to mid-level features such as edges and lines, which remain somewhat visible under mild blurring. In contrast, CNN3's performance is lower, likely because of the reliance on fine grained, high-level details, which are significantly degraded by blurring, resulting in the model to be less effective when such detail is lost.

This experiment, done for the single image discussed above, was extended across the entire test dataset using image modifications under the k-fold cross-validation framework. The models, trained solely on unmodified data, exhibited consistent trends parallel to those observed in Figure 7 when tested on modified data. As shown in Appendix B, the results across all modified test images. through iterations of horizontal and vertical stripe removal as well as increasing levels of Gaussian blur, emphasizing the patterns described earlier with more statistical significance.

3.4 Activation Map Analysis

Figure 8 displays the activation maps generated from an unmodified image, providing insight into the types of features each network is detecting. These observations align with earlier statements regarding the hierarchical feature extraction of each model. In CNN2, particularly in the earlier convolution layers, strong activation is visible along the facial outline, seen as brighter (or "hotter") regions, indicating effective edge detection and responsiveness to the general shape of the face. In contrast, CNN1 exhibits less focused activation. The absence of clearly defined brighter regions suggests that the network is responding to less informative features. While the most active regions are located within the facial area, implying that it is attending to the main object in the image, the lack of structure suggests weak feature extraction capabilities. Moreover, the final convolution output of both CNN1 and CNN2 appears highly diffuse and lacks spatial clarity, illustrating their limited ability to capture and preserve detailed local features as data moves deeper into the network. On the other hand, CNN3 shows distinct and well-structured activation maps in its final convolution layers. These maps highlight specific facial features with much greater precision, demonstrating CNN3's superior capacity for complex feature extraction. This supports the conclusion that CNN3 is better equipped for fine-grained facial recognition due to its deeper architecture and residual connections, which help retain feature information throughout the network.

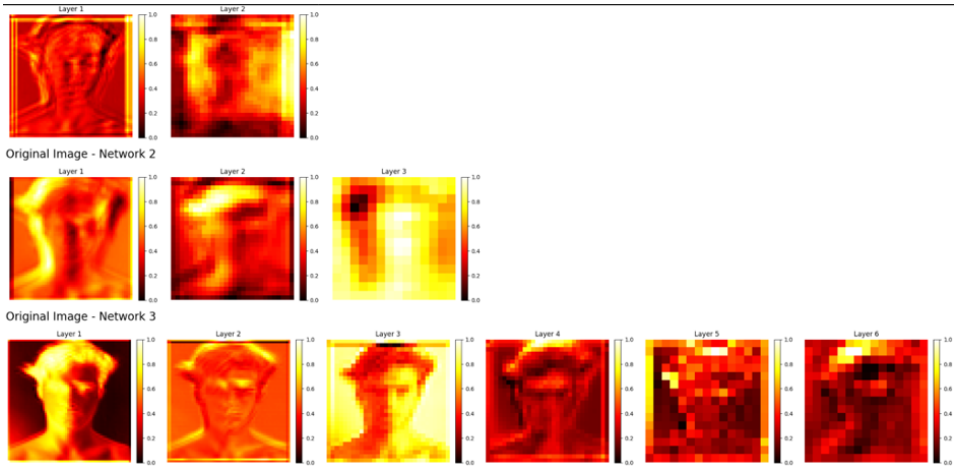

Figure 8: Final activation maps from each convolution layer in CNN1 (top), CNN2 (middle), and CNN3 (bottom) for the selected unmodified sample image.

The remaining activation maps generated from the modified images are presented in Appendix A. These maps exhibit patterns consistent with the findings discussed above, reinforcing the conclusions regarding each model's feature extraction capabilities. Activation maps for the vertical stripe modifications were excluded, as they produced results similar to those observed with the horizontal stripe alterations. Visualizations of the learned convolution filters from each network layer were not included in this report, as these filters are less interpretable in isolation, making activation maps a more effective tool for understanding the specific features being detected by the models.

4 Discussion

This project explored the extent to which CNNs replicate the biological mechanisms of visual recognition system observed in the primate inferotemporal (IT) cortex through experiments on marcque monkeys and build on those findings. By comparing networks of increasing architectural complexity (CNN1, CNN2, and CNN3) and evaluating their performance under image manipulations, biology comparisons can be made.

A key finding from these experiments is the inability of CNN1 to distinguish between faces and non-faces, as evidenced by its performance hovering around 50% and high variability across folds. This highlights CNN1's lack of specialization and selectivity, which are traits critical for modeling the visual processing hierarchy in biological systems. Its shallow architecture fails to reflect the staged, hierarchical processing observed in the primate visual cortex, particularly the transition from low-level edge detection in V1 to high-level feature integration in IT cortex. In contrast, both CNN2 and CNN3 demonstrated significantly higher accuracies, with CNN2 slightly outperforming CNN3. Despite its simpler structure, CNN2 appeared to align more closely with the vision system. This may be due to its architectural bias toward generalizable, mid-level features rather than overly complex or fine-grained detail. Such a bias mirrors human perception, where generalization and pattern recognition often take priority over sensitivity to small variations or noise [4]. CNN2's higher performance is most likely a result of the addition of a third convolution layer, which introduces another stage of feature abstraction, as compared to CNN1. The first layer captures changes in contrast, effectively detecting edges, similar to simple cells in area V1 [4]. The later layers then build upon these basic features to extract more meaningful shapes and arrangements, resembling the mid-level visual areas (such as V2/V4) where more complex receptive fields are located [4]. Importantly, the model does not continue abstracting to the point of overfitting, which is observed in CNN3. CNN3's deeper residual structure may allow it to memorize fine details, but this reduces its ability to generalize, which is a quality in biological vision systems.

Furthermore, tracking the probabilistic outputs of the networks offered a deeper understanding than binary accuracy. CNN1's constant 0.5 prediction confidence reaffirms its failure to form strong internal representations. In contrast, CNN2 and CNN3 exhibited varying confidence levels depending on the image region occluded, mimicking variable neural firing rates in response to degraded stimuli [2,3].

Overall, these findings reinforce the idea that human visual systems are highly attuned to facial stimuli, as demonstrated in prior work [2]. Consistent with findings from Issa et. al [3], the removal of key facial features significantly impaired model performance, demonstrating the importance of specific features, particularly the eye and nose region, for facial recognition. The sharp decline in accuracy observed in both CNN2 and CNN3 during horizontal stripe modifications (modifications 2–6, Figure 7) supports this, highlighting the biological relevance of the eye region, which has been shown to strongly activate face-selective neurons in the inferior temporal (IT) cortex [3]. CNN3 showed comparatively better performance when central face regions were occluded, suggesting that it may have developed an understanding of facial features outside of the nose-eye region. This aligns with hierarchical models of the visual system, in which later processing stages encode more integrated and complex features [4]. In contrast, under increasing Gaussian blur, CNN2 consistently outperformed CNN3. This suggests that CNN2 relied more on edge information, features that remain relatively preserved under blur. CNN3, by comparison, appears to have depended more heavily on fine-grained, high-frequency details, which degrade rapidly when blurred. This mirrors how different layers of the visual pathway encode information at varying stages, with earlier layers capturing coarse features and later stages representing a more detailed structure [1].

Summary of Key Findings:

1. Model Performance and Biological Correlations
   (a) Consistently poor performance of CNN1 highlights the limitations of shallow architectures in modeling facial selectivity, unsuccessfully mirroring the IT as differentiate between face and non-face stimuli.
   (b) CNN2's strong performance, even under Gaussian blur, suggests that it relies heavily on mid-level, edge-based features. This reflects the importance of lower frequency spatial information in early stages of biological vision, where neurons are tuned to detect edges and contrasts.
   (c) CNN3's more complex architecture enables hierarchical learning, allowing it to perform better under partial image removal. However, CNN3 is more prone to overfitting, potentially capturing noise rather than meaningful features.
2. Sensitivity to Facial Features: Removal of the eye region (modifications 2–4) caused the most substantial drop in model accuracy, particularly in CNN2 and CNN3. This aligns with findings from the IT cortex, where neurons are especially sensitive to the eye region of faces. Removal of other facial regions did not result in the same level of degradation, demonstrating the importance of the eye region in facial recognition.
3. Probabilistic and Spatial Feature Representations: Probabilistic outputs offers a biologically relevant parallel to neuronal firing rates. In addition, activation maps provided insight into the spatial features each model detected, revealing how deeper networks are able to extract and preserve more meaningful information.

Building on the findings of this project and prior research, future work can further deepen the understanding of facial stimulus processing and the role of specific facial features. One direction involves creating a more biologically grounded mapping of visual processing areas, such as V1, V2, and V4, within CNN architectures by incorporating known receptive field sizes and tuning properties [4]. Additionally, more precise facial feature removal, such as selectively removing or blurring the eyes, nose, or mouth (rather than vertical and horizontal lines) could be used to evaluate their individual contributions to recognition. Examining how these targeted removals impact model output probabilities can represent quantitative neuronal firing rates observed in face-selective regions, enabling more direct comparisons between artificial and biological systems.

Appendices

A Convolution Layer Activation Maps

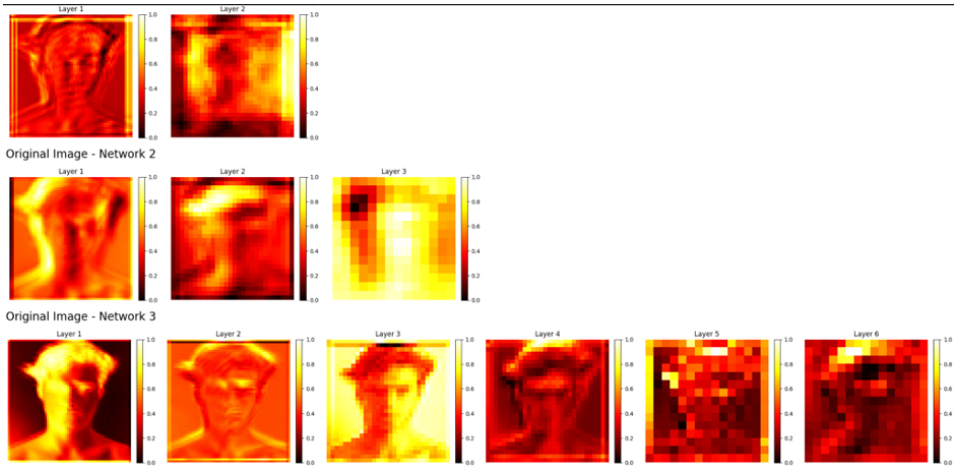

Figure A1: Final activation maps from each convolution layer in CNN1 (top), CNN2 (middle), and CNN3 (bottom) for the selected unmodified sample image.

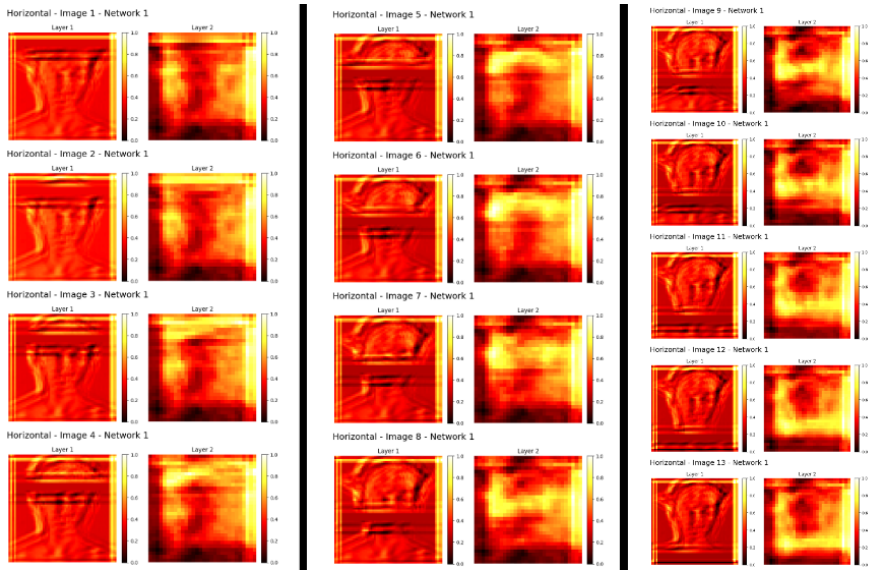

Figure A2: FFinal activation maps from each convolution layer in CNN1 for a modified sample image, where the horizontal stripe position was altered.

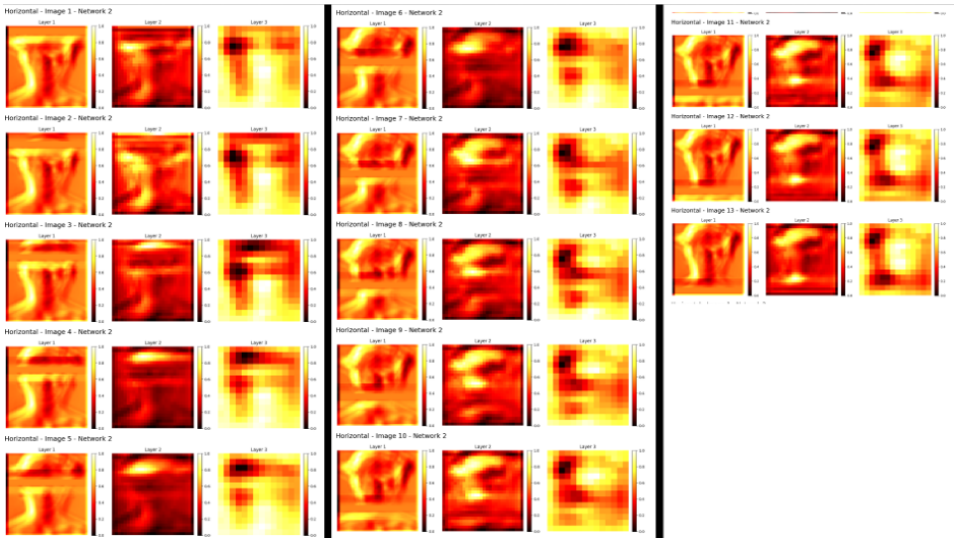

Figure A3: Final activation maps from each convolution layer in CNN2 for a modified sample image, where the horizontal stripe position was altered.

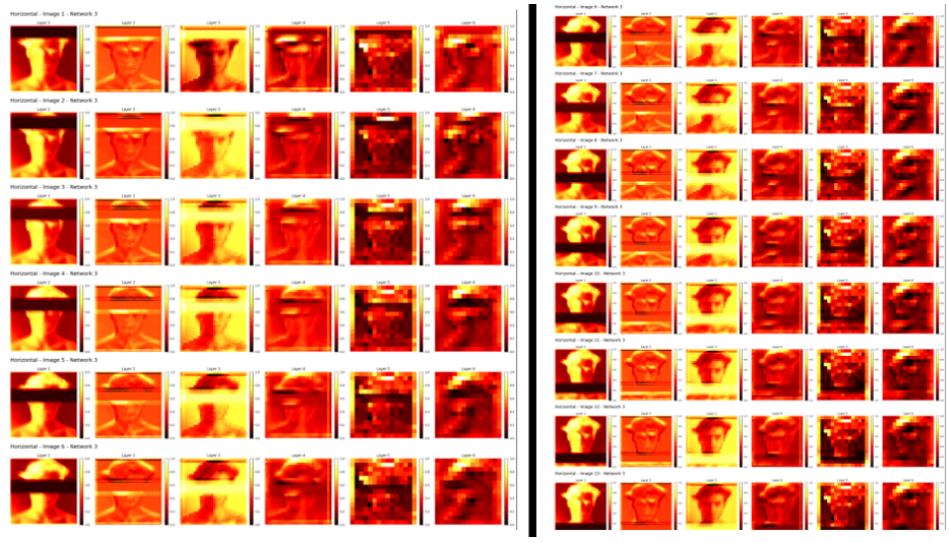

Figure A4: Final activation maps from each convolution layer in CNN3 for a modified sample image, where the horizontal stripe position was altered.

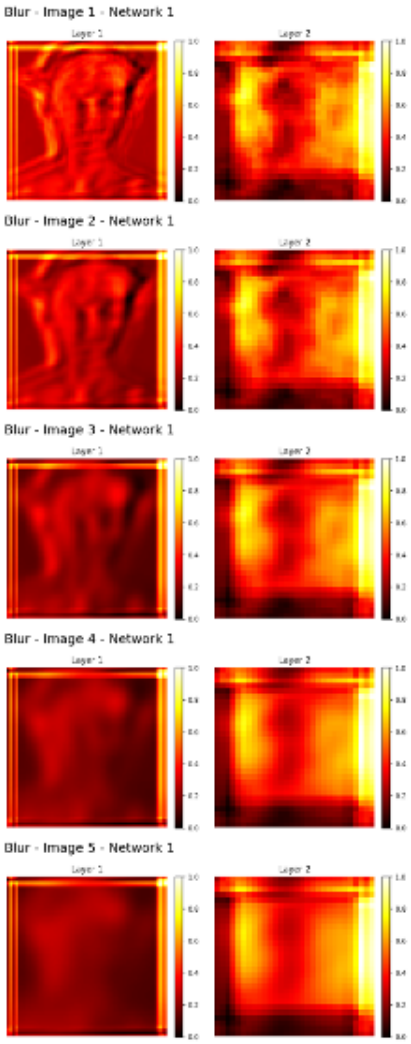

Figure A5: Final activation maps from each convolution layer in CNN1 for a modified sample image, where an increasing Gaussian blur was applied.

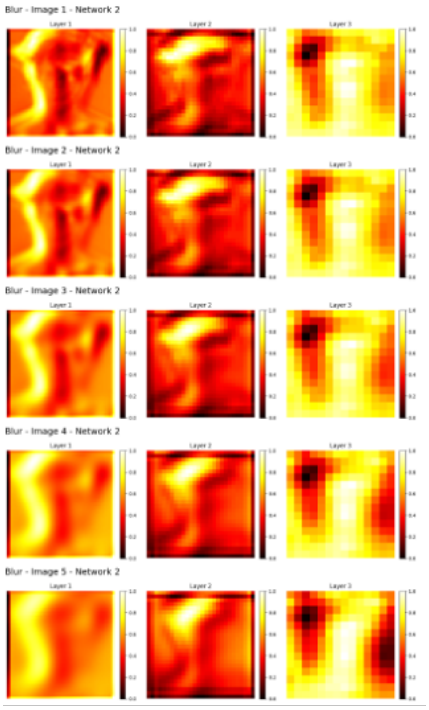

Figure A6: Final activation maps from each convolution layer in CNN2 for a modified sample image, where an increasing Gaussian blur was applied.

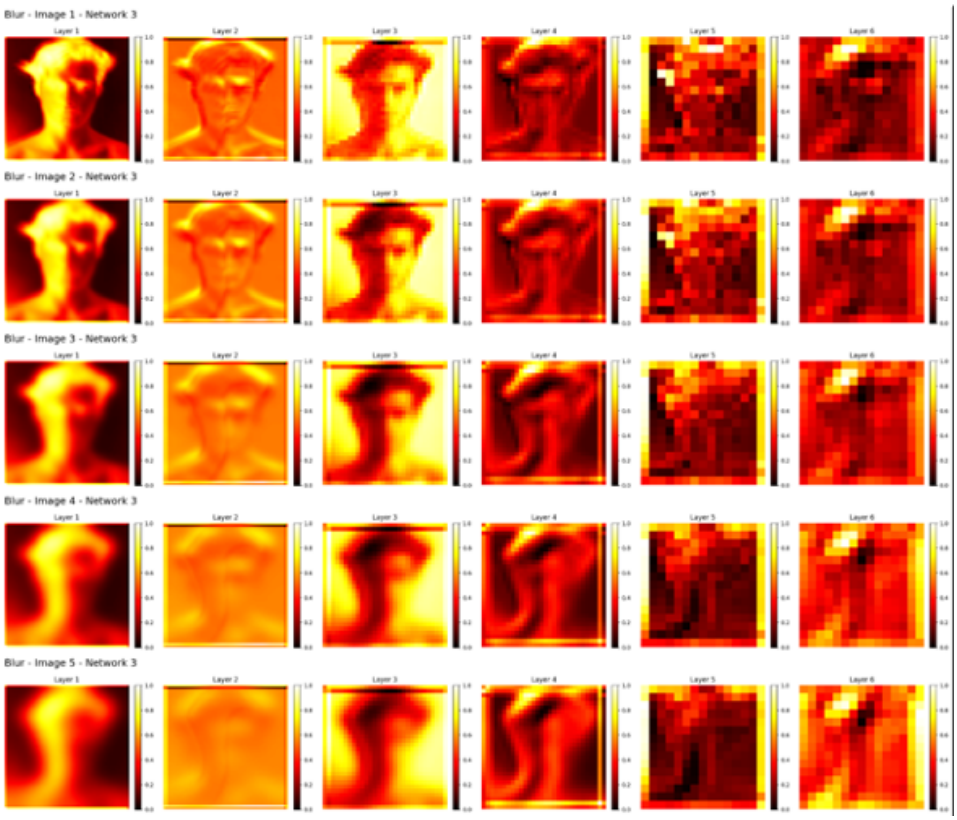

Figure A7: Final activation maps from each convolution layer in CNN3 for a modified sample image, where an increasing Gaussian blur was applied.

B Testing Accuracy for Modified Images Using Cross-Fold Validation

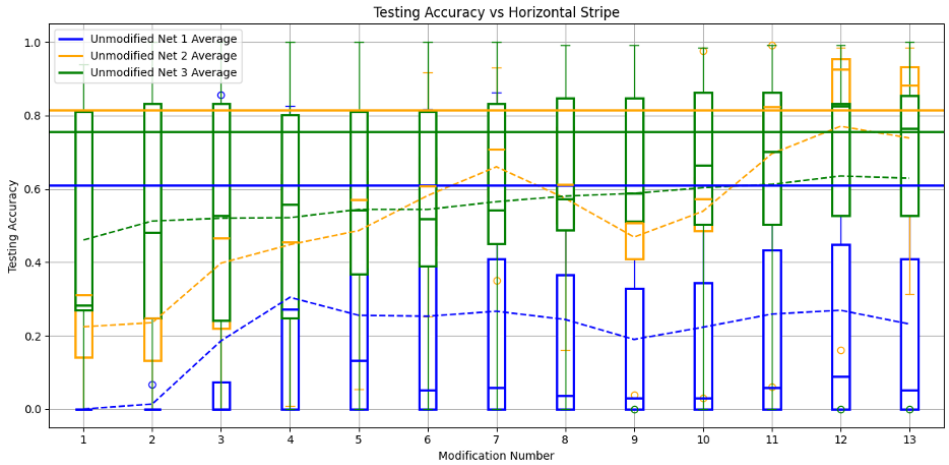

Figure B1: Model performance on all modified test images with varying positions of a horizontal line (presented in the same order as Figures 2), evaluated across five iterations for cross-validation. The solid line represents the baseline average testing accuracy for unmodified images for each model.

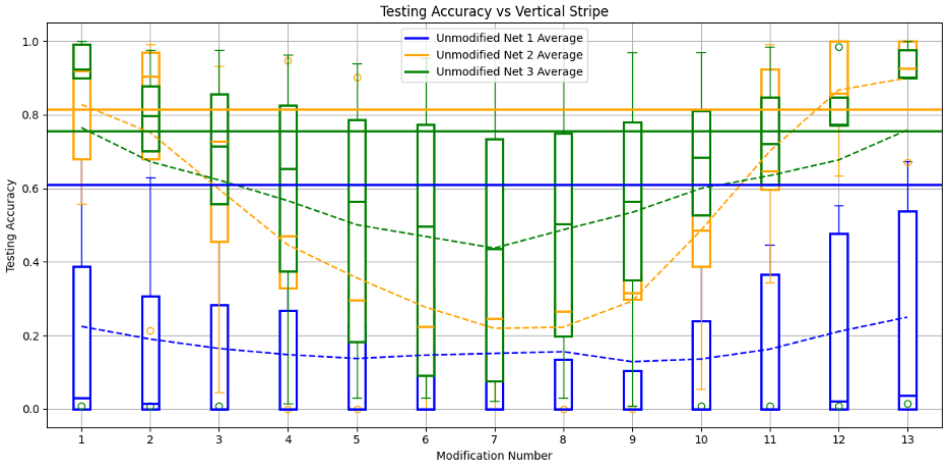

Figure B2: Model performance on all modified test images with varying positions of a vertical line (presented in the same order as Figure 3), evaluated across five iterations for cross-validation. The solid line represents the baseline average testing accuracy for unmodified images for each model.

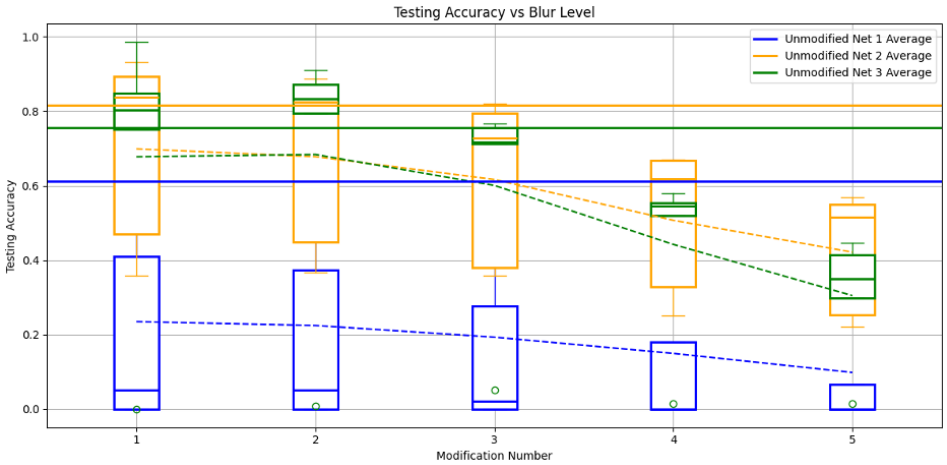

Figure B3: Model performance on all modified test images with varying increasing blur (presented in the same order as Figure 4), evaluated across five iterations for cross-validation. The solid line represents the baseline average testing accuracy for unmodified images for each model.

References

[1] D. G. W. P. McKernan *et al.*, "Neuronal responses to facial features in the macaque inferior temporal cortex," *PMC6404234*, 2019. Available: `https://pmc.ncbi.nlm.nih.gov/articles/PMC6404234/`

[2] L. E. J. Ohayon *et al.*, "The role of the IT cortex in face-selective processing," *PMC2678572*, 2009. Available: `https://pmc.ncbi.nlm.nih.gov/articles/PMC2678572/`

[3] G. Y. Tsao *et al.*, "Face recognition in monkeys: An update," *PubMed Central*, 2012. Available: `https://pubmed.ncbi.nlm.nih.gov/23175821/`

[4] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, MIT Press, 2001.

[5] Lava18, "How batch normalization and ReLU solve vanishing gradients," *Medium*, 2018. Available: `https://lava18.medium.com/how-batch-normalization-and-relu-solve-vanishing-gradients-3f1a8ace1c88`

[6] R. Arora, "Residual blocks in deep learning," *Towards Data Science*, 2020. Available: `https://towardsdatascience.com/residual-blocks-in-deep-learning-11d95ca12b00/#:`

~:text=Residual%20Block%20can%20be%20used,shortcut%20connections%20with%20additional%20gates

[7] DataCamp, "K-fold cross-validation in machine learning," 2021. Available: https://www.datacamp.com/tutorial/k-fold-cross-validation