EMERGENCE OF SURPRISE AND PREDICTIVE SIGNALS FROM LOCAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical predictive models are often used to model cortical representations. These models exploit the local or global computation of predictive signals in the neural network, but their biological plausibility is limited as it is currently unknown whether cortical circuits perform such computations at all. This paper seeks to further investigate the inverted Forward-Forward Algorithm, a biologically plausible innovative approach to learning with only forward passes, in order to demonstrate that hierarchical predictive computations can emerge from a simpler contrastive constraint on the network's representation. Through the identification of compelling similarities between our model and hierarchical predictive coding, as well as the examination of the emergent properties of resulting representations, we advance the hypothesis that the computational properties that emerge in neocortical circuits, widely acknowledged as the basis of human intelligence, may be attributed to local learning principles.

1 INTRODUCTION

The neocortex contains hierarchically layered circuits with rich feedforward and feedback connections (Chaudhuri et al., 2015; Bassett & Sporns, 2017; Siegle et al., 2021). The feedforward (or bottom-up) pathway involves the transfer of information from lower-level sensory areas to higher-level association areas, leading to the extraction of input-specific features. In contrast, the feedback (or top-down) pathway aids in the integration of high-level information by relaying signals from higher-level areas to lower-level ones. Though often assumed to propagate learning-related errors (LeCun et al., 2015), the functional role of feedback connections has been implicated in many different perceptual and cognitive abilities such as attention, efference copies, memory retrieval, etc. (Mechelli et al., 2004; Gilbert & Li, 2013).

In the Bayesian view of cortical feedback, the bidirectional flow of information enables integration of ongoing sensory inputs with existing cortical representation of prior contextual information (Badre & Nee, 2018; Khan & Hofer, 2018; Froudarakis et al., 2019). A specific way to convey such contextual information is through surprise- and familiarity-based signals. When incoming sensory input corresponds to expectations, the surprise signal is minimal. When the input deviates from expectations, the surprise signal rises, indicating novelty or unfamiliarity. It has been shown that novel stimuli elicit increased neural activities that decrease over repeated presentations (Garrett et al., 2023; Piet et al., 2023). Such surprise signals contain information for the brain to improve its internal model of the external world, which leads to more refined and accurate expectations that can direct behavior Wolpert et al. (1998); Kawato (1999); Schenck (2008).

However, the mechanisms through which feedforward and feedback connections interact and generate such surprise signals are not well understood. An influential theory in neuroscience, predictive coding (Rao & Ballard, 1999; Jiang & Rao, 2022), postulates that feedback circuits deliver top-down spatiotemporal predictions of lower-level neural activities, while feedforward circuits send bottom-up prediction errors (surprises) to higher levels. Despite its popularity, minimizing prediction error is an iterative process based on gradient descent, which requires physical time for convergence and implies the symmetry between the feedback and feedforward synaptic weights, limiting its biological plausibility (Rao & Ballard, 1999; Lillicrap et al., 2016). Additionally, computing the prediction error requires a one-to-one correspondence between the predictive neurons and error neurons, which have not been confirmed experimentally (Jordan & Keller, 2020).

Here, we present a simple and biologically plausible mechanism that captures the spatiotemporal predictive nature of cortical processing without generating explicit predictions. Our model is based on the Forward-Forward model (Hinton, 2022), a recently introduced form of contrastive learning. We inverted the original Forward-Forward objective to minimize the activity of positive training data and maximize the activity of negative training data, where we now refer to the level of activity as surprise. Such an objective promotes activity cancellation when top-down labels match bottom-up sensory input. As a consequence, different layers across the hierarchy learn to predict each other's activity to enable such minimization (or cancellation) of activities.

Our most significant contributions are:

- we demonstrate that our model reproduces both hierarchical and temporal properties of predictive computations by means of generating information flows that lead to surprise and cancellation signals (Secs. 3.1 to 3.2);
- we illustrate a mechanistic understanding of the emergence of such information fluxes by tracing their origin to the circuit's ability to implement spatiotemporal cancellation across layers of activity (Secs. 3.2 to 3.3)
- we derive an equivalence between our contrastive learning rule and a unique form of three-factor Hebbian plasticity with compelling connections to predictive coding, thereby highlighting the biological plausibility of our model (Sec. 3.4).

These results demonstrate that the application of a fundamental contrastive learning technique that integrates surprise and cancellation dynamics generates predictive spatiotemporal properties. This suggests that these properties, which are generally regarded to be distinctive characteristics of neocortical computations, can be generated by simplified learning principles.

2 Model Architecture and Learning Scheme

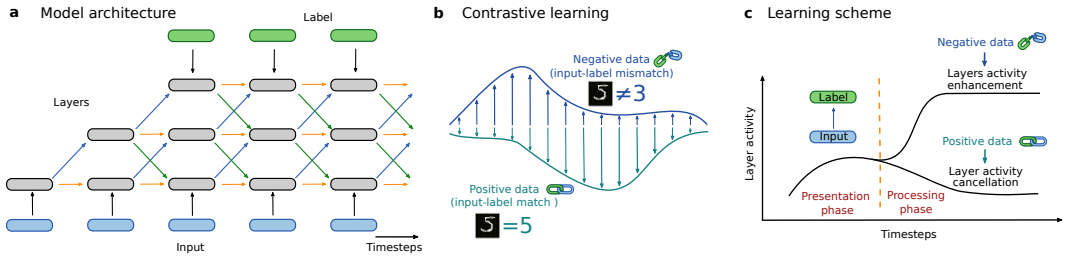

Figure 1: Simple illustrations representing model architecture and learning scheme. a) Model architecture is shown where inputs are clamped to the bottom and labels are clamped to the top of the network. b) Forward-Forward contrastive learning schematic with definition of positive and negative datasets, where the label mismatches or matches the sensory input. c) Learning scheme of the model is shown where the training phase proceeds in two steps - the presentation and processing phase respectively. In the processing phase, positive data should have a low activity, whereas negative data should have a high activity.

We extend the Forward-Forward model (Hinton, 2022; Ororbia & Mali, 2023; Ororbia, 2023), a back-propagation-free learning paradigm regarded as a form of contrastive learning. Our model consists of a hierarchical network with multiple layers, with label information clamped at the top and input at the bottom (Fig. (1a)). The information flow of the input is considered to be bottom-up, whereas the information flow of the label is considered to be top-down. Each layer receives input from a lower layer, a higher layer, and itself at the previous timestep (lateral or recurrent connections).

The training process involves defining positive and negative datasets. Positive (negative) data is one in which the label and input do (not) match. For the negative dataset, the surprise of each layer is increased, resulting in increased layer activity (Fig. (1b)). For the positive dataset, the surprise is diminished resulting in decreased activations. Surprise is defined in terms of the individual layer activations at each time $\vec{x}_{\text{layer}}(t')$:

$$\mathcal{L}_{\text{layer}} = \sum_{t'} \mathcal{S}(t') = \sum_{t'} (-1)^{\eta} \sigma(\vec{x}^{T}_{\text{layer}}(t') \vec{x}_{\text{layer}}(t') - \theta) \tag{1}$$

where the parameter $\theta \in \mathbb{R}$ is the threshold set for the surprise calculation, while η equals zero or one ($\eta = 0, 1$) for negative and positive data samples respectively. The final ingredient is the σ which is taken to be a soft-plus function.

Training involves conducting forward passes for positive and negative data simultaneously. The most essential aspect of training is that only forward passes are made at all times. This makes the Forward-Forward algorithm biologically plausible, as it avoids the non-locality of weight information inherent to backpropagation. Only intralayer weights, which are local properties of individual neurons, are modified during training. In addition, we demonstrate that this algorithm is theoretically equivalent to particular types of Hebbian learning (see Supplementary Material). The network learns to process input and label by combining bottom-up and top-down information flows to generate activations that reflect positive and negative data points, respectively, upon training.

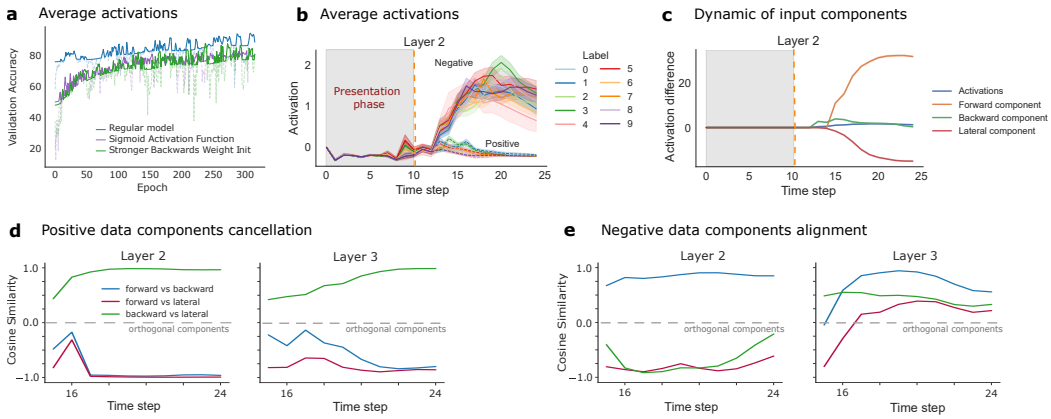

Figure 2: Figures for validation accuracy, layer-wise activation progression throughout time, layer-specific cancellation, and cancellation patterns between the components forming a layer's activation update. a) Accuracy (y-axis) over time (x-axis) is shown for various configurations of the network. We show a deterioration for both a stronger label clamped weight initialization, as well as with a sigmoid activation instead of leaky ReLU. b) For the second layer in the network, the average activations (y-axis) obtained over 1000 samples are shown per class over time (x-axis), for both positive and negative data. Negative data induces a large and sustained surprise signal with raising activities after the label is shown. Positive data has a very small surprise and returns to baseline low activity. c) For the second layer in the network, the activations for positive data averaged across 1000 samples, are broken into their pre-synaptic components (y-axis) and plotted across time (x-axis), which show strong cancellation as indicated by the resultant summed post-synaptic activity (blue). d) For the second and the third (middle) layer, the cosine similarity (y-axis) of positive data activations averaged across 1000 samples is shown between the pre-synaptic components over time (x-axis). The cancellation profile is consistently strong within these layers. e) Same as the previous plot, but for negative data.

We revise the training procedure to increase similarity to signal processing in biological networks (Fig. (1c)). First, the input is presented to the network (presentation phase), followed by the generation and introduction of the label (processing phase). Each phase is comprised of a number of timesteps (10 and 15, respectively). Training does not occur during the presentation phase; rather, training occurs only when label information is received during the processing phase. Each layer's surprise decreases (increases) based on whether the associated label matches (mismatches) the presented input. This resembles cortical processing where surprise signals, often resulting in increased activity, are thought to arise from a mismatch between our sensory inputs and top-down signals reflecting our internal beliefs or world model – here, the label information. In other words, when the input and label match each other, the activity is lower than when a mismatch, or surprise, occurs.

3 Results

We train the model on the MNIST dataset following the scheme highlighted in Fig. (1). For every iteration, a single MNIST image is selected and presented as an input to the network (presentation phase of 10 timesteps). Following this presentation phase the label is introduced, while still present-

ing the image, and the network processes both input and label information (processing phase of 15 timesteps). We first focus on the spatial integration of bottom-up and top-down information flows. Learning follows as per Eq. (1) and accuracy is computed as outlined in Hinton (2022): for each input image x all possible labels (classes 0 to 9) are introduced to the network, we deem the input image to be accurately processed if the surprise for the correct label is lower than for any other label.

We train a 5 layer network with 700 neurons per layer minimizing Eq. (1). We used RMSProp as the optimizer with learning rate $5 \cdot 10^{-5}$, batch size 500, and Leaky ReLU as the transfer function for all units. Weight initializations and further details can be found in the available repository [1]. Individual layer activations were substantially different across training, during the presentation and processing phases, and the activity appeared to be highly dynamic.

The 5-layer, 25 timestep model accuracy achieved 95% test accuracy upon training (Fig. (2a)). Better test accuracy and resilience to extended timesteps was found by amplifying the weights connecting the labels to the last hidden layer. Performance peaked at an arbitrary weight amplification, after which further increases in the factor led to negative returns in performance. Different activation functions were attempted, and sigmoid consistently showed to perform worse than ReLU derivatives, Fig. (2a). A 3-layer, rather than 5-layer, network of the best performing architecture was able to achieve 98% test accuracy on MNIST. However this 3-layer configuration will not be discussed further in the text as a 5-layer network offers richer inter-layer dynamics, and corresponds more closely to the hierarchy layers in the cortex.

3.1 Hierarchical emergence of surprise and cancellation signal

By analyzing layer activity via L2 norm over time, we were able to confirm that the model learned to dynamically suppress neural activity across both layers and time whenever the input image matched the respective label (Fig. (2b)). The difference between negative and positive activations showed a clear divergence upon label presentation, Fig. (2b). This trend was the result of the contribution of multiple input components to the layers, Fig. (2c). In order to understand how these input components were driving the decrease in activity for positive data (the cancellation upon label presentation) – and the increase in activity for negative data (surprise signal), we focused on late timesteps (16-25). We verified whether different input components were aligned or misaligned with each other, therefore issuing a cancellation in the overall activities. To this end we computed the cosine similarity (scalar product) between all pairs of the three input components (Figs. (2d) to (2e)). For positive data the forward component was anti-aligned to both the backward and lateral components, suggesting that the decrease in activity was due to the bottom-up (forward) information flow canceling the top-down (backward) and recurrent (lateral) information flows (Fig. (2d)). Conversely, for negative data, the top-down and bottom-up information flows showed a higher degree of alignment, resulting in increased activations (surprise signal) (Fig. (2e)).

This analysis shows that our model reproduces hierarchical properties of predictive computations by generating information flows that result in surprise and cancellation signals. These signals are associated with the processing of negative and positive data, respectively, and involve distinct network information flows based on the dynamic cancellation of multiple input components.

3.2 Dynamical emergence of surprise and cancellation signals

We next interrogated the temporal characteristics of the cancellation and surprise information flows. We began by plotting activation differences across all layers (as performed in Fig. (2c)) in Fig. (3a). This demonstrated that the encoding of positive versus negative data diverged more rapidly between early layers compared to later layers. To confirm this, we analyzed activations for positive data after introducing the label (Fig. (3b)). These clarified that early layer activations cancel, returning to a lower activation state, prior to late layers establishing a bottom-up cancellation ordering. We also analyzed the cosine similarity between activations of consecutive layers for both positive and negative data, as shown in Fig. (3c). In the case of negative data, we revealed that early layers aligned (cosine similarity was greater) more quickly than later layers. Together, these findings shown in Figs. (3a) to (3c) indicate that alignment and anti-alignment dynamics across layer activations,

[1] github [REDACTED FOR REVIEW]

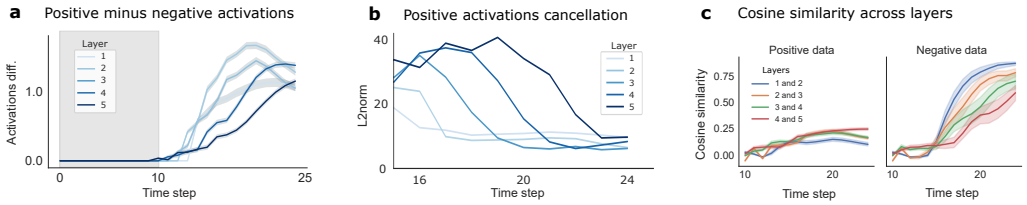

Figure 3: Various representations for activations over time. a) The negative minus positive activations (differences) over time (x-axis) are shown as a measure of the negative activation surprise signal, offset from the baseline of our positive actions. b) The L2 norm of positive activations (y-axis) is shown across time (x-axis), visualizing the cancellation cascade. c) The cosine similarity (x-axis) is shown for all adjacent layer pairs over time (y-axis) for both positive and negative data. This shows constructive signal evolving bottom-up for negative data, while a more orthogonal signal evolves for positive data.

leading to surprise and cancellation signals, originate in early layers despite the introduction of the label at the top of the hierarchy.

3.3 Interpreting latent representations which drive cancellations

In the following analysis, we seek to understand the intricate mechanics governing the latent space. By examining the average class-wise activations across various PCs, we discern a looping behavior inherent to the network mechanics for positive data. We further delve into a temporal explanation of these latent spaces, affirming the aforementioned looping dynamics. Lastly, our quantitative analysis on the representation of all layers across time yields a temporal hierarchy of representation decodability that is opposite in direction for presentation and processing timesteps.

In order to understand the latent space mechanics driving cancellations on positive data, we first plotted the average class-wise activations for various PCs in lower dimensions (Fig. (4a) and Fig. (4b)). We observe that the lower-order PCs do not offer a strong representation of the class, but they do offer a consistent path through the space that starts and ends at the same point. This is in line with the mechanics of the network under positive data, which starts from an initially low activity and recovers to a similarly low activity following all the timesteps where the label is presented. In the higher-order PCs, the same looping mechanics are shown, however now the classes are represented in a separable manner. For negative data, the looping behavior in the latent space does not occur, and the latent states drive away from the origin erratically.

A temporal analysis of the same latent space in two dimensions was conducted on the higher-order PCs with stronger class representation (Fig. (4c)). Activations start in the middle of the represented structure, before diverging and returning back to the beginning. This analysis shows behavior corroborating the above looping mechanics, driven by the recovery of the initial low activation state after initial excitement.

To understand if the above representations effectively captured the variance of the underlying data, we trained a slew of multilayer perceptrons (MLPs) on the latents of 1000 samples at every timestep, reduced in dimensionality by a sliding-window of three-paired PCs. This analysis shows high label decodability for the PCs plotted above which showed cleaner separability (4-6) (Fig. (4d)), motivating the choice of these particular PCs. Additionally, most of the variance within this data is captured within the first 20 PCs.

To form a more quantitative measure of the directionality of information flow throughout, we trained 5 MLPs on the latents of each of the 5 layer-wise activations for positive data, reduced to their first 50 PCs. High decodability indicates a strong representation among these latents, and as such we can observe two distinct cascades of decodability spikes: one for the first 10 presentation timesteps; and another for the latter 15 processing timesteps (Fig. (4e)). These decodability spikes cascade in separate directions, revealing a layer-wise temporal ordering in opposite directions. In the presentation timesteps 0-10, the decodability is high for the lower layers followed by the upper layers, indicating a bottom-up temporal ordering. By contrast, in the processing timesteps 10-25, the decodability is high for the upper layers followed by the lower layers. This analysis examines population-level coding and the mechanism of cancellation across multiple layers in greater detail. It demonstrates

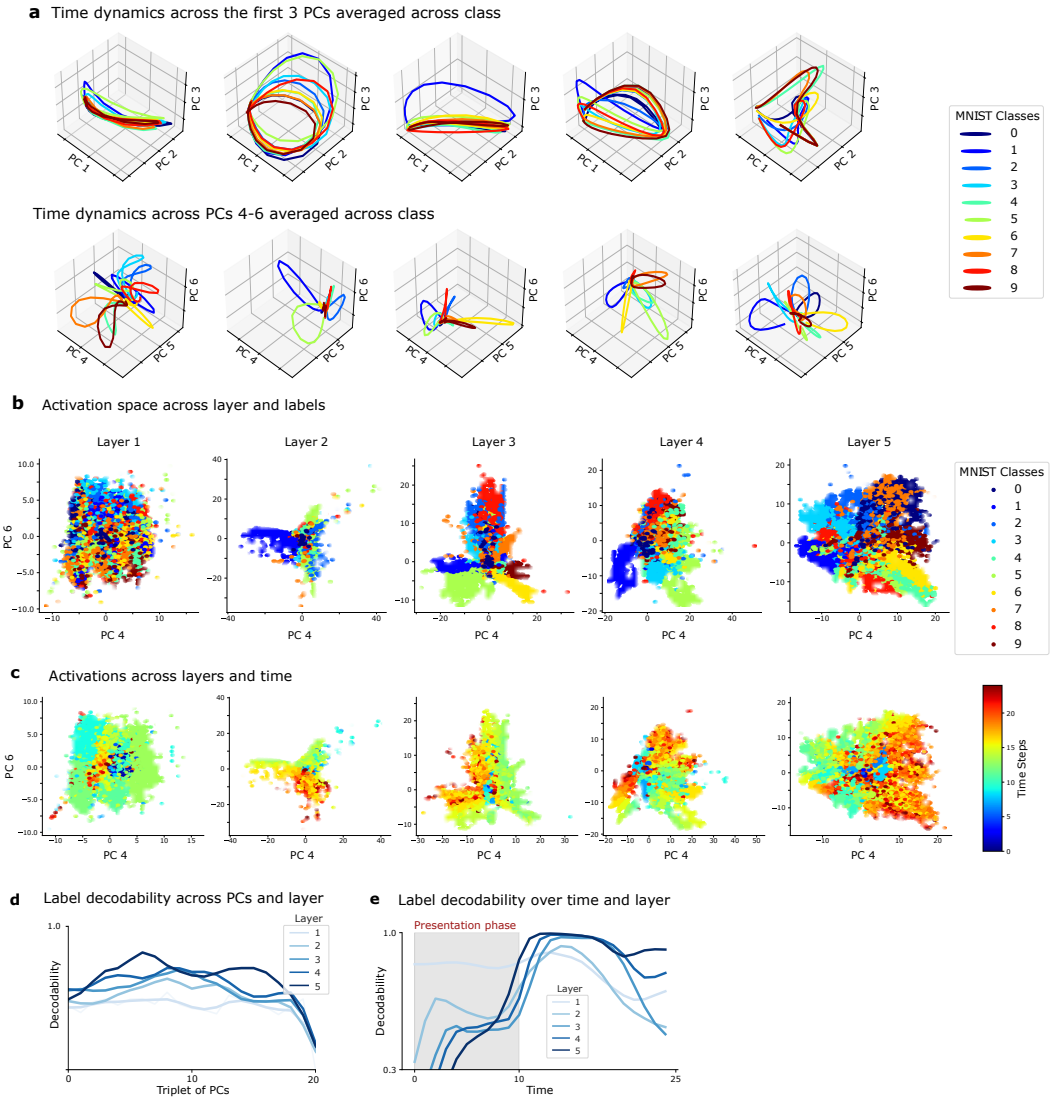

Figure 4: Latent representations and label decodability over both principal components and layer. a) Representation of the layer-wise latent spaces on three dimensions via PCA where classes are represented by color. The first three PCs are shown to indicate a lack of class separability. Higher-order PCs are shown to indicate stronger class separability. b) Representation of the layer-wise latent spaces on two dimensions via PCA where classes are represented by color. c) Same latent space representation, but color-coded based on timestep. d) Decodability (y-axis) across different layer's PCs (x-axis), indicating that PCs 4-10 capture rich representations. e) Decodability (y-axis) over time (x-axis) for different layers, indicating that the pre-label timesteps are driven by a distinct bottom-up temporal cascade, and by contrast the top layers are driven by a top-down cascade.

the significance of such cancellation in establishing the aforementioned looping dynamic and the encoding properties of the network throughout its hierarchy.

3.4 The Inverted Forward-Forward is a Contrastive Three-Factor-Learning Rule which Converges to Synaptic Drive Cancellation

In this section, we apply a simple argument to the underlying dynamics as the result of minimizing cumulative objective functions.

For a $N > 3$ Forward-Forward architecture where N is the total number of layers, the dynamics of each layer are governed by

$$\begin{aligned}\dot{\vec{r}}_1 &= \phi\left(W_1\vec{r}_1 + F_1 I(t) + B_1\vec{r}_2\right)\\ \dot{\vec{r}}_i &= \phi\left(W_i\vec{r}_i + F_i\vec{r}_{i-1} + B_i\vec{r}_{i+1}\right)\\ \dot{\vec{r}}_N &= \phi\left(W_N\vec{r}_N + F_N\vec{r}_{N-1} + B_N\ell(t)\right).\end{aligned}$$

The locally defined loss for a one-step update takes the form of

$$\mathcal{L}_{layer}(t') = (-1)^{\eta}\sigma(\vec{r}_i^T\vec{r}_i - T)$$

where $\sigma(x)$ represents the softplus as $\sigma(x) = \log(1 + e^x)$ and is a smooth version of the ReLU nonlinearity. Importantly, the dynamics of the $\eta(t')$ are governed by bistable dynamics switching between

$$\eta(t') = \delta_{L(t'),I(t')}$$

where δ_{ij} is the Kronecker delta notation. This bistable switching-like dynamics occurs with long timescales between switches and could have compelling correspondence with our understanding of neuromodulatory induced switching of the underlying dynamics. This also suggests a criterion for the selection of $\eta(t')$ on the instantaneous surprise of the stimulus against the speculative label. While beyond the scope of this work, the closure of this loop between activations and cost function may generate valuable insights into unsupervised variants of these learning rules (Ororbia & Mali, 2023).

We then execute a single step-gradient update in each parameter:

$$\begin{aligned}\partial_t W_i &= -\alpha\nabla_{W_i}\mathcal{L}(t')\\ \partial_t B_i &= -\alpha\nabla_{B_i}\mathcal{L}(t')\\ \partial_t F_i &= -\alpha\nabla_{F_i}\mathcal{L}(t').\end{aligned}$$

By iteratively minimizing this locally defined objective function, we seek a hierarchical structure that will work in concert with the other layers to minimize activations for positive data. However, when there is a data mismatch between label and image class, the representations will learn to avoid cancellation to maximize their surprise.

Indeed this single-step update for a given layer takes the form of a Hebbian learning, since

$$\nabla_{W_i}\mathcal{L}_i(t') = (-1)^{\eta}\underbrace{\sigma'(\vec{r}_i^T(t')\vec{r}_i(t') - T)(\vec{r}_i(t'))}_{\text{Gating factor}}\underbrace{\phi'(z(t-1))}_{\text{Pre-synaptic current}}\underbrace{\vec{r}_i(t'-1)}_{\text{post-synaptic firing rate}}$$

where $\vec{z}(t-1) = W_i\vec{r}_i(t'-1) + F_i\vec{r}_{i-1}(t'-1) + B_i\vec{r}_{i+1}(t'-1)$ is the input current into the nonlinearity.

This takes the form a gated Hebbian or three-factor rule (Bahroun et al.; Kuśmierz et al., 2017; Bredenberg et al., 2021; Pogodin & Latham, 2020; Portes et al., 2022) linking the locally defined objective function to the product of the pre-synaptic current and the post-synaptic activation. Crucially, this loss for a single trial is indeed the cumulative sum over the full sequence:

$$\Delta x_i = \alpha\sum_{t'}\nabla_{x_i}\mathcal{L}_i(t').$$

These gradients have important implications on the shape of the learned solutions. These learned solutions (where the gradient goes toward zero can occur under a number of conditions). These conditions include the direct cancellation of the synaptic drive currents (input components) governing the time dynamics of the hidden layer: $W_i\vec{r}_i + F_i\vec{r}_{i-1} + B_i\vec{r}_{i+1} = 0$.

In this section, we established an equivalence between the inverted Forward-Forward and a unique form of gated Hebbian plasticity. In addition, we have demonstrated that the cancellation mechanism of lateral, top-down, and bottom-up signals satisfies the stationary solution to this deduced three-factor learning rule.

4 DISCUSSION

4.1 MECHANISMS BEHIND CANCELLATION ORDER

As a result of our simulation, a compelling non-trivial logic has emerged from the model's hierarchical predictive dynamics, which are often difficult to comprehend. Here, we seek to explain the fundamental mechanistic principles underlying the network's information flow generation. We determined that despite the fact that such insights are difficult to isolate or prove, they may still be necessary to comprehend the model's inner mechanisms.

For the initial presentation phase of both positive and negative data, the image representation flows from the bottom-up. The differences are, however, quite different in the processing phases. For negative data, the processing phase maintains the initial bottom-up dynamic until the label representation percolates to the bottom layer, indicating a mismatch and evoking a large and sustained excitatory surprise. For positive data, the label representation traverses to the bottom layer without inducing cancellations, whereby the cancellations then start in a bottom-up manner, despite the top-down label representation.

As the presentation phase blends with the processing phase, it is insightful to note that, neither for the positive nor negative case is some predisposed behavior taking place. The network does not amplify or reduce its activations across layers until the label representation reaches the bottom layer. This suggests the bottom-up cancellation could be a result of the network's optimizing drive to alter activities when in a familiar state relative to training. During training, there is a continuous and consistent exposure to label-augmented activations, particularly past the early stages, which ingrains a behavior within the network. The network recognizes label saturated activations as the dominant trend it should ideally be prepared for. Given this recognition, the network is best equipped for cancellation when it encounters activation components (forward, backward, recurrent) with label information coming from all components, not just from the top. Under this concept, the bottom layer would cancel first, followed by a transmission of label-infused activities upwards to the next layer to induce the next phase of bottom-up cancellation.

It is also essential to consider the implications if the upper layers were to cancel first. The primary consequence would be the rapid removal of label information. In the Forward-Forward's design, the downward transmission of label-derived activations is essential for informing and priming the bottom layers for their own cancellation. If the upper layers were to cancel prematurely, they would be withholding vital label information, thus preventing the bottom layers from receiving the necessary cues to initiate their cancellation. This would disrupt the network's learning dynamics, leading to poorly canceling bottom layers. By preserving the label information and not canceling immediately, the upper layers play a role in ensuring that the network's lower layers are sufficiently informed and primed for the subsequent cancellation process.

4.2 BIO-PLUASIBILITY OF THE INVERTED FORWARD-FORWARD

The inverted Forward-Forward model uses activation contrast to navigate credit assignment in hierarchical architectures in a bio-plausible fashion by incorporating: the absence of weight transport (Lillicrap et al., 2016; Portes et al., 2022), online capable learning rules consistent with Hebbian plasticity, a biologically analogous separation of timescales, and the incorporation of structural hierarchy. First, feedback is separated from backpropagation and instead incorporated as a top-down signal avoiding weight transport. By disconnecting the F and B matrices, the flexible learning rule finds aligned but non-weight transported solutions.

The local update rules of the inverted Forward-Forward result in local Hebbian plasticity which is gated by the presence of a label and the signed by the data type. This learning rule is tightly associated with a third factor where now is external signal (commonly associated with neuromodulatory input) is responsible for the minimization (maximization) of the layer activity for positive (negative) labels. The slow timescales of this switching relative to both the dynamics and the plasticity opens suggests an influential role of these biologically important small molecules (Kuśmierz et al., 2017). A promising direction will be to explore the implications of linking this signal not to the label identity itself but to an unsupervised signal generated elsewhere in the network. In this fashion, the inverted Forward-Forward model brings bio-plausibility to the direct competition of top-down

and bottom-up signal processing with interesting ramifications for how we interpret the hierarchy of biological systems (Siegle et al., 2021; Garrett et al., 2023)

5 CONCLUSION

In this work, we have presented a biologically plausible mechanism that sheds light on the spatiotemporal and predictive nature of cortical processing without necessitating explicit predictions. Drawing inspiration from the Forward-Forward model, an emerging form of contrastive learning based solely on forward passes, we inverted its original objective function to minimize positive training data while maximizing negative training data surprise. This inversion incentivizes activity cancellation between information flows when top-down labels align with bottom-up sensory input. As a consequence, various layers across the hierarchy develop the ability to predict each other's activities, facilitating the minimization of layer surprise.

In conclusion, our study reveals that employing a local contrastive learning approach involving surprise and cancellation dynamics can create predictive spatiotemporal properties, typically linked to neocortical computations. This implies that these distinctive properties, previously thought to be specific to neocortical processes, can be generated using simplified learning principles. This discovery offers a new avenue to improve the computational capacities of biologically plausible models.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

David Badre and Derek Evan Nee. Frontal cortex and the hierarchical control of behavior. *Trends in cognitive sciences*, 22(2):170–188, 2018. URL https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613%2817%2930245-0?elsca1=etoc&amp%3Belsca2=email&amp%3Belsca3=1364-6613_201802_22_2_&amp%3Belsca4=Cell+Press&code=cell-site. Publisher: Elsevier.

Yanis Bahroun, Dmitri B Chklovskii, and Anirvan M Sengupta. A Normative and Biologically Plausible Algorithm for Independent Component Analysis.

Danielle S. Bassett and Olaf Sporns. Network neuroscience. *Nature neuroscience*, 20(3):353–364, 2017. URL https://www.nature.com/articles/nn.4502. Publisher: Nature Publishing Group US New York.

Colin Bredenberg, Benjamin S. H. Lyo, Eero P. Simoncelli, and Cristina Savin. Impression learning: Online representation learning with synaptic plasticity. November 2021. URL https://openreview.net/forum?id=MAorPaLqam_.

Rishidev Chaudhuri, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy, and Xiao-Jing Wang. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron*, 88(2):419–431, 2015. URL https://www.cell.com/neuron/pdf/S0896-6273(15)00765-5.pdf. Publisher: Elsevier.

Emmanouil Froudarakis, Paul G. Fahey, Jacob Reimer, Stelios M. Smirnakis, Edward J. Tehovnik, and Andreas S. Tolias. The Visual Cortex in Context. *Annual Review of Vision Science*, 5(1):317–339, September 2019. ISSN 2374-4642, 2374-4650. doi: 10.1146/annurev-vision-091517-034407. URL https://www.annualreviews.org/doi/10.1146/annurev-vision-091517-034407.

Marina Garrett, Peter Groblewski, Alex Piet, Doug Ollerenshaw, Farzaneh Najafi, Iryna Yavorska, Adam Amster, Corbett Bennett, Michael Buice, Shiella Caldejon, Linzy Casal, Florence D'Orazi, Scott Daniel, Saskia EJ de Vries, Daniel Kapner, Justin Kiggins, Jerome Lecoq, Peter Ledochowitsch, Sahar Manavi, Nicholas Mei, Christopher B. Morrison, Sarah Naylor, Natalia Orlova, Jed Perkins, Nick Ponvert, Clark Roll, Sam Seid, Derric Williams, Allison Williford, Ruweida Ahmed, Daniel Amine, Yazan Billeh, Chris Bowman, Nicholas Cain, Andrew Cho, Tim Dawe, Max Departee, Marie Desoto, David Feng, Sam Gale, Emily Gelfand, Nile Gradis, Conor Grasso, Nicole Hancock, Brian Hu, Ross Hytnen, Xiaoxuan Jia, Tye Johnson, India Kato, Sara Kivikas, Leonard Kuan, Quinn L'Heureux, Sophie Lambert, Arielle Leon, Elizabeth Liang, Fuhui Long, Kyla Mace, Ildefons Magrans de Abril, Chris Mochizuki, Chelsea Nayan, Katherine North, Lydia Ng, Gabriel Koch Ocker, Michael Oliver, Paul Rhoads, Kara Ronellenfitch, Kathryn Schelonka, Josh Sevigny, David Sullivan, Ben Sutton, Jackie Swapp, Thuyanh K. Nguyen, Xana Waughman, Joshua Wilkes, Michael Wang, Colin Farrell, Wayne Wakeman, Hongkui Zeng, John Phillips, Stefan Mihalas, Anton Arkhipov, Christof Koch, and Shawn R. Olsen. Stimulus novelty uncovers coding diversity in visual cortical circuits, February 2023. URL https://www.biorxiv.org/content/10.1101/2023.02.14.528085v2. Pages: 2023.02.14.528085 Section: New Results.

Charles D. Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, May 2013. ISSN 1471-0048. doi: 10.1038/nrn3476.

Geoffrey Hinton. The Forward-Forward Algorithm: Some Preliminary Investigations, December 2022. URL http://arxiv.org/abs/2212.13345. arXiv:2212.13345 [cs].

Linxing Preston Jiang and Rajesh P. N. Rao. Dynamic predictive coding: A new model of hierarchical sequence learning and prediction in the cortex. June 2022. doi: 10.1101/2022.06.23.497415. URL https://doi.org/10.1101/2022.06.23.497415.

Rebecca Jordan and Georg B. Keller. Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex. *Neuron*, 108(6):1194–1206.e5, December 2020. ISSN 08966273. doi: 10.1016/j.neuron.2020.09.024.

Mitsuo Kawato. Internal models for motor control and trajectory planning. *Current opinion in neurobiology*, 9(6):718–727, 1999. URL https://www.sciencedirect.com/science/article/pii/S0959438899000288?casa_token=4-yYDedK9_8AAAAA:JagChpbOID-Lg21iEfnobU0W1yZHGx_eyZ79KtM5K6j-NkoL3z69DSVRrmgzUhPvLg1aNNmzHsNj. Publisher: Elsevier.

Adil G. Khan and Sonja B. Hofer. Contextual signals in visual cortex. *Current Opinion in Neurobiology*, 52:131–138, 2018. URL https://www.sciencedirect.com/science/article/pii/S0959438818300825?casa_token=xIXteX3UtOgAAAAA:jgedYEVdm7FxHu6zV8-YxvKg_7nKGt73wzlHmrNIunlgUgKFzvmP16Socgq6cQALCeddYX6vA_U_. Publisher: Elsevier.

Łukasz Kuśmierz, Takuya Isomura, and Taro Toyoizumi. Learning with three factors: modulating Hebbian plasticity with errors. *Current Opinion in Neurobiology*, 46:170–177, October 2017. ISSN 0959-4388. doi: 10.1016/j.conb.2017.08.020. URL https://www.sciencedirect.com/science/article/pii/S0959438817300612.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539.

Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7 (1):13276, November 2016. ISSN 2041-1723. doi: 10.1038/ncomms13276. URL https://www.nature.com/articles/ncomms13276. Number: 1 Publisher: Nature Publishing Group.

Andrea Mechelli, Cathy J. Price, Karl J. Friston, and Alumit Ishai. Where bottom-up meets top-down: neuronal interactions during perception and imagery. *Cerebral cortex*, 14(11):1256–1265, 2004. URL https://academic.oup.com/cercor/article-abstract/14/11/1256/331439. Publisher: Oxford University Press.

Alexander Ororbia. Learning Spiking Neural Systems with the Event-Driven Forward-Forward Process, March 2023. URL http://arxiv.org/abs/2303.18187. arXiv:2303.18187 [cs].

Alexander Ororbia and Ankur Mali. The Predictive Forward-Forward Algorithm, April 2023. URL http://arxiv.org/abs/2301.01452. arXiv:2301.01452 [cs].

Alex Piet, Nick Ponvert, Douglas Ollerenshaw, Marina Garrett, Peter A. Groblewski, Shawn Olsen, Christof Koch, and Anton Arkhipov. Behavioral strategy shapes activation of the Vip-Sst disinhibitory circuit in visual cortex. *bioRxiv*, pp. 2023–04, 2023. URL https://www.biorxiv.org/content/10.1101/2023.04.28.538575.abstract. Publisher: Cold Spring Harbor Laboratory.

Roman Pogodin and Peter Latham. Kernelized information bottleneck leads to biologically plausible 3-factor Hebbian learning in deep networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7296–7307. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/517f24c02e620d5a4dac1db388664a63-Abstract.html.

Jacob P. Portes, Christian Schmid, and James M. Murray. Distinguishing Learning Rules with Brain Machine Interfaces, October 2022. URL http://arxiv.org/abs/2206.13448. arXiv:2206.13448 [cs].

Rajesh PN Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999. URL https://www.nature.com/articles/nn0199_79. Publisher: Nature Publishing Group.

Wolfram Schenck. *Adaptive internal models for motor control and visual prediction*. Number 20. Logos Verlag Berlin GmbH, 2008. URL https://www.google.com/books?hl=en&lr=&id=aorWUm2hbLYC&oi=fnd&pg=PA1&dq=schenck+internal+model&ots=3gmKIJ1I3M&sig=2RzLtJQ1QM0_kMsORSD7AKpUNqs.

Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K. Ramirez, Hannah Choi, and Jennifer A. Luviano. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021. URL https://www.nature.com/articles/s41586-020-03171-x. Publisher: Nature Publishing Group UK London.

Daniel M. Wolpert, R. Chris Miall, and Mitsuo Kawato. Internal models in the cerebellum. *Trends in cognitive sciences*, 2(9):338–347, 1998. URL https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(98)01221-2. Publisher: Elsevier.