ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

# Analysing the impact of brain-inspired predictive coding dynamics through gradient based explainability methods

Bhavin Choksi[1], Gionata Paolo Zalaffi[2], Giovanna Maria Dimitri[3], and Gemma Roig[1]

1- Goëthe Universität, Frankfurt, (Germany)
2- Università La Sapienza, Rome, (Italy)
3- DIISM, Università di Siena, Siena (Italy)

**Abstract**.
Multiple theories exist for the role of feedback connections in the brain and in the artificial neural networks, but remain untested using modern tools. In this work, we undertake this task by exploring the utility of explainability methods like GradCAMs[1] in investigating bio-inspired recurrent networks–provided with the *predify*[2] package–that perform hierarchical updates inspired by the predictive coding theory in neuroscience. We report an extensive search with different levels of feedforward and feedback information. Our preliminary results show that the dynamics are able to recover the GradCAMs on noisy images, providing promising avenues for future work aiming to understand the role of recurrence.

## 1 Introduction

Given that biological brains contain a large number of feedback connections, researchers in Artificial Intelligence (AI), with the aim of incorporating their desirable properties, have proposed various recurrent neural networks. While each approach relies on adding certain architectural features into the RNNs, a large number of proposals differ in the nature of the recurrent dynamics incorporated into the networks. What properties these dynamics render to modern neural networks, along with a thorough comparison between the influences of top-down and bottom-up information, remains to-date relatively under-explored. While prior studies have investigated smaller recurrent neural networks and their dynamical properties[3, 4], very few have focussed on large-scale recurrent networks performing well on complex tasks. Our work resembles that of [5] where the authors investigated three bio-inspired feedback connections–task-trained, performing surround suppression, and performing predictive coding updates–and investigated their effects on the representational spaces, albeit from the perspective of robustness.In this work, we start by using a popular method used to ascertain class-based attribution–GradCAM–to explore recurrent networks, specifically models available with the *predify* package that perform recurrent dynamics based on predictive coding. Our choice of using these models is informed due to a few specific reasons. First, predictive coding has been a prominent theory in neuroscience since the early 2000s, along with a recent upsurge in neural network-based implementations. Thus, this will allow any insights we

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933.
Available from http://www.i6doc.com/en/.

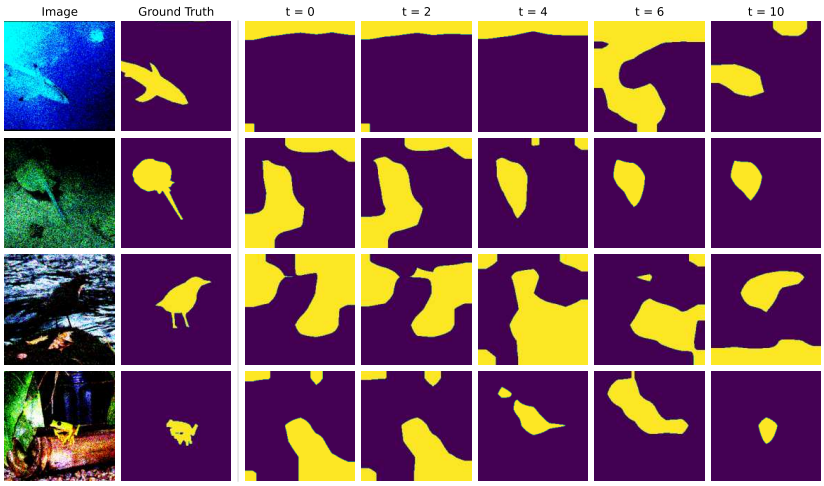



Fig. 1: GradCAM visualizations of noisy inputs (gaussian noise $\sigma = 0.6$) at different timesteps along with the ground truth masks. Over time, the GradCAMs become progressively similar to the ground truth masks. The hyperparameters for the network used are feedforward $\beta = 0.4$ and feedback $\lambda = 0.5$.

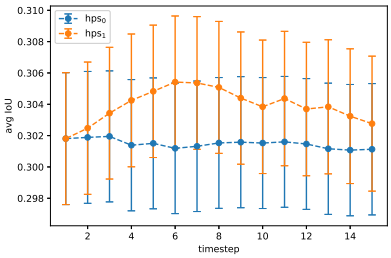
obtain to be of relevance to neuroscience. Second, [2] provides relevant architectures along with a flexible framework–allowing us to weigh the contributions of top-down and bottom-up information, while also providing the possibility of converting the network into a complete feedforward network. These connections can also be trained either using purely unsupervised (reconstruction), supervised (crossentropy), or a combination of both losses.In this paper we report our explorations and preliminary results of using GradCAM on predictive coding networks, and found that the recurrent dynamics can recover the GradCAMs obtained on noisy images over iterations (or timesteps). We also explore this behavior by altering the contributions from top-down or bottom-up information in the networks. Overall, we believe that our results provide a promising use-case of explainability methods for investigating the role of recurrence in neural networks.

## 2 Methods

### 2.1 Predify

In this work, we investigate convolutional neural networks incorporated with recurrent predictive coding dynamics using the *Predify* package[2]. We refer the interested reader to [2] for all the details. Briefly, the resulting recurrent network has N pcoders. Each pcoder consists of an encoding layer $e_n$ (a layer from the feedforward layer) and a decoding layer $d_n$ that predicts the input received. After initial instantiation, the activations across the pcoders are changed using

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.




(a)

| | $\beta$ (feedforward) | $\lambda$ (feedback) | $\alpha$ (error-correction) |
|---|---|---|---|
| $\mathbf{hps}_0$ | 0.20 | 0.05 | 0.01 |
| | 0.40 | 0.10 | 0.01 |
| | 0.40 | 0.10 | 0.01 |
| | 0.50 | 0.10 | 0.01 |
| | 0.60 | 0.00 | 0.01 |
| $\mathbf{hps}_1$ | 0.19 | 0.30 | 0.01 |
| | 0.70 | 0.30 | 0.00 |
| | 0.34 | 0.30 | 0.01 |
| | 0.70 | 0.00 | 0.01 |
| | 0.70 | 0.30 | 0.01 |

(b)

Fig. 2: IoU across timesteps calculated for the two sets of hyperparameters (denoted $\text{hps}_0$ and $\text{hps}_1$) described in Table 2b. Using a bootstrap procedure on 2000 images we evaluated the average IoU for each timestep with its standard deviation (error bars).

the equations [2]:

$$\boldsymbol{d}_n(t) = W^b_{n+1,n}\boldsymbol{e}_{n+1}(t) \tag{1}$$

$$\boldsymbol{e}_n(t+1) = \beta_n W^f_{n-1,n}\boldsymbol{e}_{n-1}(t+1)+\lambda_n\boldsymbol{d}_n(t)+(1-\beta_n-\lambda_n)\boldsymbol{e}_n(t)-\alpha_n\nabla\epsilon_{n-1}(t) \tag{2}$$

where $W^f_{n-1,n}$ are the feedforward weights connecting layer $n-1$ to layer $n$, and $W^b_{n+1,n}$ are the feedback weights. In the equations the coefficients $\beta_n$, $\lambda_n$, and $\alpha_n$ balance the contributions from the feedforward, feedback and error-correction terms respectively. In this work, we start with the pretrained PVGG network provided by the authors and investigate the effects of changing the feedforward and feedback information on GradCAM.

### 2.2 Evaluation

**Dataset:** In this work, we used the large-scale ImageNet-S dataset comprised of 1.2 million training images and 50,000 high-quality semantic masks [6].
**GradCAM:** Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique used for visualizing and interpreting CNN predictions by highlighting key regions in an image that influence the network's decision. It estimates the contribution of a neuron by using the gradient of the output class score relative to the activations. These gradients are combined with feature maps to create a heatmap, demonstrating the regions of the image the network is most sensitive to for making its prediction.[1].

## 3 Experiments

We start by investigating the pretrained predified-VGG16, or PVGG16, network made publicly available by the [2]. For each image, over timesteps we calculate the GradCAM maps on the features obtained after last convolutional layer using the ground-truth category. Visual inspection shows that, for certain configurations (i.e. after injecting some gaussian noise), the activation maps improved over timesteps (see Fig.1).

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

To quantify this behaviour, we measure the commonly used IoU (Intersection over Union) metric between the GradCAMs and ground-truth segmentation masks provided in the Imagenet-S dataset. This quantification also allowed us to automate the exploration in the hyperparameter space (co-efficients $\beta$,$\lambda$, and $\alpha$ in Eq.2); allowing us to evaluate the behaviour of the network by changing the impacts of feedforward and feedback information. Fig.2 shows mean IoU values obtained for two sets of hyperparameter configurations on 2000 samples. The purely feedforward VGG16 network (i.e., the recurrent network at $t = 0$) shows a mean IoU value of $\approx$0.3, with additional recurrent dynamics improving this over a few timesteps. Interestingly, this decreases over further timesteps. We speculate that this could be because either the hyperparameters are not optimally tuned, or that the dynamics haven't converged. The latter is also informed by the trends observed by [2] where the activations themselves converged after certain timesteps.

**Investigating the impact at different layers:** to investigate which layer contributed the most to this correction, we systematically added Pcoders to the network one-by-one. This can be easily done by setting $\beta_i = 1$ and $\lambda_i = \alpha_i = 0$ for all other pcoders. We then measured the mean IoU for each configuration (see Fig.3). As the final IoUs obtained are bound by the feedforward backbone on which the dynamics are added, we report the values normalized by the values obtained for the corresponding feedforward counterpart. This normalization facilitates comparisons across different layers and hyperparameter configurations. We observe that a range of values in the $\beta$-$\lambda$ space improve the

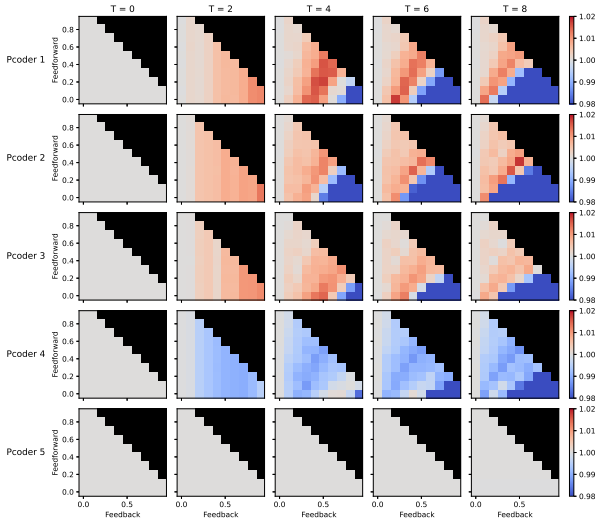



Fig. 3: Using predictive coding dynamics only on a single pcoder: Each row in the figure represents the mean IoU values obtained over timesteps by using a single Pcoder. All the values are normalized w.r.t. the values at timestep t=0. The x- and y-axes show the values of feedback and feedbforward co-efficients ($\lambda$ and $\beta$ respectively) used. The dark region in the top right of a plot is where the sum of the parameters is greater than 1, and thus skipped.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

IoU over timesteps, an effect that is more pronounced for earlier layers in the networks.

**Predictive coding dynamics help in recovering the GradCAMs of noisy images:** Given that these dynamics have been shown to improve the performance of the network on noisy images, and to further validate the approach, we investigated whether we can see this effect on the GradCAMs. To simulate this, we added gaussian noises of varying degrees ($\sigma \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$) to the input images and then calculated the mean IoU as above. For this analysis, as before, we normalized the IoU values (using the values obtained on corresponding feedforward backbones) to compare them across noise levels and restricted our evaluations to only the last layer in the network. We fixed the hyperparameters to be the same across all pcoders. We observed that for lower noise values, the dynamics relied on lower values of feedforward and feedback (and thus on more memory), but didn't help substantially (see Fig.4). On the contrary, for higher noises, the effect seems to be flipped, with higher feedback and feedforward values aiding in recovering the IoU values. This is consistent with previous reports that demonstrate that higher feedback information helps in sustaining network performance[7]

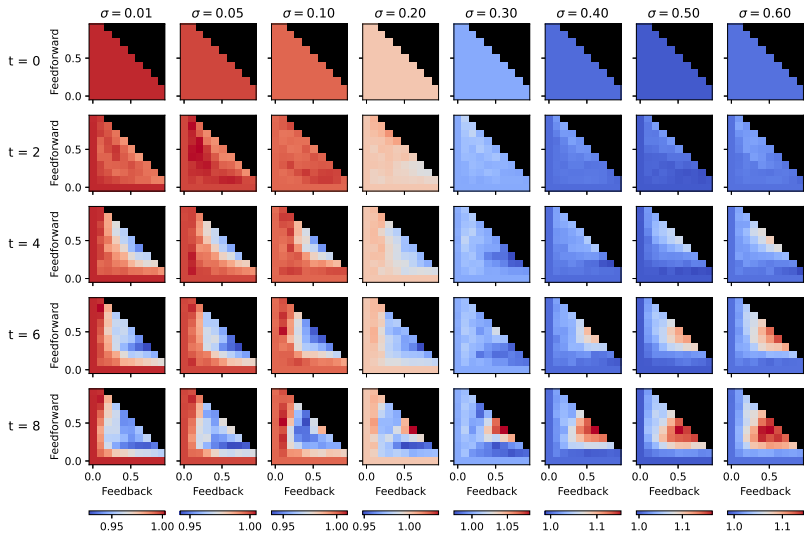



Fig. 4: Grid search with noise-injected images. The plots represent the IoU scores across timesteps (rows) after injection of gaussian noise of various levels (columns)

## 4 Discussion and Future Work

In this work, we explored the possibility of using explainability methods like GradCAMs to investigate recurrent networks performing predictive coding updates. More specifically, changing the hyperparameters, we investigated the impact of different levels of feedforward and feedback signal on the resulting

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

class activation maps. Consistently with previous reports [7], we found that, under noisy conditions, predictive coding updates were indeed helpful, and in our case able to recover the segmentation maps, demonstrating improved reliance of the network on class-specific information in the image. Given the deconvolutional nature of the feedback in such networks, and earlier efforts of using deconvolutions [8] for visualizing neural networks, one would intuit that predified networks can be directly studied using their reconstructions. But, as also argued by [1], such broad pixel-based methods do not allow one to look at "class-discriminative" information. Thus, using reconstructions, or even broad layer based correlations [2], limit the scope of the analysis warranting methods like GradCAMs. Indeed, the strength of our evaluations is contingent upon the utility of GradCAM. There has been a growing awareness about the limitations of such attribution-based methods, with novel propositions being consistently proposed. We intend to adapt and modify our toolset based on new findings and techniques. Nevertheless, our preliminary results are promising, and provide interesting avenues for future research. An immediate pursuit could be to contrast predictive feedback connections to those performing surround suppression.

## Acknowledgments

GMD was the recipient of the 2023 DAAD AInet Fellowship "Generative Models in Machine Learning" https://www.daad.de/en/the-daad/postdocnet/fellows/fellows/. GR was funded by the German Research Foundation (DFG) - DFG Research Unit FOR 5368 (GR).

## References

[1] Ramprasaath R et al. Selvaraju. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[2] Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, Benjamin Ador, Andrea Alamia, and Rufin VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. *Advances in Neural Information Processing Systems*, 34:14069–14083, 2021.

[3] David Sussillo and Omri Barak. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural computation*, 25(3):626–649, 2013.

[4] Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in neural information processing systems*, 32, 2019.

[5] Grace W Lindsay, Thomas D Mrsic-Flogel, and Maneesh Sahani. Bio-inspired neural networks implement different recurrent visual processing strategies than task-trained ones do. *bioRxiv*, pages 2022–03, 2022.

[6] Shanghua et al. Gao. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7457–7476, 2022.

[7] Andrea Alamia, Milad Mozafari, Bhavin Choksi, and Rufin VanRullen. On the role of feedback in image recognition under noise and adversarial attacks: A predictive coding perspective. *Neural Networks*, 157:280–287, 2023.

[8] MD Zeiler and R Fergus. Visualizing and understanding convolutional networks. arxiv. *arXiv preprint arXiv:1311.2901*, 2013.