SUPPLEMENTARY MATERIAL FOR "EMERGENCE OF SURPRISE AND PREDICTIVE SIGNALS FROM LOCAL CONTRASTIVE LEARNING"

Anonymous authors
Paper under double-blind review

ABSTRACT

This is the supplementary information for the "Emergence of surprise and predictive signals from local contrastive learning."

A APPENDIX A: EQUIVALENCE BETWEEN FORWARD-FORWARD ARCHITECTURE AND HEBBIAN LEARNING

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

For a $N > 3$ Forward-Forward architecture, 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 criteria 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.

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 heiarhical structure which 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 rule.

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

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) 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 hidden layer: $W_i r_i + F_i r_{i-1} + B_i r_{i+1} = 0$.

A.1 Linearizing everything and reducing the dimension of the layers to one unit each

In the limit of linear dynamics of the underlying network and linear dynamics of the locally defined objective function:

$$\mathcal{L}_\backslash(t') = \sum r_i^2 - T$$

and

$$r_i = W_i r_i + B_i r_{i+1} + F_i r_{i-1}$$

The the gradients give rise to the simple learning dynamics of the form:

$$\dot{W}_i = \alpha(-1)^\eta r_i(t+1) r_i(t)$$

$$\dot{F}_i = \alpha(-1)^\eta r_i(t+1) r_{i-1}(t)$$

$$\dot{B}_i = \alpha(-1)^\eta r_i(t+1) r_{i+1}(t)$$

With the edge cases taking the form of:

$$\dot{B}_N \alpha(-1)^\eta r_i(t+1) L(t)$$

$$\dot{F}_1 \alpha(-1)^\eta r_1(t+1) I(t)$$

These linear dynamics are determined by the discrete variable η which tells you if the data is positive data or not. We can call then one when $I(t) \approx L(t)$ where the labels are presented as classes and the images are presented as floats near the given class identity.

These nonlinear dynamics give us a rich learning sequence in which the we have fast time dynamics governing the layer population scalars r_i and the slower dynamics of the learned parameters. This separation of timescales allows us to represent the mean population activity over the trial to study the convergence of the learning dynamics to steady state solutions.

The simulation of these linearized but still nonlinear dynamics dynamics for a three layer network in an online learning setting confirms our analytic for this one-dimensional projection of population activity in that $F \to -B$ over training timesteps 1.

For a one-layer architecture with similar representations of image class and label $L(t) \approx I(t)$, the positive data equivalence trivially forces the feedforward F and feedback B matrices to converge toward the negative of each other, $B \to -F$. This can be seen in the evolution of the above dynamics of the weight vectors. The only way that the linearized dynamics can go to zero is when $\dot{B}_1 = 0 = -r_i L(t)$ and thus since $L(t)$ is fixed at a non-zero value by the supervision, the r_i must go to zero. To achieve this, we must have clean cancellation of the underlying dynamics in the hidden layer. This forms the basis of the positive data cancellation and finds solutions which are consistent with increasing activation in response to mismatch or surprise.

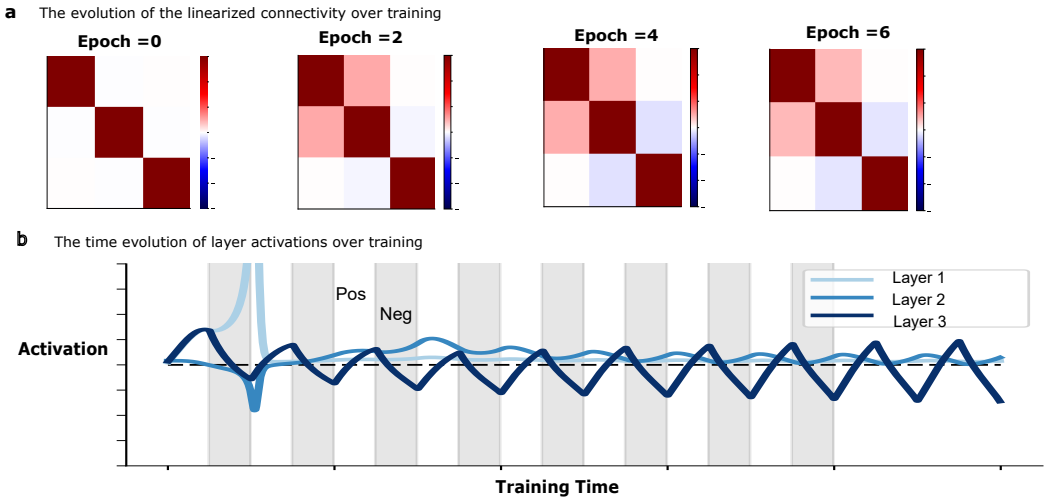

Figure 1: a) Over training the evolution of the simplified connectivity matrix develops opposing terms resulting in the cancellation of matched signals (top-down and bottom-up) into layer 2. b) The time course of training these linearized dynamics generates a system capable of switching between positive and negative data samples in an online fashion. Negative data is characterized by growing activations in layer 2 while positive data is characterized by cancelling activations in layer 2. These linearized dynamics give us a simplified playground to understand the emergence of cancellation with simple local learning rules.

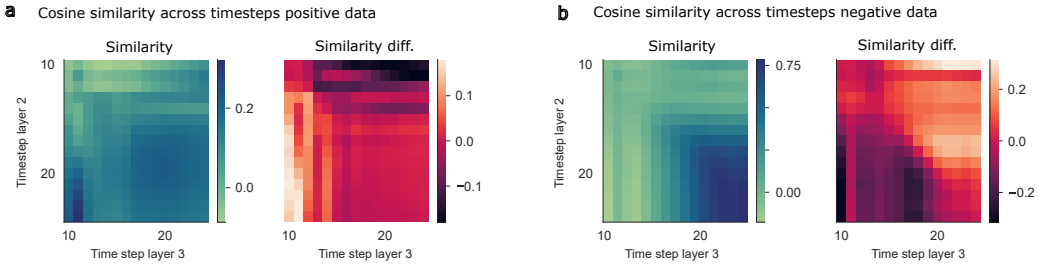

Figure 2: a) Cosine similarity analysis across all timesteps of layer 2 and 3 (left panel). Similarity Difference metric (SD) across timesteps between layer 2 and 3. b) Same as panel d but for negative data.

B Appendix B: Temporal dynamics of layer similarity

We expanded the cosine similarity analysis across all timesteps for any two consecutive layers to further investigate this phenomenon. Figs. (2a) to (2b) illustrate the cosine similarity between activations of layers 2 and 3 across any two timesteps during the processing phase for both positive (Fig. (2a) left panel) and negative (Fig. (2b) left panel) data. For positive data, cosine similarities decrease over timesteps confirming that activations across different layers decorrelate over this period. On the other hand, for negative data, similarities increase over the same period, confirming and generalizing our findings in the main text. This analysis demonstrates the emergence of a striking temporal ordering of cancellations (positive data) and activations (negative data) which reflects the structurally imposed hierarchy of the layers. This relationship emphasizes the importance of a mechanistic understanding going beyond the naive cancellation of image and label representations as their first collision.

To examine the temporal dynamics further, we analyzed the difference between such similarities: for any pair of time steps, we computed the following metric. Denoted with $cos(a_{l2}(t_1), a_{l3}(t_2))$ is the cosine similarity between the activations $a_{l2}(t_1)$ of layer 2 at time t_1 and the activations $a_{l3}(t_2)$ of layer 3 at time t_2. We also defined the Similarity Difference $\text{SD}_{l23}(t_1, t_2) = cos(a_{l2}(t_1), a_{l3}(t_2)) - cos(a_{l2}(t_2), a_{l3}(t_1))$. This quantity provides insight into the temporal dynamics because, for $t_1 < t_2$, it is positive if the similarity between earlier activations in the first layer and later activations

in the second layer is greater than the similarity between later activations in the first layer and earlier activations in the second layer. This value quantifies when current signals in one layer are analogous to subsequent signals in a second layer for any given timestep. A positive SD above the diagonal (accompanied by a negative SD below the diagonal) quantifies the influence of the first layer on subsequent timesteps in the second layer. This case, as described, is what we observed for negative data, confirming a bottom-up flow in late timesteps, Fig. (2b). For positive data, a top-down signal appears to flow into layer for a few time steps before activities across layers decorrelate and cancellation of activity occurs, Fig. (2a) right panel.

This analysis validates the presence of two information flows for positive and negative data, with distinct temporal relationships between layers. It further illustrates that such information flows have specific dynamical properties across layers, where the activity in a given layer precedes or follows the activity in others across the hierarchy, enabling the generation of predictive types of signals.

REFERENCES

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

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_.

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.

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.