

AutoInit: Analytic Signal-Preserving Weight Initialization for Neural Networks

Garrett Bingham^{1, 2} and Risto Miikkulainen^{1, 2}

¹ The University of Texas at Austin, Austin, TX 78712

² Cognizant AI Labs, San Francisco, CA 94105
bingham@cs.utexas.edu, risto@cs.utexas.edu

Abstract

Neural networks require careful weight initialization to prevent signals from exploding or vanishing. Existing initialization schemes solve this problem in specific cases by assuming that the network has a certain activation function or topology. It is difficult to derive such weight initialization strategies, and modern architectures therefore often use these same initialization schemes even though their assumptions do not hold. This paper introduces AutoInit, a weight initialization algorithm that automatically adapts to different neural network architectures. By analytically tracking the mean and variance of signals as they propagate through the network, AutoInit appropriately scales the weights at each layer to avoid exploding or vanishing signals. Experiments demonstrate that AutoInit improves performance of convolutional, residual, and transformer networks across a range of activation function, dropout, weight decay, learning rate, and normalizer settings, and does so more reliably than data-dependent initialization methods. This flexibility allows AutoInit to initialize models for everything from small tabular tasks to large datasets such as ImageNet. Such generality turns out particularly useful in neural architecture search and in activation function discovery. In these settings, AutoInit initializes each candidate appropriately, making performance evaluations more accurate. AutoInit thus serves as an automatic configuration tool that makes design of new neural network architectures more robust. The AutoInit package provides a wrapper around TensorFlow models and is available at <https://github.com/cognizant-ai-labs/autoinit>.

1 Introduction

Proper weight initialization is crucial to achieve high performance with deep networks. A common motif in such networks is repeated layers or building blocks. Thus, if a given layer amplifies or diminishes the forward or backward propagation of signals, repeated applications of that layer will result in exploding or vanishing signals, respectively (Hochreiter 1991; Hanin 2018). This phenomenon makes optimization difficult, and can even exceed machine precision. The issue persists regardless of whether the weights are sampled to be uniform, normal, or orthogonal (Saxe, McClelland, and Ganguli 2013; Hu, Xiao, and Pennington 2020).

While many initialization strategies have been proposed in the past, these strategies apply only to neural networks

with specific activation functions, topologies, or layer types. Thus, researchers designing new models or activation functions have two options. The first option is to derive weight initialization strategies manually for every architecture considered, which is generally difficult and time consuming. The second option is to use existing initialization strategies in new settings, where they may be incorrect and therefore misleading: A candidate model may appear poor when it is the suboptimal initialization that makes training difficult.

To overcome this problem, this paper proposes AutoInit, an algorithm that automatically calculates analytic mean- and variance-preserving weight initialization for neural networks. Since AutoInit is algorithmic, it relieves the researcher from a difficult but consequential step in model design. It is no longer necessary to use existing weight initialization strategies in incorrect settings: AutoInit provides an appropriate default initialization automatically, resulting in better and more reliable performance.

2 Related Work

Weight initialization strategies attempt to solve the following problem: How should weights be initialized so that the neural network does not suffer from vanishing or exploding signals? This section reviews previous research in neural network weight initialization, which has focused on stabilizing signals by accounting for specific components of neural networks such as the activation function, topology, layer types, and training data distribution.

Activation-Function-Dependent Initialization As is common in the literature, fan_in and fan_out refer to the number of connections feeding into and out of a node, respectively. LeCun et al. (2012) recommend sampling weights from a distribution with mean zero and standard deviation $\sqrt{\text{fan_in}}$. This initialization encourages propagated signals to have variance approximately one if used with an activation function symmetric about the origin, like $1.7159 \tanh(\frac{2}{3}x)$ or $\tanh(x) + \alpha x$ for some small choice of α . The standard sigmoid $f(x) = 1/(1+e^{-x})$ induces a mean shift and should not be used in this setting.

Glorot and Bengio (2010) proposed one initialization strategy to ensure unit variance in the forward-propagated signals and another to ensure unit variance for the backward-propagated gradients. As a compromise between

the two strategies, they initialized weights by sampling from $\mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{f_{\text{an.in}}+f_{\text{an.out}}}}, \frac{\sqrt{6}}{\sqrt{f_{\text{an.in}}+f_{\text{an.out}}}}\right)$. They also avoided sigmoid, and instead chose symmetric functions with unit derivatives at 0, such as \tanh or $\text{Softsign}(x) = x/(1+|x|)$.

He et al. (2015) introduced the PReLU activation function and a variance-preserving weight initialization to be used with it that samples weights from $\mathcal{N}(0, \sqrt{2/f_{\text{an.in}}})$. Similarly, Klambauer et al. (2017) introduced SELU, an activation function with self-normalizing properties. These properties are only realized when SELU is used with the initialization scheme by LeCun et al. (2012).

The above weight initialization strategies attempt to solve the same fundamental problem: How can weights be scaled so that repeated applications of the activation function do not result in vanishing or exploding signals? While these approaches solve this problem in a few special cases, the issue is more general. Manually deriving the correct scaling is intractable for complicated activation functions. One approach for an arbitrary function f is to sample Gaussian inputs x and adjust the weights according to the empirical variance $\text{Var}(f(x))$ (Brock, De, and Smith 2021). This paper proposes an alternative and potentially more accurate approach: integration by adaptive quadrature. The result is a weight initialization strategy that is compatible with any integrable activation function. Indeed, previous activation-function-dependent initializations are special cases of the AutoInit algorithm.

Topology-Dependent Initialization The activation-function-dependent initializations discussed above were designed for neural networks composed of convolutional or dense layers. After the introduction of residual networks (ResNets; He et al. 2016a,b), new weight initialization schemes had to be developed to account for the effect of shortcut connections and various types of residual branches.

Taki (2017) analyzed signal propagation in plain and batch-normalized ResNets. They developed a new weight initialization to stabilize training, but did not consider modifications like using deeper residual blocks or reordering components like the activation function or batch normalization layers. In contrast, AutoInit is topology-agnostic: It adapts to any of these changes automatically.

Zhang, Dauphin, and Ma (2019) introduced Fixup, an initialization method that rescales residual branches to stabilize training. Fixup replaces batch normalization in standard and wide residual networks (Ioffe and Szegedy 2015; He et al. 2016a,b; Zagoruyko and Komodakis 2016) and replaces layer normalization (Ba, Kiros, and Hinton 2016) in transformer models (Vaswani et al. 2017). The disadvantages of this scheme are that it only applies to residual architectures, needs proper regularization to get optimal performance, and requires additional learnable scalars that slightly increase model size.

Arpit, Campos, and Bengio (2019) proposed a new initialization scheme for weight-normalized networks (Salimans and Kingma 2016) that relies on carefully scaling weights, residual blocks, and stages in the network. Like related ap-

proaches, this technique improves performance in specific cases, but imposes design constraints, like requiring ReLU activation functions and a specific $\text{Conv} \rightarrow \text{ReLU} \rightarrow \text{Conv}$ block structure.

Just as \tanh -inspired weight initialization does not stabilize training of ReLU networks, initialization schemes designed for non-residual networks fail with ResNets (Hanin and Rolnick 2018; Bachlechner et al. 2020; Brock, De, and Smith 2021). This observation suggests that future classes of neural networks will again require developing new weight initializations. Additionally, practitioners with models that do not fit neatly within the restricted settings of existing weight initialization research are left to derive their own initialization or use a suboptimal one. For example, many initialization schemes assume that the activation function is ReLU (He et al. 2015; Taki 2017; Arpit, Campos, and Bengio 2019; Zhang, Dauphin, and Ma 2019; De and Smith 2020). Indeed, ReLU is currently the most popular activation function (Nwankpa et al. 2018; Apicella et al. 2021), but it is not the best choice in every case. ReLU prevents dynamical isometry (Saxe, McClelland, and Ganguli 2013; Pennington, Schoenholz, and Ganguli 2017), weakens adversarial training (Xie et al. 2020), and results in poorer accuracy compared to other activation functions in certain tasks (Bingham and Miikkulainen 2022). A general weight initialization strategy that does not impose architectural constraints and achieves good performance in diverse settings is needed. AutoInit is designed to meet this challenge.

Layer-Dependent Initialization Hendrycks and Gimpel (2016a) noted that dropout layers (Srivastava et al. 2014) also affect the variance of forward-propagated signals in a network. To stabilize training properly, it is necessary to take dropout layers and the specific dropout rate into account in weight initialization. In fact, pooling, normalization, recurrent, padding, concatenation, and other layer types affect the signal variance in a similar way, but current initialization schemes do not take this effect into account. AutoInit is designed to adapt to each of these layer types dynamically, and can be extended to include new layer types as they are introduced in the future.

Data-Dependent Initialization Mishkin and Matas (2015) fed data samples through a network and normalized the output of each layer to have unit variance. Krähenbühl et al. (2015) adopted a similar approach, but opted to normalize along the channel dimension instead of across an entire layer. Data-dependent weight initializations are most similar in spirit to AutoInit; they rely on empirical variance estimates derived from the data in order to be model-agnostic. However, data-dependent weight initializations introduce a computational overhead (Mishkin and Matas 2015), and are not applicable in settings where data is not available or its distribution may shift over time, such as online learning or reinforcement learning. The quality of the initialization is also dependent on the number of the data samples chosen, and suffers when the network is very deep (Zhang, Dauphin, and Ma 2019). AutoInit instead uses an analytic approach for greater efficiency and higher accuracy.

Summary Thus, previous techniques solved the initialization problem for networks with specific activation functions, topologies, and layer types. In contrast, AutoInit does not impose design constraints, depend on data samples, or incur a parameter overhead (Dauphin and Schoenholz 2019; Zhu et al. 2021) and is therefore a good starting point especially in new settings.

3 Neural Network Signal Propagation

AutoInit aims to stabilize signal propagation throughout an entire neural network. More precisely, consider a layer that shifts its input by α and scales the input by a factor of β . Given an input signal with mean μ_{in} and variance ν_{in} , after applying the layer, the output signal will have mean $\mu_{\text{out}} = \alpha + \beta\mu_{\text{in}}$ and variance $\nu_{\text{out}} = \beta^2\nu_{\text{in}}$. In a deep network in which the layer is applied L times the effect is compounded and the signal at the final layer has mean and variance

$$\mu_{\text{out}} = \beta^L \mu_{\text{in}} + \alpha(\beta^L + \beta^{L-1} + \dots + \beta + 1), \quad \nu_{\text{out}} = \beta^{2L} \nu_{\text{in}}. \quad (1)$$

If $|\beta| > 1$, the network will suffer from a mean shift and exploding signals as it increases in depth:

$$\lim_{L \rightarrow \infty} \mu_{\text{out}} = \infty, \quad \lim_{L \rightarrow \infty} \nu_{\text{out}} = \infty. \quad (2)$$

In the case that $|\beta| < 1$, the network will suffer from a mean shift and vanishing signals:

$$\lim_{L \rightarrow \infty} \mu_{\text{out}} = \alpha/(1 - \beta), \quad \lim_{L \rightarrow \infty} \nu_{\text{out}} = 0. \quad (3)$$

AutoInit calculates analytic mean- and variance-preserving weight initialization so that $\alpha = 0$ and $\beta = 1$, thus avoiding the issues of mean shift and exploding/vanishing signals.

4 The AutoInit Framework

AutoInit is a general framework that adapts to different layer types. Its implementation is outlined in Algorithm 1. A given layer in a neural network receives as its input a tensor x with mean μ_{in} and variance ν_{in} . After applying the layer, the output tensor has mean $\mu_{\text{out}} = \mathbb{E}(\text{layer}(x))$ and variance $\nu_{\text{out}} = \text{Var}(\text{layer}(x))$. The function g_{layer} maps input mean and variance to output mean and variance when the layer is applied:

$$g_{\text{layer}} : (\mu_{\text{in}}, \nu_{\text{in}}) \mapsto (\mu_{\text{out}}, \nu_{\text{out}}). \quad (4)$$

Note that g in Equation 4 depends on the type of layer; e.g. g_{Dropout} and g_{ReLU} are different functions. For layers with trainable weights, the mean and variance mapping will depend on those weights. For example, the function $g_{\text{Conv2D}, \theta}$ maps input mean and variance to output mean and variance after the application of a Conv2D layer parameterized by weights θ . Deriving g for all layers makes it possible to model signal propagation across an entire neural network. Thus, if μ_{in} and ν_{in} are known, it is natural to calculate initial weights θ such that the layer output will have zero mean and unit variance.

For example, for Conv2D layers, one possibility is

$$\theta \sim \mathcal{N}\left(0, 1/\sqrt{\text{fan_in}(\nu_{\text{in}} + \mu_{\text{in}}^2)}\right) \implies g_{\text{Conv2D}, \theta}(\mu_{\text{in}}, \nu_{\text{in}}) = (0, 1). \quad (5)$$

Algorithm 1: AutoInit

Input: Network with layers L , directed edges E
 $\text{output_layers} = \{l \in L \mid (l, l') \notin E \forall l' \in L\}$
for output_layer **in** output_layers **do**
 $\text{initialize}(\text{output_layer})$
def $\text{initialize}(\text{layer})$:
 $\text{layers_in} = \{l \in L \mid (l, \text{layer}) \in E\}$
 $i = 1$
 for layer_in **in** layers_in **do**
 $\mu_{\text{in}_i}, \nu_{\text{in}_i} = \text{initialize}(\text{layer_in})$
 $i = i + 1$
 $\mu_{\text{in}} = (\mu_{\text{in}_1}, \mu_{\text{in}_2}, \dots, \mu_{\text{in}_N})$
 $\nu_{\text{in}} = (\nu_{\text{in}_1}, \nu_{\text{in}_2}, \dots, \nu_{\text{in}_N})$
 if layer has weights θ **then**
 $\text{initialize } \theta \text{ s.t. } g_{\text{layer}, \theta}(\mu_{\text{in}}, \nu_{\text{in}}) = (0, 1)$
 $\mu_{\text{out}}, \nu_{\text{out}} = 0, 1$
 else
 $\mu_{\text{out}}, \nu_{\text{out}} = g_{\text{layer}}(\mu_{\text{in}}, \nu_{\text{in}})$
 return $\mu_{\text{out}}, \nu_{\text{out}}$

The AutoInit framework includes mean and variance mapping functions g for the majority of layers used in modern architectures. Appendix A details how these functions and the corresponding initialization strategies (e.g. Equation 5) are derived. New layers can be included by deriving g manually, or by approximating it through Monte Carlo simulation. This approach ensures that reliable estimates for μ_{in} and ν_{in} are available at all layers in a network, which in turn allows for weight initialization that stabilizes the signals to have zero mean and unit variance, avoiding the issues of mean shift and exploding/vanishing signals (Equations 2 and 3).

The main advantage of AutoInit is that it is a general method. Unlike prior work, which imposes design constraints, AutoInit adapts to different settings automatically in order to improve performance. The next seven sections demonstrate this adaptability experimentally from several perspectives: different classes of models (convolutional, residual, transformer), hyperparameter settings (activation function, dropout rate, weight decay, learning rate, optimizer), model depths (nine layer CNN to 812 layer ResNet), image sizes (ten-class 28×28 grayscale to 1,000-class 160×160 RGB), and data modalities (vision, language, tabular, multi-task, transfer learning). AutoInit also outperforms data-dependent initialization methods and stabilizes convolutional, residual, and transformer networks without normalization layers. This generality is shown to be particularly useful in neural architecture search and activation function discovery, where thousands of new designs need to be evaluated robustly. AutoInit produces specialized weight initialization strategies for each candidate, which allows for measuring their performance more accurately. As a result, better solutions are discovered. The experiments thus show that AutoInit is an effective initialization algorithm for existing networks as well as a good starting point for networks that may be developed in the future.

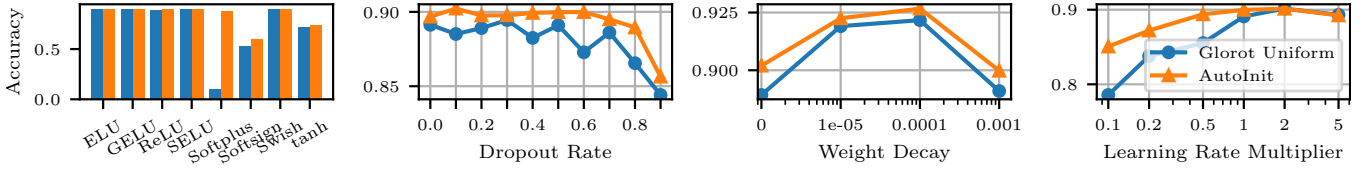


Figure 1: All-CNN-C test accuracy on CIFAR-10. AutoInit results in comparable or better performance with different activation functions, better performance across all dropout rates and weight decay settings, and is less sensitive to the choice of learning rate than the default initialization.

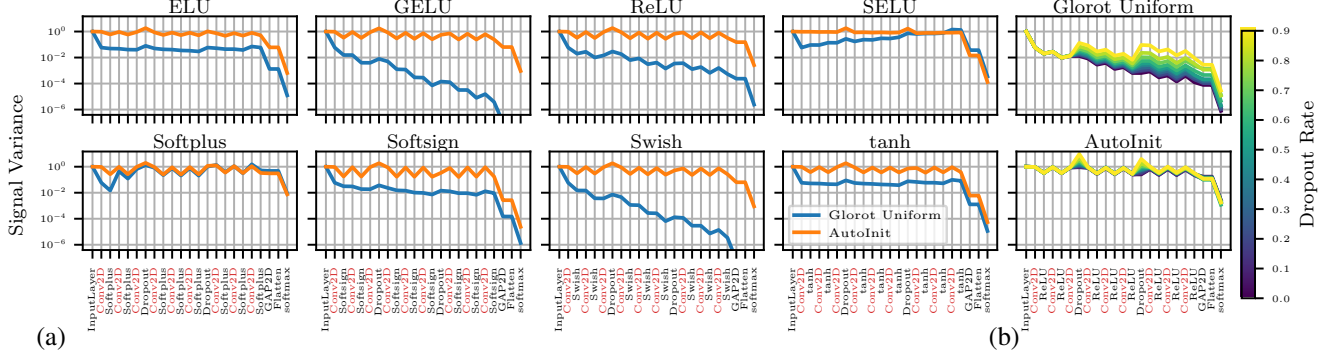


Figure 2: Signal propagation in All-CNN-C networks with different (a) activation functions and (b) dropout rates. With the default initialization, signals often vanish with depth, and their behavior is inconsistent across activation functions and dropout rates. With AutoInit, the variance fluctuates naturally as each layer modifies its input. At layers with weights (marked in red), AutoInit scales the weights appropriately to return the variance to approximately 1.0, stabilizing training in each case.

5 Hyperparameter Variation in CNNs

Experiment Setup The first experiment tests AutoInit’s performance across a range of hyperparameter values for CNNs. The experiment focuses on the All-CNN-C architecture (Springenberg et al. 2015), which consists of convolutional layers, ReLU activation functions, dropout layers, and a global average pooling layer at the end of the network. This simple design helps identify performance gains that can be attributed to proper weight initialization. The network is trained on the CIFAR-10 dataset (Krizhevsky, Hinton et al. 2009) using the standard setup (Appendix B). In particular, the baseline comparison is the “Glorot Uniform” strategy (also called Xavier initialization; Glorot and Bengio 2010), where weights are sampled from $\mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{f_{\text{an.in}}+f_{\text{an.out}}}}, \frac{\sqrt{6}}{\sqrt{f_{\text{an.in}}+f_{\text{an.out}}}}\right)$.

Hyperparameter Variation In separate experiments, the activation function, dropout rate, weight decay, and learning rate multiplier were changed. While one hyperparameter was varied, the others were fixed to the default values.

Results Figure 1 shows the performance of the network with the default initialization and with AutoInit in these different settings. In sum, AutoInit improved performance in every hyperparameter variation evaluated. As Figure 2 shows, AutoInit is adaptive. It alters the initialization to account for different activation functions and dropout rates automatically.

AutoInit is also robust. Even as other hyperparameters like learning rate and weight decay change, AutoInit still results

in a higher performing network than the default initialization. The results thus suggest that AutoInit provides an improved default initialization for convolutional neural networks.

6 Stability in Deep ResNets

This section expands the experimental analysis of AutoInit to residual networks, focusing on preactivation residual networks of various depths (He et al. 2016b). The training setup is standard unless explicitly stated otherwise (Appendix B). In particular, the initialization is “He Normal” (He et al. 2015), where weights are sampled from $\mathcal{N}(0, \sqrt{2/f_{\text{an.in}}})$.

Visualizing Signal Propagation Figure 3 shows how the signal variance changes with depth. With ResNet-56, the variance increases where the shortcut connection and residual branch meet, and the variance drops whenever ReLU is applied. Although the variance increases exponentially with the default initialization and linearly with AutoInit (note the log scale on the y axis), training is still stable because batch normalization layers return the signal to variance 1.0. Without batch normalization, the signal variance never stabilizes under the default initialization. In contrast, removing batch normalization is not an issue with AutoInit; the signal variance remains stable with depth.

With the deeper ResNet-164 and ResNet-812 networks, the conclusions are similar but more pronounced. In the case of ResNet-812 without batch normalization, the signals explode so severely that they exceed machine precision. AutoInit avoids this issue entirely.

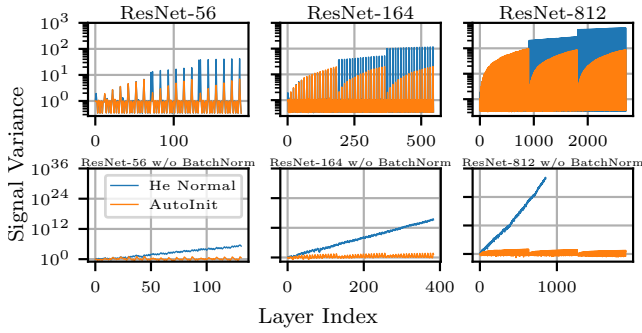


Figure 3: Signal propagation in residual networks. Gaussian input was fed to the networks and empirical variance computed at each layer. Since ReLU, BatchNormalization, and Add are counted as individual layers in this diagram, the total number of layers is different from that in the architecture name (i.e. ResNet-164 has 164 convolutional layers but over 500 total layers). The default initialization causes exploding signals, while AutoInit ensures signal propagation is stable.

Stable Initial Learning Rates Exploding or vanishing signals make optimization difficult because they result in gradient updates that are too large to be accurate or too small to be meaningful. This phenomenon can be observed when the network does not exceed chance accuracy. Therefore, a simple way to quantify whether a weight initialization is effective is to observe a network’s performance after a few epochs.

Using this metric, AutoInit was compared against the default initialization by training unnormalized versions of ResNet-56, ResNet-164, and ResNet-812 for five epochs with a variety of learning rates. With the default initialization, ResNet-56 requires a learning rate between 10^{-8} and 0.5×10^{-3} to begin training, but training was not possible with ResNet-164 or ResNet-812 because of exploding signals (Figure 4a). AutoInit stabilizes training for all three networks, and its effect does not diminish with depth. The networks remain stable with higher learning rates between 10^{-4} and 0.05. Such rates speed up learning, and also correlate with better generalization performance (Jastrzebski et al. 2017; Smith et al. 2018; Smith and Le 2018).

Full ResNet Training In the third residual network experiment, ResNet-164 was trained to completion on CIFAR-10 with different learning rate schedules. All schedules included a linear warm-up phase followed by a decay to zero using cosine annealing (Loshchilov and Hutter 2016).

Figure 4b displays the performance with a variety of such schedules. When the best learning rate schedules are used, ResNet-164 achieves comparable performance with the default initialization and with AutoInit. However, when a sub-optimal schedule is used, performance degrades more quickly with the default initialization than it does with AutoInit. Without batch normalization, the network requires proper weight initialization for stability. In this case, ResNet-164 with the default initialization fails to train regardless of the learning rate schedule, whereas AutoInit results in high accuracy for the majority of them.

Together, the experiments in this section show that Au-

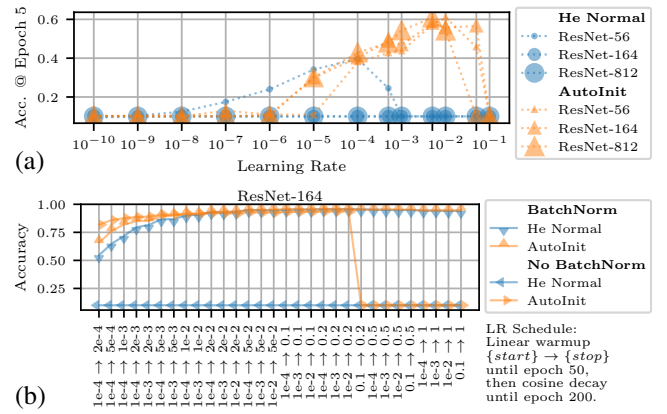


Figure 4: ResNet accuracy on CIFAR-10 with different settings. (a) Accuracy of unnormalized ResNet architectures after five epochs of training with different learning rates and weight initializations. While default initialization makes training difficult in ResNet-56 and impossible at greater depths, AutoInit results in consistent training at all depths. (b) Accuracy of ResNet-164 with a variety of learning rate schedules and initializations. AutoInit is comparable to or outperforms the default initialization in every case.

CoAtNet	w/ GELU	w/ ReLU	w/ SELU	w/ Swish	w/o Norm
Default Init.	89.38	89.22	86.09	88.69	-
Glorot Normal	91.44	91.54	87.59	90.42	85.89
Glorot Uniform	91.16	91.18	88.25	90.06	85.73
He Normal	88.48	88.05	86.11	88.36	-
He Uniform	88.66	87.87	86.37	88.41	-
LeCun Normal	91.11	90.57	87.80	90.83	-
LeCun Uniform	90.55	90.65	87.67	90.57	-
AutoInit	92.48	92.15	86.80	92.28	85.73

Table 1: CoAtNet top-1 accuracy on Imagenette, shown as median of three runs. The first four experiments vary the activation function, while the fifth removes all normalization layers from the architecture. A “-” indicates that training diverged. AutoInit produces the best model in three of the five settings, and remains stable even without normalization layers.

toInit is effective with deep networks. It prevents signals from exploding or vanishing, makes it possible to use larger learning rates, and achieves high accuracy, with and without batch normalization.

7 High-Resolution Images with Transformers

This section extends AutoInit to transformer architectures and applies them to high-resolution image classification. Specifically, AutoInit is applied to CoAtNet, a model that combines convolutional and attention layers (Dai et al. 2021). The model is trained on Imagenette, a subset of 10 classes from the ImageNet dataset (Howard 2019; Deng et al. 2009). Imagenette allows evaluating AutoInit in a high-resolution image classification task with a $132\times$ smaller carbon footprint than the full ImageNet dataset would (Appendix F). As Table 1 shows, AutoInit outperforms six commonly used initialization schemes as well as the de-

	top-1	top-5
Default Init.	74.33	91.60
AutoInit	75.35	92.03

Table 2: ResNet-50 top-1 and top-5 validation accuracy on ImageNet. AutoInit improves performance, even with large and challenging datasets.

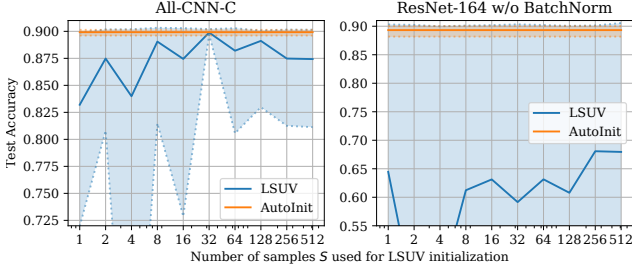


Figure 5: Mean CIFAR-10 test accuracy for AutoInit vs. LSUV with different numbers of samples S . Each evaluation is repeated 10 times; the shaded area shows the maximum and minimum accuracy among all trials. AutoInit is consistent, but LSUV struggles when S is small or the network is deep.

fault initialization, which initializes convolutional layers from $\mathcal{N}(0, \sqrt{2/\text{fan.out}})$ and fully-connected layers from $\mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{\text{fan.in} + \text{fan.out}}}, \frac{\sqrt{6}}{\sqrt{\text{fan.in} + \text{fan.out}}}\right)$. Furthermore, AutoInit stabilizes the network even when normalization layers are removed, suggesting that it is a promising candidate towards developing normalizer-free transformer architectures. Full experiment details are in Appendix B.

8 Scaling up to ImageNet

In order to compliment the results from Section 7 and demonstrate that AutoInit can scale to more difficult tasks, ResNet-50 was trained from scratch on ImageNet with the default initialization and with AutoInit. As Table 2 shows, AutoInit improves top-1 and top-5 accuracy in this task as well. Full training details are in Appendix B.

9 Contrast with Data-Dependent Initialization

The layer-sequential unit-variance (LSUV) algorithm is the most natural data-dependent initialization comparison to AutoInit because both approaches aim to scale the weights appropriately in an architecture-agnostic way. LSUV pre-initializes the weights with an existing approach, feeds S training samples through the network, and adjusts the scale of the weights so that each layer’s output variance is approximately one (Mishkin and Matas 2015).

Data-dependent initialization is time-consuming for large S (indeed, even $S = 1$ is used in practice (Kingma and Dhariwal 2018)). However, if S is too small, the samples may not reflect the statistics of the dataset accurately, leading to poor initialization. Figure 5 demonstrates this phenomenon. In

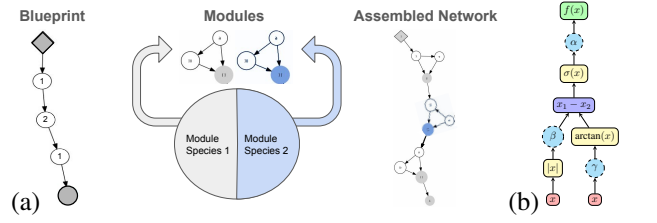


Figure 6: (a) The CoDeepNEAT method. Modules replace nodes in the blueprint to create a candidate neural network. (b) An example activation function created with the PANGAEA method. The computation graph represents the parametric function $f(x) = \alpha \cdot \sigma(\beta \cdot |x| - \arctan(\gamma \cdot x))$. CoDeepNEAT and PANGAEA generate a variety of architectures and activation functions that can be used to evaluate AutoInit’s generality and flexibility.

some training runs LSUV matches the performance of AutoInit, but in many instances the randomly selected samples do not accurately reflect the overall dataset and performance suffers. Since AutoInit is not data-dependent, it does not have this issue. Details of this experiment are in Appendix B.

10 Enabling Neural Architecture Search

Sections 5 through 9 demonstrated that AutoInit works well for convolutional, residual, and transformer networks with a variety of hyperparameter values and depths. In this section, the results are extended to a broader variety of network topologies and types of tasks, for two reasons. First, whereas custom weight initialization may be developed by hand for the most popular machine learning benchmarks, it is unlikely to happen for a variety of architectures and tasks beyond them. Second, as new types of neural network designs are developed in the future, it will be important to initialize them properly to reduce uncertainty in their performance. This section evaluates the generality of AutoInit by applying it to the variety of networks generated in a neural architecture search process with five types of tasks.

The CoDeepNEAT Architecture Search Method Neural networks are evolved using CoDeepNEAT (Liang et al. 2019; Miikkulainen et al. 2019). CoDeepNEAT extends previous work on evolving network topologies and weights (Moriarty and Miikkulainen 1997; Stanley and Miikkulainen 2002) to the level of evolving deep learning architectures. Utilizing a cooperative coevolution framework (Potter and Jong 2000), CoDeepNEAT evolves populations of modules and blueprints simultaneously (Figure 6a). Modules are small neural networks, complete with layers, connections, and hyperparameters. Blueprints are computation graphs containing only nodes and directed edges. To create a candidate neural network, CoDeepNEAT chooses a blueprint and replaces its nodes with selected modules. This mechanism makes it possible to evolve deep, complex, and recurrent structures, while taking advantage of the modularity often found in state-of-the-art models. In addition to the network structure, CoDeepNEAT evolves hyperparameters like dropout rate, kernel regularization, and learning rate. The network weights are not

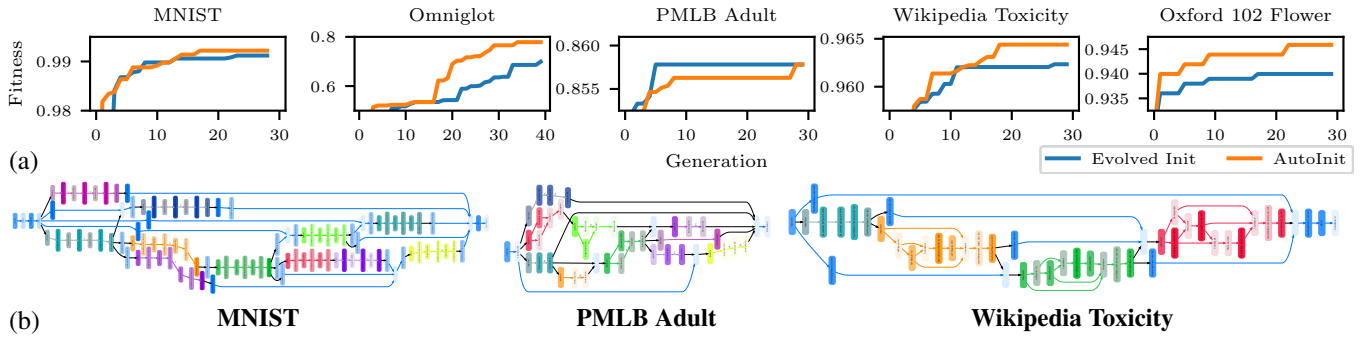


Figure 7: Evaluation of AutoInit with neural architecture search. (a) Performance improvement over generations in the five tasks. AutoInit outperforms the evolved initialization on four tasks and matches it on one. (b) Representative networks evolved with AutoInit. Although the networks are distinct, AutoInit initializes them properly, leading to good performance in each case.

evolved, but instead trained with gradient descent. The generality of CoDeepNEAT helps minimize human design biases and makes it well-suited to analyzing AutoInit’s performance in a variety of open-ended machine learning settings.

Tasks Using CoDeepNEAT, networks are evolved for their performance in vision (MNIST), language (Wikipedia Toxicity), tabular (PMLB Adult), multi-task (Omniglot), and transfer learning (Oxford 102 Flower) tasks (Appendix C).

Results Figure 7a shows how CoDeepNEAT discovers progressively better networks over time on the five tasks. Evolution often selects different weight initialization strategies for the different layers in these networks, so this scheme is already a flexible and powerful baseline. However, by accounting for each model’s unique topology and hyperparameters, AutoInit outperforms the baseline in four of the five tasks, and matches it in the fifth.

Beyond performance, three interesting phenomena can be observed. First, the mean population fitness varies greatly with the default initialization in each task, sometimes dropping significantly from one generation to the next (Figure 9 in Appendix C). Though some variation is natural in a stochastic evolutionary process like CoDeepNEAT, AutoInit makes the discovery process more reliable by stabilizing the performance of the entire population.

Second, hyperparameters play a large role in the final performance of the dense networks, in particular in the “Oxford 102 Flower” task. While CoDeepNEAT discovers good models with both initialization strategies, performance is consistently higher with AutoInit. This finding agrees with Section 5, where AutoInit was shown to be robust to different hyperparameter values.

Third, while many networks exhibit motifs popular in existing architectures, such as alternating convolution and dropout layers and utilizing residual connections, other phenomena are less common (Figure 7b). For example, the networks make use of different activation functions and contain several unique information processing paths from the input to the output. Because AutoInit provides effective initialization in each of these cases, it allows for taking full advantage of unusual design choices that might otherwise hurt performance under default initialization schemes.

The results in this section suggest that AutoInit is an effective, general-purpose algorithm that provides a more effective initialization than existing approaches when searching for new models.

11 Enabling Activation Function Discovery

As new activation functions are developed in the future, it will be important to adjust weight initialization to maintain stable signal propagation. Since AutoInit makes this adjustment automatically, it is well-suited to the task. Indeed, Figure 1 confirmed that AutoInit improves performance with several existing activation functions. This section presents a more challenging task. To simulate future research in activation function design, hundreds of novel activation functions were generated as arbitrary computation graphs and trained with a CNN. AutoInit’s ability to initialize each of these networks was then evaluated. The method for creating such activation functions is described first, followed by experimental details, and results on stability, performance, and generality.

Creating Novel Activation Functions An important area of automated machine learning (AutoML) is to discover new, better activation functions (Basirat and Roth 2018; Ramachandran, Zoph, and Le 2018; Bingham, Macke, and Miikkulainen 2020; Liu et al. 2020). Among existing approaches, PANGAEA (Bingham and Miikkulainen 2022) has the most flexible search space and is therefore used to generate new functions in this section.

PANGAEA represents activation functions as computation graphs containing unary and binary operators (Figure 6b). Creating a novel activation function involves three steps. First, a minimal computation graph is initialized with randomly selected unary and binary operators. Second, the functional form of the activation function is modified by applying three random mutations to increase diversity. Third, the function is augmented with up to three learnable parameters. These parameters are analogous to those in other parametric activation functions, such as PReLU (He et al. 2015); they are initialized to one and learned during training by gradient descent. Through this process, it is possible to understand to what extent AutoInit can improve performance with activation functions that have yet to be discovered.

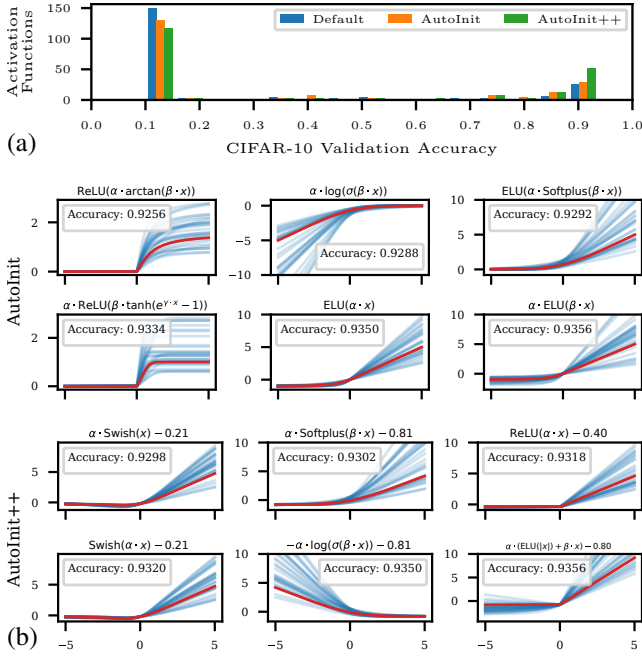


Figure 8: Evaluation of AutoInit with activation function discovery. (a) Distribution of accuracies achieved with 200 activation functions and different weight initialization strategies. AutoInit and AutoInit++ make training more stable and allow more high-performing activation functions to be discovered than the default initialization does. (b) High-performing activation functions. The red line shows the function at initialization, with $\alpha = \beta = \gamma = 1$. The blue lines show the shapes the activation function takes during training, created by sampling α, β, γ from $\mathcal{U}(0.5, 2.0)$. AutoInit’s flexibility should turn out useful for developing new activation functions in the future.

Experimental Setup An important insight in this domain is that in addition to modifying the variance of the signals in a network, activation functions can induce mean shifts. Prior work encouraged stability by reparameterizing the weights to have zero empirical mean (Huang et al. 2017b; Qiao et al. 2019; Brock, De, and Smith 2021). An alternative and more direct approach is to modify the activation function itself so that it does not cause a mean shift in the first place. Given an activation function f with Gaussian mean $\mu_f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-x^2/2} dx$, this goal can be accomplished with $\tilde{f} := f - \mu_f$, which has zero Gaussian mean. To take advantage of this idea, a version of AutoInit called AutoInit++ was created for this domain, thus extending AutoInit slightly beyond weight initialization.

Thus, three initialization strategies were compared. With the default initialization, weights were sampled from $\mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{f_{\text{an.in}} + f_{\text{an.out}}}}, \frac{\sqrt{6}}{\sqrt{f_{\text{an.in}} + f_{\text{an.out}}}}\right)$ (Glorot and Bengio 2010). With AutoInit, the weights were sampled from $\mathcal{N}(0, 1/\sqrt{f_{\text{an.in}}\mu_f})$ to account for an arbitrary activation function f ; the dropout adjustment (Section 5) was not used.

Finally, AutoInit++ takes advantage of \tilde{f} as described above, but is otherwise identical to AutoInit.

For each initialization strategy, 200 activation functions were created using the PANGAEA process. Each activation function was used with the All-CNN-C architecture on the CIFAR-10 dataset following the standard training setup. To avoid overfitting to the test set when evaluating such a large number of activation functions, the accuracy with a balanced validation set of 5000 images is reported instead.

Stability Achieving better-than-chance accuracy is a useful metric of training stability (Section 6). As shown in Figure 8a, many activation functions result in chance accuracy regardless of how the network is initialized. This phenomenon is not surprising; since the activation functions are arbitrary computation graphs, many of them will turn out to be poor. With the default initialization strategy, 149 activation functions caused training to fail in this way. With AutoInit, the number of failed activation functions dropped to 130, and with AutoInit++, it further decreased to 117. AutoInit and AutoInit++ thus make training more stable, allowing it to succeed for a greater number of activation functions.

Performance Beyond training stability, a good weight initialization should also improve performance. As a baseline, when trained with ReLU and the default initialization, All-CNN-C achieved 89.10% test accuracy. Twenty-two of 200 activation functions from the PANGAEA search space outperformed this accuracy with the default initialization. With AutoInit, this number increased to 26, and with AutoInit++, to 50—a notable improvement. Thus, with the default initialization, one can naively create a randomly generated computation graph activation function and have roughly a one in nine chance of outperforming ReLU, but with AutoInit++, this probability increases to one in four.

Indeed, the Mann-Whitney U test (Mann and Whitney 1947) concludes that the distribution of accuracies induced by AutoInit++ is *stochastically larger* than that from AutoInit ($p < 0.05$) or the default initialization ($p < 0.01$). This result means that for any level of performance, it is always more probable to discover an activation function that achieves that level of performance when initializing with AutoInit++ versus AutoInit or the default initialization. The result implies that activation function researchers who properly initialize their networks are more likely to discover state-of-the-art activation functions, while staying with the default initialization may hinder that research effort. More detailed statistical significance analyses are included in Appendix D.

Generality Figure 8b plots several activation functions from the PANGAEA search space. Many discovered functions have similar shapes to existing functions. However, others are nonmonotonic, have discontinuous derivatives, or saturate to nonzero values. These properties are less common in existing activation functions. This observation suggests that AutoInit is a general approach that does not depend on a specific type of activation function; it may therefore serve as a useful tool in developing new such functions in the future.

12 Future Work

Experiments in Other Domains The experiments in this paper demonstrate that AutoInit can improve performance in a variety of settings, suggesting that it can be applied to other domains as well. For instance in reinforcement learning, good estimates of activation statistics are usually not available due to the online nature of the algorithm. It is not possible to stabilize training using e.g. batch normalization, but it may be possible to do it with AutoInit. Similarly, training of generative adversarial networks (Goodfellow et al. 2014) is often unstable, and proper initialization may help. Applying AutoInit to such different domains should not only make them more reliable, but also lead to a better understanding of their training dynamics.

Accelerating Model Search In Sections 10 and 11, AutoInit was shown to facilitate the discovery of better neural network designs and activation functions. This ability is possible because AutoInit is a general method, i.e. not restricted to a single class of models, and it could similarly augment other meta-learning algorithms (e.g. those reviewed by Elsken, Metzen, and Hutter 2019; Wistuba, Rawat, and Pedapati 2019).

However, this finding points to an even more promising idea. As model search techniques become more prevalent in real-world applications, it will be most worthwhile to derive general principles rather than specific instantiations of those principles. For example, past weight initialization strategies improved performance with specific activation functions through manual derivation of appropriate weight scaling (Section 2). In contrast, AutoInit is a general method, leveraging Gaussian quadrature for any activation function. Similarly, AutoInit resulted in better initialization than strategies discovered by CoDeepNEAT through evolution (Section 10). Further, AutoInit++ (Section 11), rather than producing a few high-performing activation functions, introduces the general property that activation functions with zero Gaussian mean ($\tilde{f} := f - \mu_f$) tend to perform well. This property discovered a highly diverse set of powerful activation functions in the PANGAEA search space (Figure 8).

Thus, AutoInit is successful because it is not a single initialization strategy, but rather a mapping from architectures to initialization strategies. Such mappings, whether focused on initialization or some other aspect of model design, deserve increased attention in the future. They can lead to performance gains in a variety of scenarios. They also accelerate model search by focusing the search space to more promising regions. If one does not have to worry about discovering a good initialization, compute power can instead be used in other areas, like designing architectures and activation functions. Thus, general tools like AutoInit save time and resources, and lead to better models as a result.

Further technical extensions to AutoInit are outlined in Appendix E), including variations on the core AutoInit algorithm, support for new layer types, and integration with deep learning frameworks.

13 Conclusion

This paper introduced AutoInit, an algorithm that calculates analytic mean- and variance-preserving weight initialization for neural networks automatically. In convolutional networks, the initialization improved performance with different activation functions, dropout rates, learning rates, and weight decay settings. In residual networks, AutoInit prevented exploding signals, allowed training with higher learning rates, and improved performance with or without batch normalization. In transformers, AutoInit was scaled up to high-resolution image classification, and improved performance with several activation functions with and without normalization. AutoInit also improved accuracy on the ImageNet dataset. The initialization is independent of data and is therefore efficient and reliable. AutoInit’s generality proved instrumental in two types of AutoML. In neural architecture search, new architectures were evaluated more accurately, resulting in better networks in vision, language, tabular, multi-task, and transfer learning settings. In activation function discovery, AutoInit stabilized training and improved accuracy with a large diversity of novel activation functions. Thus, AutoInit serves to make machine learning experiments more robust and reliable, resulting in higher performance, and facilitating future research in AutoML.

References

- Abadi, M.; Barham, P.; Chen, J.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, 265–283.
- Apicella, A.; Donnarumma, F.; Isgrò, F.; and Prevete, R. 2021. A survey on modern trainable activation functions. *Neural Networks*.
- Arpit, D.; Campos, V.; and Bengio, Y. 2019. How to initialize your network? robust initialization for weightnorm & resnets. *arXiv:1906.02341*.
- Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv:1607.06450*.
- Bachlechner, T.; Majumder, B. P.; Mao, H. H.; Cottrell, G. W.; and McAuley, J. 2020. Rezero is all you need: Fast convergence at large depth. *arXiv:2003.04887*.
- Basirat, M.; and Roth, P. M. 2018. The quest for the golden activation function. *arXiv:1808.00783*.
- Bingham, G.; Macke, W.; and Miikkulainen, R. 2020. Evolutionary Optimization of Deep Learning Activation Functions. In *Genetic and Evolutionary Computation Conference*.
- Bingham, G.; and Miikkulainen, R. 2022. Discovering parametric activation functions. *Neural Networks*, 148: 48–65.
- Brock, A.; De, S.; and Smith, S. L. 2021. Characterizing signal propagation to close the performance gap in unnormalized ResNets. *arXiv:2101.08692*.
- Chen, M.; Pennington, J.; and Schoenholz, S. 2018. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning*, 873–882. PMLR.

- Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.
- Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *CoRR*, abs/1511.07289.
- Courbariaux, M.; Bengio, Y.; and David, J.-P. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Neural information processing systems*, 3123–3131.
- Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.
- Dai, Z.; Liu, H.; Le, Q. V.; and Tan, M. 2021. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34: 3965–3977.
- Dauphin, Y. N.; and Schoenholz, S. 2019. MetaInit: Initializing learning by learning to initialize. *Advances in Neural Information Processing Systems*, 32.
- De, S.; and Smith, S. 2020. Batch normalization biases residual blocks towards the identity function in deep networks. *Neural Information Processing Systems*, 33.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Li, F.-F. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107: 3–11.
- Elsken, T.; Metzen, J. H.; and Hutter, F. 2019. Neural Architecture Search: A Survey. *Journal of Machine Learning Research*, 20(55): 1–21.
- Epps, T.; and Singleton, K. J. 1986. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4): 177–203.
- Gilboa, D.; Chang, B.; Chen, M.; Yang, G.; Schoenholz, S. S.; Chi, H.; and Pennington, J. 2019. The Dynamics of Signal Propagation in Gated Recurrent Neural Networks.
- Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *13th international conference on artificial intelligence and statistics*, 249–256.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Neural information processing systems*, 27.
- Hanin, B. 2018. Which neural net architectures give rise to exploding and vanishing gradients? In *32nd International Conference on Neural Information Processing Systems*, 580–589.
- Hanin, B.; and Rolnick, D. 2018. How to start training: The effect of initialization and architecture. *arXiv:1803.01719*.
- Harris, C. R.; Millman, K. J.; van der Walt, S. J.; et al. 2020. Array programming with NumPy. *Nature*, 585(7825): 357–362.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE international conference on computer vision*, 1026–1034.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 770–778.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *European conference on computer vision*, 630–645. Springer.
- Hendrycks, D.; and Gimpel, K. 2016a. Adjusting for dropout variance in batch normalization and weight initialization. *arXiv:1607.02488*.
- Hendrycks, D.; and Gimpel, K. 2016b. Gaussian error linear units (gelus). *arXiv:1606.08415*.
- Hochreiter, S. 1991. Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München*, 91(1).
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Howard, J. 2019. Imagenette. <https://github.com/fastai/imagenette>. Accessed: 2022-07-27.
- Hu, W.; Xiao, L.; and Pennington, J. 2020. Provable benefit of orthogonal initialization in optimizing deep linear networks. *arXiv:2001.05992*.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017a. Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition*, 4700–4708.
- Huang, L.; Liu, X.; Liu, Y.; Lang, B.; and Tao, D. 2017b. Centered weight normalization in accelerating training of deep neural networks. In *IEEE International Conference on Computer Vision*, 2803–2811.
- Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, 448–456.
- Jastrzebski, S.; Kenton, Z.; Arpit, D.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2017. Three factors influencing minima in sgd. *arXiv:1711.04623*.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.
- Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.
- Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-normalizing neural networks. In *Neural information processing systems*, 971–980.
- Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, 202–207.

- Krähenbühl, P.; Doersch, C.; Donahue, J.; and Darrell, T. 2015. Data-dependent initializations of convolutional neural networks. *arXiv:1511.06856*.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.
- Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.
- LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2.
- LeCun, Y. A.; Bottou, L.; Orr, G. B.; and Müller, K.-R. 2012. Efficient backprop. In *Neural networks: Tricks of the trade*, 9–48. Springer.
- Liang, J.; Meyerson, E.; Hodjat, B.; Fink, D.; Mutch, K.; and Miikkulainen, R. 2019. Evolutionary neural automl for deep learning. In *Genetic and Evolutionary Computation Conference*, 401–409.
- Liu, H.; Brock, A.; Simonyan, K.; and Le, Q. V. 2020. Evolving Normalization-Activation Layers. *arXiv:2004.02967*.
- Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv:1608.03983*.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Maas, A. L.; Hannun, A. Y.; and Ng, A. Y. 2013. Rectifier nonlinearities improve neural network acoustic models. In *30th international conference on machine learning (ICML-13)*, 3.
- Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.
- Massey Jr, F. J. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253): 68–78.
- Miikkulainen, R.; Liang, J.; Meyerson, E.; Rawal, A.; Fink, D.; Francon, O.; Raju, B.; Shahrzad, H.; Navruzian, A.; Duffy, N.; et al. 2019. Evolving deep neural networks. In *Artificial intelligence in the age of neural networks and brain computing*, 293–312. Elsevier.
- Mishkin, D.; and Matas, J. 2015. All you need is a good init. *arXiv:1511.06422*.
- Moriarty, D. E.; and Miikkulainen, R. 1997. Forming neural networks through efficient and adaptive coevolution. *Evolutionary computation*, 5(4): 373–399.
- Mutch, K. 2017–2021. Studio Go Runner. <https://hub.docker.com/repository/docker/leafai/studio-go-runner>.
- Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *27th international conference on machine learning (ICML-10)*, 807–814.
- Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *The Indian Conference on Computer Vision, Graphics and Image Processing*.
- Nwankpa, C.; Ijomah, W.; Gachagan, A.; and Marshall, S. 2018. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv:1811.03378*.
- Olson, R. S.; La Cava, W.; Orzechowski, P.; Urbanowicz, R. J.; and Moore, J. H. 2017. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 10(1): 36.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Neural information processing systems*, 32: 8026–8037.
- Pennington, J.; Schoenholz, S. S.; and Ganguli, S. 2017. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *31st International Conference on Neural Information Processing Systems*, 4788–4798.
- Pennington, J.; and Worah, P. 2018. The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network. In *NeurIPS*, 5415–5424.
- Pennington, J.; and Worah, P. 2019. Nonlinear random matrix theory for deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124005.
- Piessens, R.; de Doncker-Kapenga, E.; Überhuber, C. W.; and Kahaner, D. K. 2012. *QUADPACK: A subroutine package for automatic integration*, volume 1. Springer Science & Business Media.
- Potter, M. A.; and Jong, K. A. D. 2000. Cooperative coevolution: An architecture for evolving coadapted subcomponents. *Evolutionary computation*, 8(1): 1–29.
- Qiao, S.; Wang, H.; Liu, C.; Shen, W.; and Yuille, A. 2019. Micro-batch training with batch-channel normalization and weight standardization. *arXiv:1903.10520*.
- Radosavovic, I.; Johnson, J.; Xie, S.; Lo, W.-Y.; and Dollár, P. 2019. On network design spaces for visual recognition. In *IEEE/CVF International Conference on Computer Vision*, 1882–1890.
- Ramachandran, P.; Zoph, B.; and Le, Q. V. 2018. Searching for Activation Functions. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*.
- Salimans, T.; and Kingma, D. P. 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Neural information processing systems*, 29: 901–909.
- Saxe, A. M.; McClelland, J. L.; and Ganguli, S. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120*.
- Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 1100612. International Society for Optics and Photonics.
- Smith, S. L.; Kindermans, P.-J.; Ying, C.; and Le, Q. V. 2018. Don't Decay the Learning Rate, Increase the Batch Size. In *International Conference on Learning Representations*.

- Smith, S. L.; and Le, Q. V. 2018. A Bayesian Perspective on Generalization and Stochastic Gradient Descent. In *International Conference on Learning Representations*.
- Springenberg, J.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (workshop track)*.
- Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1): 1929–1958.
- Stanley, K. O.; and Miikkulainen, R. 2002. Evolving neural networks through augmenting topologies. *Evolutionary computation*, 10(2): 99–127.
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.
- Taki, M. 2017. Deep residual networks and weight initialization. *arXiv:1709.02956*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Neural information processing systems*, 5998–6008.
- Virtanen, P.; Gommers, R.; Oliphant, T. E.; et al. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.
- Wightman, R.; Touvron, H.; and Jégou, H. 2021. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*.
- Wistuba, M.; Rawat, A.; and Pedapati, T. 2019. A Survey on Neural Architecture Search. *arXiv:1905.01392*.
- Wu, Y.; and He, K. 2018. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.
- Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *26th international conference on world wide web*, 1391–1399.
- Xiao, L.; Bahri, Y.; Sohl-Dickstein, J.; Schoenholz, S.; and Pennington, J. 2018. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, 5393–5402. PMLR.
- Xie, C.; Tan, M.; Gong, B.; Yuille, A.; and Le, Q. V. 2020. Smooth adversarial training. *arXiv:2006.14536*.
- Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.
- Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv:1605.07146*.
- Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.
- Zhang, H.; Dauphin, Y. N.; and Ma, T. 2019. Fixup initialization: Residual learning without normalization. *arXiv:1901.09321*.
- Zhokhov, P.; Denisov, A.; and Mutch, K. 2017–2021. Studio. <https://studio.ml>.
- Zhu, C.; Ni, R.; Xu, Z.; Kong, K.; Huang, W. R.; and Goldstein, T. 2021. GradInit: Learning to initialize neural networks for stable and efficient training. *Advances in Neural Information Processing Systems*, 34: 16410–16422.

A Mean and Variance Estimation for Different Layer Types

In the AutoInit framework of Algorithm 1, the mean and variance mapping function g needs to be defined for each type of layer in a given neural network. This appendix presents derivations for a majority of the most commonly used layers available in the TensorFlow package (Abadi et al. 2016) at time of writing. Extending AutoInit to support new layers in the future will require deriving the function g for those layers. Monte Carlo sampling can also be used as an approximation for g before it is manually derived.

In the following paragraphs, x denotes the input to a layer, and y is the output. The incoming and outgoing means and variances are denoted as $\mu_{\text{in}} := E(x)$, $\mu_{\text{out}} := E(y)$, $\nu_{\text{in}} := \text{Var}(x)$, and $\nu_{\text{out}} := \text{Var}(y)$. The notation $\text{Conv}\{1\text{D}, 2\text{D}, 3\text{D}\}$ is used to refer to Conv1D , Conv2D , and Conv3D , and analogously for other layer types. Inputs to each layer are assumed to be independent and normally distributed. Although these assumptions may not always hold exactly, experiments show that AutoInit models signal propagation across different types of networks well in practice.

Convolution and Dense Layers The analysis in the next paragraph applies to $\text{Conv}\{1\text{D}, 2\text{D}, 3\text{D}\}$, $\text{DepthwiseConv}\{1\text{D}, 2\text{D}\}$, and Dense layers, since convolution layers are dense layers with sparse connectivity. Notation and derivation are inspired by that of Glorot and Bengio (2010) and He et al. (2015).

A feedforward layer can be written as $y = Wx + b$, where x is the input, W is a $\text{fan_out} \times \text{fan_in}$ weight matrix, b is a vector of biases, and y is the result. Assume the elements of W are mutually independent and from the same distribution, and likewise for the elements of x . Further assume that W and x are independent of each other. The outgoing mean can then be written as $\mu_{\text{out}} = E(W)\mu_{\text{in}}$. For the outgoing variance, letting W have zero mean and expanding the product of independent random variables yields $\nu_{\text{out}} = \text{fan_in} \text{Var}(W)(\nu_{\text{in}} + \mu_{\text{in}}^2)$. Sampling the weights W according to

$$W \sim \mathcal{N}\left(0, \frac{1}{\sqrt{\text{fan_in}(\nu_{\text{in}} + \mu_{\text{in}}^2)}}\right) \quad (6)$$

or

$$W \sim \mathcal{U}\left(-\frac{\sqrt{3}}{\sqrt{\text{fan_in}(\nu_{\text{in}} + \mu_{\text{in}}^2)}}, \frac{\sqrt{3}}{\sqrt{\text{fan_in}(\nu_{\text{in}} + \mu_{\text{in}}^2)}}\right) \quad (7)$$

is sufficient to ensure that

$$\mu_{\text{out}} = 0 \text{ and } \nu_{\text{out}} = 1. \quad (8)$$

Activation Functions The analysis in the next paragraph accounts for all activation functions in TensorFlow, including `elu`, `exponential`, `gelu`, `hard_sigmoid`, `LeakyReLU`, `linear`, `PReLU`, `ReLU`, `selu`, `sigmoid`, `softplus`, `softsign`, `swish`, `tanh`, and `ThresholdedReLU` (by Clevert, Unterthiner, and Hochreiter 2015; Hendrycks and Gimpel 2016b; Maas, Hannun, and Ng 2013; He et al. 2015; Nair and Hinton 2010;

Klambauer et al. 2017; Ramachandran, Zoph, and Le 2018; Elfwing, Uchibe, and Doya 2018; Courbariaux, Bengio, and David 2015, respectively), and in fact extends to any integrable Activation function f .

Let $p_{\mathcal{N}}(x; \mu, \sigma)$ denote the probability density function of a Gaussian distribution with mean μ and standard deviation σ . By the law of the unconscious statistician,

$$\mu_{\text{out}} = \int_{-\infty}^{\infty} f(x) p_{\mathcal{N}}(x; \mu_{\text{in}}, \sqrt{\nu_{\text{in}}}) dx, \quad (9)$$

$$\nu_{\text{out}} = \int_{-\infty}^{\infty} f(x)^2 p_{\mathcal{N}}(x; \mu_{\text{in}}, \sqrt{\nu_{\text{in}}}) dx - \mu_{\text{out}}^2. \quad (10)$$

These integrals are computed for an arbitrary activation function f with adaptive quadrature, a well-established numerical integration approach that approximates integrals using adaptively refined subintervals (Piessens et al. 2012; Virtanen et al. 2020; Harris et al. 2020).

Dropout Layers Dropout layers randomly set `rate` percentage of their inputs to zero (Srivastava et al. 2014). Therefore,

$$\mu_{\text{out}} = \mu_{\text{in}}(1 - \text{rate}) \text{ and } \nu_{\text{out}} = \nu_{\text{in}}(1 - \text{rate}). \quad (11)$$

However, this analysis only applies to $\text{SpatialDropout}\{1\text{D}, 2\text{D}, 3\text{D}\}$ layers. For regular Dropout layers, TensorFlow automatically scales the values by $1/(1 - \text{rate})$ to avoid a mean shift towards zero.¹ Adjusting for this change gives

$$\mu_{\text{out}} = \mu_{\text{in}} \text{ and } \nu_{\text{out}} = \nu_{\text{in}}/(1 - \text{rate}). \quad (12)$$

Pooling Layers The same approach applies to all commonly used pooling layers, including $\text{AveragePooling}\{1\text{D}, 2\text{D}, 3\text{D}\}$, $\text{MaxPooling}\{1\text{D}, 2\text{D}, 3\text{D}\}$, $\text{GlobalAveragePooling}\{1\text{D}, 2\text{D}, 3\text{D}\}$, and $\text{GlobalMaxPooling}\{1\text{D}, 2\text{D}, 3\text{D}\}$.

Let $op(\cdot)$ be the average operation for an average pooling layer, and the maximum operation for a max pooling layer. Define K to be the pool size of the layer. For standard 1D, 2D, and 3D pooling layers, K would equal k , $k \times k$, and $k \times k \times k$, respectively. The global pooling layers can be seen as special cases of the standard pooling layers where the pool size is the same size as the input tensor, except along the batch and channel dimensions. Analytically, the outgoing mean and variance can be expressed as

$$\mu_{\text{out}} = \int \cdots \int_{\mathbb{R}^K} op(x_1, x_2, \dots, x_K) \cdot \prod_{i=1}^K p_{\mathcal{N}}(x_i; \mu_{\text{in}}, \sqrt{\nu_{\text{in}}}) dx_1 dx_2 \cdots dx_K, \quad (13)$$

$$\nu_{\text{out}} = \int \cdots \int_{\mathbb{R}^K} op(x_1, x_2, \dots, x_K)^2 \cdot \prod_{i=1}^K p_{\mathcal{N}}(x_i; \mu_{\text{in}}, \sqrt{\nu_{\text{in}}}) dx_1 dx_2 \cdots dx_K - \mu_{\text{out}}^2, \quad (14)$$

¹<https://github.com/tensorflow/tensorflow/blob/v2.5.0/tensorflow/python/keras/layers/core.py#L149-L150>

where the x_i represent tensor entries within a pooling window. Unfortunately, even a modest 3×3 pooling layer requires computing nine nested integrals, which is prohibitively expensive. In this case, a Monte Carlo simulation is appropriate. Sample $x_{1_j}, x_{2_j}, \dots, x_{K_j}$ from $\mathcal{N}(\mu_{\text{in}}, \sqrt{\nu_{\text{in}}})$ for $j = 1, \dots, S$ and return

$$\mu_{\text{out}} = \frac{1}{S} \sum_{j=1}^S \text{op}(x_{1_j}, x_{2_j}, \dots, x_{K_j}), \quad (15)$$

$$\nu_{\text{out}} = \frac{1}{S} \sum_{j=1}^S \text{op}(x_{1_j}, x_{2_j}, \dots, x_{K_j})^2 - \mu_{\text{out}}^2. \quad (16)$$

Normalization Layers BatchNormalization, LayerNormalization, and GroupNormalization normalize the input to have mean zero and variance one (Ioffe and Szegedy 2015; Ba, Kiros, and Hinton 2016; Wu and He 2018). Thus,

$$\mu_{\text{out}} = 0 \text{ and } \nu_{\text{out}} = 1. \quad (17)$$

Arithmetic Operators Assume the input tensors x_1, x_2, \dots, x_N with means $\mu_{\text{in}_1}, \mu_{\text{in}_2}, \dots, \mu_{\text{in}_N}$ and variances $\nu_{\text{in}_1}, \nu_{\text{in}_2}, \dots, \nu_{\text{in}_N}$ are independent. The following mean and variance mapping functions are derived. For an Add layer,

$$\mu_{\text{out}} = \sum_{i=1}^N \mu_{\text{in}_i} \text{ and } \nu_{\text{out}} = \sum_{i=1}^N \nu_{\text{in}_i}. \quad (18)$$

For an Average layer,

$$\mu_{\text{out}} = \frac{1}{N} \sum_{i=1}^N \mu_{\text{in}_i} \text{ and } \nu_{\text{out}} = \frac{1}{N^2} \sum_{i=1}^N \nu_{\text{in}_i}. \quad (19)$$

For a Subtract layer,

$$\mu_{\text{out}} = \mu_{\text{in}_1} - \mu_{\text{in}_2} \text{ and } \nu_{\text{out}} = \nu_{\text{in}_1} + \nu_{\text{in}_2}. \quad (20)$$

Finally, for a Multiply layer,

$$\mu_{\text{out}} = \prod_{i=1}^N \mu_{\text{in}_i} \text{ and } \nu_{\text{out}} = \prod_{i=1}^N (\nu_{\text{in}_i} + \mu_{\text{in}_i}^2) - \prod_{i=1}^N \mu_{\text{in}_i}^2. \quad (21)$$

Concatenation Layers Assume the inputs x_1, x_2, \dots, x_N with means $\mu_{\text{in}_1}, \mu_{\text{in}_2}, \dots, \mu_{\text{in}_N}$ and variances $\nu_{\text{in}_1}, \nu_{\text{in}_2}, \dots, \nu_{\text{in}_N}$ are independent, and let input x_i have C_i elements. Then for a Concatenate layer,

$$\mu_{\text{out}} = \frac{1}{\sum C_i} \sum_{i=1}^N C_i \mu_{\text{in}_i}, \quad (22)$$

$$\nu_{\text{out}} = \frac{1}{\sum C_i} \sum_{i=1}^N C_i (\nu_{\text{in}_i} + \mu_{\text{in}_i}^2) - \mu_{\text{out}}^2. \quad (23)$$

Recurrent Layers A Monte Carlo simulation can be used to estimate the outgoing mean and variance for recurrent layers, including GRU, LSTM, and SimpleRNN (Hochreiter and Schmidhuber 1997; Chung et al. 2014). Recurrent layers

often make use of activation functions like sigmoid and tanh that constrain the scale of the hidden states. Because of this practice, recurrent layers should be initialized with a default scheme or according to recent research in recurrent initialization (Chen, Pennington, and Schoenholz 2018; Gilboa et al. 2019). AutoInit will then estimate the outgoing mean and variance in order to inform appropriate weight scaling elsewhere in the network.

Padding Layers ZeroPadding{1D, 2D, 3D} layers augment the borders of the input tensor with zeros, increasing its size. Let z be the proportion of elements in the tensor that are padded zeros. Then $z = (\text{padded.size} - \text{original.size}) / \text{padded.size}$, and

$$\mu_{\text{out}} = \mu_{\text{in}}(1 - z) \text{ and } \nu_{\text{out}} = \nu_{\text{in}}(1 - z). \quad (24)$$

Shape Adjustment Layers Many layers alter the size or shape of the input tensor but do not change the distribution of the data. These layers include Flatten, Permute, Reshape, UpSampling{1D, 2D, 3D}, and Cropping{1D, 2D, 3D} layers. The same is true of TensorFlow API calls `tf.reshape`, `tf.split`, and `tf.transpose`. For these layers,

$$\mu_{\text{out}} = \mu_{\text{in}} \text{ and } \nu_{\text{out}} = \nu_{\text{in}}. \quad (25)$$

Input Layers An InputLayer simply exposes the model to the data, therefore

$$\mu_{\text{out}} = \mu_{\text{data}} \text{ and } \nu_{\text{out}} = \nu_{\text{data}}. \quad (26)$$

TensorFlow allows nesting models within other models. In this use case where the InputLayer does not directly connect to the training data,

$$\mu_{\text{out}} = \mu_{\text{in}} \text{ and } \nu_{\text{out}} = \nu_{\text{in}}. \quad (27)$$

Matrix Multiplication A call to `tf.matmul` takes input tensors x_1 and x_2 of shape $\dots \times m \times n$ and $\dots \times n \times p$ and produces the output tensor x_{out} of shape $\dots \times m \times p$ with entries

$$x_{\text{out} \dots, i, j} = \sum_{k=1}^n x_{1 \dots, i, k} x_{2 \dots, k, j}. \quad (28)$$

Assuming independent matrix entries, the output statistics can then be calculated as

$$\mu_{\text{out}} = n \mu_{\text{in}_1} \mu_{\text{in}_2}, \quad (29)$$

$$\nu_{\text{out}} = n ((\nu_{\text{in}_1} + \mu_{\text{in}_1}^2)(\nu_{\text{in}_2} + \mu_{\text{in}_2}^2) - \mu_{\text{in}_1}^2 \mu_{\text{in}_2}^2). \quad (30)$$

Reduction Operators A call to `tf.reduce_mean` reduces the size of the input tensor by averaging values along one or more axes. For example, averaging an input tensor of shape $128 \times 8 \times 8 \times 256$ along axes 1 and 2 would produce an output tensor of shape $128 \times 1 \times 1 \times 256$. Let D represent the product of the length of the axes being averaged over (in the example above, $8 \times 8 = 64$). The output tensor has

$$\mu_{\text{out}} = \mu_{\text{in}} \text{ and } \nu_{\text{out}} = \nu_{\text{in}} / D. \quad (31)$$

The function `tf.reduce_sum` performs similarly, summing entries instead of averaging them. In this case,

$$\mu_{\text{out}} = D \mu_{\text{in}} \text{ and } \nu_{\text{out}} = D \nu_{\text{in}}. \quad (32)$$

Maintaining Variance $\neq 1$. In Algorithm 1, AutoInit initializes weights so that the output signal at each layer has mean zero and variance one. Although signal variance $\nu = 1$ is sufficient for stable signal propagation, it is not a necessary condition. Indeed, other values for the signal variance ν could be utilized, as long as ν remains consistent throughout the network. If a different ν is desired, weights can be initialized according to Equation 6 or 7 and then multiplied by $\sqrt{\nu}$. For instance, such a modification was done for the CoAtNet model in Section 7, resulting in slightly improved final performance.

B Convolutional, Residual, and Transformer Network Experiment Details

This appendix contains implementation details for the experiments in Sections 5-9.

All-CNN-C The training setup follows that of Springenberg et al. (2015) as closely as possible. The network is trained with SGD and momentum 0.9. The dropout rate is 0.5 and weight decay as L2 regularization is 0.001. The data augmentation involves featurewise centering and normalizing, random horizontal flips, and random 32×32 crops of images padded with five pixels on all sides. The initial learning rate is 0.01 and is decreased by a factor of 0.1 after epochs 200, 250, and 300 until training ends at epoch 350.

Because Springenberg et al. (2015) did not specify how they initialized their weights, the networks are initialized with the “Glorot Uniform” strategy (also called Xavier initialization; Glorot and Bengio 2010), where weights are sampled from $\mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{f_{\text{an.in}}+f_{\text{an.out}}}}, \frac{\sqrt{6}}{\sqrt{f_{\text{an.in}}+f_{\text{an.out}}}}\right)$. This initialization is the default setting in TensorFlow,² and is sufficient to replicate the results reported by Springenberg et al. (2015).

Residual Networks The networks are optimized with SGD and momentum 0.9. Dropout is not used, and weight decay is 0.0001. Data augmentation includes a random horizontal flip and random 32×32 crops of images padded with four pixels on all sides.

CoAtNet A smaller variant of the CoAtNet architecture³ was used in order to fit the model and data on the available GPU memory. The architecture has three convolutional blocks with 64 channels, four convolutional blocks with 128 channels, six transformer blocks with 256 channels, and three transformer blocks with 512 channels. This architecture is slightly deeper but thinner than the original CoAtNet-0 architecture, which has two convolutional blocks with 96 channels, three convolutional blocks with 192 channels, five transformer blocks with 384 channels, and two transformer blocks with 768 channels (Dai et al. 2021). The models are otherwise identical.

The training hyperparameters were inspired by Wightman, Touvron, and Jégou (2021) and are common in the literature.

²<https://github.com/tensorflow/tensorflow/blob/v2.5.0/tensorflow/python/keras/layers/convolutional.py#L608-L609>

³https://github.com/leondgarse/keras_cv_attention_models/blob/v1.3.0/keras_cv_attention_models/coatnet/coatnet.py#L199

Specifically, images were resized to 160×160 . The learning rate schedule was increased linearly from $1e^{-4}$ to $4e^{-4}$ for six epochs and it then followed a cosine decay until epoch 105. Weight decay was set to 0.02 times the current learning rate at each epoch. The model was trained with batch size 256 and optimized with AdamW (Loshchilov and Hutter 2017). Data augmentation included RandAugment applied twice with a magnitude of six (Cubuk et al. 2020). Mixup and Cutmix were also used with alpha 0.1 and 1.0, respectively (Zhang et al. 2017; Yun et al. 2019). The training images were augmented with random resized crops (Szegedy et al. 2015) that were at minimum 8% of the original image; after training the model was evaluated on 95% center crops.

As discussed in Appendix A, AutoInit maintains signal variance $\nu = 1$, but it is also possible to adjust ν if desired. In the CoAtNet experiments, $\nu = 0.01$ was found to give the best performance among $\nu = \{1, 0.1, 0.01, 0.001\}$. The experiment removing normalization layers used the default of $\nu = 1$.

ImageNet The experiment in Section 8 used the same training setup as the experiments with CoAtNet on Imagenette in the previous paragraph except for two changes. The batch size was 2,048 (512 per GPU across four GPUs), and the maximum learning rate was $3.2e^{-2}$.

Data-Dependent Initialization Comparison In the experiments in Section 9, a learning rate schedule inspired by superconvergence (Smith and Topin 2019) was used to save time. The learning rate increases linearly to 0.1 during the first five epochs, and then decreases linearly for 20 epochs, after which training ends. The weight decay for All-CNN-C was also decreased by a factor of 10. This modification is common when networks are trained with superconvergence (Smith and Topin 2019).

Initialization Time It is important to note that AutoInit does not incur a significant overhead. Each layer must be visited once to be initialized, so the complexity is $O(L)$ where L is the number of layers. For example: All-CNN-C, ResNet-56, and ResNet-164 take 1, 33, and 106 seconds to initialize. The costs are hardware-dependent, but only spent once, and are small compared to the cost of training.

C Neural Architecture Search Experiment Details

This appendix contains implementation details for the experiments in Section 10. The progress of CoDeepNEAT in these experiments was shown in Figure 7 in the main text. Figure 9 shows this same data but in greater detail. Table 3 contains the training hyperparameters, neural network layers, evolutionary hyperparameters, and mutation probabilities used in each of the five tasks. The five tasks used with CoDeepNEAT are:

Vision The MNIST dataset contains 28×28 grayscale images of handwritten digits 0-9. There are 60,000 training images (5,000 of which were used for validation) and 10,000 test images (LeCun, Cortes, and Burges 2010). MNIST is

	MNIST	Omniglot	PMLB Adult	Wikipedia Toxicity	Oxford 102 Flower
Training Hyperparameters					
Activation			All domains: {ReLU, Linear, ELU, SELU}		
Batch Size	128	1000 iterations	32	128	8
Dropout Rate	[0.0, 0.7]	[0.0, 0.7]	[0.0, 0.9]	[0.0, 0.5]	[0.0, 0.9]
Epochs	5	3	25	3	30
Filters/Units	[16, 96]	[16, 96]	[64, 192]	[64, 192]	[64, 192]
Kernel Reg.	L2: [10 ⁻⁹ , 10 ⁻³]	L2: [10 ⁻⁹ , 10 ⁻³]	{L1, L2}: [10 ⁻⁹ , 10 ⁻³]	{L1, L2}: [10 ⁻⁹ , 10 ⁻³]	{L1, L2}: [10 ⁻⁷ , 10 ⁻³]
Kernel Size	{1, 3}	{1, 3}	N/A	{1, 3, 5, 7}	N/A
Learning Rate	[10 ⁻⁴ , 10 ⁻²]	[10 ⁻⁴ , 10 ⁻³]	[10 ⁻⁴ , 10 ⁻²]	[10 ⁻⁴ , 10 ⁻²]	[10 ⁻⁴ , 10 ⁻¹]
Optimizer	Adam	Adam	Adam	Adam	SGD (Nesterov=0.9)
Weight Init.	All domains: {Glorot Normal, Glorot Uniform, He Normal, He Uniform} or AutoInit				
Neural Network Layers					
Add	✓		✓	✓	
Concatenate			✓	✓	
Conv1D				✓	
Conv2D	✓	✓			
Dense			✓		✓
Dropout	✓	✓	✓		✓
GRU				✓	
LSTM				✓	
MaxPooling1D				✓	
MaxPooling2D	✓	✓			
SpatialDropout1D				✓	
WeightedSum		✓			
Evolutionary Hyperparameters					
Elitism (B)	0.4	0.4	0.4	0.2	0.1
Elitism (M)	0.4	0.4	0.4	0.2	0.4
Evaluations (B)	4	4	4	4	1
Generations	30	40	30	30	30
Population Size (B)	22	22	22	22	20
Population Size (M)	56	56	56	56	20
Preserved Networks	12	12	12	12	1
Species (B)	1	1	1	1	1
Species (M)	4	4	4	4	2
Mutation Probabilities					
Change Hyperparam.	0.25	0.25	0.25	0.5	0.5
New Connection (B)	0.12	0.12	0.12	0.2	0.12
New Connection (M)	0.08	0.08	0.08	0.2	0.08
New Layer (M)	0.08	0.08	0.08	0.2	0.08
New Node (B)	0.16	0.16	0.16	0.2	0.16

Table 3: Configuration of neural architecture search experiments in the five tasks. Entries with a (B) or (M) suffix apply to CoDeepNEAT blueprints or modules, respectively. These values were found to work well in preliminary experiments. When AutoInit is applied to an evolved network, it replaces the weight initialization method selected by evolution, but the setup otherwise remains unchanged. The neural architecture search experiments were designed to show that AutoInit improves performance in a wide variety of settings, including those with different data modalities, network topologies, computational complexities, and hyperparameter configurations.

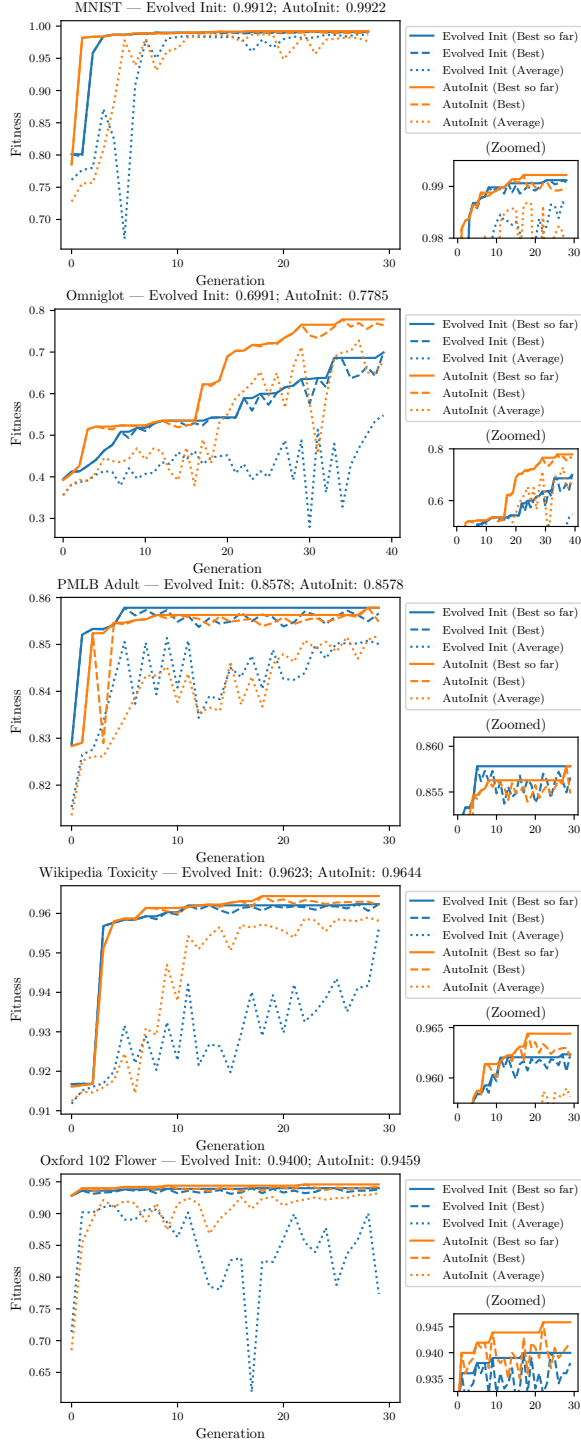


Figure 9: Progress of neural architecture search in the five tasks. The data is the same as that in Figure 7, but this plot also shows how AutoInit can stabilize mean population performance, leading to more reliable discovery of powerful models.

used under the Creative Commons Attribution-Share Alike 3.0 license.

Language In the Wikipedia Toxicity dataset, the task is to classify English Wikipedia comments as toxic or healthy contributions (Wulczyn, Thain, and Dixon 2017). The dataset contains 92,835, 31,227, and 30,953 comments in the training, validation, and test sets, respectively.

Tabular In the Adult dataset (Kohavi 1996) from the Penn Machine Learning Benchmarks repository (PMLB; Olson et al. 2017)) the task is to predict whether an individual makes over \$50K per year based on 14 features. Out of 48,842 total instances, 20% are randomly separated to create a test set. The dataset is used under the MIT License.

Multi-Task The Omniglot dataset contains handwritten characters in 50 different alphabets (Lake, Salakhutdinov, and Tenenbaum 2015); classifying characters in each alphabet is a natural multi-task problem. The characters are 105×105 grayscale images, and there are 20 instances of each character. To save compute resources, 20 of the 50 alphabets were randomly selected for experiments. A fixed training, validation, and testing split of 50%, 20%, and 30% was used with each task. The learning rate decays as $\text{learning_rate} = 0.1^{\text{epoch}/10} * \text{initial_learning_rate}$ during training. The dataset is used under the MIT License.

Transfer Learning A DenseNet-121 network is first pre-trained on the ImageNet dataset (Deng et al. 2009; Huang et al. 2017a). Models are then evolved to utilize its embeddings to classify images in the Oxford 102 Flower dataset, consisting of 102 types of flowers found in the United Kingdom (Nilsback and Zisserman 2008). Each class has between 40 and 258 images; the training and validation sets have 10 images per class, and the test set contains the remaining images from the dataset. During training, the weight decay (L2 loss) is scaled by the current learning rate. Images are also augmented to improve generalization performance. Images are randomly flipped horizontally, rotated up to 40 degrees, shifted up/down and left/right up to 20%, and shear intensity and zoom range varied up to 20%.

D Statistical Significance of Results in Activation Function Meta-Learning

Sampling activation functions from the PANGAEA search space results in a distribution of possible models for each weight initialization strategy. Comparing the empirical distribution functions (EDFs) induced by each initialization strategy makes it possible to quantify the importance of the initialization (Radosavovic et al. 2019).

Given n activation functions with errors $\{e_i\}$, the EDF $F(e) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[e_i < e]$ gives the fraction of activation functions that result in error less than e . Let F_{default} , F_{AutoInit} , and $F_{\text{AutoInit++}}$ be the EDFs for the three initialization strategies. Figure 10 plots these EDFs along with the Kolmogorov-Smirnov test statistic $D = \sup_e |F_1(e) - F_2(e)|$, which measures the maximum vertical discrepancy between two EDFs (Massey Jr 1951). This statistic shows that (1) AutoInit outperforms the default initialization ($D = 0.105$); (2) Au-

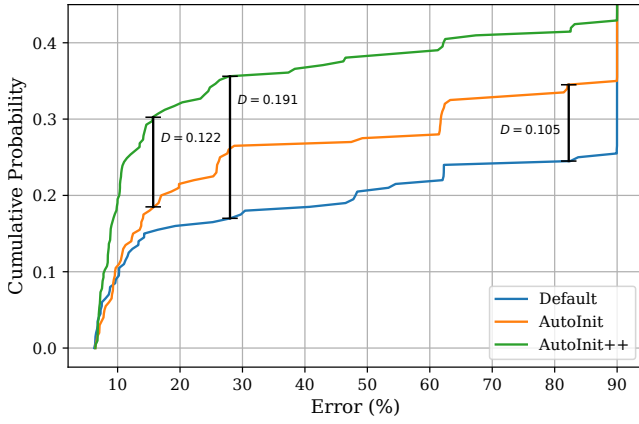


Figure 10: Error EDFs for PANGAEA activation functions when using different weight initialization strategies. The Kolmogorov-Smirnov statistic D quantifies the maximum vertical distance between the EDFs, and shows that proper initialization provides a measurable increase in expected performance. Notice that the x -axis shows percent error, and not accuracy as in Figure 8.

toInit++ delivers an even greater boost in performance over the default initialization ($D = 0.191$); and (3) AutoInit++ is measurably better than AutoInit ($D = 0.122$), confirming that having zero Gaussian mean is a useful property for activation functions to have.

Other ways of measuring statistical significance lead to similar conclusions. For instance, consider the null hypothesis that $F_{\text{default}} = F_{\text{AutoInit}}$. In other words, this null hypothesis states that AutoInit provides no benefit and that the accuracies obtained come from the same underlying distribution. With the Epps-Singleton test (Epps and Singleton 1986) this null hypothesis is rejected with $p < 0.05$. Similarly, the test rejects the null hypothesis that $F_{\text{default}} = F_{\text{AutoInit++}}$ with $p < 0.001$. Even stronger statements can be made in the case of AutoInit++. With the Mann-Whitney U test (Mann and Whitney 1947), the null hypothesis that $F_{\text{default}}(e) \geq F_{\text{AutoInit++}}(e)$ for some e is rejected ($p < 0.01$) in favor of the alternative that $F_{\text{default}}(e) < F_{\text{AutoInit++}}(e)$ for all e . Similarly, the null hypothesis that $F_{\text{AutoInit}}(e) \geq F_{\text{AutoInit++}}(e)$ for some e is rejected ($p < 0.05$) in favor of the alternative that $F_{\text{AutoInit}}(e) < F_{\text{AutoInit++}}(e)$ for all e . As discussed in the main text, this result states that the distribution of accuracies induced by AutoInit++ is stochastically larger than that from AutoInit or the default initialization.

E Future Technical Extensions

AutoInit is based on understanding and utilizing the training dynamics of neural networks, leading to higher and more robust performance, and facilitating further advances in meta-learning. It can be improved and its scope broadened in several ways in the future, as outlined below.

Initial Weight Distributions AutoInit calculates appropriate weight scaling, but it does not impose a distribution from which weights are drawn (Equation 6). All experiments in

this paper used a truncated normal distribution. In preliminary experiments, AutoInit also used untruncated normal, uniform, and orthogonal distributions, but no clear trends were observed. Indeed, assuming weights are scaled appropriately, whether training is stable depends only on the architecture and not the distribution from which weights are sampled (Hanin 2018). However, this conclusion applies only in limited theoretical settings; in other settings, orthogonal initialization was found to be beneficial (Saxe, McClelland, and Ganguli 2013; Hu, Xiao, and Pennington 2020). Whether there is a single distribution that is optimal in every case, or whether certain distributions are better-suited to different models, tasks, or layers, remains an open question, and a compelling direction for future research.

Variations of AutoInit Several variations of the core AutoInit algorithm can be devised that may improve its performance. For example, AutoInit stabilizes signals by analyzing the forward pass of activations from the input to the output of the network. It is possible to similarly model the backward pass of gradients from the output to the input. Indeed, past weight initialization strategies have sometimes utilized signals in both directions (Glorot and Bengio 2010; Arpit, Campos, and Bengio 2019). It would be interesting to find out whether AutoInit could similarly benefit from analyzing backward-propagating signals.

Alternative objectives beyond mean and variance stabilization could also be considered. Two promising objectives are tuning the conditioning of the Fisher information matrix (Pennington and Worah 2018) and achieving dynamical isometry (Xiao et al. 2018). Mean field theory and nonlinear random matrix theory (Pennington and Worah 2019) could potentially be used to implement these objectives into AutoInit.

Support for New Layer Types AutoInit calculates outgoing mean and variance estimates for the majority of layer types available in current deep learning frameworks (Appendix A). If AutoInit encounters an unknown layer, the default behavior is to assume that the mean and variance are not changed by that layer: $g_{\text{layer}}(\mu_{\text{in}}, \nu_{\text{in}}) = \mu_{\text{in}}, \nu_{\text{in}}$. This fallback mechanism tends to work well; if there are only a few unknown layers, then the variance estimation will be incorrect only by a constant factor and training can proceed. However, mean and variance estimation functions g can be derived for new types of layers as they are developed, either analytically or empirically with Monte Carlo sampling, thus taking full advantage of AutoInit’s ability to stabilize training in the future as well.

Tighter Integration with Deep Learning Frameworks Using AutoInit is simple in practice. The AutoInit package provides a wrapper around TensorFlow models. The wrapper automatically traverses the TensorFlow computation graph, calculates mean and variance estimations for each layer, and reinstantiates the model with the correct weight scaling. However, this implementation can be streamlined. The most effective approach would be to integrate AutoInit natively with deep learning frameworks like TensorFlow (Abadi et al. 2016) and PyTorch (Paszke et al. 2019). Native integration would not just make AutoInit easier to use,

it would also make it more accessible to general machine learning practitioners. For example, TensorFlow provides a few initialization strategies that can be leveraged by changing the `kernel_initializer` keyword in certain layers. However, implementing other weight initialization strategies requires subclassing from the `Initializer` base class, which is both time-consuming and complicated, especially for non-experts. Native integration would ensure that the benefits of smarter initialization are available immediately to the wider machine learning community.

F Computing Infrastructure

Experiments in this paper were run in a distributed framework using StudioML software (Mutch 2017–2021; Zhokhov, Denisov, and Mutch 2017–2021) to place jobs on machines with NVIDIA GeForce GTX 1080 Ti and RTX 2080 Ti GPUs. The CoAtNet experiments (Section 7) were run on an AWS g5.48xlarge instance with NVIDIA A10G GPUs. Training CoAtNet on Imagenette required an average of 0.91 GPU hours per run. Training CoAtNet on ImageNet once took 119.89 GPU hours. Training on Imagenette instead of ImageNet therefore emit approximately $119.89/0.91 \approx 132$ times less carbon. The AutoInit package is available at <https://github.com/cognizant-ai-labs/autoinit>.