Fine-tuning hierarchical circuits through learned stochastic co-modulation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Attentional gating is a core mechanism supporting behavioral flexibility, but its biological implementation remains uncertain. Gain modulation of neural responses is likely to play a key role, but simply boosting relevant neural responses can be insufficient for improving behavioral outputs, especially in hierarchical circuits. Here we propose a variation of attentional gating that relies on *stochastic* gain modulation as a dedicated indicator of task relevance. We show that targeted stochastic modulation can be effectively learned and used to fine-tune hierarchical architectures, without reorganization of the underlying circuits. Simulations of such networks demonstrate improvements in learning efficiency and performance in novel tasks, relative to traditional attentional mechanisms based on deterministic gain increases. The effectiveness of this approach relies on the availability of representational bottlenecks in which the task relevant information is localized in small subpopulations of neurons. Overall, this work provides a new mechanism for constructing intelligent systems that can flexibly and robustly adapt to changes in task structure.

1 Introduction

Constructing neural representations is a fine balancing act between stability and plasticity: the brain must quickly adapt to new task demands, while maintaining performance in previous contexts. How is this accomplished? Resource constraints prevent the construction of *de novo* representations for each new task, while reorganization of existing synapses to satisfy the demands of each task runs the risk of catastrophic loss of previous capabilities [1–3]. Instead, the brain seems to achieve its balance by dynamically but reversibly altering the flow of information through circuits. The neural mechanisms of this process are not fully understood, but dynamic gain modulation is a ubiquitous aspect of neural activity [4, 5] and seems likely to play a critical role.

Gain modulation has been proposed to underlie improvements in task performance that arise from attentional focus. Specifically, experimental evidence suggests that attention selectively boosts the firing rate of neurons throughout the visual hierarchy, temporarily improving the quality (signal-to-noise ratio) of their representation [6–11]. These effects appear to be focused on those neurons most relevant for the current task, which can be either spatially localized (the 'spotlight' of attention) [12], or spatially distributed but informative for specific relevant attributes (e.g., color) [13, 1]. The origin of these adjustments, their impact on later stages of hierarchical processing, and their effect on decision-making is less clear [14].

A more recently documented form of gain modulation relies on the existence of shared fluctuations in neural populations that arise from low-dimensional modulatory signals [15–18]. The strength of these fluctuations varies across neurons in a cortical population, and appears to be largest for those neurons that are most relevant for the current task [19, 20]. These observations form the basis for a recently

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

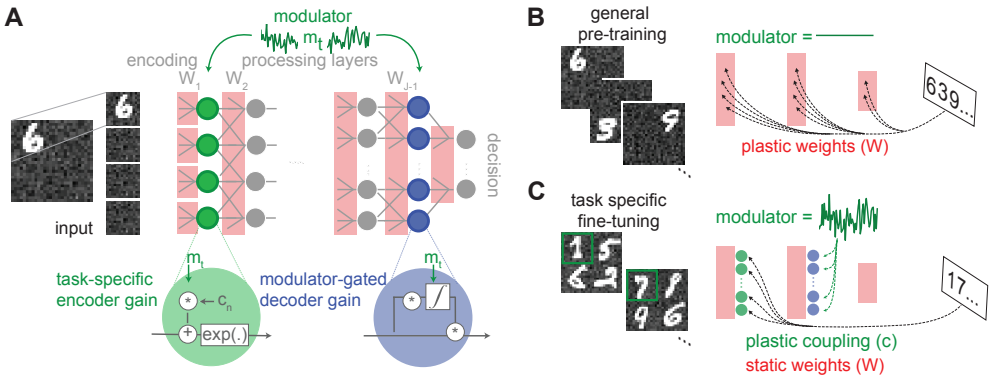

Figure 1: Model of task-specific fine-tuning through stochastic modulation. A) A feedforward base network with J layers is trained to map input images into categorical outputs. Neurons in the initial encoding layer have receptive fields localized to lie within one of 4 spatial quadrants, while all other layers are all-to-all connected. A stochastic modulator induces correlated gain fluctuations in the encoding layer, with neuron-specific coupling strengths c_i (green circles). Activities of neurons in the last layer are adaptively gated based on within-trial correlations between the modulator and their stimulus-driven responses (blue circles, Eq. 3). B) During pretraining, feedforward weights W_j are optimized (via gradient descent) on a general categorization task (here, position-invariant MNIST digit classification), with the modulator disabled (i.e., all c_i's set to zero). C) During fine-tuning, the feedfoward weights W_j are held fixed, and the modulator coupling strengths c_i are optimized (again via gradient descent) to support binary classification of a specific pair of digits, localized within a specific spatial quadrant (here, '1' vs. '7' in the upper left quadrant), in the presence of distractors. Output gains (blue) are automatically adjusted based on correlation with the modulator (Eq. 3), without task feedback.

proposed computational theory, in which the rapidly fluctuating common modulation of task-relevant neurons provides a 'label' that enables downstream circuits to learn which neurons to 'listen' to, based on local correlations [21]. Although these signals are clearly present in neural data [20], and can change their targeting on a timescale of tens of trials when the task changes, it is not clear how this adaptation occurs, or how these modulatory fluctuations can be used to improve performance of tasks that rely on decoding the activity of downstream neurons in a hierarchical architecture.

Here, we develop a model that solves these two problems. A feedforward hierarchical network is augmented to include encoder gain modulation in an early stage, and a modulator-gated decoder in the last stage. We demonstrate the effectiveness of this solution in the context of a "base" network that is initially trained to solve a standard MNIST 10-digit classification task [22]. In a subsequent phase, the targeting of the encoder gain modulation is trained to optimize performance on a new specialized task (binary discrimination of a pair of digits, at a particular location), thus fine-tuning the network for the new task without any reorganization of the feedforward weights. This modulatory labelling signal is task-specific and ephemeral, allowing the network to instantly revert to the initially-trained state once task demands are removed. We show that the learning of the stochastic modulator is substantially faster than retraining the full network, and more effective than using targeted constant ("attentional") encoder gain modulation. Finally, empirical exploration of the effects of injecting the modulator at different stages of the network reveals that its labelling is most effective when applied to layers in which task-specific information is concentrated in a small subpopulation - an "informativeness bottleneck".

2 Learned stochastic modulation targeting in a hierarchical network

We build on the model of stochastic co-modulation introduced in Ref. [21], which provides a theoretical framework for decoding information from large neural populations, only a small fraction of which carry task-relevant information. It postulates that fast low-dimensional co-fluctuations targeting the task-informative subset serve to label the information for use by a decoder. A "modulator-

guided" decoder can then use these fluctuations to estimate the correct decoding weights for the task, achieving high levels of performance within a handful of trials, with minimal explicit task feedback.

Here, we generalize this framework to a hierarchical feedforward neural network, in which neurons linearly combine their inputs, add a bias term, and pass the result through a nonlinear activation function ($\exp(\cdot)$ for the encoding layer in accordance with [21], and $\mathrm{ReLU}(.)$ thereafter). We incorporate a stochastic modulator which fluctuates on a significantly faster timescale (indexed $t = 1, \ldots, T$) than the stimulus presentation trials (indexed $k = 1, \ldots, K$), with two distinct effects on the network. First, the modulator, m_{kt}, controls the gains of all neurons in the encoding layer, via learnable coupling strengths $\mathbf{c}$ (Fig. 1A, "task-specific encoder gain"):

$$\mathbf{h}_{kt}^{(1)} = \exp\left(\mathbf{W}_1 \mathbf{s}_k + m_{kt}\mathbf{c} - b\right), \tag{1}$$

where $\mathbf{h}_{kt}^{(1)}$ is a vector of activities of the encoding layer for trial k and time t, $\mathbf{s}_k$ the multi-dimensional stimulus vector and $\mathbf{W_1}$ the weight matrix of the first network layer. Unlike the model of Ref. [21], this modulator affects both the mean and the variance of the neural responses, combining traditional deterministic gain boosting with stochastic labeling. This formulation facilitates fast learning of the coupling strengths, and is also more biologically realistic (since modulation of mean and co-variability covary in the cortex), and provides an opportunity to directly compare to models of attention that rely on deterministic gain boosts ($m_{kt} = 0, \forall k, t$).

As in Ref. [21], we assume the modulator is available at the decision stage of the network, and can be used to guide decoding. This is implemented as an adaptable gain, $\mathbf{g}_k$, on the neurons in the final processing layer, which directly map into the network output (Fig. 1A, "modulator-gated decoder gain"):

$$\mathbf{h}_{kt}^{(J)} = \mathbf{g}_k F\left(\mathbf{W_J}\mathbf{h}_{kt}^{(J-1)}\right), \tag{2}$$

with F a rectifying nonlinearity. The strength of both gating mechanisms adapts over time to fine-tune the network's operation on a new task. First, the coupling strengths $\mathbf{c}$ in the first, 'labelled' layer are optimized based on explicit feedback so as to maximize network performance on the task (using backpropagation). Second, in the final layer the neural gains g are adjusted based on the correlation of neural activity with the modulator, following the modulator-guided estimation rules proposed in [21]:

$$\mathbf{g}_k = \frac{1}{T}\sum_{t}^{T} \bar{m}_{k-1,t}\bar{\mathbf{h}}_{k-1,t}^{(J-1)}, \tag{3}$$

where $\bar{m}_{kt}$ and $\bar{\mathbf{h}}_{kt}^{(J)}$ denote the mean-subtracted modulator and neural activity, respectively.[1] Importantly, this rule is independent of stimulus or reward, and only requires the modulator as a 'key' to identify responses of task-relevant neurons in the last layer. All feedforward weights remain unchanged throughout this task-specific learning – their values are assumed to reflect a slower optimization process on a general set of tasks. The task-specific adaptation is only applied to the modulation coupling strengths, c_i, in the encoding layer – a parameter set of size N_1, compared to retraining of the full set of $\sum_{j=0}^{J-1} N_j N_{j+1}$ network weights (where N_0 denotes the input dimension). By concentrating the learning on this small parameter set, the model requires fewer training examples to improve performance on the current task.

3 Numerical results

Fine-tuning MNIST digit recognition in the presence of distractors. To validate the idea of a stochastic modulator guiding task-specific information flow in hierarchical networks, we used task variations built around MNIST digit recognition. We first defined a location-invariant digit recognition task, in which downscaled MNIST images are embedded in a noisy background, at different spatial locations (Fig. 1B; full image size 28×28) and must be identified regardless of their position. We pretrained the network on this 'general task', then used stochastic modulation to fine-tune it for 'specific' binary classification tasks between two digits confined to one image quadrant,

[1] Correlations are computed from the fluctuations of the modulator and the neural responses at the fast time scale t, integrated over the time scale of single trials, during which the stimulus is constant.

in the presence of randomly chosen distractor digits that appear in the other three quadrants (Fig. 1C). Different instances of this specific task vary in the choice of the relevant digit pair and quadrant, but all are subtasks of the general digit classification problem, and thus the information needed to solve them should be present within the pre-trained network. However, since the network only experiences digits in isolation during training, decision layer neurons will respond to the combination of task relevant and distractor digits. The objective of the stochastic modulation refinement is to focus the readout on those neurons that carry the specific task-relevant information.

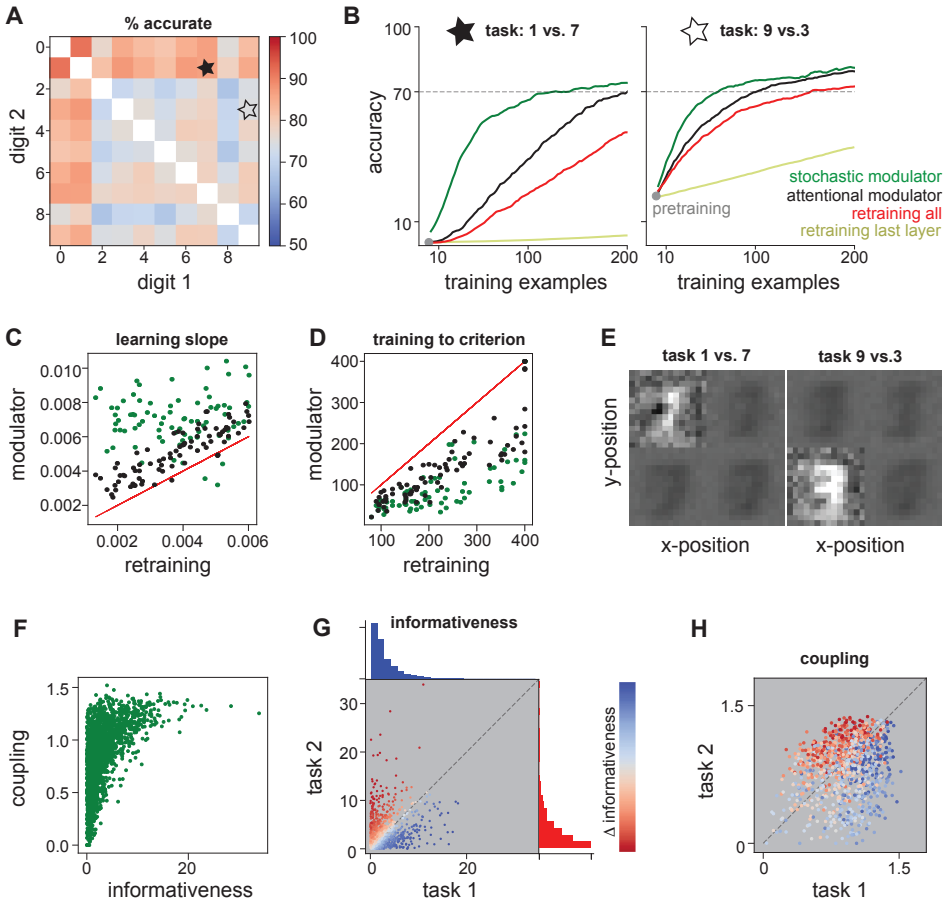

Figure 2: Analysis of model performance. A) Average performance (% correct) of the pretrained network in discriminating digit pairs at any location without distractors. B) Classification accuracy for two different specific tasks. Grey dot indicates baseline performance of the pretrained network. Grey dotted line indicates a performance criteria of 70%. Lines represent averages over 10 simulations for each learning procedure. C) Initial slope of performance improvement during learning over different two-digit classification tasks, relative to that of retraining. Slopes are estimated by linear regression on performance over the initial 50 training samples. D) Number of training examples required to reach 70% accuracy, compared to those needed for retraining. E) Learned coupling strengths mapped back to the input space for the two tasks indicated in B; coupling strengths are standardized (z-scored) before averaging. F) Comparison of modulator coupling strength and informativeness ($|d'|$) for all first-stage neurons with receptive fields in the task-relevant input quadrant. G) Comparison of task informativeness of first-stage neurons in the task-relevant input quadrant for two tasks that involve different digit pairs within the same quadrant. H) Comparison of coupling strengths (same neurons, tasks, and colormap as G).

We use a 3-layer feedforward network with stochastic modulation of activity in the initial (encoding) layer, followed by an all-to-all connected hidden layer, and a final readout (decision) layer that maps into a categorical output (softmax). The encoding layer receives local information about the input, with 'receptive field'-like weights limited to one of four image quadrants. The modulator has to

isolate the relevant subset of neurons – those that encode the task-relevant input quadrant and whose responses differentiate the task-specified digit pair. The hypothesis is that learning will target the modulation specifically towards these neurons, and that the modulatory labelling of their responses will propagate through the densely connected layer so that the modulation-guided readout can use it to perform the task.

The unmodulated network was pretrained on a 'generic' recognition problem: identify a single digit at an arbitrary location within the image. Weights were optimized using conventional backpropagation with Adam [23], using the MNIST training set, using 20 image minibatches (with modulator $m_{kt} = 0$ and the gain terms $\mathbf{g} = \mathbf{1}$).[2] The location and background noise of each image were drawn independently for each image, uniform for location and using additive i.i.d. Gaussian pixel noise (std=0.1, for image pixels in the range $[0, 1]$) for the background (training dataset includes 4000 images). The extent of pretraining ensures that the network reaches good performance on the 10-class digit classification; the trained network also exhibits good performance on two-digit categorization at any location, in the absence of distractors (Fig.2A), but falls to near-chance levels when distractor digits are introduced. During task-refinement, learning alternates between updating the modulator coupling $\mathbf{c}$ by backpropagation, and updating the readout gains $\mathbf{g}$ using correlations estimated within a single trial ($T = 100 - 500$) according to Eq. 3. To simplify the comparison to other forms of feedback-based learning and avoid any interactions between intrinsic network noise and feedback-based learning, the network dynamics are deterministic (modulator is held constant) during the backpropagation steps, and modulator stochasticity is only introduced in the second step (m_{kt} drawn i.i.d. from $\mathcal{N}(1, 0.1)$). Hence, the effects of modulation on the coupling gradients are indirect, via its effects on the readout gain.

To assess the effectiveness of this combined learning, labelling, and decoding procedure on specialized task performance, we compared it to three alternatives. The first uses backpropagation to relearn all feedforward weights (initializing from the pretrained weights), which we term 'retraining' ($m = 0$, $\sigma_m^2 = 0.0$). The second uses an attentional modulator to boost only the encoder gain, the feedforward weights are fixed, and the responses in the encoding layers are amplified by scale factors (the task-specific encoder gain) learned via backpropagation ($m = 1$, $\sigma_m^2 = 0.0$); since our procedure also includes a boost in mean responses, this control provides a natural lower bound on the benefits of stochastic gain modulation. The third method adjusts the weights of the final layer ('retraining last layer'), leaving the rest of the network unchanged.

Modulator label allows for efficient and effective fine-tuning. Evaluating the performance of different learning algorithms on example digit pair tasks reveals systematic differences in the speed of learning, with the stochastic modulator outperforming its competitors by a substantial margin (Fig. 2B). We quantified the systematicity of these observations across all tasks by measuring the initial slope of learning, estimated with linear regression over the first 50 measurements (Fig. 2C), and by the number of training examples required to reach a criterion accuracy level of 70% (Fig. 2D).[3] These measures confirm that using a learned stochastic modulator to fine-tune the network to the requirements of the new task is faster / requires less data than the other three methods. The complete retraining is generally much slower, presumably because it needs to tune a much larger number of parameters. Importantly, the stochastic modulation scheme generally improves over deterministic gain boosts, despite the fact that the feedback-based part of learning is identical in the two conditions.

To better understand the nature of the adapted modulation solution, we linearly projected the modulation strengths (after training on 400 examples) back into the pixel space, as a means of visualizing which features in the input are enhanced via modulation (Fig. 2E). The modulation is seen to preferentially affect localized patterns that reflect both common and distinct features of the two task-relevant digits, within the task-relevant quadrant. This spatial specificity is expected, given that the task requires quadrants containing distractors to be ignored, and that the quadrant structure is explicitly mirrored in the encoding layer. The fine structure of the modulation effects reflects the shape of the digits to be discriminated, showing that the modulation targets neurons with task-relevant feature selectivity. In fact, further analysis shows that within the subgroup of spatially relevant neurons there is a strong positive relationship between task-specific coupling strength and informativeness measured

[2] For details on hyperparameters and their optimization, please see Suppl. Info.

[3] Since retraining the last layer performed very poorly overall, we did not include it here.

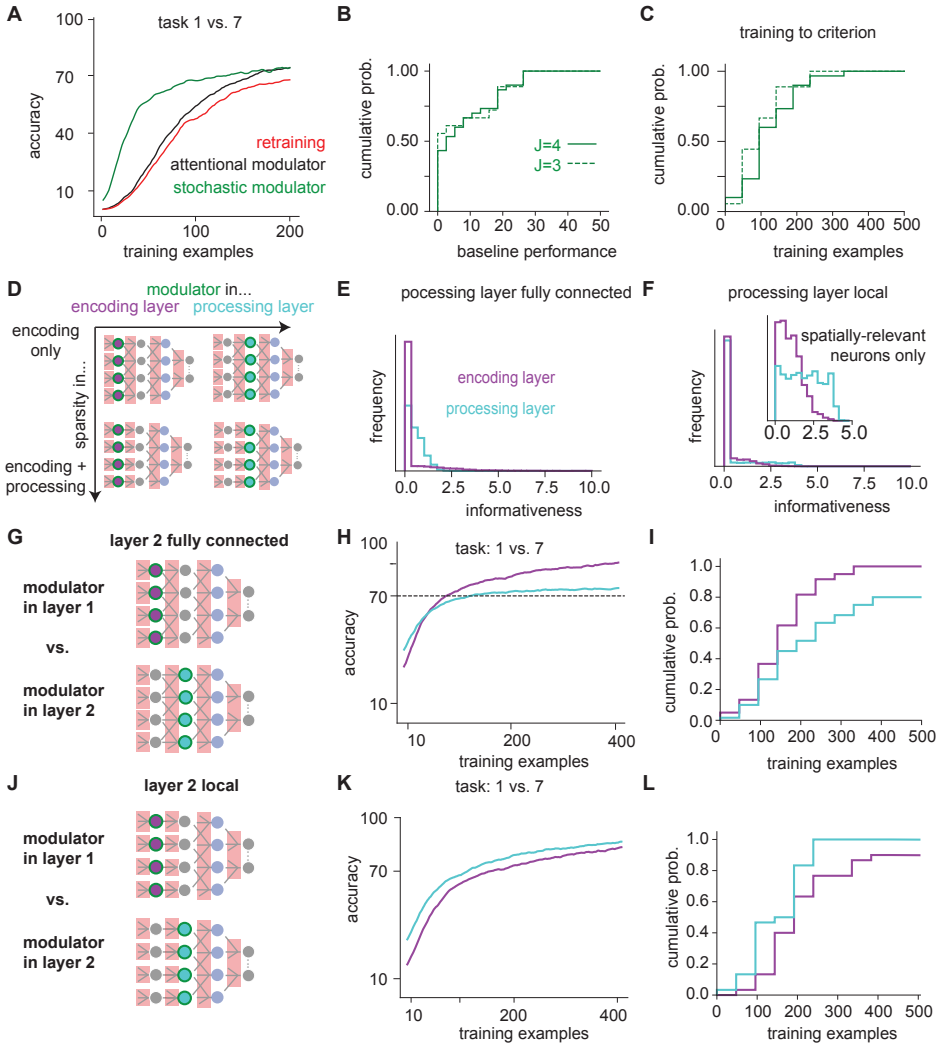

Figure 3: Stochastic modulation improvements are robust to changes in architecture. A) Performance comparison for architecture with two all-to-all intermediate layer. B) Baseline performance distribution for $J = 3$ vs. $J = 4$ layers. C) Corresponding distribution for the number of training examples needed to reach the 70% performance criterion. D) Variations of model architecture along two dimensions, modulator placement and sparsity. E) Informativeness ($|d'|$) distributions for neurons in layer 1 (purple) and 2 (cyan) when layer 2 is all-to-all connected; all neurons included. F) Same as D, when layer 2 is locally connected. Inset: Same as E, but only considering neurons with receptive fields within the task-relevant quadrant. G) Comparing effects of directing modulation in either the first (solid) or the second (dashed) layers for the network with fully connected layer 2 with respect to H) accuracy on an example task, and I) number of training examples needed to reach criterion. J-L) Same as G-I, when layer 2 is locally (within-quadrant) connected.

by $|d'|$ (Fig. 2F).[4] As in the original model [21], the informativeness distribution is skewed, with task informativeness concentrated in relatively small subpopulations of neurons that are distinct across tasks (Fig. 2G). The learned coupling correspondingly changes across tasks to reflect these differences in task informativeness (Fig. 2H: neurons that are more informative in one compared to another task tend to also have higher coupling strength in that task). This confirms that the task-specific targeting of stochastic modulation posited in the original theory can be directly learned from experience.

[4]Note that although $|d'|$ is easy to compute across layers, it only provides a coarse measure of informativeness by ignoring the effects of network nonlinearities.

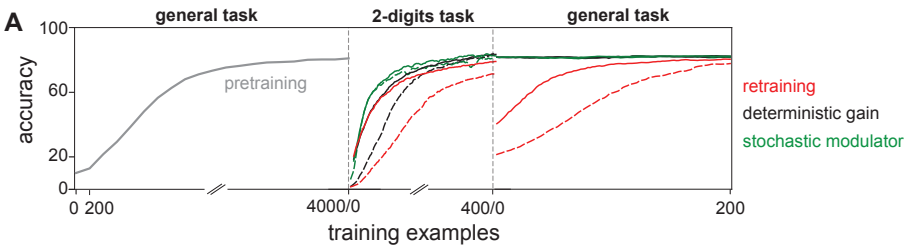

Figure 4: Task switching. A) Evolution of the network's categorization accuracy over learning, from pretraining, to specific task, and returning to the general task; solid and dashed lines show results for two example tasks, respectively.

Results generalize to deeper architectures. One concern is that the initial modulator label will lose its specificity as it propagates through additional layers of distributed nonlinear processing. To test how intermediate levels of processing affect learned stochastic modulation we extended the network by incorporating an additional all-to-all connected layer. Repeating the experiments above, we confirmed that our learned stochastic modulation still functions, even when the label needs to propagate further (Fig.3A). The experiments on the new architecture qualitatively reproduced the speed and sample efficiency improvements of the stochastic modulator over alternative learning procedures, but a direct comparison between stochastic modulation in the 3- vs. 4-layer architecture does reveal a modest slow-down of learning, despite the fact that baseline performance was statistically matched between the two (Fig.3B, C).

Modulation should target layers in which task information is concentrated in a subpopulation. Thus far, we have assumed that the primary effects of modulation are directed towards the first (encoding) layer of the network, but does that need to be the case? The theory assumes only a small fraction of the modulated population carries task-relevant information. In contrast, if this information were uniformly distributed across the entire layer, stochastic modulation should not help at all. Hence, the presence of task-specific localized information seems to be a critical consideration for effective targeting of stochastic modulation.

Different choices of neural architecture will localize information about features in the input in different ways (Fig. 3D). Intuitively, we expect that sparse and localized connectivity will result in features (e.g. locations) being represented in subsets of neurons, whereas broad or all-to-all connectivity is likely to distribute feature information across neurons. Indeed, this is the case in our simple networks: the informativeness distribution is substantially broader in the second (all-to-all connected) layer, compared to the first (Fig. 3E; $|d'|$ estimated using the task digit pair, for the pretrained network). In contrast, when the first two layers have local connectivity, the additional nonlinearities allow for more specific feature selectivity in the second layer (Fig. 3F), especially within the subset of neurons coding for the task-relevant quadrant (inset in Fig. 3F).

We hypothesize that applying the modulation to the second layer should negatively affect the ability of the stochastic modulator to fine-tune the network when that layer is all-to-all connected. The opposite should hold when its weights are localized. To avoid potential confounds caused by across-layer differences in the neural nonlinearities, we use $\mathrm{ReLU}(\cdot)$ as the activation function in all layers, and modify first-stage modulation in Eq. 1 as follows: $\mathbf{h}_{kt}^{(1)} = \mathrm{ReLU}\left(\mathbf{W}_1 \mathbf{s}_k\right) \exp\left(m_{kt}\mathbf{c} - b\right)$. When the second layer is all-to-all connected, we find that directing the modulator towards the encoding layer yields faster learning and better end performance (Fig.3, H). Quantifying the number of training examples required to reach criterion performance across tasks reveals a systematic shift in the distribution across the two scenarios, confirming that early modulation is preferable for this architecture (Fig.3I). Repeating the same analysis with an architecture in which both layers are spatially localized leads to the opposite conclusion (Fig.3J-L). Here we find that performance is better for late modulation (Fig.3K), with results robustly reproduced across task instances (Fig.3L). Overall, these results confirm our expectation that stochastic modulation is most effective when directed towards bottlenecks of task-relevant information.

Seamless switching back to the general task and continual learning. One of the immediate appeals of gain modulation (either stochastic or deterministic) as a mechanism for task-specific information routing is that the feedforward weights stay unchanged. Returning to original performance in the general task is instantaneous, while a retrained network with good performance on the specific task takes about 100-200 trials to reach the original, pretraining performance on the general task again (Fig. 4A). The combination of this capability and the increase in speed of learning makes stochastic modulation a remarkably effective mechanism for adapting (and unadapting) to specific tasks. In contrast, weight retraining alters the entire network in a way that cannot be easily undone. The extent of these changes during task retraining depends on the task itself and the training duration. In extreme cases, parameter retraining to restore initial capabilities may take just as long as the original pretraining.

4 Discussion

Brains exhibit a remarkable capacity for flexible behavior, refocusing their computational resources on specific tasks as needed. We have proposed a framework enabling such adaptive behavior, in which stochastic modulation adaptively and transiently fine-tunes the hierarchical processing of task-relevant features, while retaining a stable network 'backbone' for general computation. The key idea is that rapidly varying modulatory noise injected into the task-relevant subset of neurons in an early processing layer accompanies the stimulus information as it propagates through the network, and guides the readout at the final decision stage. The selection of targeted neurons is learned from trial feedback on the current task. We explored the properties of this mechanism in multilayer feedforward networks trained on variants of MNIST digit classification. We found that task specific targeting of the modulator can be learned from small numbers of examples, yielding substantially more efficient task adaptation than traditional deterministic gain modulation, or retraining of the feedforward network as a whole. Moreover, we found that modulation is most effective when injected into the layers in which task-specific information is concentrated in a small fraction of the neurons.

In the numerical experiments presented here, we have used spatial locality to create and control informativeness bottlenecks in the network. Nonetheless, the modulatory mechanism was also able to exploit informativeness in the shape of the relevant digits, whose feature locality emerged solely from pretraining on the recognition task. This confirms the generality of task-relevant feature targeting.

Two important principles guide placement of the modulator. First, targeted layers should be task-specific representational bottlenecks (i.e. informativeness should be sparsely distributed in the population). When multiple layers exhibit such structure (as in our local layer 2 example), then placing the modulator closer to the decision yields better performance, likely because the stochastic modulator has to propagate through fewer layers, and/or because the learning signals are also stronger when backpropagated through fewer layers [24]. Since the sparsity of feature representations is determined through complex interactions between network architecture, statistics of the training data, and details of pretraining (including the algorithm and choices of regularization), we expect that these same issues will affect the optimal targeting of the modulator, and its overall success in learning to fine-tune the network. Nonetheless, networks whose feature representations are inherently localized in space, and across distinct channels may particularly benefit from stochastic modulation. Biologically, such feature maps are ubiquitous in cortical sensory processing. Spatially localized receptive fields with selectivity to specific image features have been discovered and studied throughout the visual hierarchy, and the sparsity of the associated neural responses appears to be conserved throughout [25].

For the stochastic mechanism to achieve its goals, noise needs to be directed towards the relevant neurons, and this targeting has to be learned from experience. In the examples presented here, this learning arises through backpropagation, which is not only biologically-unrealistic but also has the practical disadvantage that the learning signals get weaker as the depth of the network increases [24]. This is a common problem in training of CNNs, where clever architectural additions such as skip connections provide a way of speeding up learning [26, 27]. In the specific context of stochastic co-modulation we have the additional advantage that we only need to update the modulator couplings, and the intermediate backpropagation signals do not need to be represented explicitly. As such, it should be possible to train a separate network to directly generate the required signals in parallel to pretraining (since the backpropagation operations are architecture-specific, but not task-specific), similar to synthetic gradients [28, 29]. Once the learning signal is available in the modulated circuit, the update of individual modulator strengths is local and Hebbian in form, and thus amenable to

implementation via synaptic plasticity. Since the modulator varies on a time scale independent of and much faster than the stimulus presentations, learning the modulator-guided decoder gain is data efficient. However, there is an increased computational cost that comes with updating the decoder gain whenever the encoder gain changes. Hence, there is a trade-off between how much data is available for learning and the computation time required for fine-tuning the network.

We found that stochastic modulation signals injected early in a hierarchical network remains reasonably effective in labeling and guiding decoding several stages later. This was not a foregone conclusion: models of attention in deep convolutional networks for object categorization have documented instances when attentionally-induced increases in activity of early layers fail to propagate to decision circuits [14]. Intuitively, deterministic gain modulation intermingles with information about the stimulus (the stimulus-driven responses) with information about task relevance (the gains), making it difficult or even impossible to disentangle the two at later stages of processing [30]. In contrast, the stochastic modulation signal is essentially orthogonal to the stimulus information, and thus can serve as an accessible label for the relevant stimulus information, analogous to the role of the carrier signal in FM radio transmission. These theoretical considerations suggest that the primary mechanism for task-specific information routing in the brain could be structured covariability, rather than increases in response amplitudes. This may explain some puzzling experimental observations in which attentional effects on behavior could be disrupted, despite intact attentional boosts in neural responses [31]. It is likely that both processes contribute to the behavioral improvements we typically attribute to attention, with boosts in mean responses serving to improve the initial encoding of the stimulus, and targeted covariability facilitating task-specific signal transmission and decoding.

Attention mechanisms inspired by the brain have already been shown to improve performance of deep learning models [32], but both continual learning and few-shot learning after distribution shifts remain key open problems. Current machine learning algorithms approach these problems from many different angles, from optimizing the network's initial conditions for subsequent training (as in MAML [33]), to probabilistically detecting changes in the input distribution [34], or learning segregate representations across tasks to begin with [1, 2, 35]. Our model adds the paradigm of stochastic modulation to this list, opening the door for new biologically-inspired advances in machine learning.

References

[1] Nicolas Y Masse, Gregory D Grant, and David J Freedman. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, 115(44):E10467—-E10475, 2018.

[2] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[3] Stefano Fusi, Patrick J Drew, and Larry F Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, 2005.

[4] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[5] David A McCormick, Dennis B Nestvogel, and Biyu J He. Neuromodulation of brain state and behavior. *Annual review of neuroscience*, 43:391–415, 2020.

[6] Jeffrey Moran and Desimone Robert. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784, 1985.

[7] Carrie J McAdams and John HR Maunsell. Effects of attention on the reliability of individual neurons in monkey visual cortex. *Neuron*, 23(4):765–773, 1999.

[8] Stefan Treue and John HR Maunsell. Attentional modulation of visual motion processing in cortical areas mt and mst. *Nature*, 382(6591):539–541, 1996.

[9] Marlene R. Cohen and John H.R. Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12(12):1594, 2009.

[10] Jude F. Mitchell, Kristy A. Sundberg, and John H. Reynolds. Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron*, 63(6):879–888, 2009.

[11] Douglas A. Ruff and Marlene R. Cohen. Attention can either increase or decrease spike count correlations in visual cortex. *Nature neuroscience*, 17(11):1591–7, 2014.

[12] Richard J Krauzlis, Lee P Lovejoy, and Alexandre Zénon. Superior colliculus and visual spatial attention. *Annual review of neuroscience*, 36:165–182, 2013.

[13] Leonardo Chelazzi, Earl K Miller, John Duncan, and Robert Desimone. A neural basis for visual search in inferior temporal cortex. *Nature*, 363(6427):345–347, 1993.

[14] Grace W. Lindsay and Kenneth D. Miller. How biological attention mechanisms improve task performance in a large-scale visual system model. *elife*, pages 1–29, 2018.

[15] RLT Goris, JA Movshon, and EP Simoncelli. Partitioning neuronal variability. *Nature Neuroscience*, 17(6):858–865, 2014.

[16] Alexander S Ecker, Philipp Berens, R James Cotton, Manivannan Subramaniyan, George H Denfield, Cathryn R Cadwell, Stelios M Smirnakis, Matthias Bethge, and Andreas S Tolias. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1):235–248, 2014.

[17] Marieke L. Schölvinck, Aman B. Saleem, Andrea Benucci, Kenneth D. Harris, and Matteo Carandini. Cortical state determines global variability and correlations in visual cortex. *Journal of Neuroscience*, 35(1):170–178, 2015.

[18] I. Chun Lin, Michael Okun, Matteo Carandini, and Kenneth D. Harris. The nature of shared cortical variability. *Neuron*, 87(3):644–656, 2015.

[19] N Rabinowitz, RL Goris, M Cohen, and EP Simoncelli. Attention stabilizes the shared gain of v4 populations. *Elife*, pages 1–24, 2015.

[20] Caroline Haimerl, Douglas A Ruff, Marlene R Cohen, Cristina Savin, and Eero P Simoncelli. Targeted comodulation supports flexible and accurate decoding in V1. *bioRxiv*, 2021.

[21] Caroline Haimerl, Cristina Savin, and Eero P Simoncelli. Flexible information routing in neural populations through stochastic comodulation. *Advances in Neural Information Processing Systems*, 32:14402–14411, 2019.

[22] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[23] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

[24] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. *arXiv preprint arXiv:1507.06228*, 2015.

[25] Nicole C. Rust and James J. DiCarlo. Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *Journal of Neuroscience*, 32(30):10170–10182, 2012.

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[28] Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In *International Conference on Machine Learning*, pages 1627–1635. PMLR, 2017.

[29] Owen Marschall, Kyunghyun Cho, and Cristina Savin. A unified framework of online learning algorithms for training recurrent neural networks. *Journal of Machine Learning Research*, 21(135):1–34, 2020.

[30] Taosheng Liu, Jared Abrams, and Marisa Carrasco. Voluntary attention enhances contrast appearance. *Psychological science*, 20(3):354–362, 2009.

[31] A. Zènon and R. Krauzlis. Attention deficits without cortical neuronal deficits. *Nature*, 489:434––437, 2012.

[32] Grace W Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14:29, 2020.

[33] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[34] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. TENT: Fully test-time adaptation by entropy minimization. *ICLR*, 2021.

[35] Lea Duncker, Laura Driscoll, Krishna V Shenoy, Maneesh Sahani, and David Sussillo. Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Fine-tuning hierarchical circuits through learned stochastic co-modulation — Supplement —

Anonymous Author(s)
Affiliation
Address
email

1 Training

The loss for pretraining is defined by the crossentropy function with 10-categories and parameters were optimized using backpropagation with Adam **?**. The same loss function and optimization is used for the task-learning. The learning rate for pretraining is $1\mathrm{e}{-4}$ with a batch size of 200 images and 20 batches used for training (resulting in a total of 4000 images).

For the task-training, the batch size is reduced to 2 images, to allow testing performance in the low-sample regime. The total number of batches may vary and is specified in each main text figure. The modulator-coupling learning is stable for a learning rate of $1\mathrm{e}{-3}$ to $1\mathrm{e}{-4}$ (Fig.**??**A). We use the slower learning rate of $1\mathrm{e}{-4}$. Given the small batch size and the many parameters that need to be adjusted, similar learning rates lead to unstable learning trajectories for retraining. We optimized the retraining learning rate hyperparameter so as to achieve stable learning, measured by the variance of the across-runs final performance. We used a grid search with log-spacing and found that a learning rate of $1\mathrm{e}{-6}$ provided a low-variance learning performance similar to that of pretraining and modulator based learning at their respective learning rates (measured in $\%$ correct). For the task-learning there is an L_1-norm penalty term applied to weights and coupling ($\lambda = 0.1$).

For the modulator learning, the coupling parameters where initialized i.i.d. from the uniform distribution $[0.9, 1.1]$.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.