

# ACCELERATING TRAINING WITH NEURON INTERACTION AND NOWCASTING NETWORKS

Boris Knyazev<sup>1</sup> Abhinav Moudgil<sup>2,4</sup> Guillaume Lajoie<sup>3,4</sup>

Eugene Belilovsky<sup>2,4</sup> Simon Lacoste-Julien<sup>3,4</sup>

<sup>1</sup>Samsung - SAIT AI Lab, Montreal <sup>2</sup>Concordia University <sup>3</sup>Université de Montréal <sup>4</sup>Mila

b.knyazev@samsung.com

<https://github.com/SamsungSAILMontreal/nino>

## ABSTRACT

Neural network training can be accelerated when a learnable update rule is used *in lieu of* classic adaptive optimizers (e.g. Adam). However, learnable update rules can be costly and unstable to train and use. A simpler recently proposed approach to accelerate training is to use Adam for most of the optimization steps and periodically, *only every few steps*, nowcast (predict future) parameters. We improve this approach by Neuron interaction and Nowcasting (NiNo) networks. NiNo leverages neuron connectivity and graph neural networks to more accurately nowcast parameters by learning in a supervised way from a set of training trajectories over multiple tasks. We show that in some networks, such as Transformers, neuron connectivity is non-trivial. By accurately modeling neuron connectivity, we allow NiNo to accelerate Adam training by up to 50% in vision and language tasks.

## 1 INTRODUCTION

Modern deep learning models, such as large language models (Touvron et al., 2023), are trained using classic adaptive optimizers, such as Adam<sup>1</sup> (Kingma & Ba, 2014; Loshchilov & Hutter, 2017). These optimizers update neural network parameters  $\Theta \in \mathbb{R}^n$  using gradient descent at step  $t$  as  $\Theta_{t+1} = \Theta_t - \Delta\Theta_t$ ,  $\forall i \in [1, n]$ , where  $\Delta\Theta_t$  is the update computed analytically based on the history of parameter values, gradients and the learning rate. Recently, Jang et al. (2023); Sinha et al. (2017) showed that parameters  $\Theta$  follow a predictable trend so that optimization can be accelerated by **nowcasting** (predicting future) parameters using a learnable function  $f^\phi$ :  $\hat{\Theta}_{t+k}^i = \Theta_t^i + f^\phi(\Theta_t^i, \Theta_{t-1}^i, \Theta_{t-2}^i, \dots)$ , where  $k \gg 1$  is the future horizon. More popular learnable approaches to speed up optimization, such as “learning to optimize” (L2O), are recurrently applied at every step  $t$  (Andrychowicz et al., 2016; Metz et al., 2022). Compared to L2O, the parameter nowcaster  $f^\phi$  is applied **very rarely** reducing its overhead, e.g. a base optimizer such as Adam is run for 1k steps followed by the prediction step (Fig. 1). Moreover, training such  $f^\phi$  is simpler than L2O, since a supervised loss can be used instead of a more challenging meta-learning loss with recurrent inner steps. However, akin to Adam,  $f^\phi$  from prior works does not directly leverage the structural information of  $\Theta$ , such as connectivity between neurons and layers. This structure has been shown critical for many parameter representation tasks, such as property prediction (Navon et al., 2023; Zhou et al., 2023; Kofinas et al., 2024).

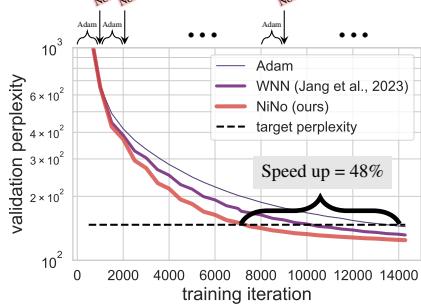


Figure 1: Adam without and with nowcasting using our NiNo vs WNN (Jang et al., 2023) on a language task that NiNo and WNN have not seen during their training.

<sup>1</sup>We use Adam throughout the paper, but our discussion and methods are in principle applicable to any optimizers that produce a trajectory of parameters, including SGD with/without momentum, AdamW, Adagrad, etc.

In this work, we propose **Neuron Interaction and Nowcasting (NiNO)** networks making better predictions of future parameters to **accelerate training with a base optimizer such as Adam** and make the following contributions:

1. We introduce NiNO, which directly uses the structure of neural network parameters by leveraging recently proposed *neural graphs* (graphs of neurons) (Kofinas et al., 2024; Lim et al., 2023).
2. We improve neural graphs for transformers by more accurately modeling the *neuron permutation symmetry* of multi-head self-attention. Our neural graphs combined with graph neural networks create a strong inductive bias to make rare but accurate predictions.
3. NiNO is conditioned on the future horizon  $k$  to allow for prediction in the near and far future without retraining it, facilitating its usage in diverse tasks and at different optimization stages.
4. We demonstrate that NiNO accelerates training with Adam for convnets and transformers reducing the number of steps to achieve the target performance of Adam by up to 50%. We release our source code and models at <https://github.com/SamsungSAILMontreal/nino>.

## 2 RELATED WORK

**Learning to optimize (L2O).** The L2O literature has offered many approaches to learn a neural network (*optimizer*) that optimizes the parameters of other neural nets (*optimizees*) (Andrychowicz et al., 2016; Chen et al., 2021; 2022; Amos, 2022). Among them, Safeguarded L2O (Heaton et al., 2023; Premont-Schwarz et al., 2022) is most related to ours as it switches between an L2O optimizer and SGD/Adam. While Safeguarded L2O alleviates the *meta-generalization* challenge (Thérien et al., 2024), training an L2O optimizer remains costly and unstable due to the use of meta learning methods and its long inner loop unrolls required at each meta-training iteration (Metz et al., 2019; 2022). Moreover, overheads of L2O add up at each iteration making it more computationally intensive than Adam. In contrast, our approach follows *weight nowcaster networks* (WNNs) (Jang et al., 2023), where the nowcaster model is applied very rarely (e.g. once per 1k steps of Adam) making the total overhead negligible, yet still speeding up training significantly.

**Parameter prediction and generation.** This area has grown significantly recently, primarily aiming at reducing computational costs of training neural nets (Peebles et al., 2022; Ashkenazi et al., 2022; Schürholz et al., 2022; 2024; Knyazev et al., 2021b; 2023; Zhou et al., 2024c; Soro et al., 2024; Wang et al., 2024). Most related to us are IntrospectionMLP (Sinha et al., 2017) and WNNs (Jang et al., 2023) serving the basis of our work. These methods train simple MLPs to periodically (e.g. every few epochs of Adam) predict future parameter values of a neural net given its past parameters (Section 3.1). However, their MLPs predict parameter coordinate-wise (for each parameter independently) similar to optimization methods without leveraging the structure of neural networks. Moreover, their MLPs predict parameters only for a predefined future horizon  $k$ , whereas different tasks and different optimization stages can have different parameter evolution trends (Morchdi et al., 2022; Guille-Escuret et al., 2023; Lange et al., 2023). We address these shortcomings by conditioning the prediction on  $k$ .

**Representation learning of neural network parameters.** This area has also developed fast recently (Navon et al., 2023; Schürholz et al., 2021; 2024; Ashkenazi et al., 2022; Andreis et al., 2023; Zhou et al., 2023; 2024b;a). One of the main goals in these works is to model *neuron permutation symmetry* – the property of neural networks that if the neurons in one layer are permuted in the same way as the neurons in the next layer, the neural network preserves its function (Hecht-Nielsen, 1990). Accurate modeling of this symmetry allows for better estimation of network properties or encoding implicit neural representations. To model neuron permutation symmetry, Kofinas et al. (2024); Lim et al. (2023) proposed *neural graphs* (graphs of neurons) enabling the usage of graph neural networks (GNNs) (Kipf & Welling, 2016; Corso et al., 2020). Remarkably, as GNNs can digest graphs of any size and connectivity, the synergy of neural graphs and GNNs enables processing diverse parameters from different architectures and tasks using a *single* GNN. We leverage both the neural graphs and GNNs to learn a single NiNo model that can accelerate optimization in diverse tasks.

**Dynamic models.** Our work is also related to dynamic interactive models, where message passing networks and GNNs are used to learn the interaction within relational temporal networks (Trivedi et al., 2019; Knyazev et al., 2021a) and physical systems (Kipf et al., 2018; Sanchez-Gonzalez et al., 2018; 2020; Zambaldi et al., 2019). However, developing a strong parameter prediction model requires many specific design choices and careful modeling of neuron permutation symmetry, motivating our work.

### 3 BACKGROUND

We consider a neural network parameterized by  $\Theta_t \in \mathbb{R}^n$  trained for  $t$  iterations with Adam (or another optimizer). Our goal is to accelerate optimization by  $k \gg 1$  using a meta-network  $f_k^\phi$ , parameterized by  $\phi$ , predicting future parameters  $\hat{\Theta}_{t+k}$  given its past values:  $\hat{\Theta}_{t+k} = \Theta_t + f_k^\phi(\Theta_t, \Theta_{t-1}, \Theta_{t-2}, \dots)$ .

#### 3.1 WEIGHT NOWCASTER NETWORKS (WNNs)

WNNs (Jang et al., 2023) model  $f_k^\phi$  as an MLP predicting the delta (update difference) of the  $i$ -th parameter:  $\Delta\hat{\Theta}_{\tau+k}^i = f_k^\phi(\tilde{\Theta}_\tau^i, \tilde{\Theta}_{\tau-1}^i, \dots, \tilde{\Theta}_{\tau_c}^i)$ , where  $\tau_c = \tau - c + 1$  and  $c$  is the context length. Here,  $\tau, \tau-1, \dots$  are **epoch indices** and  $\tilde{\Theta}$  are the parameters scaled based on the range of values in  $\Theta_\tau^i, \dots, \Theta_{\tau_c}^i$  (see details in Section A.1). The predicted parameter value is obtained by unscaleing the predicted delta:  $\hat{\Theta}_{\tau+k}^i = \Theta_\tau^i + \text{unscale}(\Delta\hat{\Theta}_{\tau+k}^i)$ . WNNs are trained in a supervised way by collecting a training dataset of parameter trajectories  $\{[\Theta_1, \Theta_2, \dots]\}_1^C$  obtained with Adam and applying the  $l_1$ -loss:

$$\text{argmin}_\phi \|\Delta\tilde{\Theta}_{\tau+k} - \Delta\hat{\Theta}_{\tau+k}\|_1, \quad (1)$$

where  $\Delta\tilde{\Theta}_{\tau+k}$  are the scaled target parameter deltas. Sinha et al. (2017) proposed the approach originally, however WNNs are applied **periodically**, once per every  $c$  epochs of a base optimizer (Adam) and  $k = c$ , which better accelerates optimization. For example, to use a trained  $f_k^\phi$  on a new task, first Adam is run for  $c$  epochs ( $c = 5$  by default) after which  $f_k^\phi$  is applied to predict future (10-th epoch) parameters, after which the procedure repeats (the base optimizer is continued for another 5 epochs followed by parameter prediction) until convergence. So the WNN is applied very rarely, e.g. in the case of 200 Adam steps per epoch, the WNN is applied only once per 1,000 steps.

#### 3.2 (NAIVE) NEURAL GRAPH OF TRANSFORMERS

WNNs apply an MLP for parameter  $\Theta^i$  given only its past values. Such an MLP is inherently limited, since future parameter values depend on many factors, including connectivity of neurons. Modeling neuron connectivity in a way that generalizes to arbitrary network structures to make parameter prediction as general as possible is challenging. Simple parameter representations, e.g. flattening of parameters into a vector (Schürholt et al., 2021), are not general and do not model *neuron permutation symmetry* (the neurons can be permuted without changing the network output). Recently, Kofinas et al. (2024); Lim et al. (2023) proposed to represent  $\Theta$  using a *neural graph*  $\mathcal{G}(\Theta) = (\mathbf{V}, \mathbf{E})$  with node features  $\mathbf{V} \in \mathbb{R}^{|\mathbf{V}| \times d_V}$  and edge features  $\mathbf{E} \in \mathbb{R}^{|\mathbf{V}| \times |\mathbf{V}| \times d_E}$ , where  $d_V, d_E$  are the node and edge feature dimensionalities, respectively. For example, equation 2 shows the edge features for an  $L$ -layer MLP (we ignore biases for simplicity):  $\Theta = \{\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L)}\}$ , where  $\mathbf{W}^{(l)} \in \mathbb{R}^{d_{l-1} \times d_l}$  are the weights for  $l \in [1, L]$ .

Neural graphs can theoretically represent the parameters of any neural network taking into account neuron permutation symmetry as the rows and columns of weight matrices correspond to the nodes in the neural graph. This way, neural graphs impose a strong inductive bias similarly to using convolution for images. So a model operating on neural graphs, such as a graph neural network (GNN) and our NiNo model, can be used in diverse tasks and should be able to learn parameter prediction rules that generalize better with fewer samples than, for example, MLPs. However, to fully leverage the neural graphs in practice, we need to accurately construct them, which is not trivial for such networks as transformers, motivating us in this work.

**Transformers.** For input  $\mathbf{x} \in \mathbb{R}^{N \times d}$ , where  $N$  is a sequence length, a multi-head self-attention (MSA) layer (Vaswani et al., 2017) for each head  $h \in [1, H]$  projects  $\mathbf{x}$  using  $\mathbf{W}_h^q, \mathbf{W}_h^k, \mathbf{W}_h^v$  followed by self-attention and projection using another weight matrix  $\mathbf{W}_h^o$  (see Section A.4 for details):

$$\mathbf{A}_h = \frac{(\mathbf{x}\mathbf{W}_h^q)(\mathbf{x}\mathbf{W}_h^k)^T}{\sqrt{d}} \quad \text{and} \quad \mathbf{y}_h = \text{softmax}(\mathbf{A}_h)(\mathbf{x}\mathbf{W}_h^v) \quad \text{and} \quad \text{MSA}(\mathbf{x}) = \sum_{h=1}^H \mathbf{y}_h \mathbf{W}_h^o. \quad (3)$$

$$\mathbf{E} = \left[ \begin{array}{ccccccccc} d_0 & d_1 & d_2 & \cdots & d_L & & & & \\ \mathbf{W}^{(1)} & & & & & & & & \\ & \mathbf{W}^{(2)} & & & & & & & \\ & & \ddots & & & & & & \\ & & & \mathbf{0} & & & & & \\ & & & & \ddots & & & & \\ & & & & & \mathbf{0} & & & \\ & & & & & & \ddots & & \\ & & & & & & & \mathbf{W}^{(L)} & \\ & & & & & & & & \vdots \\ & & & & & & & & d_0 \\ & & & & & & & & d_1 \\ & & & & & & & & d_2 \\ & & & & & & & & \vdots \\ & & & & & & & & d_L \end{array} \right] \quad (2)$$

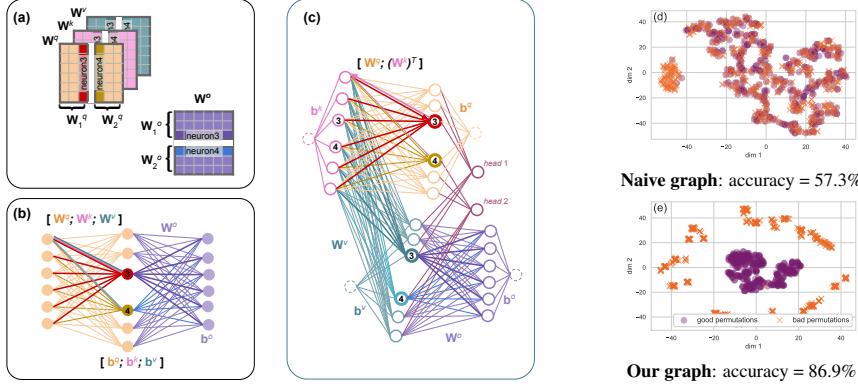


Figure 2: (a) MSA weights with  $d = 6$  and  $H = 2$  heads; (b) its naive (Kofinas et al., 2024) and (c) our neural graph with color-coded edge/parameter types. (d-e) Neural graph embeddings obtained using a GNN and projected to 2d using tSNE for 1k **good** and **bad** permutations of the MSA weights: (d) naive graph, (e) our graph. Accuracy is computed by fitting logistic regression for neural graph embeddings with labels being if the MSA output changes or not as described in Section A.6.

To construct neural graphs of MSA, Kofinas et al. (2024); Lim et al. (2023) use 3-dimensional edge features  $\mathbf{e}_{ij} = \left( (\mathbf{W}^q)_{ij}, (\mathbf{W}^k)_{ij}, (\mathbf{W}^v)_{ij} \right)$ , assuming that the multi-head case is automatically handled by the neural graph. However, as we show in Section 4.1, this neural graph does not model the neuron permutation symmetry correctly, therefore we refer to it as a *naive neural graph* (Fig. 2a,b).

## 4 METHODS

In this section, we describe the details of our neural graphs for an MSA layer of transformers and the Neuron Interaction and Nowcasting (NiNo) networks operating on such graphs (see a model overview in Fig. 4 and a graph example in Fig. 7). Following WNNs, NiNo is also applied very rarely during the optimization process (Section 3.1). However, NiNo leverages the representation power of neural graphs and GNNs to speed up optimization more significantly than WNNs, as we show in Section 5.

### 4.1 NEURAL GRAPH OF TRANSFORMERS

To construct neural graphs that accurately model neuron permutation symmetry in MSA, we (1) restrict neuron permutations across heads and (2) relax permutations of weights  $\mathbf{W}^q$ ,  $\mathbf{W}^k$  (Fig. 2).

**(1) Restricting permutations across heads.** We first observe that splitting computations into  $H$  parallel heads (equation 3) breaks neuron permutation symmetry across the heads, so shuffling neurons across the heads may change the overall output of  $\text{MSA}(\mathbf{x})$ . Consider, for example, an MSA layer with  $d = 6$  and  $H = 2$  (Fig. 2a) and the dot product between outputs  $\mathbf{x}_1^q$  and  $(\mathbf{x}_1^k)^T$  obtained after projecting  $\mathbf{x}$  using the weights of the first head  $\mathbf{W}_1^q$  and  $\mathbf{W}_1^k$ :  $\mathbf{x}_1^q(\mathbf{x}_1^k)^T = [\mathbf{x}_1^q, \mathbf{x}_2^q, \mathbf{x}_3^q][\mathbf{x}_1^k, \mathbf{x}_2^k, \mathbf{x}_3^k]^T$ . The result of this dot product does not change when permuting neurons in both  $\mathbf{W}_1^q$  and  $\mathbf{W}_1^k$  using the same  $\pi$ , since it results in permuting the outputs:  $\pi(\mathbf{x}_1^q)\pi(\mathbf{x}_1^k)^T = \mathbf{x}_1^q(\mathbf{x}_1^k)^T$ . However, consider shuffling neurons across head 1 and 2 in  $\mathbf{W}^q$  and  $\mathbf{W}^k$ , e.g. neurons 3 and 4 (highlighted in different colors in Fig. 2a) resulting in:  $[\mathbf{x}_1^q, \mathbf{x}_2^q, \mathbf{x}_4^q][\mathbf{x}_1^k, \mathbf{x}_2^k, \mathbf{x}_4^k]^T$ . Now the dot product no longer equals  $\mathbf{x}_1^q(\mathbf{x}_1^k)^T$  unless  $\mathbf{x}_4^q \mathbf{x}_4^k = \mathbf{x}_3^q \mathbf{x}_3^k$ . So self-attention matrix  $\mathbf{A}_1$  for head 1 and  $\text{MSA}(\mathbf{x})$  can change.

To take into account this restriction, we allow for neuron permutations **only within a head** by adding a separate node for each head that connects the appropriate neurons so that permuting neurons 3 and 4 from our example changes the graph while permuting within a head does not change it (Fig. 2c).

**(2) Relaxing permutations for  $\mathbf{W}^q$ ,  $\mathbf{W}^k$ .** We also observe that neurons in  $\mathbf{W}_h^q$  and  $\mathbf{W}_h^k$  can be permuted in a different way ( $\pi'$ ) from  $\mathbf{W}_h^v$  and  $\mathbf{W}_h^o$  without changing the MSA output:

$$(\mathbf{x}\mathbf{W}_h^q)(\mathbf{x}\mathbf{W}_h^k)^T = (\mathbf{x}\pi'_c(\mathbf{W}_h^q))(\mathbf{x}\pi'_c(\mathbf{W}_h^k))^T \quad \text{and} \quad (\mathbf{x}\mathbf{W}_h^v)\mathbf{W}_h^o = (\mathbf{x}\pi_c(\mathbf{W}_h^v))\pi_r(\mathbf{W}_h^o), \quad (4)$$

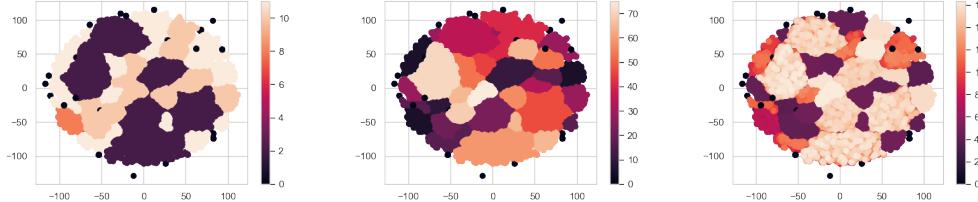


Figure 3: Laplacian Positional Encoding (LPE) of our neural graph’s nodes color-coded by the layer type (left), layer index (middle) and neuron index (right) for a transformer with 6 layers and 384 hidden units. We show tSNE projections from the 8-dimensional LPE to 2d.

where  $\pi_c$  and  $\pi_r$  denote permutation  $\pi$  of the columns and rows in the matrix respectively. In the naive graph the neurons in  $\mathbf{W}_h^q, \mathbf{W}_h^k, \mathbf{W}_h^v, \mathbf{W}_h^o$  are always permuted in the same way ( $\pi_c = \pi'_c$ ) making the neural graph unnecessary restrictive (Fig. 2b). To address this issue, we keep  $\mathbf{W}_h^q, \mathbf{W}_h^k, \mathbf{W}_h^v$  as **separate 1-dimensional edge features** in a neural graph instead of stacking them as 3-dimensional edge features (Fig. 2c). Since  $\mathbf{W}_h^q$  and  $\mathbf{W}_h^k$  share the input and output neurons while  $\mathbf{W}_h^k$  is transposed in equation 3, it is important to preserve the edge direction, so that for example  $e_{3,4}$  corresponds to the weights of  $\mathbf{W}_h^q$  while  $e_{4,3}$  corresponds to the weights of  $\mathbf{W}_h^k$ .

**Edge features.** In our neural graph, all the model parameters  $\Theta$  and auxiliary connections  $\Theta'$ , such as residual and head connections introduced in Section 4.1, are represented using the edge features  $\mathbf{E} \in \mathbb{Z}^{|\mathbf{V}| \times |\mathbf{V}| \times d_E}$  so that  $\|\mathbf{E}\|_0 = |\Theta| + |\Theta'|$  (Fig. 2c). To differentiate between  $\Theta$  and  $\Theta'$ , we associate an integer edge type with each edge:  $\mathbf{E}^{\text{type}} \in \mathbb{Z}^{|\mathbf{V}| \times |\mathbf{V}| \times 1}$ , so our neural graph  $\mathcal{G}(\Theta) = (\mathbf{V}, \mathbf{E}^{\text{type}}, \mathbf{E})$ .

**Node features.** For the node features we use the Laplacian Positional Encoding (LPE) (Belkin & Niyogi, 2003; Dahl et al., 2023) that allow graph neural networks to capture structural information more easily. For example the LPE implicitly embeds nodes of the same layer and same type close to each other even though this information is not explicitly provided (Fig. 3). To compute the LPE, we use *unweighted edges*, transform the graph to an *undirected* one and extract 8 smallest non-trivial eigenvectors, so our node features are  $\mathbf{V}^{\text{lpe}} \in \mathbb{R}^{|\mathbf{V}| \times 8}$ . In addition, for transformer’s word embedding layers, we found it beneficial to leverage a positional feature  $\mathbf{V}^w \in \mathbb{Z}^{|\mathbf{V}| \times 1}$ , since these layers often have the same size with ordered neurons. For other layers, this feature is set to zero.

**Sequence of parameters as a neural graph.** We consider a history of  $c$  past parameter vectors following WNNs (Section 3.1):  $\Theta_{\tau:\tau_c} = [\Theta_\tau, \Theta_{\tau-1}, \dots, \Theta_{\tau_c}] \in \mathbb{R}^{n \times c}$ , where  $\tau_c = \tau - c + 1$  as in Section 3.1. Since each parameter vector in  $\Theta_{\tau:\tau_c}$  represents the same neural network structure, we transform the entire  $\Theta_{\tau:\tau_c}$  into a single neural graph  $\mathcal{G}(\Theta_{\tau:\tau_c}) = (\mathbf{V}^{\text{lpe}}, \mathbf{V}^w, \mathbf{E}^{\text{type}}, \mathcal{E}_{\tau:\tau_c})$ , where  $\mathcal{E}_{\tau:\tau_c}$  is obtained by stacking edge features:  $\mathcal{E}_{\tau:\tau_c} = [\mathbf{E}_\tau, \mathbf{E}_{\tau-1}, \dots, \mathbf{E}_{\tau_c}] \in \mathbb{R}^{|\mathbf{V}| \times |\mathbf{V}| \times d_E \times c}$ , whereas node features  $\mathbf{V}^{\text{lpe}}$ ,  $\mathbf{V}^w$  and edge types  $\mathbf{E}^{\text{type}}$  do not change over time.

## 4.2 NEURON INTERACTION AND NOWCASTING NETWORKS (NiNo)

Given an input neural graph  $\mathcal{G}(\Theta_{\tau:\tau_c})$ , our NiNo model processes it using layerwise scaling layer, node and edge embedding layers and GNN layers with a hidden size  $D$ . NiNo then predicts edge features for multiple steps in the future  $\hat{\mathcal{E}}_{1:K}$  that are mapped back to the parameter space  $[\Delta\hat{\theta}_{\tau+1}, \dots, \Delta\hat{\theta}_{\tau+K}]$  using the inverse neural graph construction  $\Delta\hat{\theta}_{\tau+k} = \mathcal{G}^{-1}(\hat{\mathcal{E}}_k)$  (Fig. 4).

**Layerwise scaling.** Since parameter values can vary in scale significantly across architectures and tasks, it is important to scale them appropriately. Compared to WNNs (Jang et al., 2023) scaling each parameter independently using min-max, we extend a layerwise scaling (Schürholt et al., 2022) to a sequence of parameters. Specifically, given weights  $\mathbf{W}_{\tau:\tau_c}^{(l)} \in \mathbb{R}^{d_{l-1} \times d_l \times c}$  of the  $l$ -th layer of the input  $\mathcal{E}_{\tau:\tau_c}$ , we obtain scaled weights as  $\tilde{\mathbf{W}}_{\tau:\tau_c}^{(l)} = (\mathbf{W}_{\tau:\tau_c}^{(l)} - \mu_\tau^{(l)}) / \sigma_\tau^{(l)}$ , where  $\mu_\tau^{(l)}, \sigma_\tau^{(l)}$  are the scalar mean and standard deviation of  $\mathbf{W}_{\tau:\tau_c}^{(l)}$ . After repeating this step  $\forall l \in [1, L]$ , we obtain scaled  $\tilde{\mathcal{E}}_{\tau:\tau_c}$ .

**Embedding layers and GNN.** We use linear  $\phi_{\text{lpe}}, \phi_e$  and embedding  $\phi_w, \phi_{\text{type}}$  layers to project node and edge features to the  $D$ -dimensional space followed by  $M$  GNN layers and the output layer  $\phi_{\text{DMS}}$ :

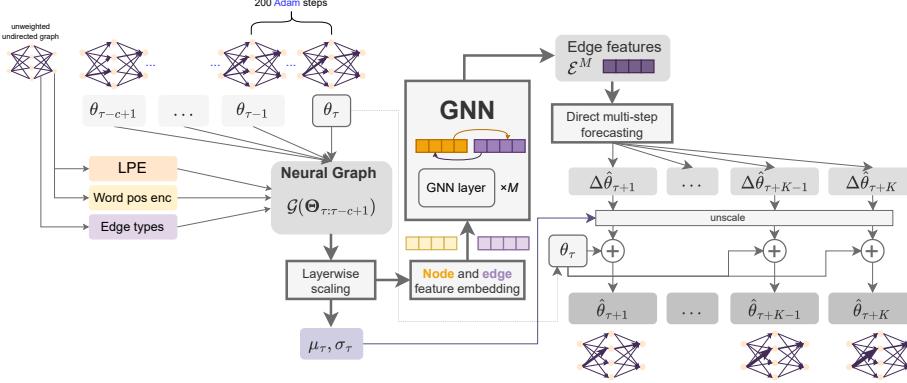


Figure 4: Our Neuron interaction and Nowcasting (NiNo) model. We feed  $c$  past parameter states as input and predict future  $K$  states leveraging our improved neural graph structure (Section 4.1) using a graph neural network. For a new optimization task, NiNo is applied rarely over time: only once per 1k steps of Adam (Section 5).

$$\mathbf{V}^0 = \phi_{\text{lpe}}(\mathbf{V}^{\text{lpe}}) + \phi_w(\mathbf{V}^w), \quad \mathbf{V}^0 \in \mathbb{R}^{|\mathbf{V}| \times D}, \quad (5)$$

$$\mathcal{E}^0 = \phi_e(\tilde{\mathcal{E}}) + \phi_{\text{type}}(\mathbf{E}^{\text{type}}), \quad \mathcal{E}^0 \in \mathbb{R}^{|\mathbf{V}| \times |\mathbf{V}| \times D}, \quad (6)$$

$$\mathbf{V}^m, \mathcal{E}^m = \text{GNN}_{\phi_m}(\mathbf{V}^{m-1}, \mathcal{E}^{m-1}), \quad m = 1 \dots M, \quad (7)$$

$$\hat{\mathcal{E}} = \phi_{\text{DMS}}(\mathcal{E}^M), \quad \hat{\mathcal{E}} \in \mathbb{R}^{|\mathbf{V}| \times |\mathbf{V}| \times K}. \quad (8)$$

Our GNN layers are based on [Kofinas et al. \(2024\)](#), but to enable better efficiency and generalization when working with large models, we use a simple mean aggregation (see Section A.5).

**Direct multi-step forecasting (DMS).** The final layer  $\phi_{\text{DMS}}$  outputs  $K$  values for each edge corresponding to future steps from  $\tau + 1$  to  $\tau + K$ . This is motivated by direct multi-step forecasting performing well in time series forecasting by avoiding error accumulation effects of autoregressive forecasting ([Chevillon, 2007](#); [Zeng et al., 2023](#)). To train the model given a training dataset of parameters trained with Adam, we use equation 1, but applied for  $k = [1, \dots, K]$  instead of fixing  $k$ :

$$\operatorname{argmin}_\phi \frac{1}{K} \sum_{k=1}^K (\|\Delta \tilde{\theta}_{\tau+k} - \Delta \hat{\theta}_{\tau+k}\|_1), \quad (9)$$

where  $\Delta \hat{\theta}_{\tau+k} = \mathcal{G}^{-1}(\hat{\mathcal{E}}_k)$  and  $\tilde{\theta}_{\tau+k}$  are the target parameters scaled using  $\mu_\tau, \sigma_\tau$ . The predicted parameters are obtained by unscaling the predicted delta using  $\mu_\tau, \sigma_\tau$ :  $\hat{\theta}_{\tau+k} = \theta_\tau + \text{unscale}(\Delta \hat{\theta}_{\tau+k})$ .

**Inference with  $k$ -decay.** Once our model is trained, it can predict future parameters for  $k \in [1, K]$ . While  $k$  can be treated as a hyperparameter and kept fixed during optimization of the target task, we found that in the initial optimization stage using very large  $k$  is beneficial, whereas in the later optimization stages  $k$  should decrease fast, since the parameter values do not change significantly. Therefore, we propose to decay  $k$  during optimization as  $k \approx K((T-t)/T)^p$ , where  $T$  is the maximum number of optimization steps and  $p$  controls the decay speed (Fig. 10). Although  $p$  can be tuned, we found that  $p = 2$  works well in our experiments.

## 5 EXPERIMENTS

We experiment with nine tasks, each defined by a dataset and a neural network architecture in the vision or language domains (Table 1). Four of the tasks, the *in-distribution tasks*, are of a relatively smaller scale and used to train our *meta-models* (NiNo, WNN and their variants). The other five tasks, the *out-of-distribution tasks*, differ from the in-distribution tasks in the architecture and/or dataset.

**Indexing parameters.** Since the notion of epoch used for  $\tau$  in WNNs (Section 3.1) can be ill-defined in some tasks (e.g. in language tasks the models are often trained only for 1 epoch), in our experiments we found  $\tau$  and  $\tau+1$  indexing parameters at step  $t$  and  $t+200$  to be a well performing strategy in all the tasks, so for our default context length  $c = 5$  we apply the meta-models every  $5 \times 200 = 1,000$  steps.

Table 1: In-distribution and out-of-distribution tasks.

|                            | IN-DISTRIBUTION TASKS |                |                 |                 | OUT-OF-DISTRIBUTION TASKS |        |         |            |            |
|----------------------------|-----------------------|----------------|-----------------|-----------------|---------------------------|--------|---------|------------|------------|
|                            | FM/16                 | C10/16         | LM1B/3-24       | LM1B/2-32       | FM/32                     | C10/32 | C100/32 | LM1B/3-64  | WIKI/3-64  |
| Final training loss        | $0.25 \pm 0.06$       | $0.91 \pm 0.1$ | $5.94 \pm 0.03$ | $5.85 \pm 0.04$ | —                         | —      | —       | —          | —          |
| #models                    | 300                   | 300            | 200             | 200             | —                         | —      | —       | —          | —          |
| #params                    | 14K                   | 15K            | 1.2M            | 1.6M            | 56K                       | 57K    | 63K     | 3.3M       | 3.3M       |
| Target (validation) metric | Acc                   | Acc            | Perplexity      | Perplexity      | Acc                       | Acc    | Acc     | Perplexity | Perplexity |
| Target value               | 89.5%                 | 66.0%          | 352             | 319             | 90.5%                     | 72.5%  | 39%     | 181        | 147        |
| Adam #steps                | 8606                  | 8732           | 23000           | 23500           | 8269                      | 8607   | 8341    | 23500      | 13500      |
| NiNo #steps                | 4582                  | 3775           | 11500           | 12000           | 4395                      | 4323   | 4646    | 12000      | 7000       |

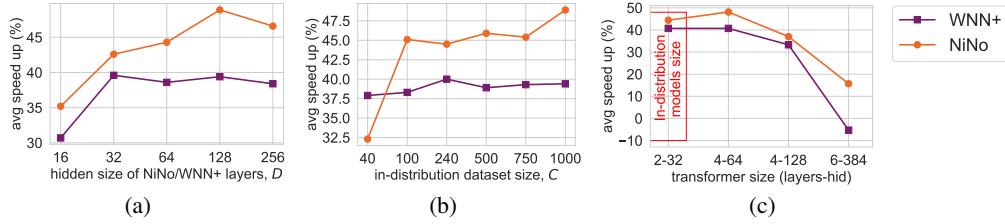


Figure 5: Scaling (a,b) and generalization (c) trends for NiNo vs WNN+.

### Training and evaluation pipeline:

1. Train the meta-models with the supervised parameter prediction loss (equation 1 for WNN and equation 9 for NiNo) using training checkpoints from the four *in-distribution tasks*.
2. Train task-specific neural networks in all the tasks using the Adam (vision) and AdamW (language) optimizers with the meta-model applied every 1k steps given  $c = 5$  past parameters.
3. Define a target performance level for each task based on running the baseline Adam optimization (without the meta-model applied) and reporting the relative reduction of the number steps to achieve that performance by other methods.

### 5.1 SETUP

**Vision tasks.** We use the FashionMNIST (F), CIFAR-10 (C10) and CIFAR-100 (C100) datasets and three layer convolutional neural networks with 16, 32 and 32 channels (e.g. task FM/16) or 32, 64 and 64 channels (e.g. task C100/32). The convolutional architectures are the same as in the L2O experiments of Kofinas et al. (2024) to enable fair comparison. In all cases, these tasks are optimized using Adam (Kingma & Ba, 2014) without weight decay, with a constant learning rate of 6e-3 (for FashionMNIST) or 3e-3 (for CIFAR) with a batch size of 128 for  $T=10k$  steps.

**Language tasks.** We use the LM1B (Chelba et al., 2014) and WikiText103 (WIKI) (Merity et al., 2016) datasets and train GPT2 style transformers (Radford et al., 2019) with 3 layers, 24 hidden units and 3 attention heads (denoted as 3-24); 2 layers, 32 units and 2 heads (2-32) or 3 layers, 64 units and 4 heads (3-64). These tasks are optimized for the next token prediction loss with AdamW (Loshchilov & Hutter, 2017), weight decay 1e-2, learning rate 2e-4, batch size of 32, sequence length of 1024 for either 1 epoch (for LM1B) or 4 epochs (for WIKI) corresponding to around 24k or 14k steps respectively. We use a predefined GPT2 tokenizer in all the cases with a fixed vocabulary.

**Meta-training dataset.** We use FM/16, C10/16, LM1B/3-24 and LM1B/2-32 as the in-distribution tasks training 300, 300, 200 and 200 models in each task respectively ( $C = 1000$  models in total). We save the checkpoints of intermediate steps, having in total around  $1.6 \times 10^6$  checkpoints. Our dataset is large and diverse (Fig. 8), yet it is relatively cheap to be collected as the tasks are small.

**Baselines.** As a reference optimization method we use Adam for vision and AdamW for language tasks with a constant learning rate that we tune independently on each task based on the validation metric. Following WNNs (Jang et al., 2023) we use *Linefit* as another baseline (Section A.2). We further improve it with a simple scaling term to promote more recent parameters denoted as *Linefit+* (see Section A.3). As the strongest baseline, we use WNN with our layerwise scaling and  $k$ -decay which we denote as WNN+. Finally, we use the learning to optimize (L2O) model from Kofinas et al. (2024), which showed very strong results on similar tasks. We use their L2O (NG-GNN) pretrained on FM/16 as retraining their model on all our in-distribution tasks is expensive. For a fair comparison

Table 2: **Reduction (%) of the number of steps until a target performance w.r.t. Adam.** We observe that on both in-distribution and out-of-distribution tasks our models improve over the base optimizer and other acceleration methods.

|                    | IN-DISTRIBUTION TASKS |             |             |             | OUT-OF-DISTRIBUTION TASKS |             |             |             | AVG SPEED UP |
|--------------------|-----------------------|-------------|-------------|-------------|---------------------------|-------------|-------------|-------------|--------------|
|                    | FM/16                 | C10/16      | LM1B/3-24   | LM1B/2-32   | FM/32                     | C10/32      | C100/32     | LM1B/3-64   |              |
| Linefit            | 27.0                  | 26.8        | 32.6        | 31.9        | 16.6                      | 26.5        | 20.7        | 29.8        | 33.3         |
| WNN                | 29.3                  | 31.4        | 26.1        | 23.4        | 22.2                      | 27.9        | 25.0        | 27.7        | 25.9         |
| Linefit+           | 27.3                  | 28.3        | 32.6        | 29.8        | 23.3                      | 27.0        | 25.0        | 25.5        | 33.3         |
| WNN+               | 33.8                  | 45.1        | 45.7        | 44.7        | 31.6                      | 44.0        | 37.8        | 38.3        | 39.8         |
| NiNo (naive graph) | <b>48.7</b>           | 53.1        | 33.7        | 12.8        | 42.1                      | 49.3        | 37.8        | 23.4        | 33.3         |
| NiNo               | 46.8                  | <b>56.8</b> | <b>50.0</b> | <b>48.9</b> | <b>46.8</b>               | <b>49.8</b> | <b>44.3</b> | <b>48.9</b> | <b>48.1</b>  |

Table 3: **Ablations and analysis of hyperparameters. Average speed up (%) is reported.**

| (a) Varying scaling                  | (b) Varying $p$ in $k$ -decay | (c) Ablating embeddings (equation 5)  |
|--------------------------------------|-------------------------------|---|
| layerwise ( $\mu, \sigma$ per layer) | <b>48.9</b>                   | LPE + word pos enc ( $\mathbf{V}^w$ )   |
| $\mu, \sigma$ per param              | 41.5                          | no LPE  |
| min-max per param                    | 38.8                          | no word pos enc ( $\mathbf{V}^w$ )  |
| no scaling                           | 32.1                          | no edge type  |
| (d) Varying context length           | (e) Varying NiNo size         | (f) Other hyperparameters ( $b, \gamma$ are the batch size and learning rate used to meta-train NiNo) |
| $c = 3$                              | 29.8                          | $C = 10^3, b = 4, \gamma = 3e - 3$  |
| $c = 5$                              | <b>48.9</b>                   | $C = 10^2, b = 4, \gamma = 3e - 3$  |
| $c = 7$                              | 44.0                          | $C = 10^3, b = 4, \gamma = 1e - 3$  |
| $c = 10$                             | 34.0                          | $C = 10^3, b = 2, \gamma = 3e - 3$  |
|                                      | $M = 3, D = 128$              | <b>48.9</b>   |
|                                      | $M = 3, D = 32$               | 45.8  |
|                                      | $M = 1, D = 128$              | 42.6  |
|                                      | $M = 2, D = 128$              | 39.7  |
|                                      |                               | 40.8  |
|                                      |                               | 44.7  |

with L2O, we also trained NiNo on FM/16 only (denoted as NiNo/FM16). We train WNN+ and NiNo with  $M \in [1, 2, 3]$  layers and  $D \in [16, 32, 64, 128, 256]$ , where  $M = 3, D = 128$  are used unless otherwise stated (see training details in Section A.7).

**Evaluation.** We follow Dahl et al. (2023) when choosing the approach to compare training methods and set our target based on the validation set performance using tuned Adam, which we consider as the main baseline (Table 1). To reduce evaluation uncertainty we train all the tasks multiple times (10 for vision and 3 for language) and use the median number of steps. We report a relative reduction of the number of steps in %, e.g. for task WIKI/3-64 if the method achieves perplexity 147 in 7000 steps (as our NiNo does), the reduction is 48.1% based on Table 1. For language tasks due to more expensive evaluation we only evaluate every 500 steps, whereas for vision we evaluate every step.

## 5.2 RESULTS

**Main results.** Our NiNo model shows consistently better performance than Linefit and WNN and their improved variants on all the nine tasks (Table 2). On average we speed up Adam optimization by 48.9% reducing the number of steps to achieve the target performance roughly by half. NiNo is followed by WNN+ and NiNo with a naive graph (Section 4.1). While the latter achieves slightly better performance on the FM/16 in-distribution task, in the out-of-distribution tasks this model significantly underperforms to NiNo, especially in the language tasks. This performance gap is expected since NiNo represents neural graphs for transformers more accurately as we further validated qualitatively (Fig. 2d,e) and quantitatively (Section A.6). Other baselines perform poorly, but we highlight that both Linefit baselines compare favorably to WNN indicating that the latter could have learned a very simple rule despite training on a lot of data.

**Comparison to L2O.** L2O performs the best in-distribution on the same task (Fig. 6a), however when applied to an unseen task it overfits to the training set (Fig. 6b). In contrast, NiNo/FM16 trained on the same task as L2O performs well on the validation set of unseen tasks, presumably because Adam is used for most of the steps and for the input NiNo uses the trajectory of parameters of the target task to make the prediction. In addition, L2O has an overhead at every optimization step making the total computation cost noticeable (Table 6 in Section A.7). It is also more computationally challenging to meta-train it on multiple tasks, therefore we only use a pretrained L2O from Kofinas et al. (2024). Since L2O does not reach the target validation scores in most tasks, we compare it only using the curves.

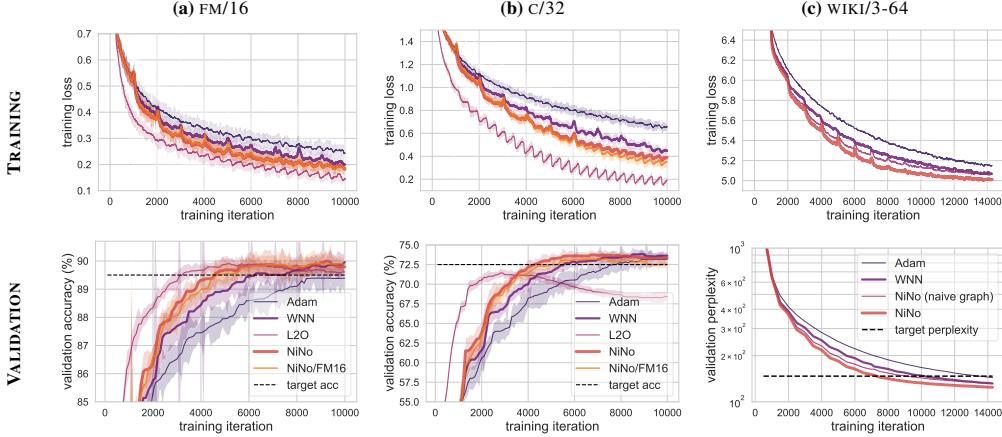


Figure 6: Train loss and validation performance on FashionMNIST, CIFAR-10 and WIKI.

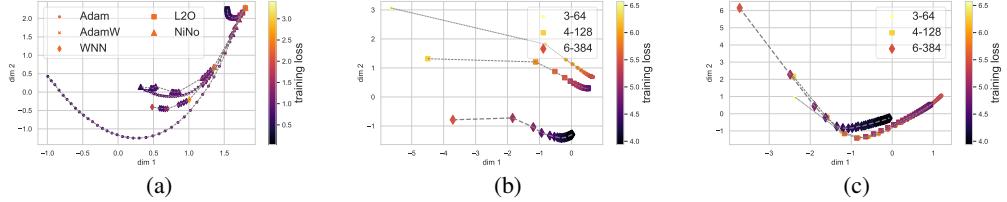


Table 4: (a) Comparison of optimization methods by projecting **parameters** to 2d using PCA at every training step. (b,c) Comparing the embedding quality of WNN+ (b) and NiNo (c) during optimization by projecting **neural graph embeddings** to 2d using PCA every 1,000 steps.

**Ablations.** We ablate NiNo components and hyperparameters and found that our layerwise parameter scaling and  $k$ -decay are the most important components (Table 3a,b). Among the hyperparameters, we found that context length  $c = 5$  is a sweet spot (Table 3d), because small  $c$  may not have enough information to capture the parameter trend, while larger  $c$  leads to the NiNo applied too rarely (for  $c = 10$  it is applied every  $10 \times 200 = 2,000$  steps). Using deeper (Table 3e) and wider (Fig. 5a) GNNs is also important. However we found that depth  $M > 3$  or width  $D > 128$  make it more challenging to train a performant model. Increasing the number of models ( $C$ ) helps NiNo achieve better speed ups (Table 3f, Fig. 5b). In contrast, WNN+ saturates fast when  $D$  or  $C$  is increased.

**Evaluating on larger models.** We evaluated the ability of NiNo to speed up training of much larger models than it was trained on. Specifically, we trained 4 and 6 layer transformers on WIKI with 128 and 384 hidden units having around 7M and 29M parameters respectively. The latter is around 18 times bigger than the largest in-distribution architecture. Moreover, the WIKI dataset is different from LM1B used in meta-training. Despite these challenges, NiNo was able to speed up training by around 40% and 15% for the 7M and 29M models respectively outperforming WNN+ in all cases (Fig. 5c).

### 5.3 ANALYSIS

**Parameter space.** We show how the parameters evolve during training on C10/32 by projecting them to 2d using PCA (Fig. 4a). We show the first 8,000 steps for Adam/AdamW and 4,000 steps for other methods with 200 steps between points. In the beginning of training, both WNN and NiNo make big steps along the Adam/AdamW when predicting the parameters, while L2O diverges early. At later optimization stages, NiNo makes better predictions than WNN as the latter leads to significant loss spikes. Interestingly, NiNo’s trajectory closely corresponds to a  $2\times$  faster AdamW, even though NiNo’s base optimizer is Adam in this task. This indicates that NiNo may implicitly regularize optimization.

**Graph embedding space.** We optimize three different transformers on the language task and visualize how the neural graph embeddings evolve during training (Fig. 4b,c). We construct embeddings by computing the average edge/parameter features after the last NiNo/WNN+ layer by feeding the 5 past

parameter states every 1,000 iterations as in our other experiments. We then project the embeddings to 2d using PCA for visualization. Surprisingly, even WNN+ can encode parameter in a meaningful way – the parameters of different architectures and at different optimization are visually distinct in general (Fig. 4b). However, WNN+’s embeddings virtually collapse at later training stages (e.g. for 6-384), while NiNo’s embeddings are visually more distinct for lower loss values (Fig. 4c).

**Computational costs.** NiNo is applied only every 1,000 steps of the base optimizer, so the wallclock time overhead of using our default NiNo is <0.5% for a 7M transformer compared to around 5% by L2O (Table 6). As for the memory required for a prediction with NiNo, in our implementation NiNo’s peak memory usage is similar to the peak memory usage of a training step of the transformer with a batch size of 16-32. We expect that in future work, we can parallelize NiNo’s inference on large neural graphs using approaches similar to large scale model inference in image generation (Li et al., 2024).

## 6 CONCLUSION AND FUTURE WORK

We proposed a novel NiNo model to accelerate training of neural networks, including transformers. Our work makes a step in a relatively underexplored area of periodic parameter prediction, which as we show has a lot of potential in reducing training costs. In addition to reducing training costs, we reveal potentially useful byproducts of NiNo, such as low-dimensional encoding of network parameters during training, which can be used to analyze models and their training dynamics. We believe that our work can promote interesting future research. In particular, scaling up our approach appears to be intriguing as the scaling trends show better speed ups with more data and larger models.

### ACKNOWLEDGMENTS

The experiments were in part enabled by computational resources provided by Calcul Quebec and Compute Canada. Simon Lacoste-Julien is a CIFAR Associate Fellow in the Learning in Machines & Brains program as well as a Canada CIFAR AI Chair. Guillaume Lajoie is a Canada CIFAR AI chair as well a Canada Research Chair in neural computations and interfacing.

### REFERENCES

- Brandon Amos. Tutorial on amortized optimization for learning to optimize over continuous domains. *arXiv preprint arXiv:2202.00665*, 2022. [2](#)
- Bruno Andreis, Soro Bedionita, and Sung Ju Hwang. Set-based neural network encoding. *arXiv preprint arXiv:2305.16625*, 2023. [2](#)
- Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016. [1](#), [2](#)
- Maor Ashkenazi, Zohar Rimon, Ron Vainshtein, Shir Levi, Elad Richardson, Pinchas Mintz, and Eran Treister. Nern–learning neural representations for neural networks. *arXiv preprint arXiv:2212.13554*, 2022. [2](#)
- Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. [5](#)
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Philipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling, 2014. [7](#)
- Can Chen, Xi Chen, Chen Ma, Zixuan Liu, and Xue Liu. Gradient-based bi-level optimization for deep learning: A survey. *arXiv preprint arXiv:2207.11719*, 2022. [2](#)
- Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *arXiv preprint arXiv:2103.12828*, 2021. [2](#)

- Guillaume Chevillon. Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4):746–785, 2007. 6
- Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal Neighbourhood Aggregation for Graph Nets. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. 2, 15
- George E Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, et al. Benchmarking neural network training algorithms. *arXiv preprint arXiv:2306.07179*, 2023. 5, 8
- Charles Guille-Escuret, Hiroki Naganuma, Kilian Fatras, and Ioannis Mitliagkas. No wrong turns: The simple geometry of neural networks optimization paths. *arXiv preprint arXiv:2306.11922*, 2023. 2
- Howard Heaton, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Safeguarded learned convex optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7848–7855, 2023. 2
- Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. In *Advanced Neural Computers*, pp. 129–135. Elsevier, 1990. 2
- Jinhyeok Jang, Woo-han Yun, Won Hwa Kim, Youngwoo Yoon, Jaehong Kim, Jaeyeon Lee, and ByungOk Han. Learning to boost training by periodic nowcasting near future weights. In *International Conference on Machine Learning (ICML)*, 2023. 1, 2, 3, 5, 7, 14
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1, 7
- Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pp. 2688–2697. PMLR, 2018. 2
- Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2
- Boris Knyazev, Carolyn Augusta, and Graham W Taylor. Learning temporal attention in dynamic graphs with bilinear interactions. *Plos one*, 16(3):e0247936, 2021a. 2
- Boris Knyazev, Michal Drozdzal, Graham W Taylor, and Adriana Romero Soriano. Parameter prediction for unseen deep architectures. *Advances in Neural Information Processing Systems*, 34: 29433–29448, 2021b. 2
- Boris Knyazev, Doha Hwang, and Simon Lacoste-Julien. Can we scale transformers to predict parameters of diverse imagenet models? *arXiv preprint arXiv:2303.04143*, 2023. 2
- Miltiadis Kofinas, Boris Knyazev, Yan Zhang, Yunlu Chen, Gertjan J. Burghouts, Efstratios Gavves, Cees G. M. Snoek, and David W. Zhang. Graph Neural Networks for Learning Equivariant Representations of Neural Networks. In *12th International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 3, 4, 6, 7, 8, 14, 15, 16
- Richard D Lange, Devin Kwok, Jordan Kyle Matelsky, Xinyue Wang, David Rolnick, and Konrad Kording. Deep networks as paths on the manifold of neural representations. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 102–133. PMLR, 2023. 2
- Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Kai Li, and Song Han. Distrifusion: Distributed parallel inference for high-resolution diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7183–7193, 2024. 10
- Derek Lim, Haggai Maron, Marc T Law, Jonathan Lorraine, and James Lucas. Graph metanetworks for processing diverse neural architectures. *arXiv preprint arXiv:2312.04501*, 2023. 2, 3, 4, 14

- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1, 7
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. 7
- Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, and Jascha Sohl-Dickstein. Understanding and correcting pathologies in the training of learned optimizers. In *International Conference on Machine Learning*, pp. 4556–4565. PMLR, 2019. 2
- Luke Metz, James Harrison, C Daniel Freeman, Amil Merchant, Lucas Beyer, James Bradbury, Naman Agrawal, Ben Poole, Igor Mordatch, Adam Roberts, et al. Velo: Training versatile learned optimizers by scaling up. *arXiv preprint arXiv:2211.09760*, 2022. 1, 2
- Chedi Morchdi, Yi Zhou, Jie Ding, and Bei Wang. On the mysterious optimization geometry of deep neural networks. 2022. 2
- Aviv Navon, Aviv Shamsian, Idan Achituv, Ethan Fetaya, Gal Chechik, and Haggai Maron. Equivariant architectures for learning in deep weight spaces. *arXiv preprint arXiv:2301.12780*, 2023. 1, 2
- William Peebles, Ilija Radosavovic, Tim Brooks, Alexei A Efros, and Jitendra Malik. Learning to learn with generative models of neural network checkpoints. *arXiv preprint arXiv:2209.12892*, 2022. 2
- Isabeau Premont-Schwarz, Jaroslav Vitkuu, and Jan Feyereisl. A simple guard for learned optimizers. *arXiv preprint arXiv:2201.12426*, 2022. 2
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 7
- Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. In *International conference on machine learning*, pp. 4470–4479. PMLR, 2018. 2
- Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pp. 8459–8468. PMLR, 2020. 2
- Konstantin Schürholz, Dimche Kostadinov, and Damian Borth. Self-supervised representation learning on neural network weights for model characteristic prediction. *Advances in Neural Information Processing Systems*, 34:16481–16493, 2021. 2, 3
- Konstantin Schürholz, Boris Knyazev, Xavier Giró-i Nieto, and Damian Borth. Hyper-representations as generative models: Sampling unseen neural network weights. *Advances in Neural Information Processing Systems*, 35:27906–27920, 2022. 2, 5
- Konstantin Schürholz, Michael W Mahoney, and Damian Borth. Towards scalable and versatile weight space learning. *arXiv preprint arXiv:2406.09997*, 2024. 2
- Abhishek Sinha, Mausoom Sarkar, Aahitagni Mukherjee, and Balaji Krishnamurthy. Introspection: Accelerating neural network training by learning weight evolution. *arXiv preprint arXiv:1704.04959*, 2017. 1, 2, 3, 14
- Bedionita Soro, Bruno Andreis, Hayeon Lee, Song Chong, Frank Hutter, and Sung Ju Hwang. Diffusion-based neural network weights generation. *arXiv preprint arXiv:2402.18153*, 2024. 2
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013. 14
- Benjamin Thérien, Charles-Étienne Joseph, Boris Knyazev, Edouard Oyallon, Irina Rish, and Eugene Belilovsky. mulo: Compute-efficient meta-generalization of learned optimizers. *arXiv preprint arXiv:2406.00153*, 2024. 2

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1
- Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *International conference on learning representations*, 2019. 2
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 14
- Kai Wang, Zhaopan Xu, Yukun Zhou, Zelin Zang, Trevor Darrell, Zhuang Liu, and Yang You. Neural network diffusion. *arXiv preprint arXiv:2402.13144*, 2024. 2
- Vinicio Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. Deep reinforcement learning with relational inductive biases. In *International conference on learning representations*, 2019. 2
- Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023. 6
- Allan Zhou, Kaien Yang, Kaylee Burns, Yiding Jiang, Samuel Sokota, J Zico Kolter, and Chelsea Finn. Permutation equivariant neural functionals. *arXiv preprint arXiv:2302.14040*, 2023. 1, 2
- Allan Zhou, Chelsea Finn, and James Harrison. Universal neural functionals. *arXiv preprint arXiv:2402.05232*, 2024a. 2
- Allan Zhou, Kaien Yang, Yiding Jiang, Kaylee Burns, Winnie Xu, Samuel Sokota, J Zico Kolter, and Chelsea Finn. Neural functional transformers. *Advances in neural information processing systems*, 36, 2024b. 2
- Xinyu Zhou, Boris Knyazev, Alexia Jolicoeur-Martineau, and Jie Fu. Logah: Predicting 774-million-parameter transformers using graph hypernetworks with 1/100 parameters. *arXiv preprint arXiv:2405.16287*, 2024c. 2

## A APPENDIX

### A.1 WNN DETAILS

In WNNs (Jang et al., 2023), parameter scaling is performed by first computing the minimum and maximum values in the input parameter values:  $s_\tau^i = \max(\theta_\tau^i, \dots, \theta_{\tau_c}^i) - \min(\theta_\tau^i, \dots, \theta_{\tau_c}^i)$ ,  $\forall i \in [1, n]$ . Each  $i$ -th parameters scaling factor  $s_\tau^i$  is then used to scale both the inputs:  $\tilde{\theta}_\tau^i = \theta_\tau^i / (s_\tau^i + \epsilon)$ , and, only for training WNNs, the targets:  $\tilde{\theta}_{\tau+k}^i = \theta_{\tau+k}^i / (s_\tau^i + \epsilon)$ . Jang et al. (2023) also subtract the last element from the sequence, however we found this to be redundant. Unscaling is performed by multiplying the prediction by  $s_\tau^i$ . Scaling and unscaling are performed per parameter (for each parameter independently).

### A.2 LINEFIT

Sinha et al. (2017); Jang et al. (2023) introduced a “linefit” baseline to predict future parameter values by extrapolating the line to a future point:

$$\begin{aligned} \forall i \in [1, n] : \\ \hat{\theta}_{\tau+k}^i &= 2a^i c + b^i, \end{aligned} \quad (10)$$

where  $a^i, b^i$  are obtained by optimizing the following objective:

$$\operatorname{argmin}_{a^i, b^i} \|a^i \mathbf{x} + b^i - \Theta_{\tau:\tau_c}^i\|_2, \quad (11)$$

where  $\mathbf{x} = [1, 2, \dots, c]$  and  $\Theta_{\tau:\tau_c}^i = [\theta_\tau^i, \theta_{\tau-1}^i, \dots, \theta_{\tau_c}^i] \in \mathbb{R}^c$ . This baseline fits a line  $(a^i, b^i)$  for each parameter  $i$  given only the past values of the same parameter without collecting a training dataset. Despite its simplicity, this baseline was shown to improve on using Adam/SGD only.

### A.3 LINEFIT+

Linefit outlined in Section A.2 can be viewed as a form of momentum that promotes the parameters to be consistent with their global trend. However, Linefit weighs all past values uniformly when fitting the line, while in the momentum more recent values have more effect on the weight update (Sutskever et al., 2013). Motivated by this observation, we introduce the *Linefit+* baseline that penalizes more the errors for later parameter values in the trajectory:

$$\begin{aligned} \operatorname{argmin}_{a^i, b^i} &\|\boldsymbol{\mu}(a^i \mathbf{x} + b^i - \Theta_{\tau:\tau_c}^i)\|_2, \\ \text{where } \boldsymbol{\mu} &= [1/c, 2/c, \dots, 1]. \end{aligned} \quad (12)$$

### A.4 NAIVE NEURAL GRAPH OF TRANSFORMERS

A multi-head self-attention (MSA) layer of transformers (Vaswani et al., 2017) consists of three weight matrices  $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v$  applied to the input  $\mathbf{x} \in \mathbb{R}^{N \times d}$  and another weight matrix  $\mathbf{W}^o$  returning the output, where  $N$  is a sequence length and all the four weight matrices are  $d \times d$ . To compute MSA with  $H$  heads, the weight matrices are typically split into  $H$  groups across columns for  $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v$  and rows for  $\mathbf{W}^o$  (Fig. 2a). The heads are processing the input in parallel,  $\forall h \in [1, H]$  using equation 3.

Neural graphs defined by Kofinas et al. (2024) and Lim et al. (2023) are conceptually similar, however Lim et al. (2023) model biases  $\mathbf{b}^{(l)}$  and normalization layers as extra nodes connected to the neurons in the  $l$ -th layer, hence the node features of Lim et al. (2023) do not include any parameters (Fig. 2c). We build on this variant, since it is more easily to implement it for the layers with complicated neuron permutation symmetry, such as multi-head self-attention. Compared to Kofinas et al. (2024), Lim et al. (2023) also provide a slightly different description for the MSA neural graph, however the  $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v$  edges are also stacked. Given the lack of detailed description for the multi-head case and no implementation currently available publicly, it makes it challenging to use their neural graphs directly. Therefore, we use (Kofinas et al., 2024) as the base (*naive*) neural graph when constructing MSA layers.

## A.5 GRAPH NEURAL NETWORK

A Graph Neural Network (GNN) layer with edge features can be formulated using the mean aggregation as follows (Corso et al., 2020):  $\mathbf{v}_i = \phi_a \left( \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij} \right)$ , where  $\mathbf{m}_{ij}$  is computed using message passing  $\phi_m$  given node and edge features  $\mathbf{v}_i, \mathbf{v}_j, \mathbf{e}_{ij}$ ;  $\phi_a$  is an MLP;  $\mathcal{N}(i)$  are the neighbors of node  $i$ . Kofinas et al. (2024) modify how  $\mathbf{m}_{ij}$  is computed to better incorporate edge features and introduce an edge update step to make better per edge predictions (Table 5). Stacking such layers form a GNN that (1) does not change the size of the input graph and (2) is permutation equivariant. (1) means that for the neural graphs there are  $|\Theta'| \times d_E$  input edge features and  $|\Theta'| \times 1$  predictions, where  $|\Theta'|$  is the number of trainable parameters in the input model plus auxiliary non-trainable parameters (e.g. residual and head connections). (2) means that any permutation of input nodes results in the same permutation of the output nodes and corresponding edges. These properties make GNNs a suitable model to predict future parameters and to work with neural graphs where nodes (neurons) can be permuted in the ways described in Section A.6.

Table 5: GNN layer comparison.

| Step               | Typical GNN (Corso et al., 2020)   | Neural Graph GNN (Kofinas et al., 2024)   |
|--------------------|--|---|
| 1. Message passing | $\mathbf{m}_{ij} = \phi_m ([\mathbf{v}_i, \mathbf{v}_j, \mathbf{e}_{ij}])$ | $\mathbf{m}_{ij} = \phi_{\text{scale}} (\mathbf{e}_{ij}) \odot \phi_m ([\mathbf{v}_i, \mathbf{v}_j]) + \phi_{\text{shift}} (\mathbf{e}_{ij})$ |
| 2. Aggregation     |  | $\mathbf{v}_i = \phi_a \left( \frac{1}{ \mathcal{N}(i) } \sum_{j \in \mathcal{N}(i)} \mathbf{m}_{ij} \right)$                                 |
| 3. Edge update     | $\mathbf{e}_{ij}^{(k+1)} = \mathbf{e}_{ij}^{(k)}$                          | $\mathbf{e}_{ij}^{(k+1)} = \phi_e^{(k+1)} \left( [\mathbf{v}_i^{(k)}, \mathbf{e}_{ij}^{(k)}, \mathbf{v}_j^{(k)}] \right)$                     |

## A.6 NEURON PERMUTATION SYMMETRY

**Permutation symmetry in neural graphs.** Let  $\pi^{\text{good}}(\Theta)$  be a permutation of neurons in  $\Theta$  that does not change the network function:  $f(\mathbf{x}, \Theta) = f(\mathbf{x}, \pi^{\text{good}}(\Theta))$ . Likewise, let  $\pi^{\text{bad}}(\Theta)$  be a permutation that changes the function:  $f(\mathbf{x}, \Theta) \neq f(\mathbf{x}, \pi^{\text{bad}}(\Theta))$ . Denoting  $\cong$  and  $\not\cong$  as “isomorphic” and “non-isomorphic” operators on graphs respectively, for neural graphs  $\mathcal{G}$  the following equations are satisfied under some assumptions (Kofinas et al., 2024):  $\mathcal{G}(\Theta) \cong \mathcal{G}(\pi^{\text{good}}(\Theta))$  and  $\mathcal{G}(\Theta) \not\cong \mathcal{G}(\pi^{\text{bad}}(\Theta))$ . Here, by definition, the neural graphs are assumed to be constructed correctly, however, achieving or validating that in practice is not trivial for such networks as transformers.

**Neuron permutation symmetry experiment.** To validate how well our neural graphs correspond to true neuron permutations in transformers, we run a simple neuron permutation experiment. We use a single transformer layer with  $d = 12$  and  $H = 4$ , permute rows and columns in its MSA weight matrices  $\Theta$  using a uniformly sampled permutation  $\pi$ , and label each permutation as  $y_\pi = 1$  for  $\pi^{\text{good}}$  or  $y_\pi = 0$  for  $\pi^{\text{bad}}$  depending if the transformer output  $f(\mathbf{x}, \Theta)$  changes or not:

$$y_\pi = \begin{cases} 1, & \text{iff } f(\mathbf{x}, \Theta) = f(\mathbf{x}, \pi(\Theta)) \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

We generate 1,000 such permutations with about 500 of good and bad permutations. Then we extract a neural graph embedding for each permutation. For correctly constructed neural graphs the embeddings corresponding to  $y_\pi = 1$  should be close to the non-permuted  $\Theta$ . To extract embeddings we use a randomly initialized graph neural network with 3 layers and 32 hidden units based on Section A.5. We then visualize the embeddings in 2d (Fig. 2d,e) and color-code each embedding with  $y_\pi$ . We also train a logistic regression model  $\psi$  that takes a graph embedding  $\mathbf{h}$  as input and predicts if it is a good or bad permutation:  $\hat{y}_\pi = \psi(\mathbf{h})$ , and we compute classification accuracy between  $\hat{y}_\pi$  and  $y_\pi$ . We use all 1,000 samples for training and evaluation  $\psi$  to estimate how easily can the graph embeddings be separated based on  $y_\pi$  (random guess accuracy is around 50%). We repeated the above experiment for the naive and our neural graphs (Fig. 2, d,e). The visualization and classification results indicate that our neural graph construction respects good and bad permutations well, while naive neural graphs confuse them.

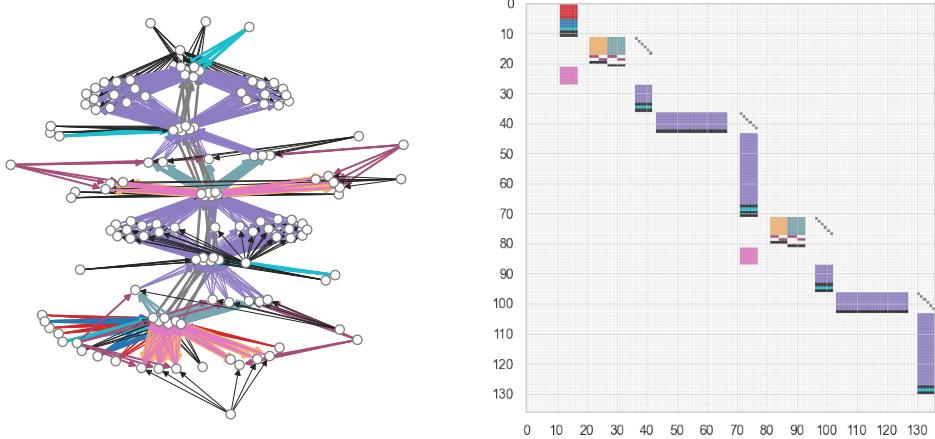


Figure 7: Graph and adjacency matrix of a 2 layer transformer with  $d = 6$  and  $H = 2$  with head restrictions and separate  $q, k, v$  edge features introduced in Section 4.1.

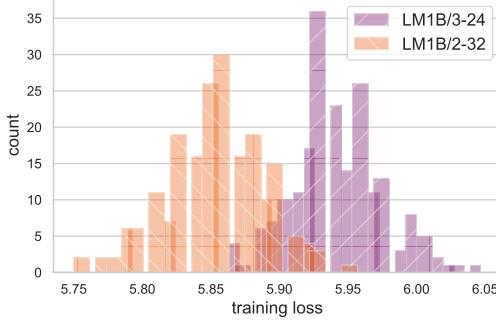


Figure 8: Histogram of final training losses in the language meta-training tasks. See Table 1 for more details.

#### A.7 NiNo TRAINING DETAILS

For convolutional networks,  $d_E = h \times w$  (Kofinas et al., 2024), where  $h, w$  are convolution kernel size, so edge features representing  $c$  states are  $|\theta'| \times chw$ . To make the feature dimensionality the same for different architectures, we use zero-padding. We train meta-models for 20k training iterations using AdamW, learning rate 3e-3 with cosine decay and weight decay 0.01. We sample a batch of 4 checkpoints in each training iteration and use automatic mixed precision. Training of the NiNo and WNN+ meta-models completes in under 7 and 6 hours respectively on a single NVIDIA RTX8000 with 48GB of memory. Meta-training behavior can give important highlights about the model and data. Therefore we explore meta-training for different hidden sizes (128 vs 32) and numbers of dataset samples (checkpoints from the 1000 vs just 4 models). We found that NiNo achieves a lower meta-training loss than WNN+ despite having a comparable number of parameters indicating the importance of leveraging neural graphs (Fig. 9). However our NiNo model is still in a severe underfitting regime when trained on all the training checkpoints, since when the number of checkpoint models is reduced to just 4 the loss is much lower for NiNo (with 128 hidden units). This result suggests that further scaling up NiNo could help fit all the training data better and achieve stronger results.

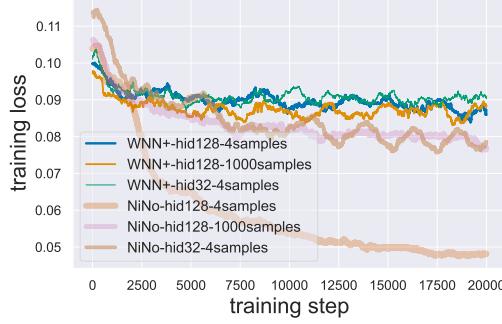
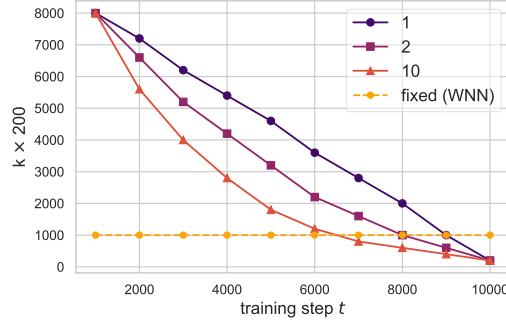


Figure 9: Meta-training curves.

Figure 10: Different  $k$ -decay schedules.Table 6: Computational costs after running for 1k steps with batch size  $b = 32$ , sequence length=1024.

| Method             | 3-64 (3.3M) |               |                     | 4-128 (7M) |               |                     |
|--------------------|-------------|---------------|---------------------|------------|---------------|---------------------|
|                    | Peak Mem    | Time          | Total time overhead | Peak Mem   | Time          | Total time overhead |
| Adam               | 31GB        | 166ms/step    | 0%                  | 40GB       | 232ms/step    | 0%                  |
| WNN+ ( $D=32$ )    | 9GB         | 33ms/predict  | 0.02%               | 12GB       | 55ms/predict  | 0.02%               |
| WNN+ ( $D=128$ )   | 16GB        | 86ms/predict  | 0.05%               | 24GB       | 129ms/predict | 0.06%               |
| NiNo ( $D=32$ )    | 9GB         | 151ms/predict | 0.09%               | 12GB       | 320ms/predict | 0.14%               |
| NiNo ( $D=128$ )   | 22GB        | 280ms/predict | 0.17%               | 24GB       | 729ms/predict | 0.31%               |
| L2O (mlp, $D=32$ ) | 9GB         | 171ms/step    | 3.0%                | 12GB       | 243ms/step    | 4.74%               |