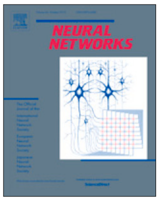
Neural Networks 157 (2023) 280–287

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



On the role of feedback in image recognition under noise and adversarial attacks: A predictive coding perspective

Andrea Alamia [a,b,*,1], Milad Mozafari [a,c,1], Bhavin Choksi [a], Rufin VanRullen [a,b]

[a] CerCo, CNRS, 31052 Toulouse, France
[b] ANITI, Université de Toulouse, 31062, Toulouse, France
[c] IRIT, CNRS, 31062, Toulouse, France

ARTICLE INFO

Article history:
Received 25 April 2022
Received in revised form 5 October 2022
Accepted 18 October 2022
Available online 27 October 2022

Keywords:
Predictive coding
Deep learning
Feedback
Corruption robustness

ABSTRACT

Brain-inspired machine learning is gaining increasing consideration, particularly in computer vision. Several studies investigated the inclusion of top-down feedback connections in convolutional networks; however, it remains unclear how and when these connections are functionally helpful. Here we address this question in the context of object recognition under noisy conditions. We consider deep convolutional networks (CNNs) as models of feed-forward visual processing and implement Predictive Coding (PC) dynamics through feedback connections (predictive feedback) trained for reconstruction or classification of clean images. First, we show that the accuracy of the network implementing PC dynamics is significantly larger compared to its equivalent forward network. Importantly, to directly assess the computational role of predictive feedback in various experimental situations, we optimize and interpret the hyper-parameters controlling the network's recurrent dynamics. That is, we let the optimization process determine whether top-down connections and predictive coding dynamics are functionally beneficial. Across different model depths and architectures (3-layer CNN, ResNet18, and EfficientNetB0) and against various types of noise (CIFAR100-C), we find that the network increasingly relies on top-down predictions as the noise level increases; in deeper networks, this effect is most prominent at lower layers. All in all, our results provide novel insights relevant to Neuroscience by confirming the computational role of feedback connections in sensory systems, and to Machine Learning by revealing how these can improve the robustness of current vision models.

1. Introduction

Feed-forward deep convolutional networks (DCNs) reached remarkable accuracy in several visual tasks, including image classification. Inspired by biological visual systems (Fukushima, 1980), they share several similarities with them. For example, both systems have a hierarchical structure, in which neurons in the higher (lower) levels of the hierarchy have larger (smaller) receptive field sizes and respond to more complex (simpler) stimuli (Hubel & Wiesel, 1959). Further, representational (Khaligh-Razavi & Kriegeskorte, 2014) and functional similarities (Bashivan, Kar, & DiCarlo, 2019) between the feed-forward DCNs and the brain's feed-forward visual pathway have provided novel opportunities to study the brain through the lens of DCNs.

However, contrary to biological visual systems, DCNs blunder significantly when confronted with noisy images and adversarial attacks, revealing an important deficit in robustness (Hendrycks & Dietterich, 2019; Nguyen, Yosinski, & Clune, 2014; Szegedy et al., 2013). One main difference with their biological counterpart consists in the lack of recurrent or feedback connections. It has been shown that the brain relies on feedback pathways for robust object recognition under challenging conditions (Choksi et al., 2021; Kar & DiCarlo, 2021; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Kietzmann et al., 2019; Li, Bradshaw, & Sharma, 2019; Rajaei, Mohsenzadeh, Ebrahimpour, & Khaligh-Razavi, 2019; Schott, Rauber, Bethge, & Brendel, 2019; Wyatte, Jilk, & O'Reilly, 2014). In recent years, several approaches aimed to introduce feedback connections in deep networks to improve not only biological plausibility but also model robustness, and accuracy (Huang et al., 2020; Kubilius et al., 2018; Nayebi et al., 2018; Yan et al., 2019). Importantly, feedback connections can be trained either in a supervised fashion to optimize the task objective (e.g., object recognition) or in an unsupervised way to minimize the reconstruction errors (i.e., prediction errors). In the latter case, feedback connections are trained to predict the activity of lower layers, and the network can be described as a hierarchical generative model. More generally, top-down predictions represent expectations about lower layers activity, updated based on the

[*] Corresponding author at: CerCo, CNRS, 31052 Toulouse, France.
E-mail address: andrea.alamia@cnrs.fr (A. Alamia).

[1] Equal contribution.

https://doi.org/10.1016/j.neunet.2022.10.020
0893-6080/© 2022 Elsevier Ltd. All rights reserved.

incoming sensory evidence over iterations. This interpretation about the role of top-down connections finds its natural place in a prominent framework in Neuroscience, namely Predictive Coding (Huang & Rao, 2011; Rao & Ballard, 1999).

The Predictive Coding (PC) paradigm in Neuroscience is endorsed by a large body of neuroscientific experimental evidence (Baldeweg, 2006; Garrido, Kilner, Stephan, & Friston, 2009; Hohwy, Roepstorff, & Friston, 2008; Kilner, Friston, & Frith, 2007; Pawel Zmarz, 2016), but see Kevin S. Walsh and McGovern (2020) for a critical review of experimental evidence in favor and against the PC framework. It characterizes perception as an inference process in which sensory information is combined with prior expectations to attain the final percept. Accordingly, PC postulates two fundamental terms: predictions and prediction errors (PEs). Considering the visual system as a hierarchical structure, these two signals interact between subsequent brain regions in an iterative process. Ideally, the interplay between feedback predictions and feed-forward PEs converges over iterations into a state in which predictions fully represent the sensory information and PE falls to zero. Although several models implemented and described this dynamic in different conditions (Alamia & VanRullen, 2019; Friston & Kiebel, 2009; Spratling, 2010), the functional role of these two main actors remains largely unexplored.

Here, we address this question by taking a computational perspective and leveraging current state-of-the-art deep neural networks used in visual object recognition. The key insight in our approach consists in hyper-parameter optimization, based on the functional benefit of each connection; we then evaluate the outcome across various experimental (noise) conditions. Our approach significantly extends our previously developed framework for visual perception (Choksi et al., 2021) by systematically optimizing hyper-parameters to investigate the functional benefit of each connection under various noise conditions. On the one hand, from a Neuroscience point of view, our results supported the hypothesis that feedback plays a crucial role in the cortical processes involved in biological vision. On the other hand, from a machine learning perspective, our simulations demonstrated a more robust class of models based on an established biologically inspired framework.

2. Methods

2.1. Predictive coding dynamics

Irrespective of the considered architecture, we implemented the proposed predictive coding dynamics through a stack of modules called *PCoders*. . As in Choksi et al. (2021), the activity of each PCoder m_i at time-step t is driven by four terms, as described in the following equation:

$$m_i(t+1) = \mu m_i(t) + \gamma \mathcal{F}_i(m_{i-1}(t+1), \theta_i^{ff}) + \beta \mathcal{B}_{i+1}(m_{i+1}(t), \theta_{i+1}^{fb}) - \alpha \nabla \epsilon_i(t), \tag{1}$$

$$\epsilon_i(t) = MSE(\mathcal{B}_i(m_i(t), \theta_i^{fb}), m_{i-1}(t)), \tag{2}$$

where $\mathcal{F}_i$ computes the feed-forward drive of the ith PCoder with parameters θ_i^{ff}, and $\mathcal{B}_{i+1}$ computes the feedback drive (prediction) with parameters θ_{i+1}^{fb} given m_{i+1}. The gradient $\nabla \epsilon_i(t)$ is calculated with respect to the activity of the higher layer ($m_i(t)$) as suggested in predictive coding theory.

A specific hyper-parameter modulates each term. First, each PCoder's activity is initialized by a feed-forward pass, i.e., without considering memory or top-down connections, in line with experimental observations in biological visual systems (VanRullen & Thorpe, 2001a, 2001b). Then, at successive time-steps, the activity is determined by several terms. First, a memory term, regulated by the μ hyper-parameter, that retains information from previous time-steps, essentially acting as a time constant. The γ and β hyper-parameters modulate the feed-forward drive and feedback error terms, which reflect information from the lower and higher layers, respectively. The modulation of the first three terms is normalized, i.e. $\beta + \gamma + \mu = 1$. Lastly, the α hyper-parameter modulates the feed-forward error term, which aims at reducing the prediction-error, i.e. the mean squared error (MSE) between the prediction by a PCoder and the activity of the lower one (or the "input stimuli" in case of the first PCoder). As an implementation detail, we multiply α by a scaling factor (see supplementary section A.2) to remove the effect of batch, layer, and (de)convolution kernel size. A generic schematic of the proposed PC dynamics is illustrated in Fig. 1B. As postulated by predictive coding formulation, the feedback and feed-forward error terms regulate each PCoder's activity to reduce prediction-errors over time. Importantly, the dynamic described above is equivalent to the one proposed by Rao and Ballard (1999), with the only difference being the feed-forward term (for the mathematical proof see Choksi et al., 2021).

2.2. Architectures

Shallow model. We first implemented a shallow three-layer CNN with two additional dense layers having 120 and 10 neurons, respectively. As shown in Fig. 1A, the convolutional layers have 12, 18 and 24 channels and a kernel size equal to 5 × 5. Max-pooling operations with stride equal to 2 were applied from lower to higher layers. In this network, we consider each convolutional layer as the feedforward module ($\mathcal{F}$) of a separate PCoder. Each PCoder predicts the lower one's activity through a bilinear upsampling operation with scale factor equal to 2, to approximate reversal of the max-pooling operation, followed by a transposed convolutional layer with window size equal to 3 × 3 (i.e., feedback modules $\mathcal{B}$). The number of channels for the transposed convolution is set in accordance to the prediction target.

Extending to deep architectures. Given a very deep architecture, it is not computationally efficient to assign every layer to a separate PCoder's feedforward drive. Instead, we assigned a segment of feedforward network's layers to each PCoder (i.e. each PCoder's $\mathcal{F}$ is a sequence of backbone's layers). We took advantage of "Predify", a python package introduced in Choksi et al. (2021), that allows to introduce PC dynamics in pre-trained feed-forward networks. In the present paper, we introduce PResNet18 and PEffNetB0 by adding the proposed PC dynamics to feed-forward ResNet18 and EfficientNetB0 architectures, respectively.

To explore more diversity over input images and network depth, we examined PResNet18 and PEffNetB0 on CIFAR100 and ImageNet, respectively. For PEffNetB0 we used the original EfficientNetB0 architecture with pretrained weights on ImageNet as the feed-forward backbone; However, in order to improve ResNet18 performance on small CIFAR100 images, we lowered the kernel size of the first convolutional layer to 3 × 3 and omitted its following max-pooling layer to prevent information loss in early layers.

We implemented the block-wise PC dynamics into ResNet18 and EfficientNetB0 by splitting their layers into five and eight PCoders, respectively (see supplementary section A.1). Regardless of the feed-forward architecture, we used a general procedure to define the feed-forward ($\mathcal{F}$) and feedback ($\mathcal{B}$) drive modules. Assume that there are n blocks of layers in the feed-forward network. Let $y = f_i(x)$ denote the computation done by block i where x and y have the size (c_{in}, h_{in}, w_{in}) and $(c_{out}, h_{out}, w_{out})$,

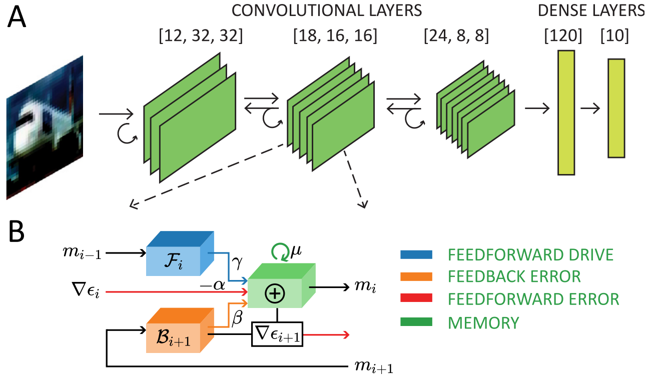


Fig. 1. Predictive Coding dynamics. (A) Architecture of the shallow model, composed of three convolutional layers and two fully connected ones. (B) Generic block diagram for updating PCoder's activity. Each PCoder's activity (m_i) is a combination of four terms at each time-step; Feedforward drive which calls a particular segment of the feedforward backbone's layers ($\mathcal{F}_i$) using the activity of the previous (hierarchically lower) PCoder (m_{i-1}) as the input, Feedback drive which calls a particular network ($\mathcal{B}_{i+1}$) on the following (hierarchically higher) PCoder's activity (m_{i+1}), Feedforward error, which is the error's gradient between the activity of the previous PCoder and its prediction generated by the current PCoder, and Memory term which is the activity of the current PCoder at the previous time-step. Each term is modulated by a specific hyper-parameter.

respectively. Then, $\mathcal{F}_i$ is f_i and $\mathcal{B}_i$ is a 2D up-scaling operation by the factor of $(h_{in}/h_{out}, w_{in}/w_{out})$ followed by a transposed convolutional layer with c_{out} channels and 3 × 3 window size.

2.3. Training parameters

Supervised feed-forward. In both shallow and deep models we trained the feed-forward (θ_i^{ff}) and feedback (θ_i^{fb}) parameters separately with different loss functions. First, we trained θ_i^{ff} to optimize the cross-entropy loss (classification) without using the iterative PC dynamics (i.e., in one forward pass). Accordingly, we used a cross-entropy loss with Stochastic Gradient Descent (SGD) optimizer for the shallow model with learning rate 0.01 and momentum 0.9. In the case of deep networks, we trained the modified ResNet18 on CIFAR100 training images for 200 epochs using SGD optimizer with initial learning rate 0.1, momentum 0.9, and weight decay 5e-4. We applied learning rate decay factor 0.2 at epochs 60, 120, and 160. For PeffNetB0, we used the pretrained ImageNet model described in Tan and Le (2019).

Unsupervised feedback. Next, we optimized θ_i^{fb}s with reconstruction objectives, that is the MSE between the activity of PCoders and their top-down reconstruction on the next time-step. This unsupervised approach is akin to a generative process, in which higher layers predict the activity of lower layers, in line with the predictive coding framework. For the shallow network we used an SGD optimizer with learning rate 0.01 and momentum 0.9. While for both of the deep architectures, we employed Kingma and Ba (2014) optimizer with learning rate 0.001 and weight decay 5e-4 for 50 epochs.

Supervised feedback. In the shallow model, we also explored the role of the top-down connections when their parameters are trained for classification rather than reconstruction (as in the previous case). In this case both the θ_i^{ff} and θ_i^{fb} are optimized simultaneously for 10 time-steps to minimize the cross-entropy loss. We used an SGD optimizer with learning rate = 0.005 and momentum = 0.9. Since the learning takes place over time-steps, the network optimizes the weights given the PC dynamics described in Eq. (1). Importantly, during learning we kept the hyper-parameters values to $\gamma = \beta = \mu = 1/3$ and $\alpha = 0.01$.

2.4. Training hyper-parameters

After the training of the network's parameters, we froze them (including the statistics of batch normalization layers) and optimized uniquely the hyper-parameters γ, β and α (with μ constrained to be $1 - \beta - \gamma$, see supplementary section A.2). Particularly, we repeated the optimization multiple times with different noise types and levels, to investigate the role of each term given different levels of perturbation. We considered a Cross-Entropy loss function averaged across time-steps. In the shallow model we used an Adam optimizer with learning rate equal to 0.001, a weight decay equal to 5e-4 and a batch size of 128 images. For each noise type and level, we repeated the experiment with 10 random initializations of each hyper-parameter drawn from the uniform probability distribution in the interval [0, 1]. We used Adam optimizer with the same weight decay for deep models; however, we employed two separate learning rates equal to 0.01 for γ and λ, and 0.0001 for α. We set batch-size to 128 and 16 for PResNet18 and PEffNetB0, respectively. All the scripts and the trained parameters of the main experiments are available on GitHub.[2]

2.5. Stimuli

The parameters of both the shallow and the deeper networks were trained on clean images, using the training set of CIFAR-10, CIFAR-100 and ImageNet, while the test-set was used to compute each network's accuracy. The hyper-parameters were optimized using different levels and types of noise. Regarding the shallow model, we used additive Gaussian and Salt & Pepper noise, spanning 3 different levels (Gaussian: $\sigma = 0.2$, 0.4 and 0.8; Salt & Pepper: pixel percentage = 2%, 4% and 8%). We used CIFAR100-C, a dataset containing five levels of 19 different corruption types (Hendrycks & Dietterich, 2019) to train PResNet18's hyper-parameters. Finally, in order to train hyper-parameters of the deep PEffNetB0, we used the ImageNet validation set and applied five levels of Gaussian ($\sigma = 0.5$, 0.75, 1, 1.25, and 1.5) and Salt & Pepper (percentage = 5%, 10%, 15%, 20%, and 30%) noise.

3. Results

3.1. Three-layer model

We first tested our hypothesis on a shallow model composed of three convolutional and three dense layers (see panel A of Fig. 1). The advantage of choosing a smaller network consists in promptly exploring several approaches before replicating in deeper state-of-the-art networks. Specifically, we investigated the role of each term in Eq. (1) when training feedback weights for reconstruction or classification (unsupervised vs supervised), (1) via some ablation simulations, and (2) regarding the robustness to adversarial attacks .

3.1.1. Influence of feedback connections: reconstruction vs. classification

We first assessed the role of the feedback and each term in Eq. (1) when the top-down parameters were optimized for reconstruction. After having trained the forward weights for classification (Supplementary Figure 5A), we trained the feedback weights optimizing the reconstruction loss of each PCoder (Figure 5C). This approach is in line with the PC interpretation, in which top-down connections generate predictions to explain lower layers' activity (i.e., minimize prediction errors, or the reconstruction

[2] https://github.com/artipago/Role_of_Feedback_in_Predictive_Coding

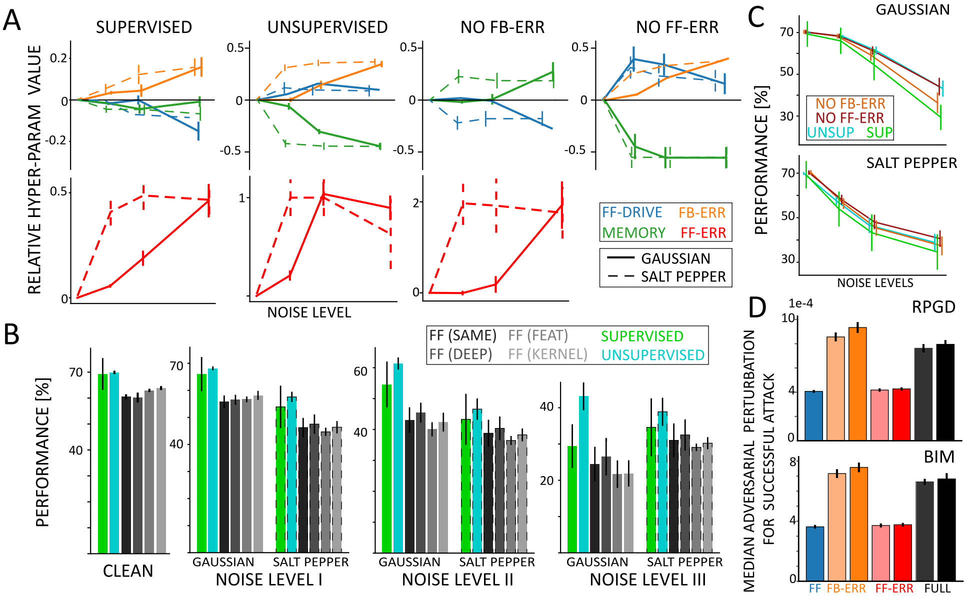


Fig. 2. Shallow model results. (A) The plots show the hyper-parameters (HPs) value relative to the clean images as a function of the noise levels. Each column shows the relative HPs trained in different conditions: supervised, unsupervised, without feedback error, or forward error. The first row shows the feedback error term, the memory, and the forward drive term, the second row shows the forward error term on a separate scale, for Gaussian (solid line) and Salt & Pepper (dashed lines) noise. In all conditions, the feedback-error and forward-error terms increase with the noise levels. (B) Performance of different models at the last time-step. The models implementing PC dynamics (in green and cyan) perform better than equivalent feed-forward networks, especially when trained in an unsupervised fashion (cyan). (C) Performance of the PC models, as a function of the noise levels, measured at the last time-step. Contrasting supervised (SUP) and unsupervised (UNSUP) models reveals the effects of feedback training objective, whereas comparing the ablation models with UNSUP shows the effect of each error term on accuracy. (D) The graph shows the median perturbation to obtain a successful attack using different HPs. Orange and red bars have higher feedback and forward error terms (paler colors correspond to smaller error terms), and blue and black bars represent the feed-forward and the full model, respectively. Our simulations reveal that PC models with higher feedback values (orange, black bars) are more robust to adversarial perturbations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

loss). In this case backward weights are trained in an unsupervised fashion. Once both forward and backward connections were optimized (for classification and reconstruction, respectively), we froze all parameters and trained only the hyper-parameters (γ, β and α in Eq. (1)). As shown in Fig. 2A, with both Gaussian and Salt & Pepper noise the hyper-parameter modulating the top-down feedback (i.e., β in Eq. (1)) increases as a function of the noise level, supporting the hypothesis that top-down connections are crucial for visual processing in noisy conditions. Remarkably, also α, which modulates the amount of bottom-up prediction-error, increases with the noise level for both types of noise. Similar results were obtained when training the top-down parameters for classification rather than reconstruction (i.e., supervised approach). As in the unsupervised case, when freezing the parameters and optimizing exclusively the hyper-parameters for different noise levels, we observed an increase of both bottom-up (α) and top-down (β) errors as a function of the noise level. Yet, Fig. 2C shows that top-down parameters trained for reconstruction proved more robust to noisy images than those trained for classification. Next, we compared the networks' performance with equivalent forward networks. First, we trained (on clean images) four types of forward networks: either having the same forward architecture as the shallow network (labeled "same" in Fig. 2B, and resulting in a slightly smaller number of parameters), or having a larger number of parameters by increasing either the kernel size, or the number of features, or the layers (labeled "kernel", "feat" and "deep", respectively). As summarized in Fig. 2B, both networks implementing predictive coding dynamics (in cyan and green in the figure) perform systematically better than all the forward networks, irrespective of the noise type and level. This result demonstrates that feedback connections, and specifically predictive coding dynamics, can improve overall classification accuracy, and specifically that recurrent connections trained for reconstructions improved network robustness to noise.

3.1.2. Ablation studies

We then investigated how selectively removing the top-down or the bottom-up error term influences the results. Importantly, we focused specifically on the unsupervised network, whose top-down parameters are trained for reconstruction, and that better represents the PC dynamics. In each condition we trained the hyper-parameters after the ablation, including the case of noiseless input. Supplementary Figure 4B shows the actual values of the hyper-parameters in all conditions. As shown in Fig. 2A, when removing the top-down error term, the forward error hyper-parameter increases with the noise levels and doubles its value as compared to the full model (labeled "unsupervised" in the figure). On the other hand, when removing the forward error term, we observed an increase of the feedback term with the noise levels, as in the full model. Concerning the networks performance, Fig. 2C reveals that removing the top-down feedback degrades the accuracy with higher noise levels (especially with Gaussian noise), confirming the conclusion that top-down feedback plays a crucial role in the processing of degraded images.

3.1.3. Adversarial attacks

To further confirm the hypothesis that top-down feedback is important for robustness, we froze the networks (feedforward, full predictive coding, or ablated networks) with manual configurations of the hyper-parameters and then tested their robustness against targeted L_∞ Random Projected Gradient Descent (RPGD) (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2017) and Basic Iterative Method (BIM) (Goodfellow, Shlens, & Szegedy, 2014) attacks, after unrolling them for 10 time-steps to keep their depths constant. We use Foolbox API 2.4.0 (Rauber, Brendel, & Bethge, 2017) and measure the median perturbation required to successfully fool the networks. As shown in Fig. 2D, we observe that networks with higher top-down feedback (two orange bars in the figure have $\alpha = 0$ and $\gamma = \beta = \mu = 0.33$; and

$\beta = 0.5$, $\gamma = 0.3$, $\mu = 0.2$, respectively) reveal better robustness to the attacks as compared to the equivalent forward network (in blue in the figure, with $\gamma = 1$ and all other hyper-parameters set to zero). Interestingly, a forward network leveraging only the feed-forward error shows a similar (lack of) robustness to the attack as the forward network (in red in the figure, both networks having $\gamma = 1$, and $\alpha = 1$ and $\alpha = 2$, respectively; all other hyper-parameters set to zero). Additionally, adding the feed-forward error to the model with top-down connections, slightly reduces its robustness (black bars in the picture, both networks with $\alpha = 1$ and $\gamma = 0.3$, while $\beta = \mu = 0.33$ and $\beta = 0.5$ $\mu = 0.2$, respectively). These results confirm that top-down connections are useful for adversarial robustness (as shown on a different dataset with a different PC implementation by Huang and colleagues Huang et al., 2020), but also suggest that feedforward error correction does not help adversarial robustness. This is likely because the feedforward prediction errors emphasize the input perturbation, which the generative feedback was not trained to account for.

3.2. Deep models

3.2.1. Shared hyper-parameters

Similar to the three-layer network, we examined PResNet18 with a single set of α, β, and γ, that is shared between all the PCoders. In this experiment, we followed the unsupervised training approach explained for the three-layer network using the CIFAR100 dataset. After having optimized the top-down connections for reconstruction, we froze the weights and trained the hyper-parameters to minimize the average cross-entropy loss over five time-steps. We performed this optimization independently on each noise type and noise level of the CIFAR100-C dataset.

Fig. 3A shows the average hyper-parameter values across all 19 noise types relative to those learned using "clean" images. Confirming the results of the shallow model, we observed that the roles of feedback and feed-forward error become more crucial as the noise level increases. Importantly, for all levels of noise, the average accuracy change across time-steps reveals a very robust (but marginal) improvement with respect to the feed-forward ResNet18 (at time step $t = 0$).

3.2.2. Separate hyper-parameters

Encouraged by the results in the "shared" approach described above, we decided to provide each PCoder with a separate set of hyper-parameters. Our reasoning was that different stages of the hierarchical visual processing would benefit differently from the combination of top-down and bottom-up information, thus granting to the network more flexibility in accounting for different representations across different layers.

As in the previous experiment, we trained PResNet18's hyper-parameters on CIFAR100-C images. Moreover, in order to validate our previous results on a more complex dataset, we trained PEffNetB0's hyper-parameters on the ImageNet2012 validation set for five levels of Gaussian and Salt & Pepper noises.

Introducing a separate set of hyper-parameters in each PCoder resulted in a very significant boost in recognition accuracy of both networks, under all conditions. As illustrated in the last column of Fig. 3B, PResNet18 consistently improved the recognition accuracy across time-steps on all noise types and levels, revealing an average improvement around 6% in the most noisy condition, and already after the first time-step. Remarkably, we could replicate these results using the deeper network PEffNetB0 with eight PCoders. As shown in Fig. 3C, PC dynamics with different hyper-parameters per PCoder yielded an impressive increase in accuracy above 20% and above 15% in the worst condition of Gaussian and Salt & Pepper noise, respectively.

In all of our experiments with deep architectures (including those with shared hyper-parameters), we used the same set of noisy stimuli to train hyper-parameters and evaluate the recognition accuracy. Notably, we notice an accuracy drop in the last time-step. We argue that this drop relates to using a cross-entropy loss averaged over time steps, together with the well-known vanishing gradients problem (Bengio, Simard, & Frasconi, 1994; He, Zhang, Ren, & Sun, 2016). As a result, the optimal decrease in the loss value is obtained by improving the accuracy in the intermediate time-steps.

We then investigated the trend of hyper-parameters across PCoders. This analysis shed some light on the role of the hyper-parameters as a function of their hierarchical stage in the network. Remarkably, we obtained very consistent results on both networks, and across different noise types. The first column in panels B and C of Fig. 3 shows the values of hyper-parameters as a function of PCoders for the medium noise level (level 3, results do not change across noise levels, see supplementary Figures 10–12). Regardless of the considered model, we found that the PCoder with the largest amount of feedback error hyper-parameter (indicated by a circle in the figure), is consistently situated at the lower layers of the network, whereas the feedback tends to zero at higher layers. This suggests that the beneficial effects of top-down connections are best achieved at lower layers of the visual hierarchy, where high-level expectations shape low level features to maximize the final classification.

In addition, the second and third columns of Fig. 3B, C confirmed our previous results, revealing how the feedback-error term increases as a function of the noise levels in the PCoder with its highest values (i.e., the second for PResNet18, and either the first or the second in PEffNetB0, depending on the noise type). This result confirms once again the hypothesis that robust object recognition requires more top-down influence (i.e., feedback and feed-forward error terms) as the level of noise increases.

4. Discussion

4.1. Summary of the results

Starting from an established framework in Neuroscience, namely Predictive Coding (PC), we investigated the role of top-down feedback connections in models of vision. The significance of our work spans across Neuroscience and Machine Learning, providing novel contributions to both fields. First, our results demonstrated how predictive coding dynamics increase the network's robustness to various types of noisy stimuli compared to equivalent feed-forward networks. Additionally, systematic optimization of hyper-parameters revealed how the feedback contribution increases with the noise severity, especially in the early stages of the network, providing important information about the role of top-down processes in visual processing. Compared with prior studies, which also showed the benefit of adding generative feedback to forward models (Choksi et al., 2021; Huang et al., 2020; Li et al., 2019; Schott et al., 2019), one original aspect of our approach is our empirical procedure, in which we let the optimization process converge to the optimal solution in each noise level.

4.2. Previous work

Previous studies explored the supervised approach to train feedback connections for classification rather than reconstruction objectives. Feedback Networks (Zamir et al., 2017) introduced top-down and temporal skip connections in a recurrent convolutional module, demonstrating an increase in performance

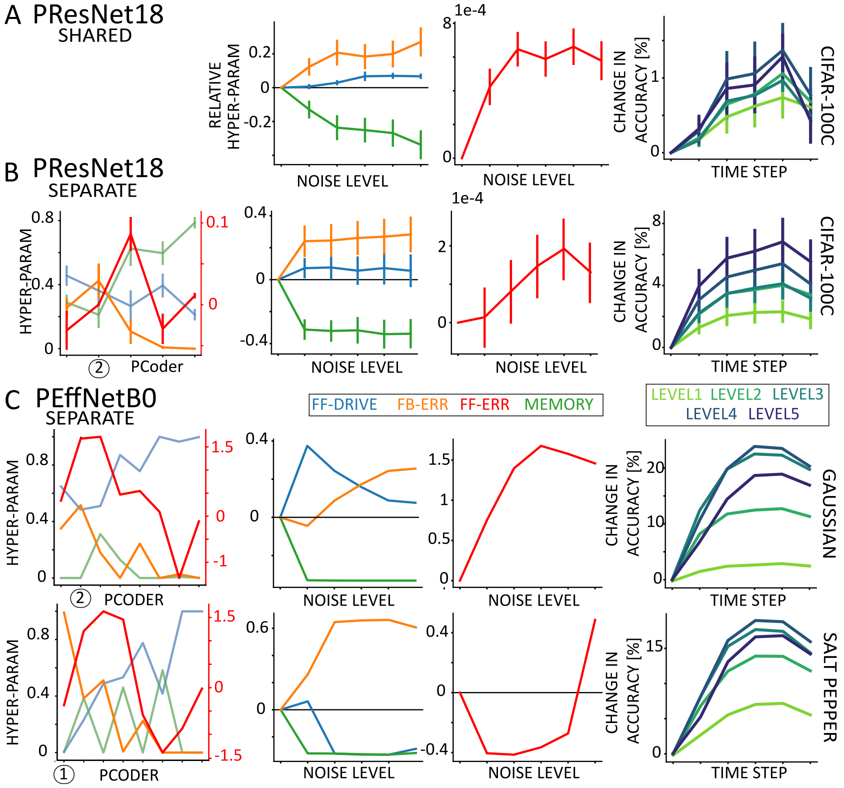


Fig. 3. Values of hyper-parameters and accuracy of the deep predictive coding networks. (A) PResNet18 with shared hyper-parameters that are trained on CIFAR100-C images. (B) PResNet18 and (C) PEffNetB0 with separate hyper-parameters that are trained respectively on CIFAR100-C and ImageNet under Gaussian and Salt & Pepper noise. Plots in the first column show the hyper-parameters as a function of PCoders from input to top layer under medium noise level. The circles indicate PCoders with maximum feedback error. In middle columns, relative values of hyper-parameters are plotted across noise levels. In case of separate hyper-parameters, the PCoder with maximum feedback error is shown. For each noise level, the accuracy difference to the first time-step (i.e. feedforward backbone) is depicted in the last column. Error bars show standard error of the mean (SEM) over 19 CIFAR100-C noise types. In all cases, the networks achieve accuracy gain by utilizing more feedback and forward error as the noise severity increases. See supplementary Figures 7–12 for the absolute values of hyper-parameters and changes in recognition accuracy per noise type and level.

followed by improvements in early features representation, taxonomic predictions, and curriculum learning. Similarly, Nayebi and colleagues (Nayebi et al., 2018) proposed a ConvRNN architecture, incorporating gating and skip connections, which significantly improved object recognition performance. Considering models advocating more explicitly for biological plausibility, Linsley and colleagues (Linsley, Kim, Veerabadran, & Serre, 2018) suggested another recurrent vision model, equipped with horizontal and gated recurrent units (hGRU). Its performance improves specifically in recognition tasks involving long-range spatial dependencies. Supported by experimental studies (Kar et al., 2019), Kubilius and colleagues also proposed a brain-inspired architecture named CorNet, which includes feedback and skip connections. Interestingly, it reveals high neural similarity to cortical visual areas such as V4 and IT (Kubilius et al., 2018).

In the PC domain, Chalasani and Principe (2013) proposed a hierarchical, generative model based on PC dynamics, including context-sensitive priors on the latent representations. Their architecture demonstrated how top-down connections from higher layers are instrumental in solving lower layers ambiguities, providing some noise robustness. The model proposed in Wen et al. (2018) is the closest one to ours. Despite following PC dynamics and the principal similarities, their model presents some critical limitations. More specifically, all weights are trained for object recognition only at the last time step, resulting in a biologically implausible behavior, in which near-chance performance is observed until the final iteration. A more in-depth comparison between this work and our proposed method is presented in Choksi et al. (2021). In this previous work (Choksi et al., 2021), we introduced the Predify package used to augment any forward network with predictive coding dynamics, and we investigated the robustness of different networks' to adversarial attacks (considering a fixed and pre-determined set of hyper-parameters). Finally, Huang and colleagues (Huang et al., 2020) implemented unsupervised feedback connections by optimizing for "self consistency" between the input image features, latent variables and label distribution. Despite a different dynamics, PC principles inspired their implementation, which also provided some robustness against gradient-based adversarial attacks on Fashion-MNIST and CIFAR10.

4.3. Insights for and from neuroscience

It is possible to implement the role of top-down feedback either as an unsupervised, generative process which predicts lower layers' activities, or as a supervised, discriminative process to optimize classification. Besides being more biologically plausible, our simulations with the shallow model revealed that the unsupervised approach is more robust to noise than the supervised one, as shown in Fig. 2B. However, when trained with supervision, feedback connections do not converge to the unsupervised solution, as shown in Figure 5C which compares the reconstruction errors in shallow models optimized for classification (supervised) or reconstruction (unsupervised).

Interestingly, when we independently optimized each PCoder in deeper networks (roughly equivalent to different brain regions across the hierarchy of visual processes), we observed consistently higher modulation of top-down activity in lower regions, and relatively less top-down feedback in higher areas. Choksi et al. (2021) further demonstrate that the proposed biologically-inspired feedback dynamics iteratively project the noisy inputs towards the learned data manifold, similar to previous studies using different approaches (Jalal, Ilyas, Daskalakis, & Dimakis, 2017; Meng & Chen, 2017; Samangouei, Kabkab, & Chellappa, 2018; Shen, Jin, Gao, & Zhang, 2017). Future research may test this prediction directly in biological brains by recording cortical activity at different stages of the visual hierarchy, and validate the hypothesis that the influence of top-down connections (as measured in increased spike rate or synaptic efficacy) increases with noise level in early visual areas (Oude Lohuis et al., 2022).

Our results demonstrated how top-down and bottom-up processes influence perception in different challenging conditions. However, how does the brain modulate each term's contribution (i.e., each hyper-parameter) during natural vision? Attention mechanisms may be responsible for the regulation of top-down processes by increasing feedback response during noisy conditions (Baluch & Itti, 2011; Feldman & Friston, 2010). Accordingly, it could be possible to envision a model inspired by current transformer architectures where an attention system modulates hyper-parameters based on input features or top-down expectations (for example based on the match between keys and queries, in which queries are either inferred from the data or set arbitrarily based on some context-based prior) (VanRullen & Alamia, 2021). In fact, top-down and bottom-up streams may be helpful or detrimental in different conditions. Our results reveal that feed-forward error corrections are beneficial for noisy images, but do not improve accuracy in the case of adversarial attack. One may speculate that, in the latter case, the attack perturbs the sensory information such that it is more unreliable than a noisy input, making it necessary to rely more on top-down expectations rather than on bottom-up features. Further study may be needed to test this hypothesis directly. On the other hand, top-down information may be harmful in certain conditions, as in visual hallucinations: in this case top-down priors may lead to perceive items that are not present in the image, as we showed in a previous study (Pang, O'May, Choksi, & VanRullen, 2021). De facto, expectation is another important process that modulates top-down feedback in the human brain (De Lange, Heilbron, & Kok, 2018; Summerfield & De Lange, 2014). In our model, the forward pass initializes the activity in each layer based on the first processing of the input (i.e., without the recurrent PC dynamic). However, it is possible to initialize the network's activity based on top-down beliefs, according to PC dynamics: the last layer of the hierarchy encodes the predictions of the expected input (i.e., a given class in a classification dataset), and propagates such predictions to initialize the activity of lower layers, similarly to the brain processes involved in sensory expectations (Kok & de Lange, 2015; Summerfield & Egner, 2009). Future work could explore how such expectations may influence the network behavior and accuracy. Besides, there has been a raising excitement in looking at the advantage of learning in Predictive Coding, over the prevailing back-propagation method (Millidge et al., 2022; Song, Lukasiewicz, Xu, & Bogacz, 2020; Song et al., 2022). In addition to this excitement, our results and others reveal a new perspective in the robustness of dealing with noise and adversarial attacks. It would be exciting in future research to see how these efforts together push towards learning paradigms based on Predictive Coding rather than back-propagation ones.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

RV is supported by an ANITI (Artificial and Natural Intelligence Toulouse Institute) Research Chair (grant ANR-19-PI3A-0004), and two ANR grants AI-REPS (ANR-18-CE37-0007-01) and OSCI-DEEP (ANR-19-NEUC-0004).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2022.10.020.

References

Alamia, A., & VanRullen, R. (2019). Alpha oscillations and traveling waves: Signatures of predictive coding? *PLoS Biology*, *17*(10), Article e3000487.

Baldeweg, T. (2006). Repetition effects to sounds: Evidence for predictive coding in the auditory system. *Trends in Cognitive Sciences*.

Baluch, F., & Itti, L. (2011). Mechanisms of top-down attention. *Trends in Neurosciences*, *34*(4), 210–224.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439).

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Chalasani, R., & Principe, J. C. (2013). Deep predictive coding networks. arXiv preprint arXiv:1301.3541.

Choksi, B., Mozafari, M., Biggs O'May, C., Ador, B., Alamia, A., & VanRullen, R. (2021). Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *Advances in Neural Information Processing Systems*, *34*, 14069–14083.

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *364*(1521), 1211–1221.

Fukushima, K. (1980). Neocognitron:A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193—202.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology*, *120*(3), 453–463.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701.

Huang, Y., Gornet, J., Dai, S., Yu, Z., Nguyen, T., Tsao, D., et al. (2020). Neural networks with recurrent generative feedback. arxiv, cs. arXiv preprint arXiv:2007.09200.

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Jalal, A., Ilyas, A., Daskalakis, C., & Dimakis, A. G. (2017). The robust manifold defense: Adversarial training using generative models. arXiv preprint arXiv:1712.09196.

Kar, K., & DiCarlo, J. J. (2021). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, *109*(1), 164–176.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983.

Kevin S. Walsh, A. C. R. G. O., & McGovern, D. P. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*, 242–268.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), Article e1003915.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kok, P., & de Lange, F. P. (2015). Predictive coding in sensory cortex. In *An introduction to model-based cognitive neuroscience* (pp. 221–244). Springer.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. Article 408385, BioRxiv, Cold Spring Harbor Laboratory.

Li, Y., Bradshaw, J., & Sharma, Y. (2019). Are generative classifiers more robust to adversarial attacks? In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of machine learning research*: *vol. 97*, *Proceedings of the 36th international conference on machine learning* (pp. 3804–3814). PMLR.

Linsley, D., Kim, J., Veerabadran, V., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated-recurrent units. arXiv preprint arXiv:1805.08315.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Meng, D., & Chen, H. (2017). Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security* (pp. 135–147).

Millidge, B., et al. (2022). A theoretical framework for inference and learning in predictive coding networks. arXiv:2207.12316.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., et al. (2018). Task-driven convolutional recurrent models of the visual system. arXiv preprint arXiv:1807.00053.

Nguyen, A., Yosinski, J., & Clune, J. (2014). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. arxiv, cs. arXiv preprint arXiv:1412.1897.

Oude Lohuis, M., Pie, J., Marchesi, P., et al. (2022). Multisensory task demands temporally extend the causal requirement for visual cortex in perception. *Nature Communications*, *13*(2864).

Pang, Z., O'May, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. arXiv preprint arXiv:2102.01955.

Pawel Zmarz, G. K. (2016). Mismatch receptive fields in mouse visual cortex. *Neuron*, *4*, 766—772.

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS Computational Biology*, *15*(5), Article e1007001.

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.

Rauber, J., Brendel, W., & Bethge, M. (2017). Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable machine learning in the wild workshop, 34th international conference on machine learning*. URL http://arxiv.org/abs/1707.04131.

Samangouei, P., Kabkab, M., & Chellappa, R. (2018). Defense-GAN: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605.

Schott, L., Rauber, J., Bethge, M., & Brendel, W. (2019). Towards the first adversarially robust neural network model on MNIST. In *International conference on learning representations*. URL https://openreview.net/forum?id=S1EHOsC9tX.

Shen, S., Jin, G., Gao, K., & Zhang, Y. (2017). APE-GAN: Adversarial perturbation elimination with GAN. arXiv preprint arXiv:1707.05474.

Song, Y., Lukasiewicz, T., Xu, Z., & Bogacz, R. (2020). Can the brain do backpropagation?—Exact implementation of backpropagation in predictive coding networks.. *Advances in Neural Information Processing Systems*, *33*, 22566–22579.

Song, Y., et al. (2022). Inferring neural activity before plasticity: A foundation for learning beyond backpropagation. BioRxiv.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. *Journal of Neuroscience*, *30*(9), 3531–3543.

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

VanRullen, R., & Alamia, A. (2021). GAttANet: Global attention agreement for convolutional neural networks. In *International conference on artificial neural networks, 281-293*.

VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655–668.

VanRullen, R., & Thorpe, S. J. (2001b). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., & Liu, Z. (2018). Deep predictive coding network for object recognition. In *International conference on machine learning* (pp. 5266–5275). PMLR.

Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, *5*, 674.

Yan, S., Fang, X., Xiao, B., Rockwell, H., Zhang, Y., & Lee, T. S. (2019). Recurrent feedback improves feedforward representations in deep neural networks. arXiv preprint arXiv:1912.10489.

Zamir, A. R., Wu, T.-L., Sun, L., Shen, W. B., Shi, B. E., Malik, J., et al. (2017). Feedback networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1308–1317).