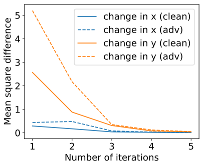
We thank the reviewers for their thoughtful feedback. We are encouraged that all the reviewers found our work to be creative in porting predictive coding and Bayesian brain theories in neuroscience to deep learning models using a mathematically rigorous framework. We want to assure the reviewers that we will heed their advice such as making the figures more informative and interpretable as well as stating the theorems rigorously. We want to thank R2 and R3 for their insightful questions which actually align with some future works that we have envisioned based on this paper. We will add a session in the revised paper to answer them. Due to space limit, we address the main concerns here.

Model description R4 **Analogous version of AdaReLU, AdaPool, and losses in a purely feedforward model** We would like to clarify this potential misunderstanding. The fundamental difference between a CNN-F and a CNN is recurrent generative feedback; the CNN-F's initial feedforward step is equivalent to a feedforward CNN. The CNN-F's feedback and subsequent feedforward steps use the adaptive layers and multiple losses. AdaReLU and AdaPool are our own creations and are required to perform Bayesian inference. They modulate the top-down feedback and change the ReLU and MaxPool units in the feedforward pathway. Since AdaReLU, AdaPool, and generative losses (reconstruction and conditional latent likelihood loss) all need to act on feedback signals, there are no analogies in a purely feedforward model.

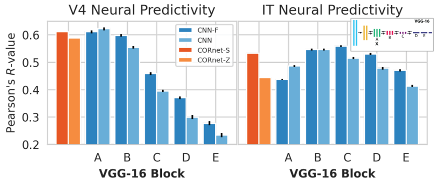
R4 **How do losses affect performance?** We included the ablation study in Appendix (line 522).

R4 **Convergence of CNN-F** Empirically, we find that 5 iterations lead to a stable solution. We compute the mean square difference between successive iterations to show convergence. The changes in reconstructed images x and logits y are shown in Figure 1. As expected, clean images converge faster than adversarial images.

R3 **Restriction to the DGM model** Thanks for noticing that the ideas can be more general. However, given CNN as the architecture for classification, DGM is required to enforce Bayes rule. For other architectures, we can use self-consistency to define iterative inference accordingly.



Neural comparison R1 **Comparison with CORnet** CORnet proposes recurrent connections within each cortical area while CNN-F proposes feedback across areas. We included the CORnet-S and CORnet-Z neural similarity scores. The CNN-F with the VGG-16 architecture outperforms the CORnet-S, and CORnet-Z in V4 and IT neural similarity.



R4 **The predictivity scores do not match brain-score.org.** Due to computational limitations, we were previously not able to perform an extensive hyperparameter search on which layer to predict neural activity for V4 and IT. We ran this study with the results in Figure 2.

Neural predictivity discussion (R4) According to Schrimpf et al. (2018), the correlation between accuracy and Brain-Score becomes insignificant models with ImageNet top-1 accuracies greater than 70%. This can be expected as V4 and IT perform other roles besides strictly object classification. As for the decreased classification performance of CNN-F, please also refer to a related question of R3 on line 50.

Adversarial robustness (AR) R1, R4 **Compare against adversarially robust models** The arguably most established method to improve the AR of neural networks is adversarial training (AT) (Madry et al., 2018). In the paper, we show that CNN-F can further improve the robustness of adversarially trained CNNs (line 229).

R4 **Trade off between clean and adversarial accuracy** We fully understand your concern. Trade-off between robustness and accuracy is still an active research area (Hongyang Zhang et al., 2019, Yao-Yuan Yang et al., 2020). In fact, we think that CNN-F provides a new lens to analyze this trade-off. We find that more iterations are needed for adversarial (harder) images and less iterations are needed for clean (easier images) (Fig 6. (f)). By varying the number of iterations, a CNN-F can fit different function classes to clean and adversarial images while achieving high accuracy.

R4 **CNN-F v.s. regularization** Indeed, regularization is a useful way to improve AR. Yuxin Wen et al. (2020) shows that AT regularizes NNs to concentrate samples around decision boundaries. Inspired by predictive coding, we propose self-consistency as a different mechanism for robustness. Our approach is compatible with other regularization such as AT. We show that feedback improves upon AT (Fig 6. (d)) and also generalizes better to unseen attacks (Fig 6. (e)).

R4 **Why not stick to ImageNet for AR** MNIST, Fashion-MNIST and CIFAR-10 are standard benchmarks across the AR community and haven't been solved yet. We also conducted experiments on CIFAR-10 and found that CNN-F has higher adversarial accuracy than CNN in later iterations and maintains as high clean accuracy in earlier iterations. Due to the space limit, we omit the results here but will put into the revised paper.

Training difficulties *(*R3*)* **Classification performance** There are three reasons for the decreased classification performance. First, generative classifier has higher asymptotic error compared to discriminative classifier when there is a mismatch between data and model (Ng et al., 2002). Second, it is harder to optimize a recurrent model like CNN-F than a feedforward model like CNN. Third, CNN-F achieves higher accuracy for clean images in early iterations (Fig. 6 (f), $\epsilon = 0.0$), indicating shorter processing time for easy images.

Training from scratch We need to use Instance Normalization (IN) to train CNN-F from scratch. We also used IN in CNN for fair comparison. The accuracies are in Table 1. For the pretrained model, we were not using any layer normalization in both CNN and CNN-F.

	Top-1	Top-5
CNN	58.85	77.50
CNN-F	56.68	74.00