Deep Feedback Models

David Calhas
INESC-ID
Instituto Superior Tecnico
University of Lisbon
&
Faculty of Computer and Information Science
University of Ljubljana

Arlindo L. Oliveira
INESC-ID
Instituto Superior Tecnico
University of Lisbon

Abstract

Deep Feedback Models (DFMs) are a new class of stateful neural networks that combine bottom up input with high level representations over time. This feedback mechanism introduces dynamics into otherwise static architectures, enabling DFMs to iteratively refine their internal state and mimic aspects of biological decision making. We model this process as a differential equation solved through a recurrent neural network, stabilized via exponential decay to ensure convergence. To evaluate their effectiveness, we measure DFMs under two key conditions: robustness to noise and generalization with limited data. In both object recognition and segmentation tasks, DFMs consistently outperform their feedforward counterparts, particularly in low data or high noise regimes. In addition, DFMs translate to medical imaging settings, while being robust against various types of noise corruption. These findings highlight the importance of feedback in achieving stable, robust, and generalizable learning. Code is available at github.com/DCalhas/deep_feedback_models.

1 INTRODUCTION

Although artificial neural networks draw inspiration from biology, they are typically static once a prediction is made, unable to revisit or refine their output based on subsequent processing. This contrasts with the dynamic and adaptive nature of the human brain, which continually updates its internal model of the world to minimize the mismatch between expectations and sensory data. The predictive coding framework offers a computational account of this process Rao et al. [1999], Clark [2024], positing that the brain operates as a hierarchical inference system. In this view, higher cortical regions generate anticipations about sensory input, while lower regions compute the discrepancy, or prediction error, between those expectations and the actual signal. These error signals drive updates to the internal state, allowing perception to evolve iteratively. Inspired by this model, we propose a feedback based neural architecture that integrates prediction error into its internal state over time. By recurrently incorporating its own output into subsequent computations, the model refines its internal representations through multiple cycles of inference. This iterative scheme contrasts with conventional feedforward networks, which operate in a single computational pass. Our approach mimics perceptual inference in the brain, where interpretation emerges gradually by reducing the gap between top-down expectations and bottom-up evidence Rao et al. [1999], Alamia et al. [2023]. In particular, this inference process does not follow a fixed path, but instead adapts dynamically to the input, enabling convergence to a coherent perceptual state Bogacz [2017].

We introduce *Deep Feedback Models* (DFMs), a new class of neural architectures in which high-level predictions are fed back into earlier processing stages, enabling internal states to evolve dynamically over time. DFMs are inspired by a core principle in neuroscience, *predictive coding*, which posits that the brain continually refines its internal representations by minimizing the mismatch between expected and actual sensory inputs Rao et al. [1999], Friston [2005]. Unlike prior recurrent models that repeat layers or use shal-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

low feedback in the final stages Spoerer et al. [2017], Linsley et al. [2020], Goetschalckx et al. [2023], DFMs implement full network feedback loops: the evolving state is propagated through the network and recursively informs future updates, as illustrated in Figure 1. This mechanism enables internal representations to be refined toward consistent predictions, an essential characteristic of predictive coding. In this work, we evaluate DFMs in tasks requiring robustness and generalization, and report several key findings:

- DFMs consistently outperform feedforward baselines under increasing levels of input noise (Section 5.1);
- DFMs generalize from fewer labeled examples, achieving accurate predictions in few-shot settings (Section 5.2);
- DFMs translate to medical imaging applications, which are characterized by limited observations and high levels of noise, outperforming their feedforward counterparts in these settings (Section 5.3);
- DFMs consistently outperform feedforward baselines in different types of noise (Section 5.3).

Our contributions are as follows:

- We propose a biologically inspired feedback architecture (Section 3) that enables stable learning on large scale data through an orthogonalization step (Gram-Schmidt, see Section 3.2);
- We extend this framework to image segmentation by introducing a spatially dissipative exponential decay layer (Section 3.1);
- We present extensive experiments comparing DFMs and their feedforward counterparts across noise, data efficiency, and medical imaging settings (Section 4);
- We offer an analysis of the internal dynamics and stability of DFMs (Section 6), and position our work within the broader context of biologically inspired and equilibrium based models (Section 7).

2 PROBLEM DESCRIPTION

Traditional artificial neural networks are *static* systems: given an input $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ and parameters θ, the network, $F : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{N\times h\times w}$, produces a fixed internal signal $F(\mathbf{x};\theta)$ and a final prediction $G(F(\mathbf{x};\theta))$, where $G : \mathbb{R}^{N\times h\times w} \to \mathbb{R}^{L}$ making $G \circ F : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{L}$. $H \times W$ represents the height and width of an image, whereas $h \times w$ represents height and width of a latent representation. Despite efforts to make artificial networks more biologically plausible Kubilius et al. [2018], Dapello et al. [2020], Rajalingham et al. [2022], most architectures lack one crucial feature of biological systems, *feedback*, the ability for high level predictions to recursively influence earlier layers. To incorporate feedback, we introduce a dynamical system with a time dependent state vector $\mathbf{h}(t) = [\mathbf{v}(t), \mathbf{u}(t)] \in \mathbb{R}^{(B+N)\times H\times W}$, which is iteratively updated by a deep neural function $F' : \mathbb{R}^{(C+B)\times H\times W} \to \mathbb{R}^{(N+B)\times h\times w}$, similar to F but the final representation is upsampled. The state is split into a feedback component $\mathbf{v}(t) \in \mathbb{R}^{B\times h\times w}$, which is softmax normalized and concatenated with the input, and an output component $\mathbf{u}(t) \in \mathbb{R}^{N\times h\times w}$, which is decoded by a classifier G:

$$\delta(\mathbf{x}, \mathbf{h}, t) = F'\left(\left[\mathbf{x}, \frac{\exp\left(\mathbf{v}(t)\right)}{\sum_i \exp\left(\mathbf{v}_i(t)\right)}\right]; \theta\right) \tag{1}$$

$$\hat{\mathbf{y}}(t) = G\left(\mathbf{u}(t)\right). \tag{2}$$

Let $\mathcal{D} = \bigcup_i \{\mathbf{x}_i, \mathbf{y}_i\}$ be a dataset where each input $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$ is an image, and the corresponding ground truth $\mathbf{y}_i$ is either a label vector $\in \mathbb{R}^{L}$ (for classification tasks) or a segmentation mask $\in \mathbb{R}^{L\times H\times W}$. Here, C is the number of input channels, $H \times W$ is the spatial resolution, and L is the number of classes. Our objective is to learn model parameters, θ, by minimizing the cross entropy loss over the dataset as

$$\mathcal{L}(\theta) = -\sum_{i}^{|\mathcal{D}|} \mathbf{y}_i \log \circ G \circ F'\left(\left[\mathbf{x}_i, \frac{\exp\left(\mathbf{v}(T)\right)}{\sum_j \exp\left(\mathbf{v}_j(T)\right)}\right]; \theta\right). \tag{3}$$

We evaluate two key capabilities of the model: robustness to noise and generalization from few examples. These are tested under the following two experimental settings:

- **Noise setting**, where Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is added to each input, i.e., $\mathbf{x}_i + \epsilon$, for a fixed dataset size $|\mathcal{D}|$. This simulates degradation in sensory input and challenges the model's ability to extract reliable features under corruption;
- **Few-shot setting**, where the number of labeled examples per class is limited to $D \ll |\mathcal{D}|/L$, with no added noise (i.e., $\sigma = 0$). We sample D instances per class uniformly from the dataset $\mathcal{D}$. This setting tests the model's ability to generalize from limited supervision.

These two conditions are mutually exclusive: noise is applied only when all training data are used, and few-

shot generalization is evaluated on clean data. Formally, the combined training objective in either case becomes

$$\mathcal{L}(\theta) = - \sum_{i}^{D \times L} \mathbf{y}_i \log \circ G \circ F' \left(\left[\mathbf{x}_i + \epsilon, \frac{\exp\left(\mathbf{v}(T)\right)}{\sum_j \exp\left(\mathbf{v}_j(T)\right)} \right]; \theta \right), \tag{4}$$

where $\epsilon = 0$ in the few-shot case, and $D \times L = |\mathcal{D}|$ in the noisy case. We perform few-shot experiments only for classification tasks, since segmentation requires nontrivial control over per-pixel class distributions. Throughout, we apply this setup to evaluate both feedforward and DFMs under controlled perturbations and limited supervision, in order to measure their relative strengths.

3 FEEDBACK

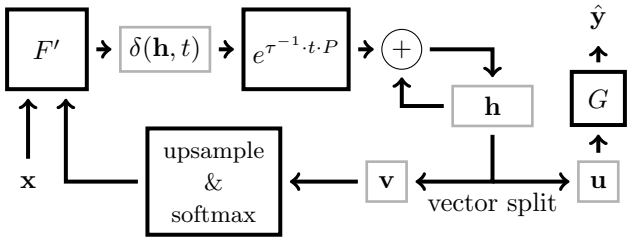

Figure 1: Neural network circuit with high level feedback.

DFMs operate as dynamical systems, where internal representations evolve over time. However, in contrast to classical recurrent cells or shallow feedback loops Spoerer et al. [2017], Linsley et al. [2020], Goetschalckx et al. [2023], we apply feedback across a full deep neural network, as illustrated in Figure 1. This introduces significant challenges because deep architectures are highly non linear and difficult to stabilize, which makes ensuring convergence or contractivity of the underlying dynamics non trivial, especially in large-scale settings. To guarantee that the system stabilizes, i.e. the internal state $\mathbf{h}(t)$ does not diverge, we introduce an exponential decay mechanism defined as

$$\mathbf{h}(T) = \mathbf{h}(0) + \sum_{t=0}^{T-1} \delta(\mathbf{h}, t)^\top \cdot Q \cdot e^{\tau^{-1} \cdot t \cdot \Sigma} \cdot Q^{-1}, \tag{5}$$

where $\delta(\mathbf{h}, t)$ is the signal produced by F' defined in equation 1, Q represents the matrix with eigenvectors in the columns, and Σ are the eigenvalues of the system. In order for the system to decay over time (see Figure 2), we define $\Sigma = -I$. Note that the projection matrix is $P = Q \cdot \Sigma \cdot Q^{-1}$, and its exponential becomes trivial, $e^P = Q \cdot e^\Sigma \cdot Q^{-1}$. This matrix defines the iterative feedback dynamics of the neural network as $\frac{\mathrm{d}\mathbf{h}}{\mathrm{d}t} = \delta(\mathbf{h}, t)^\top \cdot e^{\tau^{-1} \cdot t \cdot P}$. In practice we use a forward-Euler discretization ($\Delta t = 1$), producing the discrete update shown in equation 5. This decay mimics a dissipative process, gradually damping out feedback energy over time. Biologically, this reflects the natural temporal smoothing in cortical circuits Gros [2007]; mathematically, it introduces an attractor like behavior that leads to asymptotic stability.

3.1 Convolution with exponential decay

While the exponential decay mechanism described above enforces temporal stability by damping feedback signals, it applies uniformly across spatial locations, treating each pixel independently. However, in dense prediction tasks like semantic segmentation, spatial interactions between neighboring pixels are essential for accurate boundary localization and region coherence. A convolutional layer is parametrized by $\mathbf{w} \in \mathbb{R}^{k_h \times k_w \times C_{in} \times C_{out}}$. This operation can be modeled for exponential decay taking into account the spatial dimension considered in a convolution. Consider a single kernel window associated to an input and output channel, j, such that $\mathbf{w}^j \in \mathbb{R}^{k_h \times k_h}$ is square, i.e. $k_h = k_w$, then $\mathbf{w}^j = Q_j \cdot \Sigma_j \cdot Q_j^{-1}$ is the eigendecomposition of the kernel window where columns of Q are eigenvectors and the diagonal of Σ contains the eigenvalues. For simplicity, we refer to a convolution simply as $\mathbf{x} * \mathbf{w}$. We can now define a system that evolves in time by doing spatial exponential decay as

$$\frac{\mathrm{d}\mathbf{h}}{\mathrm{d}t} = \delta(\mathbf{h}, t) * e^{\tau^{-1} \cdot t \cdot \mathbf{w}}, \tag{6}$$

whose Euler discretization is defined by

$$\mathbf{h}(T) = \mathbf{h}(0) + \sum_{t=0}^{T-1} \delta(\mathbf{h}, t) * e^{\tau^{-1} \cdot t \cdot \mathbf{w}}. \tag{7}$$

This system naturally dissipates pixel's magnitude in the spatial dimensions defined by the kernel and is easily solved by exponentiating the eigenvalues as

$$\mathbf{h}(T) = \mathbf{h}(0) + \sum_{t=0}^{T-1} \delta(\mathbf{h}, t) * \left(Q \cdot e^{\tau^{-1} \cdot t \cdot \Sigma} \cdot Q^{-1} \right), \tag{8}$$

whose limit is naturally $\lim_{t \to \infty} \mathbf{h}(t+1) - \mathbf{h}(t) = 0$.

Normalization. In the backward pass, the gradient flows through a pixel of the input image k_h^2 times, corresponding to the total number of pixels in the kernel window. This gradient accumulation is normalized by a constant $Z = \frac{1}{k_h}$, such that $h(t+1) = h(t) + \frac{1}{Z} \frac{\mathrm{d}\mathbf{h}}{\mathrm{d}t}$.

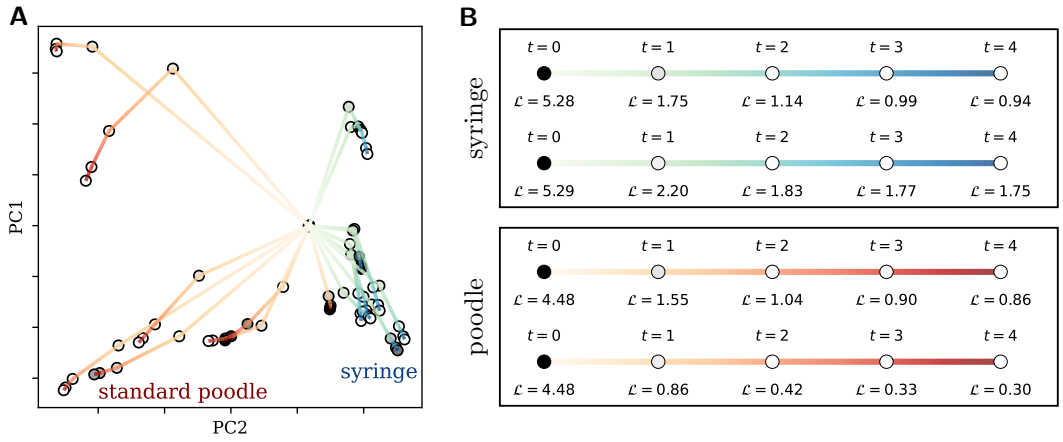

Figure 2: **A** illustrates the stability of the proposed system over time. Each line represents the trajectory of a single instance vector softmax $(G(v(t)))$. Red lines represent instances of the class *standard poodle*, which is a dog breed, and blue lines represent the class *syringe*. The trajectories are projected onto its two principal components using PCA. To compute this projection, we collect all vectors softmax $(G(v(t)))$ from all validation instances of the TinyImageNet dataset Le et al. [2015] at each timestep t, and perform PCA on this collection to obtain global directions of maximal variance. The same transformation is applied to all trajectories for consistency. The subplots represent the evolution from timestep $t = 0$ to $t = 4$. The corresponding loss value $\mathcal{L}$ is represented by how dark a marker is at that timestep, the darker the higher the loss. Early in training ($t = 0$), the trajectories exhibit large step changes, as the system begins from random initializations $\mathbf{h}(0) \sim \mathcal{N}(0, 0.001)$. **B** shows the specific loss values at each time steps for two examples of each class each.

3.2 Orthogonality

Until now, we are considering a recurrent system, whose Jacobian is defined as

$$\mathbf{J_h} = 1 + \frac{\partial \delta(\mathbf{h}, t)}{\partial \mathbf{h}} * \left[Q \cdot e^{\tau^{-1} \cdot t \cdot \Sigma} \cdot Q^{-1} \right]. \quad (9)$$

The fact that the Jacobian contains a neural network with a high level of nonlinearities makes this problem difficult and more susceptible to the nature of F', which may result in vanishing or exploding gradients. Jacobian regularizations Bai et al. [2019], Goetschalckx et al. [2023], Linsley et al. [2020] are suitable when F' is a simple cell. However when it contains a high level of complex nonlinearities, it becomes unfeasible to apply them. In equation 9, the exponential decay already guarantees a stable property for the system. If we wish to train Q and Q^{-1}, these can neither amplify nor shrink the gradient/response. For this, we introduce an orthogonal correction step after each gradient update. This goes in accordance with Riemann optimization Bécigneul et al. [2018], but instead of doing Householder projections (is computationally expensive and therefore does not allow training in large scale data), we perform a simple QR decomposition, a.k.a. Gram-Schmidt algorithm Leon et al. [2013], defined as

$$U(s) = Q(s) \cdot R(s), \quad (10)$$

where U is the non orthogonal representation of Q, Q is an orthogonal matrix, R is an upper triangular matrix, and s denotes the sth optimization step. Every time there is a gradient update to Q, as

$$U(s) = Q(s-1) + \eta \nabla_{Q(s-1)} \mathcal{L}, \quad (11)$$

equation 10 is applied to correct U to Q, allowing us to write $Q^{-1} = Q^{\top}$. This ensures stability in training and enables the model to learn under super fast convergence settings Smith et al. [2019], which would not be possible otherwise because of numerical underflow or overflow.

4 EXPERIMENTAL SETTING

We assess the robustness and generalization capabilities of our feedback models under two experimental conditions: input noise (denoted by σ) and number of examples per class (denoted by D), as described in Section 2. We consider three datasets: ImageNet-1k Deng et al. [2009], used to test both robustness to noise and generalization from limited samples; COCO-Stuff Lin et al. [2014], used for semantic segmentation under the noise condition; and MedMNIST Yang et al. [2023] where we test on the original test sets (Retina, Blood, Derma, Pneumonia, and Breast) and on the RetinaMNIST-C Di Salvo et al. [2024] to assess performance under different types of noise (brightness down, defocus blur, jpeg compression, pixelate, contrast down, gaussian noise, motion blur, speckle

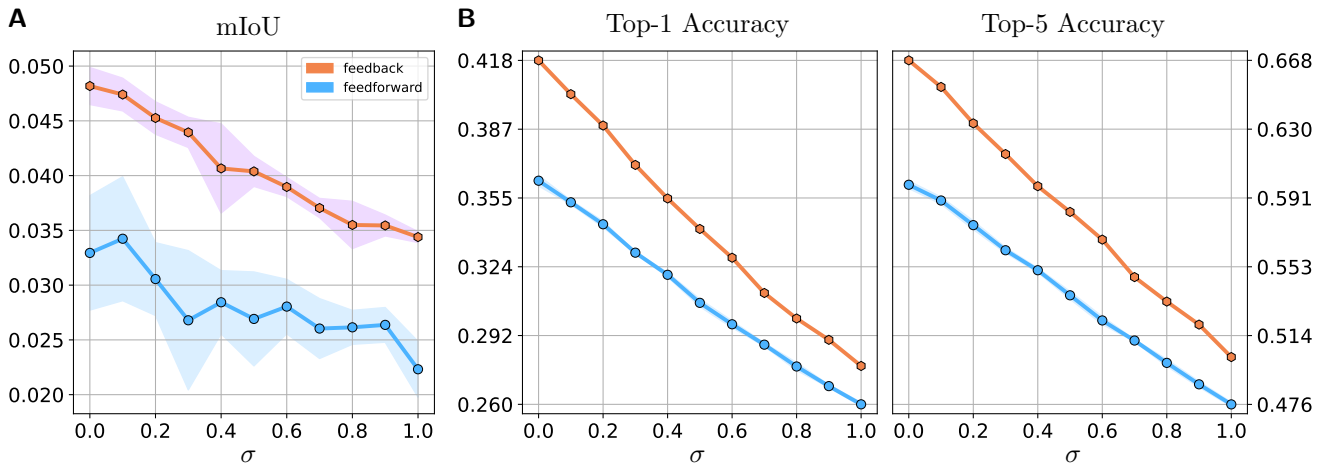

Figure 3: Performance of feedback versus feedforward in noisy settings. **A** shows the noise impact of these models in the COCO-Stuff dataset. **B** shows the noise impact in the Imagenet dataset.

noise). Different neural network architectures are used depending on the task:

- For ImageNet classification: a ResNet-50 is used for the low-data regime experiments; and a ResNet-18 is used for the noise experiments for faster training.
- For semantic segmentation (COCO-Stuff), we use DeepLabV3+ with a ResNet-18 encoder.
- For MedMNIST classification, we use a ResNet-50.

Input preprocessing differs per setting. In the ImageNet and COCO-Stuff noise experiments inputs are randomly cropped and resized to 64×64 (naturally this decreases the expected performance of the models, but speeds up training). Downsampling significantly the resolution for the COCO-Stuff emulates an experiment of noise with lower class density as the number of pixels available is decreased. For the ImageNet generalization experiments, inputs are resized to 224×224 with random cropping. In the MedMNIST, we perform a random crop followed by a resize to 224×224. We set the time constant $\tau = 1$, which controls the rate of decay. We unroll the feedback loop for a fixed number of steps $T = 5$, so that the final prediction, as described in equation 1, is based on $\mathbf{h}(5)$. All models are trained using stochastic gradient descent with cross entropy loss, with 0.1 label smoothing, and batch size 32. We use the one cycle learning rate scheduler Smith [2017] with initial learning rate of 0.05, maximum learning rate 1.0, and final learning rate of $5e-5$. This aggressive schedule enables superconvergence Smith et al. [2019], which helps expose instabilities in feedback dynamics.

5 RESULTS

We report results for two main settings, noise and number of examples per class, that address the robustness and generalization abilities, respectively. These are some of the traits that separate computer from human vision, the ability to generalize and robustness to noise. We believe that future model development has to look beyond data and model sizes, and directly address these issues.

5.1 Robustness

We evaluated the robustness of DFMs against input perturbations by adding Gaussian noise of increasing magnitude to the ImageNet and COCO-Stuff datasets, see Figure 3. In the COCO-Stuff, DFMs dominate the performance for all levels of noise. With no noise, DFMs surpass feedforward models by +1.4pp, with DFMs and feedforward claiming 0.047 and 0.032 mIoU, respectively. As noise increases the performance gap decreases. At $\sigma = 1.$, DFMs had 0.034, while feedforward 0.022 mIoU. In the ImageNet, across all noise levels, the feedback model consistently outperformed the feedforward baseline, with its performance curve serving as a strict upper bound. Without noise ($\sigma = 0$), the feedback model achieved an average top-1 accuracy of 41.83%, compared to 36.31% for the feedforward model, a notable +5.52pp improvement. At the highest noise level tested, accuracy decreased to 27.80% for feedback and 26.04% for feedforward. As expected, the performance gap narrows under extreme noise, since classification approaches random guessing and becomes equally difficult for all models. Despite this narrowing, the feedback model maintained higher top-1 and top-5 accuracy throughout. Notably, the

slope of the feedback degradation curve is more steep than that of the feedforward model, but DFMs dominated feedforward for all values indicating stronger robustness to noise. All reported values are averaged over 5 random seeds, and standard deviations were negligible, suggesting consistent behavior across runs. These results reinforce the claim that DFMs offer a more noise resilient alternative for object recognition tasks.

5.2 Generalization

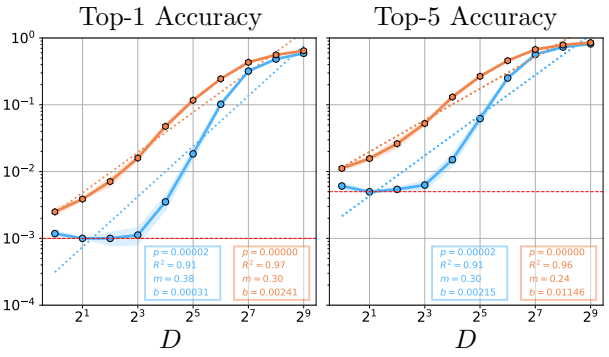

Figure 4: Performance of feedback (orange) versus feedforward (cyan). On the left, we plot the top-1 accuracy in the ImageNet dataset. On the right, we plot the top-5 accuracy. We tested 10 different values for D (see Section 4), which were $\in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$, all represent in base 2 in the plots. For each D we ran 5 seeds to plot the error. The horizontal red dotted line represents the performance of a random classifier. We also plot the linear regression (orange and cyan dotted lines) parameters of both feedback and feedforward settings.

We evaluated how well the models generalize under limited supervision by varying the number of training examples per class, D, on the ImageNet dataset. Across all conditions, the feedback model consistently outperformed the feedforward baseline in both top-1 and top-5 accuracy. As shown in Figure 4, both models exhibit a power law scaling behavior with respect to the number of training examples. While the slope of the power law is steeper for the feedforward model, indicating faster gains with more data, the feedback model consistently starts from a higher intercept. This suggests that feedback enables better generalization in extremely low-data regimes. The largest performance gap occurred at $D = 8$, where the DFM achieved a top-1 accuracy of 1.60%, compared to only 0.12% for feedforward. Even in the extreme case of $D = 1$, the DFM doubled the performance of the baseline, scoring 0.25% versus 0.12%. As the amount of data increases, the performance difference narrows: with $D = 512$, DFM reached 64.43% top-1 accuracy, while feedforward reached 59.00%. When trained on the full dataset ($D \approx 1280$), both models surpassed 72% top-1 accuracy. A similar trend holds for top-5 accuracy, with DFM maintaining consistent advantages across the full range of D. These results demonstrate that DFMs are particularly advantageous in low-data regimes, offering stronger generalization from very limited supervision.

5.3 Application in medical imaging

We evaluate DFMs on multiple *MedMNIST* datasets (Table 1) and on *RetinaMNIST-C*, a corruption benchmark derived from RetinaMNIST (Table 2). All models are trained on clean data ($\sigma = 0$) and evaluated under eight distribution shifts (blur, compression, pixelation, contrast, additive noise, motion blur, speckle, and brightness changes). Across nearly all corruptions, DFMs outperform a standard ResNet-50 backbone with and without additional augmentation techniques (brightness, contrast, saturation, and hue), yielding consistent AUROC gains. In particular, DFMs improve over the feedforward baseline by +4pp under *brightness-down* and by +1.5pp under *pixelate*, perturbations that substantially degrade the feedforward model. We compare against *Deep Equilibrium Models* (DEQs) Bai et al. [2019], which compute gradients at an equilibrium rather than through the entire trajectory; importantly, DEQs lack explicit top-down feedback. We also include *C-RBP* (Stable Recurrent Vision Models) instantiated on a ResNet-50 *backbone* Linsley et al. [2020], and *CORNet-R* Kubilius et al. [2018], a biologically motivated *local* recurrence baseline. While both DFMs and DEQs reach equilibrium states, the mechanisms differ: DFMs inject a global top-down signal during unrolling, whereas DEQs do not employ feedback. Overall, DFMs are competitive on most datasets, which we hypothesize stems from the additional top-down information shaping the gradient signal during unrolling. To isolate the role of feedback from mere iteration count, we evaluate a feedback-masked control (DFM-ResNet-50 w. $v = 0$) that zeros the top-down vector for $t > 0$ while matching T, FLOPs, and parameters. Iterating alone provides some benefit, but using feedback is advantageous on balance. This robustness under covariate shift supports the biological motivation for feedback: refining internal representations when inputs are ambiguous or degraded.

6 DISCUSSION

Ablation studies. We performed ablation studies on the role of exponential decay, orthogonality, and the

	Retina	Blood	Derma	Pneumonia	Breast
ResNet-50	0.707 ± 0.016	0.996 ± 0.001	0.919 ± 0.004	0.978 ± 0.007	0.771 ± 0.054
ResNet-50 w. augmentation	0.685 ± 0.019	$\mathbf{0.999 \pm 0.001}$	$\underline{0.927 \pm 0.008}$	$\underline{0.988 \pm 0.002}$	0.685 ± 0.140
C-RBP-ResNet-50 Linsley et al. [2020]	0.694 ± 0.049	0.998 ± 0.001	$\mathbf{0.936 \pm 0.008}$	$\mathbf{0.990 \pm 0.004}$	0.850 ± 0.019
CORNet-R Kubilius et al. [2018]	0.704 ± 0.011	0.998 ± 0.001	0.914 ± 0.005	0.984 ± 0.004	0.650 ± 0.084
DEQ-ResNet-50 Bai et al. [2019]	0.693 ± 0.073	0.995 ± 0.004	0.889 ± 0.058	0.980 ± 0.013	0.634 ± 0.140
DFM-ResNet-50 w. $\mathbf{v} = \mathbf{0}$	$\mathbf{0.725 \pm 0.014}$	$\mathbf{0.999 \pm 0.000}$	0.915 ± 0.023	0.980 ± 0.005	$\underline{0.862 \pm 0.025}$
DFM-ResNet-50 (Ours)	$\underline{0.722 \pm 0.012}$	$\mathbf{0.999 \pm 0.000}$	0.892 ± 0.020	0.983 ± 0.004	$\mathbf{0.881 \pm 0.034}$

Table 1: AUC of MedMNIST test set. All models were trained with $\sigma = 0$ and $D \times L = |\mathcal{D}|$. Results are averaged over 5 seeds.

Model	BD	DB	JC	P	CD	GN	MB	SN
ResNet-50	0.623	0.704	0.705	0.707	0.633	0.701	0.703	0.691
ResNet-50 w. augmentation	**0.678**	0.707	0.706	0.707	<u>0.665</u>	0.702	0.705	0.706
C-RBP-ResNet-50 Linsley et al. [2020]	0.633	**0.716**	**0.717**	<u>0.717</u>	0.631	<u>0.713</u>	**0.715**	**0.713**
CORNet-R Kubilius et al. [2018]	0.638	0.700	0.702	0.700	**0.680**	0.703	0.697	0.703
DEQ-ResNet-50 Bai et al. [2019]	0.605	0.708	0.711	<u>0.717</u>	0.623	**0.715**	0.708	<u>0.710</u>
DFM-ResNet-50 w. $\mathbf{v} = \mathbf{0}$	0.643	0.712	0.709	0.716	0.626	0.692	0.710	0.696
DFM-ResNet-50 (Ours)	<u>0.663</u>	<u>0.714</u>	<u>0.714</u>	**0.722**	0.648	0.691	**0.715**	0.707

Table 2: AUC of RetinaMNIST in the corrupted instances (MedMNIST-C). The various types of noise: brightness down (**BD**), defocus blur (**DB**), jpeg compression (**JC**), pixelate (**P**), contrast down (**CD**), gaussian noise (**GN**), motion blur (**MB**), and speckle noise (**SN**). All models were trained with $\sigma = 0$ and $D \times L = |\mathcal{D}|$. Results are averaged over 5 seeds.

convolution operation proposed. We used the ResNet-18 and trained the models for 70 epochs. These results are shown in Table 3. Removing exponential decay results in a notable drop in performance (from 0.425 to 0.401 top-1 accuracy), confirming its role in stabilizing the recurrent dynamics by ensuring convergence. Similarly, removing the orthogonalization step during training degrades accuracy (0.425 to 0.399), which aligns with our hypothesis that uncontrolled weight updates can lead to instabilities. Lastly, extending decay to the spatial domain via convolutional exponential decay further improves performance (from 0.425 to 0.432), suggesting that spatial dissipation enhances coherence in dense prediction tasks. These results highlight that all three components are beneficial and work in synergy to support robust learning in DFMs. While each mechanism contributes incrementally on its own, their combination yields the most stable and accurate behavior. We found no single component to be redundant, justifying their inclusion in our model.

In the COCO-Stuff experiments, we observed an effect of limited class density, while corrupting images with Gaussian noise. All our results validate the hypothesis that DFMs handle noise better and are able to generalize in few shot learning settings. As such, the superiority of DFMs in the segmentation task is due to the latter, as noise is analyzed while the class density is limited.

	✗	✓
exponential decay (equation 5)	0.401	**0.425**
orthogonality (equation 10)	0.399	**0.425**
convolution (equation 8)	0.425	**0.432**

Table 3: Ablation study performed with DFM on TinyImageNet (Top-1 accuracy). The entries without convolution, with orthogonalization, and with exponential decay are the same model.

Feedback is more robust than feedforward. Our experiments show that feedback models consistently outperform their feedforward counterparts in noisy settings, for both classification and segmentation tasks. This suggests that feedback mechanisms are especially well suited for environments with degraded input quality. Moreover, feedback models compute gradients through a recurrent unfolding of the internal state. Because the final representation $\mathbf{h}(T)$ is an accumulation over time, the backward pass implicitly aggregates gradients from all previous steps. The latter is also why we take the loss w.r.t. $\mathbf{y}(T)$. This

results in a gradient estimate that is more aligned with the true gradient of the underlying objective, as argued in energy based learning theory Scellier et al. [2019]. In this sense, feedback acts as a form of structured optimization, not just inference.

Feedback generalizes better with fewer examples. In low data regimes, feedback models show stronger generalization ability than feedforward models. On ImageNet, for example, feedback consistently dominates the performance curve across all numbers of training examples per class. Fitting a power law to the accuracy curves, with a statistical significant fit with p-values < 0.01 and $R^2 > 0.90$, confirms that both models follow the expected scaling behavior. While feedforward models exhibit a steeper slope, indicating more rapid performance gains as data increases, the feedback models achieve much higher intercepts. In the limit of very few examples, this translates to significantly better performance, sometimes more than $10\times$ higher than feedforward. This reinforces the idea that feedback enables better inductive biases for generalization, particularly in sample efficient learning.

Is it the iterations or feedback? Feedback performs multiple iterative passes, allowing it to refine internal representations over time. However, feedforward models rely on a single pass and lack the capacity to revise their interpretation of ambiguous stimuli. To test whether the advantage comes from the iterative nature or from the feedback, we define $\Delta_i = \text{DFM}_i - (\text{DFM w. } \mathbf{v} = 0)_i$, where DFM_i are the performances averaged over the seeds of DFM on the ith dataset, and test the null hypothesis $H_0 : \mathbb{E}[\Delta] \leq 0$. We found that DFM was statistically significant against DFM with $\mathbf{v} = 0$, using a one-sided paired t-test across datasets, with a p-value of 0.015 (confirming normality via Wilcoxon test). DFMs are competitive or superior on *Retina*, *Blood*, *Pneumonia*, and *Breast* (Table 1), and lead on several realistic corruptions in *RetinaMNIST-C*, notably brightness-down, pixelate, contrast-down, and motion blur (Table 2).

Global versus local recurrence. While prior recurrent architectures such as CORnet Kubilius et al. [2018] and C-RBP Linsley et al. [2020] rely on local recurrence, refining representations layer-by-layer or through Jacobian-constrained updates, our results suggest that global recurrence provides a distinct advantage. By injecting high-level predictions back to the input, DFMs enable top-down corrections that propagate throughout the entire network, rather than remaining confined to shallow feature refinements. Empirically, this yields consistent gains in robustness and generalization on medical imaging benchmarks, even when controlling for iterative depth (see table 2). In terms of complexity, DFMs are **not** substantially more expensive than other recurrent approaches: FLOPs, parameter counts, and batch processing times are comparable to CORnet C-RBP (see supplementary material), while remaining within the same order of magnitude as DEQ-based models. Thus, although DFMs incur higher cost than a single-pass feedforward baseline, they offer a favorable trade-off against alternative recurrent designs, achieving stronger robustness to corruptions and distribution shifts, which is particularly valuable in domains such as medical imaging where reliability outweighs raw throughput.

7 RELATED WORK

Biologically inspired feedback and recurrence. Although feedback as defined in this work, explicit high level to low level signal flow, has not been widely explored, there exists a rich literature on models incorporating recurrence. Classical examples include Hopfield networks Hopfield [1982], PredNet Lotter et al. [2017], NeuralODEs Chen et al. [2018], DEQs Bai et al. [2019], and Feasibility based Fixed Point Models Heaton et al. [2021]. These approaches typically repeat layers or update a latent state through iterative processes (or differential equation solvers), but do not integrate top down feedback across hierarchical levels. Notably, DEQs perform fixed point iteration until the system reaches equilibrium and then proceeds to compute the gradient at that point. We believe that this process inevitably loses information of the trajectory made during the forward process and its performance is hurt as a consequence. More biologically grounded efforts such as CORnet Kubilius et al. [2018] and predictive reconstruction frameworks Alamia et al. [2023], Fein-Ashley [2024] propose layer wise recurrence, where each layer predicts its previous and next layers' activations. However, these models rely on local feedback between adjacent layers and lack long range, hierarchical feedback that modulates early representations using high level beliefs. Our approach differs significantly: the entire neural network acts as a recurrent cell, where a global feedback vector is fed back to the input over time. This formulation poses unique challenges, particularly for training stability (see Section 3.2), and requires dedicated mechanisms such as exponential decay and orthogonality enforcement.

Zamir et al. [2017] are an important early example of feedback in deep vision models, but at the same time, our approach differs in several fundamental ways. First, as opposed to local feedback, where the output of ResNet block is fed back to itself, we perform feedback at the global architecture level. The equivalents of Zamir et al. [2017] are the ConvLSTM model used in

Linsley et al. [2020] and the gating mechanism in Lotter et al. [2017]. Importantly, DFMs use an ungated feedback pathway, that deliberately avoids gates because they can block information that is necessary for iterative refinement. DFMs propagate all high-level features back to the early layers, aligning more closely with predictive-coding-style refinement and avoiding hand-designed gating dynamics. This yields architectural and behavioral differences beyond simply adding recurrence.

Stability and robustness in recurrent networks. Several works have highlighted the benefits of recurrence for robustness, particularly under noisy or adversarial settings Spoerer et al. [2017]. Stability in such models is often enforced via Jacobian regularization Linsley et al. [2020] or by limiting the number of unrolled iterations Wang et al. [2019]. In contrast, our model achieves stable convergence through architectural constraints, notably using matrix exponentiation with negative real eigenvalues, without resorting to truncation or gradient clipping.

Connections to recurrent backpropagation. Our method also shares similarities with recurrent backpropagation (RBP), where gradients are computed at equilibrium by backpropagating through fixed point dynamics Pineda [1987], Almeida [1990], Liao et al. [2018]. In RBP, gradient convergence depends on the spectral properties of the underlying dynamics, typically requiring matrices with negative real eigenvalues. In our case, we mirror this principle during the forward pass by constructing a computation graph using a matrix P with eigenvalues $\Sigma < 0$ (see Equation 5), ensuring fast convergence to a stable state. Unlike classical RBP, which is designed for constant memory usage, our models store the full trajectory to enable learning through automatic differentiation, at the cost of higher memory usage. This enables training of complex systems that would otherwise be intractable with Jacobian based regularization.

8 CONCLUSION

It is hypothesized that the brain performs decision making by solving a differential equation over time Busemeyer et al. [1993]. However, most state of the art neural networks used in object recognition and segmentation are static, lacking biologically plausible feedback mechanisms. This limitation makes them brittle in domains such as medical imaging, where data is noisy or scarce. In this work, we introduced DFMs, a novel class of neural architectures that integrate top down feedback into the computational graph. Unlike conventional feedforward models, DFMs update internal representations through a recurrent process that refines predictions over time. This leads to more robust gradients, improved stability, and better generalization, especially under noise and limited data conditions. Our experiments demonstrate that DFMs consistently outperform feedforward baselines in both robustness and data efficiency, confirming the benefits of hierarchical feedback dynamics and making DFMs a natural candidate in medical imaging candidates.

Acknowledgments

This work was supported by the Center for Responsible AI project with reference C628696807-00454142, financed by the Recovery and Resilience Plan, and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025).

This research was supported by the Slovenian Research and Innovation Agency (ARIS) under the Artificial Intelligence and Intelligent Systems Programme (Grant No. P2–0209).

This publication is co-funded the European Union's Horizon Europe research and innovation program under the Marie Sklodowska-Curie COFUND Postdoctoral Programme grant agreement No.101081355-SMASH and by the Republic of Slovenia and the European Union from the European Regional Development Fund.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Andrea Alamia et al. On the role of feedback in image recognition under noise and adversarial attacks: A predictive coding perspective. *Neural Networks*, 2023.

Luis B Almeida. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In *Artificial neural networks: Concept learning*. 1990.

Shaojie Bai et al. Deep equilibrium models. *NeurIPS*, 2019.

Gary Bécigneul et al. Riemannian adaptive optimization methods. *arXiv*, 2018.

Rafal Bogacz. A tutorial on the free-energy framework

for modelling perception and learning. *Journal of Mathematical Psychology*, 2017.

Jerome R Busemeyer et al. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 1993.

Ricky TQ Chen et al. Neural ordinary differential equations. *Neurips*, 2018.

Andy Clark. *The experience machine: How our minds predict and shape reality*. Random House, 2024.

Joel Dapello et al. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *NeurIPS*, 2020.

Jia Deng et al. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Francesco Di Salvo et al. MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. *arXiv*, 2024.

Jacob Fein-Ashley. Contextual backpropagation loops: Amplifying deep reasoning with iterative top-down feedback. *arXiv*, 2024.

Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 2005.

Lore Goetschalckx et al. Computing a human-like reaction time metric from stable recurrent vision models. *NeurIPS*, 2023.

Claudius Gros. Neural networks with transient state dynamics. *New Journal of Physics*, 2007.

Howard Heaton et al. Feasibility-based fixed point networks. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 1982.

Jonas Kubilius et al. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 2018.

Yann Le et al. Tiny ImageNet visual recognition challenge. *CS 231N*, 2015.

Steven J Leon et al. Gram-Schmidt orthogonalization: 100 years and more. *Numerical Linear Algebra with Applications*, 2013.

Renjie Liao et al. Reviving and improving recurrent back-propagation. In *ICML*, 2018.

Tsung-Yi Lin et al. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

D Linsley et al. Stable and expressive recurrent vision models. *arXiv*, 2020.

William Lotter et al. Deep predictive coding networks for video prediction and unsupervised learning. *ICML*, 2017.

Fernando Pineda. Generalization of back propagation to recurrent and higher order neural networks. In *NeurIPS*, 1987.

Rishi Rajalingham et al. Recurrent neural networks with explicit representation of dynamic latent variables can mimic behavioral patterns in a physical inference task. *Nature Communications*, 2022.

Rajesh PN Rao et al. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 1999.

Benjamin Scellier et al. Equivalence of equilibrium propagation and recurrent backpropagation. *Neural Computation*, 2019.

Leslie N Smith. Cyclical learning rates for training neural networks. In *WACV*, 2017.

Leslie N Smith et al. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, 2019.

Courtney J Spoerer et al. Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, 2017.

Wei Wang et al. Recurrent U-Net for resource-constrained segmentation. In *ICCV*, 2019.

Jiancheng Yang et al. MedMNIST V2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 2023.

Amir R Zamir et al. Feedback networks. In *CVPR*, 2017.

Reproducibility checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**, the methods section has a clear description of the framework. We also include an algorithm in the supplementary material
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**, we provide complexity analysis of parameters, FLOPs, and batch processing time in the supplementary material.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes**.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**
   (b) Complete proofs of all theoretical results. **No**
   (c) Clear explanations of any assumptions. **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**, all of these details are described in the experimental setting section.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes**
   (b) The license information of the assets, if applicable. **Not applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not applicable**
   (d) Information about consent from data providers/curators. **Not applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not applicable**

A Feedback neurons

The dimension N depends on the architecture one is using. If it is the ResNet-18, $N = 512$, if it is the ResNet-50, $N = 2048$. These are simply the neurons that are then forwarded to the last classification layer. As for B, the neurons that will be fed back to the input of the network, we chose to set them to L, the number of classes in the problem. In early experiments, we found that with $B < L$, the neural network was not able to learn, i.e. the feedback signal was too compressed to carry meaningful class-specific information back to early layers. Since $B > L$ would impact memory we refrained from increasing it.

We hypothesize that $B = L$ represents a minimal viable feedback mechanism, sufficient to encode class-level predictions but potentially limiting for richer contextual modulation. Our analysis (see Figure 5) shows that there is a clear peak with B around L. Nevertheless, it remains an open question on how sparser representations would influence the model.

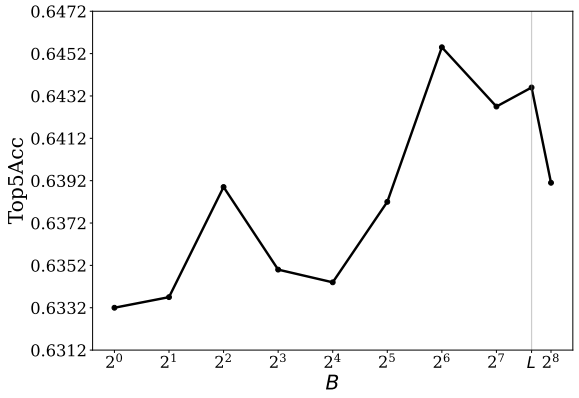

Figure 5: Impact of different values for B. The plot reports the top-5 accuracy on the TinyImageNet dataset.