A Mechanistic Perspective of Face Perception Latency: Predictive Coding

William Pugsley*[1], Junteng Zheng*[2], Jeff Orchard[2], and Roxane Itier[3]
[1]Computational Mathematics, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[2]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[3]Department of Psychology, University of Waterloo , Waterloo, ON N2L 3G1, Canada
* Contributed equally

Abstract

Face processing is widely regarded in cognitive science as the integration of individual features into a holistic percept. However, recent neuroscience research highlights a more nuanced interplay between holistic and featural mechanisms, with specific facial features receiving greater emphasis during early perception. Event-related potential studies reveal that the number and type of parafoveal features significantly influence neural response delays, yet the underlying mechanistic model remains unclear. This paper examines these phenomena through the lens of the predictive coding network, a biologically plausible alternative to traditional deep neural networks. Our findings show that predictive coding networks accurately simulate the influence of parafoveal features on neural response times while upholding the saliency hierarchy of facial features. These results provide a computational explanation for the observed neural delays and highlight the potential of predictive coding as a robust framework for understanding face perception in the human brain.

Keywords: Face Perception; Neural Network; Predictive Coding; Biological Plausibility

Introduction

Face processing is a core perceptual function essential for human recognition and social interaction (Pascalis et al., 2011; Itier & Batty, 2009). Prevailing theories suggest that face perception operates predominantly as a holistic process (Richler & Gauthier, 2014), wherein individual features (e.g., eyes, nose, mouth) are integrated into a unified perceptual representation rather than being processed in isolation. The perception of a face would occur regardless of where one is looking on it.

While the holistic perspective supports rapid and efficient recognition of faces as unique entities, recent neuroscience research (Parkington & Itier, 2018, 2019; Winward et al., 2022) suggests that early face processing is subject to facial features. A complete facial presentation may not be necessary for optimal responses (Issa & DiCarlo, 2012), and different facial features contribute variably to the early stages of face perception (Itier, 2015; Nemrodov et al., 2014).

The Event Related Potential (ERP) studies of Parkington & Itier (2019) demonstrated that the face-sensitive N170 ERP component is modulated by factors such as the presence of a face outline, the number of parafoveal facial features, or the specific type of feature in the parafovea. These results, obtained after carefully monitoring gaze position on particular features using an eye tracker, revealed that a reduction in the number of parafoveal features leads to a greater delay in the onset of the N170 and of the P1, the preceding visual ERP component, relative to stimulus presentation. For example, as illustrated in Fig.1, while a single eye within the face outline may be sufficient to elicit early neural responses associated with face perception, the N170 peak is delayed compared to when additional parafoveal features are present.

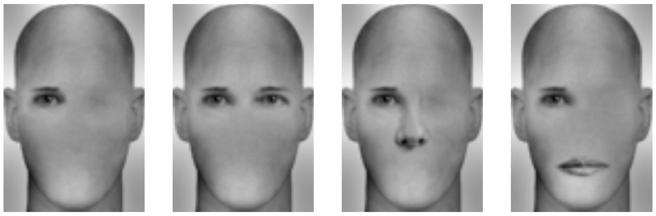

Figure 1: Illustration of facial samples with varying parafoveal features. The face processing may be understood as a malleable process, wherein variations in the number or type of parafoveal features result in differing delays in perception.

Potential explanations (Parkington & Itier, 2018; Winward et al., 2022) have been proposed for understanding the role of facial features during early face perception. But significant gaps remain in a mechanistic explanation of these processes.

Modern computational models increasingly utilize deep neural networks (DNNs) (LeCun et al., 2015), which have demonstrated impressive performance in object recognition and are now widely adopted and applied in neuroscience research (Cox & Dean, 2014; Richards et al., 2019). But they may not always serve as suitable models for all perceptual processes in the human brain, at least for capturing such delay phenomena observed in facial perception. A key question emerges: how can neural networks explain the delay phenomena observed in human facial perception, given that they typically operate as direct mappings from input images to higher-level abstractions, without any temporal aspect?

We can leverage the power of hierarchical neural networks by turning to more biologically plausible neural networks (Lillicrap et al., 2020). These models aim not only to compete on benchmark datasets but also to advance neuroscience by providing mechanistic insights and uncovering potential mechanisms underlying observed phenomena. Predictive coding networks (Friston, 2018), in particular, emerge as promising candidates for this endeavor.

Predictive coding (Mumford, 1992) has demonstrated success as a model of the visual system (Rao & Ballard, 1999)

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

and has even been proposed as a framework for introspection in the brain (Seth, 2013). Notably, predictive coding networks perform inference by iteratively refining their neuron activity until reaching an equilibrium state. In the context of face recognition, this process involves the network establishing a belief about whether it is perceiving a face. Unlike conventional neural networks, such as feed-forward networks, the generative, bi-directional operation of predictive coding (Sun & Orchard, 2020) involves an inference process that takes time. This allows us to probe its activity during inference and determine the point at which the network "makes up its mind", marked by the stabilization of its activity. Hence, the inference process offers a natural window to investigate whether delays arise in reaching equilibrium.

In this paper, we use predictive coding networks to effectively replicate the observed phenomenon in N170, where the number and type of parafoveal features modulate the timing of neural responses. The inference process in predictive coding is delayed when a reduced number of features are available. The experimental findings further indicate that different types of facial features are not weighted equally, aligning with the saliency hierarchy (eyes > mouth > nose) (Shepherd, 1981). This work provides a potential computational perspective to understand the visual system based on neuronal dynamics, and also offers further evidence supporting the theory that predictive coding networks serve as effective models for the human visual system.

Method: Predictive Coding Network

Predictive coding theory has long been proposed as a framework for brain function (Mumford, 1992) in neuroscience, positing that predictions flow from higher-level brain areas to lower levels, inferring expected activity. Simultaneously, mismatches between these predictions and actual activity generate error signals that flow back up, guiding adjustments to connection weights. There is also neurobiological evidence that the wiring within, and between, cortical columns is consistent with predictive coding (Bastos et al., 2012).

Beyond its background in neuroscience, predictive coding can be mathematically formalized and implemented as an artificial neural network. Specifically, we aim to define a biologically plausible neural network that adheres to the principles of: (1) local computation, and (2) local plasticity (Bogacz, 2017; Whittington & Bogacz, 2019). From a biological perspective, local computation means that neural activity is influenced only by input from other neurons—as is typically the case in most artificial neural networks.

Biologically plausible plasticity refers to the principle that weight changes depend on the activity of both pre-synaptic and post-synaptic neurons, commonly known as Hebbian learning rules (Hebb, 2005). In contrast, backpropagation has long been criticized as biologically implausible (Lillicrap et al., 2020), partly due to issues such as weight copying (Richards et al., 2019). Consequently, classic neural networks fail to meet the criterion of local plasticity. However, as we will demonstrate, predictive coding satisfies both local computation and plasticity conditions. This is another reason for selecting predictive coding, as it allows us to impose more biologically plausible constraints on the model.

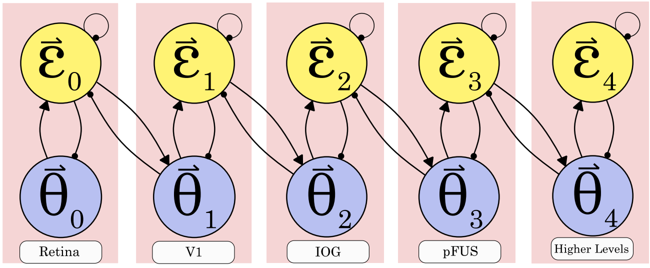

Figure 2: Architecture diagram for the predictive coding network. Each layer represents an abstraction of the hierarchical structure in the brain. Paths terminating in circles indicated inhibitory connections and those ending in arrows indicate excitatory ones.

Predictive Coding Network as a Bayesian Brain

Predictive coding can be regarded as a manifestation of the Bayesian Brain hypothesis (Spratling, 2017). Mathematically, each layer of the network is represented by a vector of activities, where the ith layer has activities $\vec{\theta}_i$. In a network, the 0th layer serves as the input layer that receives sensory input, while the Nth layer functions as the output layer, encoding high-level beliefs. The entire network can be viewed as a Bayesian chain, with the relationships between hidden layers defined by

$$p(\vec{\theta}_0,\ldots,\vec{\theta}_N)=p(\vec{\theta}_N)\prod_{i=0}^{N-1}p(\vec{\theta}_i|\vec{\theta}_{i+1}). \tag{1}$$

Here, we use $\vec{\theta}_i$ rather than $\vec{x}_i$ to emphasize that the activity at each layer is a *variable* that the network must infer, rather than a constant determined directly by the input. On top of this, not all network states are equally likely; a trained network will preferentially adopt certain states based on its learned connection weights.

A common formulation of predictive coding assumes that each layer predicts the mean of the layer below it, and we measure the likelihood of the lower layer's state from a normal distribution (Bogacz, 2017),

$$p(\vec{\theta}_i|\vec{\theta}_{i+1})=\mathcal{N}(\vec{\theta}_i;W_ih_i(\vec{\theta}_{i+1}),\Sigma_i)\ , \tag{2}$$

where the activation function h_i is a nonlinear scalar function applied element-wise to its vector input. The matrix W_i holds the connection weights from the $(i+1)$th layer to the ith layer, and Σ_i represents the covariance matrix of the distribution.

Inference in a Predictive Coding Network

In predictive coding, the network seeks to infer the most probable state. Mathematically, this means finding the values of each layer's activity, $\vec{\theta}_1, \ldots, \vec{\theta}_{N-1}$, that maximize the log joint probability given the data (in our case, images of faces),

$$\max_{\vec{\theta}_1,\ldots,\vec{\theta}_{N-1}} \left[\ln p(\vec{\theta}_N) + \sum_{i=1}^{N-1} \ln p(\vec{\theta}_i | \vec{\theta}_{i+1}) \right] . \tag{3}$$

If we substitute (2) into the objective function, it can be rewritten in terms of the negative Free Energy, F (Friston, 2010), as

$$F = \ln p(\vec{\theta}_N) - \sum_{i=1}^{N-1} \left(\vec{\theta}_i - W_i h_i(\vec{\theta}_{i+1}) \right)^{\mathrm{T}} \Sigma_i \vec{\varepsilon}_i \,, \tag{4}$$

where

$$\vec{\varepsilon}_i = \vec{\theta}_i - W_i h_i(\vec{\theta}_{i+1}) \tag{5}$$

is the difference or mismatch between the activity of the ith layer ($\vec{\theta}_i$) and the prediction $W_i h_i(\vec{\theta}_{i+1})$ from the $(i+1)$th layer. Similar to how $\vec{\theta}_i$ represents the activity vector of a neural network layer, the prediction error $\vec{\varepsilon}_i$ can be interpreted as the error vector between layers. Both $\vec{\theta}_i$ and $\vec{\varepsilon}_i$ can be thought of as representing neurons within each layer of the network.

To infer the most probable state at each layer, gradient ascent can be performed on F with respect to θ and ε. This process utilizes simple Euler time-stepping to iteratively update these parameters, starting from an initial estimate

$$\vec{\theta}_i^{(t+1)} = \vec{\theta}_i^{(t)} + \kappa \nabla_{\vec{\theta}_i} F; \quad \varepsilon_i^{(t+1)} = \varepsilon_i^{(t)} + \kappa \nabla_{\varepsilon_i} F, \tag{6}$$

where κ represents the inference rate, which controls the step size during the inference process.

More importantly, the number of inference steps naturally serves as a metric for quantifying response time. In this study, we define the response time of predictive coding as how long it takes for changes in each network state to drop below a predefined threshold, indicating convergence to a stable equilibrium and a confident decision. Consequently, the response time is quantized by the total number of inference steps required to reach this equilibrium.

During training, the values of $\vec{v}_0$ and $\vec{v}_N$ are fixed to the input images and target labels, denoted $\vec{\theta}_0$ and $\vec{\theta}_N$, respectively. Consequently, $p(\vec{\theta}_N)$ remains constant, and the derivatives of F with respect to the network states and errors are

$$\nabla_{\vec{\theta}_i} F = -\vec{\varepsilon}_i + h'_{i-1}(\vec{\theta}_i) \odot (W_{i-1}^{\mathrm{T}} \vec{\varepsilon}_{i-1}) \tag{7a}$$

$$\nabla_{\vec{\varepsilon}_i} F = \vec{\theta}_i - W_i h_i(\vec{\theta}_{i+1}) - \Sigma_i \vec{\varepsilon}_i \tag{7b}$$

where $\odot$ is element-wise multiplication, also called the Hadamard product, and i ranges from 1 to $N-1$.

In a nutshell, while the network infers the most probable state, it simultaneously minimizes Free Energy, which can also be viewed as reducing the prediction error layer by layer. Thus, the entire network achieves coherence.

Learning In a Predictive Coding Network

Similarly, we can derive the gradients of F with respect to the weights W_i and the Σ_i, enabling the network to *learn*,

$$\nabla_{W_i} F = \vec{\varepsilon}_i h_i(\vec{\theta}_{i+1})^{\mathrm{T}}, \tag{8a}$$

$$\nabla_{\Sigma_i} F = \frac{1}{2} (\vec{\varepsilon}_i \vec{\varepsilon}_i^{\mathrm{T}} - \Sigma_i^{-1}), \tag{8b}$$

where learning occurs after the network completes inference, reaching the equilibrium solution of (4). However, in practice, Σ is typically fixed to the identity matrix rather than being learned; we do the same in this paper. Recall the error is defined as the discrepancy between the prediction generated by a layer and the observed activity in the layer below it. Consequently, this learning rule can be deemed biologically plausible, since it relies solely on local information available directly from the layers it connects.

In an ideally trained network, the higher layers accurately predict the lower layers, leading to negligible or zero error. When a discrepancy arises (i.e., ε is not zero), this mismatch drives the learning process.

Neural Network Implementation

All the equations described above can be readily implemented within a neural network architecture, as illustrated in Fig. 2. This architecture can also be interpreted as an abstraction of the facial perception pathway in the human brain.

Training a predictive coding neural network for facial perception constitutes a supervised learning problem. During training, the two ends of the network are clamped with input (facial images) and output (labels, such as whether the image contains a face or not). The network then performs inference based on (7) and (6) until it reaches equilibrium. Finally, any remaining errors are used to update the weights according to (8), facilitating the learning process.

Experimental Setup

Our purpose is to emulate human face perception under varying numbers and types of parafoveal features. Before evaluating the model on face images with different feature configurations, we must first train it so that it functions like an individual participant in the experiment. This section describes the data and specifies the model settings used to prepare the network for subsequent testing.

Dataset

When human participants are involved in experiments, they have already been bombarded with countless classes of objects and variations, and they recognize human faces and features such as noses or mouths. To replicate this learning process, we prepare a dataset to train the predictive coding network accordingly. Examples of the images used in this study are shown in Fig. 3.

Face and Non-face We use 32 unique images of human faces, evenly split between male and female, identical to those employed by Winward et al. (2022). These images are

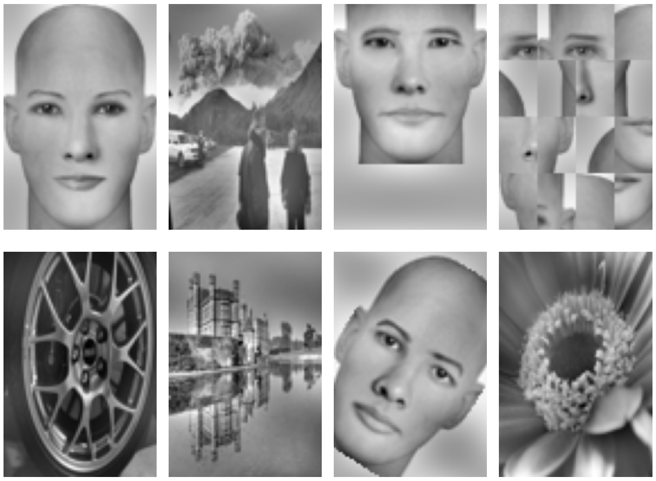

Figure 3: Samples of training images after processing and augmentation

centered with uniform head outlines and serve as the basis for training the network.

Since human brains learn to recognize faces by observing them from various positions and orientations, we augment the face images through rotations, translations, and reflections. For each face image, we include its horizontal reflection (a reflection across the vertical axis of symmetry). From these 64 images, we generate additional variations by applying three random rotations, uniformly sampled from $[-45°, 45°]$, and three vertical translations randomly selected uniformly from $[-15, 30]$ pixels. These ranges were carefully chosen to ensure that all parafoveal features remain visible during training.

Additionally, we want the network to learn to recognize faces as distinct from other objects. Similar to human learning, where the brain is exposed to countless object classes and variations, we replicate this process by incorporating a random subset of images from ImageNet (Deng et al., 2009) into our dataset.

Parafoveal features We aim to ensure that the predictive coding network learns to distinguish faces based on parafoveal features or facial contours. However, there is no guarantee that the network will rely on this information for classification. It may instead base its predictions on irrelevant cues, such as the blank background associated with the face images or the distribution of pixel intensities.

Hence, we standardize our images by resizing them to 68 $\times 100$ pixels, converting color images to grayscale, and normalizing pixel values to the range [0, 1]. We also apply a two-dimensional Fourier transform to the images and set the smallest frequencies within a 2 $\times 3$ rectangle at the center of the spectrum to 0. This process removes low-frequency content, such as overall brightness and brightness gradients. As a result, the network is encouraged to focus on learning object shapes and contours, enabling it to identify facial features effectively.

To further ensure that the network learns to recognize faces based on the presence and orientation of features, we include shuffled face images. This shuffling is performed both pixel-wise and block-wise. For pixel-wise shuffling, we randomly permute all the pixels within the face images. For block-wise shuffling, we divide each image into sixteen 25×17 blocks and shuffle these blocks (an example is shown in Fig. 3, top-right). The shuffled images are labeled as non-face images. Importantly, the resulting images retain the same pixel intensity distributions as the original faces, ensuring that the network must recognize the spatial configuration of the features in order to recognize a face.

Model

We train a population of 45 predictive coding neural networks, replicating the number of participants in the study conducted by Winward et al. (2022), which also serves as a repeated experiment to ensure that the findings are not coincidental or the result of accidentally training an exceptionally well-performing model.

We employed a five-layer neural network comprising an input layer (retina), an output layer (higher-level brain regions), and three hidden layers representing V1, IOG, and pFUS. The input layer encodes image pixels; the hidden layers contain 300, 200, and 100 neurons, respectively, and the output layer comprises 2 neurons. Bidirectional connections enable top-down and bottom-up processing, with the tanh activation function applied throughout.

Each network is trained on the aforementioned dataset for 18 epochs with a batch size of 64. The activities and errors are updated using a step size of $\kappa = 0.05$. For each sample, the network performs up to 500 inference steps, which is sufficient to reach equilibrium.

We utilize the Adam optimizer (Diederik & Jimmy, 2015) with a learning rate of 10^{-4} and exponential decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to update the weight matrices. It is important to note that using Adam does not imply backpropagation; rather, it is employed to efficiently utilize the gradients calculated in Eq.8.

Results and Discussion

Once the neural-network “participant” is trained, eight different feature combinations are used as treatments to test the response time. The network is presented with images missing certain parafoveal features to observe how it perceives them. The specific combinations of features are detailed in Table 1, and some sample images can be seen in Fig. 1.

Human eyes perform saccades — rapid eye movements — which introduce experimental noise in EEG studies. To mitigate this effect, face images are displayed with one eye consistently anchored in the participant’s foveal vision, ensuring stable gaze fixation throughout the experiment. Consequently, the number of additional features presented alongside this fixed eye is referred to as the number of parafoveal

Number	Feature Combinations	Short Notation
0	One Eye	E1
1	Two Eyes One Eye + Nose One Eye + Mouth	E2 E1N E1M
2	Two Eyes + Nose Two Eyes + Mouth One Eye + Nose + Mouth	E2N E2M E1NM
3	Full Face	FF

Table 1: Number of parafoveal features used in the experiment, corresponding feature combinations, and their short notations.

features. Besides, in our neural network, fixation is inherently maintained because there is no distinction between 'in-focus' and 'out-of-focus' inputs. Nevertheless, we retain the parafoveal feature notation for consistency.

Accuracy

Are the trained predictive coding networks suitable for our experiment? Can they reliably recognize faces based on facial features? To answer this, it is essential to evaluate their accuracy and determine their effectiveness for the study.

After training, the 45 networks achieve an average accuracy of 0.879 on the training set, with a standard deviation of 0.009. This level of performance is sufficient for the purposes of this experiment. Rather than prioritizing higher accuracy, our primary aim of training is to develop networks that can reliably identify whether an image represents a face. Remarkably, the networks successfully classify all face images, including those with varied feature combinations, without error.

Additionally, the networks successfully classify all pixel-shuffled images as non-faces and correctly identify block-shuffled faces as non-faces with accuracies ranging from 0.95 to 1.0. These results align with our expectations and validate the effectiveness of the trained networks.

Effects of Feature number and Type

As demonstrated in the pioneering studies (Parkington & Itier, 2019, 2018; Winward et al., 2022), latency exhibits a monotonic decline as the number of parafoveal features increases, with the presentation of a full face eliciting the most rapid response.

In this experiment, each of the 45 trained networks is treated as an individual participant to assess whether it reproduces the same observation. For each network, response times are measured for 32 face images presented under 8 distinct feature combinations, with the mean calculated to mitigate noise. Recall that response time t is defined as the number of inference steps required for the network to reach equilibrium, particularly when the output layer converges to a stable state, indicating the network's decision on whether the image represents a face or not.

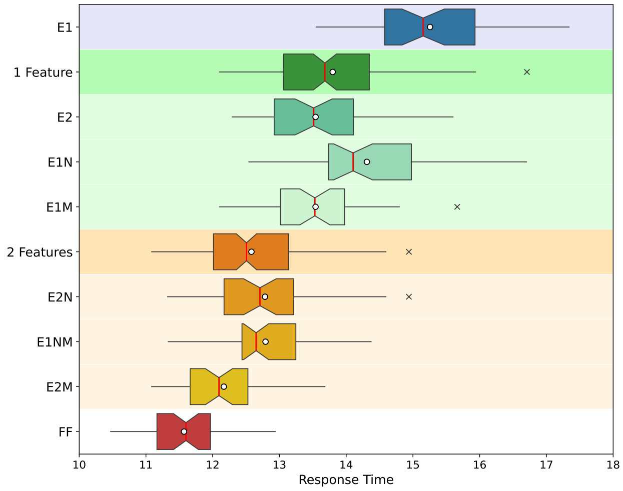

Figure 4: Response time variations across different configurations of parafoveal features. Configurations with fewer features generally result in longer response times, while the influence on response time varies depending on the specific features included.

Feature Number How do the number of facial features influence network response times, irrespective of feature type? As illustrated in Fig. 4, response times vary notably with different numbers of parafoveal facial features. Networks exhibit longer response times when fewer features are present, whereas the "Full Face" configuration consistently produces the shortest response times. This pattern aligns with expectations, as an increase in the number of features results in a monotonic decrease in response time.

Interestingly, after applying a simple linear mapping to response times, $2.1t + 124.87$, the resulting latency from the predictive coding model shows a decay pattern similar to that observed in the EEG signals adopted from (Parkington & Itier, 2019), as illustrated in Fig. 5.

Feature Type The combination of feature types also influences response times, not just the number of features. Crucially, this result highlights that features are not equally weighted in their contribution. For instance, networks respond faster when presented with two eyes compared to a combination of one eye and a nose. This finding aligns with previous studies showing that feature saliency modulates the N170 response, with a saliency hierarchy observed as eyes $>$ mouth $>$ nose (Winward et al., 2022).

Activity Phase Space

Why does the response time in predictive coding exhibit patterns similar to those observed in event-related potentials? Fortunately, we can analyze the activity within each layer of the predictive coding network to investigate this question in detail. We randomly select one of the 45 networks to act as a participant and present it with various combinations of features under the same face identity. The only variation is the presence or absence of parafoveal features.

In Fig. 6, we plot the L^2 norms of each layer's activity throughout the inference process, excluding the 0th layer (input layer), as it is held constant. The L^2 norm is defined for any vector $\vec{x}$ of size d as $\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$.

The activities of each layer exhibit a similar fluctuating pattern before approaching equilibrium. The primary differences between the full-face condition and the partial-feature conditions lie in the magnitudes of the activities. As the number of features decreases, the magnitude of the norms also decrease. Additionally, we include the network's response when presented with just the outline of a face (i.e., one without any features). This condition results in the smallest amplitude and the longest response time.

The final layer is an exception to the fluctuating pattern, as it lacks error feedback to regulate its growth. This final layer integrates the processed image representation from the layer below it, and gathers evidence to determine if the image is a face.

For comparison, the predictive coding network exhibited a different pattern of activity when it was given a random non-face input. The process of inference, in which the layers try to find a low-energy state, resulted in a low value in the top layer, correctly identifying the input as a non-face (the dashed line in Fig. 6).

The neurons in our predictive coding neural networks are less active when presented with faces containing fewer parafoveal features compared to a full face. Because the neurons are less active, it takes longer for the top layer to reach the threshold required to register a decision. It is also interesting to note that, as expected, deeper layers of the network (further from the input) exhibit delayed responses compared to shallower layers.

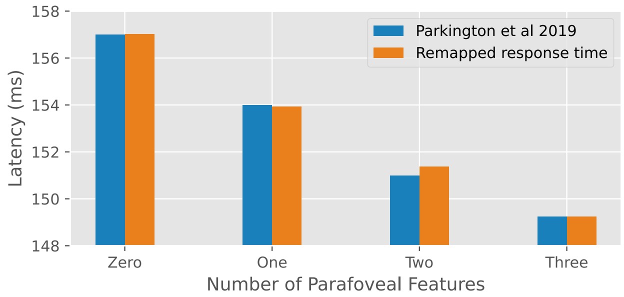

Figure 5: Comparison between data from (Parkington & Itier, 2019) and a linear least-squares fit of our response-time results.

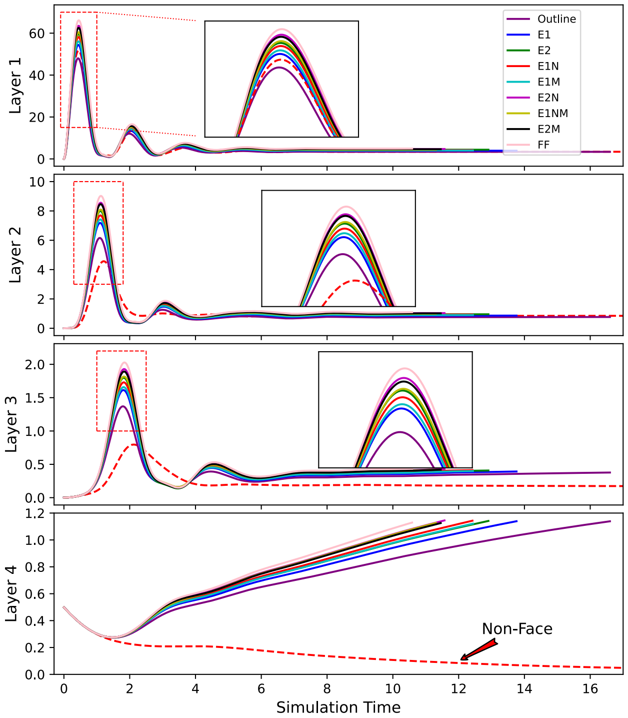

Figure 6: L^2 norm of each layer's activity in a network under different combinations of features, with each line representing a specific case. The red dashed line represents the activity of the network when the input is a non-face image. The first peak is magnified to highlight differences in amplitude. Note that the L^2 norm should be compared within each layer, as different layers have varying numbers of units.

Conclusion

This paper reproduces the phenomenon that the presence and type of parafoveal facial features modulates latency during face perception, and it does so using a predictive coding network model of a rudimentary visual system. The results show that networks exhibit longer response times (i.e., number of inference steps until convergence) when provided with fewer parafoveal features and, similarly, respond more quickly to a full face than to partial configurations. Furthermore, the type of feature proves critical; eyes evoke faster responses than mouth or nose alone, consistent with established saliency hierarchies. These trends were corroborated through analyses of the evolving activity patterns in each network layer, where conditions with fewer features produced smaller amplitudes and slower convergence.

Overall, this work highlights the utility of predictive coding in modeling delay phenomena and feature-specific saliency in face perception, and offers further evidence that the visual system generally follows a predictive-coding architecture.

Acknowledgments

We gratefully acknowledge financial support from the University of Waterloo, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the China Scholarship Council.

References

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, *76*, 198–211.

Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision. *Current Biology*, *24*(18), R921–R929.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Diederik, P. K., & Jimmy, L. B. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations.*

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Friston, K. (2018). Does predictive coding have a future? *Nature neuroscience*, *21*(8), 1019–1021.

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology press.

Issa, E. B., & DiCarlo, J. J. (2012). Precedence of the eye region in neural processing of faces. *Journal of Neuroscience*, *32*(47), 16666–16682.

Itier, R. J. (2015). Attention to eyes in face perception. *The handbook of attention*, 369–388.

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: the core of social cognition. *Neuroscience & Biobehavioral Reviews*, *33*(6), 843–863.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346.

Mumford, D. (1992). On the computational architecture of the neocortex: II The role of cortico-cortical loops. *Biological cybernetics*, *66*(3), 241–251.

Nemrodov, D., Anderson, T., Preston, F. F., & Itier, R. J. (2014). Early sensitivity for eyes within faces: A new neuronal account of holistic and featural processing. *NeuroImage*, *97*, 81–94.

Parkington, K. B., & Itier, R. J. (2018). One versus two eyes makes a difference! Early face perception is modulated by featural fixation and feature context. *Cortex*, *109*, 35–49.

Parkington, K. B., & Itier, R. J. (2019). From eye to face: The impact of face outline, feature number, and feature saliency on the early neural response to faces. *Brain Research*, *1722*, 146343.

Pascalis, O., de Martin de Viviés, X., Anzures, G., Quinn, P. C., Slater, A. M., Tanaka, J. W., & Lee, K. (2011). Development of face processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 666–675.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological bulletin*, *140*(5), 1281.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, *17*(11), 565–573.

Shepherd, J. W. (1981). Studies of cue saliency. *Perceiving and remembering faces*, 105–131.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and cognition*, *112*, 92–97.

Sun, W., & Orchard, J. (2020). A predictive-coding network that is both discriminative and generative. *Neural Computation*, *32*, 1836-1862.

Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in cognitive sciences*, *23*(3), 235–250.

Winward, S. B., Siklos-Whillans, J., & Itier, R. J. (2022). Impact of face outline, parafoveal feature number and feature type on early face perception in a gaze-contingent paradigm: A mass-univariate re-analysis of ERP data. *Neuroimage: Reports*, *2*(4), 100148.