
Synergy and Symmetry in Deep Learning: Interactions between the Data, Model, and Inference Algorithm

Lechao Xiao¹ Jeffrey Pennington¹

Abstract

Although learning in high dimensions is commonly believed to suffer from the curse of dimensionality, modern machine learning methods often exhibit an astonishing power to tackle a wide range of challenging real-world learning problems without using abundant amounts of data. How exactly these methods break this curse remains a fundamental open question in the theory of deep learning. While previous efforts have investigated this question by studying the data (\mathcal{D}), model (\mathcal{M}), and inference algorithm (\mathcal{I}) as independent modules, in this paper we analyzes the triplet $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ as an integrated system and identify important synergies that help mitigate the curse of dimensionality. We first study the basic symmetries associated with various learning algorithms $(\mathcal{M}, \mathcal{I})$, focusing on four prototypical architectures in deep learning: fully-connected networks (FCN), locally-connected networks (LCN), and convolutional networks with and without pooling (GAP/VEC). We find that learning is most efficient when these symmetries are compatible with those of the data distribution and that performance significantly deteriorates when any member of the $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ triplet is inconsistent or suboptimal.

1. Introduction

Statistical problems with high-dimensional data are frequently plagued by the *curse of dimensionality*, in which the number of samples required to solve the problem grows rapidly with the dimensionality of the input. Classical theory explains this phenomenon as the consequence of basic geometric and algebraic properties of high-dimensional spaces; for example, the number of ϵ -cubes inside a unit cube in

¹Google Research, Brain Team. Correspondence to: Lechao Xiao <xlc@google.com>, Jeffrey Pennington <jpen-nin@google.com>.

\mathbb{R}^d grows exponentially like ϵ^{-d} , and the number of degree r polynomials in \mathbb{R}^d grows like a power-law d^r . Since for real-world problems d is typically in the hundreds or thousands, classical wisdom suggests that learning is likely to be infeasible. However, starting from the groundbreaking work AlexNet (Krizhevsky et al., 2012), practitioners in deep learning have tackled a wide range of difficult real-world learning problems (Vaswani et al., 2017; Devlin et al., 2018; Silver et al., 2016; Senior et al., 2020; Kaplan et al., 2020) in high dimensions, once believed by many to be out-of-scope of current techniques. The astonishing success of modern machine learning methods clearly contradicts the curse of dimensionality and therefore poses the fundamental question: mathematically, how do modern machine learning methods break the curse of dimensionality?

To answer this question, we examine the most fundamental ingredients of machine learning methods. They are the data (\mathcal{D}), the model (\mathcal{M}), and the inference algorithm (\mathcal{I}).

Data (\mathcal{D}) is of course central in machine learning. In the classical learning theory setting, the learning objective usually has a power-law decay $m^{-\alpha}$ as the function of the number of training samples m . The theoretical bound on α is usually small (e.g. $\alpha = d^{-1}$ (von Luxburg & Bousquet, 2004; Bach, 2017; Bahri et al., 2021)) and is of limited practical utility for high-dimensional data. On the other hand, empirical measurements of α in state-of-the-art deep learning models typically reveal values of α that are significantly larger (e.g. $\alpha = 0.43$ for ResNets trained on ImageNet in Fig.5) even though d is quite large (e.g. $d \sim 10^5$ for ImageNet). This example suggests that the learning curve must have important functional dependence on \mathcal{M} and \mathcal{I} . Indeed, as we will observe later, many of the best performing methods exhibit learning curves for which $\alpha = \alpha(m)$ actually *increases* as m becomes larger, i.e. data makes the usage of data more efficient. We call this phenomenon DIDE, for **d**ata **i**mproves **d**ata **e**fficiency; see Fig.5.

Designing machine learning models (\mathcal{M}) that maximize data-efficiency is critical to the success of solving real-world tasks. Indeed, breakthroughs in machine learning are often driven by novel architectures such as LeNet (LeCun et al., 1998), AlexNet(Krizhevsky et al., 2012), Transformer (Vaswani et al., 2017), etc. While some of the inductive

biases of these methods are clear (e.g. translation symmetries of CNNs), others tend to build off of prior empirical success and are less well-understood (e.g. the implicit bias of SGD). To build our understanding of these biases and how they affect learning, we conduct a theoretical analysis of them in the infinite-width setting (Neal, 1994; Poole et al., 2016; Daniely et al., 2016; Jacot et al., 2018b; Lee et al., 2019), which preserves most salient aspects of the architecture while enabling tractable calculations.

The inference procedure (\mathcal{I}) is what enables *learning* in machine learning methods. It is widely believed that modern inference methods, specifically gradient descent and variants, “implicitly” bias the solutions of the networks towards those that generalize well and away from those that generalize poorly (Neyshabur, 2017; Gunasekar et al., 2018; Soudry et al., 2018). The effects of the inference algorithm are intimately tied to the specifics of the model (e.g. weight-sharing) and the data (e.g. augmentation), and might not be fully understood with a fixed-data, fixed-model analysis. Indeed, good performance may derive from interactions between $(\mathcal{M}, \mathcal{I})$, or $(\mathcal{D}, \mathcal{I})$, or even $(\mathcal{D}, \mathcal{M}, \mathcal{I})$. In Sec. 6, we demonstrate the DIDE effect for a particular choice of $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ and show that this effect disappears if any one of \mathcal{D} , \mathcal{M} , or \mathcal{I} is altered.

The above discussion highlights the insufficiency of treating \mathcal{D} , \mathcal{M} , and \mathcal{I} as separate non-interacting modules. They must be considered as an integrated system. Throughout this paper, we will refer to the triplet $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ as a (machine) learning system and the tuple $(\mathcal{M}, \mathcal{I})$ as the learning algorithm of the system that operates on \mathcal{D} .

We focus our study on four prototypical deep learning architectures whose similarities and differences provide a rich test bed for analysis: full-connected networks (FCN), locally-connected networks (LCN), convolutional networks with read-out vectorization (VEC), and convolutional networks with read-out global average pooling (GAP). We consider both finite-width (FCN_n , LCN_n , VEC_n , GAP_n) and infinite-width variants of these architectures (FCN_∞ , LCN_∞ , VEC_∞ , GAP_∞). We examine the basic symmetries of the $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ triplets associated to these architectures and find that better architectures break spurious symmetries. We also identify a symmetry breaking effect due to finite-width and carefully examine the impact of this effect to the performance of the system.

2. Related Work

The study of infinite networks dates back to the seminal work of Neal (1994) who showed the convergence of single hidden-layer networks to Gaussian Processes (GPs). Recently, there has been renewed interest in studying random, infinite, networks starting with concurrent work on “con-

jugate kernels” (Daniely et al., 2016; Daniely, 2017) and “mean-field theory” (Poole et al., 2016; Schoenholz et al., 2017), taking a statistical learning and statistical physics view of points, respectively. Since then this analysis has been extended to include a wide range for architectures (Lee et al., 2018; Matthews et al., 2018; Xiao et al., 2018a; Novak et al., 2019b; Yang, 2019; Hron et al., 2020b). The inducing kernel is often referred to as the Neural Network Gaussian Process (NNGP) kernel. The neural tangent kernel (NTK), first introduced in Jacot et al. (2018a), along with followup work (Lee et al., 2019; Chizat et al., 2019) showed that the distribution of functions induced by gradient descent for infinite-width networks is a Gaussian Process with NTK as the kernel. Since then, NNGP and NTK have become extremely useful and popular tools to study various properties of neural networks (Arora et al., 2019; Adlam & Pennington, 2020; Bordelon et al., 2021; Mei et al., 2021; Bietti, 2021; Favero et al., 2021; Xiao, 2021) and many others.

The implicit bias of gradient descent has been the focus of a number of recent works (Soudry et al., 2018; Lyu & Li, 2020; Ji & Telgarsky, 2019a;b; Chizat & Bach, 2020), leading to a variety of noteworthy conclusions, including the convergence of GD to the maximal margin solution for logistic-type losses during late-time training (Soudry et al., 2018). Nakkiran et al. (2019); Hu et al. (2020); Rahaman et al. (2019); Xu (2018); Xu et al. (2019); Su & Yang (2019); Yang & Salman (2019) study the early-time dynamics and spectral biases of neural networks, leading to the conclusion that simpler functions are usually learned before more complex functions.

Understanding and exploiting the structural information in natural are central aspects of designing machine learning systems. Li et al. (2018); Goldt et al. (2020); Pope et al. (2021) study the low-dimensional structure of natural data while Bruna & Mallat (2013); Petrini et al. (2021) investigate the role of deformation stability of natural data. Designing networks that maximally respect the symmetries of natural data (e.g. translational invariance/equivalence of images (Cohen & Welling, 2016; Zaheer et al., 2017)) is widely considered a principled approach in practice. Several works also demonstrate the possibility of learning such symmetries from scratch using natural or synthetic data (Neyshabur, 2020; Ingrosso & Goldt, 2022). Nevertheless, recent breakthrough in applying attention-based models (Dosovitskiy et al., 2020; Tolstikhin et al., 2021; He et al., 2021) to computer vision has fundamentally challenged the significance that symmetries play in model design. Attention-based models have weaker inductive biases (lacking even translation equivariance) than those of convolutional networks, yet their performance is comparable in the large data regime (Zhai et al., 2022). Our analysis of DIDE (Fig. 5) sheds some light on why such models are able to reach good performance with the help of more data.

3. Preliminaries and Notation

We focus our presentation on the supervised learning setting and more concretely, on image recognition. Let $\mathcal{D} \subseteq (\mathbb{R}^d)^3 \times \mathbb{R}^k \equiv \mathbb{R}^{3d} \times \mathbb{R}^k$ denote the data set (training and test) and $\mathcal{X} = \{x : (x, y) \in \mathcal{D}\}$ and $\mathcal{Y} = \{y : (x, y) \in \mathcal{D}\}$ denote the input space (images) and label space, respectively. Here d is the spatial dimension (e.g. $d = 32 \times 32$ for CIFAR-10) of the images and 3 is the total number of channels (i.e. RGB). We assume $(x, y) \in \mathcal{D}$ is obtained from some data generating process with unknown distribution $\mu_{\mathcal{D}}$ and the learning task is to recover $\mu_{\mathcal{D}}$.

3.1. Neural Networks

We use FCN_n to denote an L -hidden layer fully-connected network with identical hidden widths $n_l = n \in \mathbb{N}$ for $l = 1, \dots, L$ and with readout width $n_{L+1} = k$ (the number of logits). For each $x \in \mathbb{R}^{3d}$, we use $h^l(x), x^l(x) \in \mathbb{R}^{n_l}$ to represent the pre- and post-activation functions at layer l with input x . The recurrence relation FCN_n is given by

$$\begin{cases} h^{l+1} &= x^l W^{l+1} \\ x^{l+1} &= \phi(h^{l+1}) \end{cases} \quad \text{and } W_{i,j}^{l+1} = \frac{1}{\sqrt{n_{l+1}}} \omega_{ij}^{l+1} \quad (1)$$

where ϕ is a point-wise activation function, $W^{l+1} \in \mathbb{R}^{n_l \times n_{l+1}}$ are the weights and ω_{ij}^l are the trainable parameters, drawn i.i.d. from a standard Gaussian $\mathcal{N}(0, 1)$ at initialization. For simplicity of the presentation, the bias terms and the hyperparameters (the variances of the weights) are omitted; including them as hyperparameters will not significantly change any of the main conclusions.

For convolutional networks, the inputs are treated as tensors in $(\mathbb{R}^d)^3$. The recurrence relation of convolutional networks can be written as

$$\begin{cases} x_{\alpha,j}^{l+1} &= \phi(h_{\alpha,j}^{l+1}) \\ h_{\alpha,j}^{l+1} &= \frac{1}{\sqrt{(2k+1)n^l}} \sum_{i=1}^{n^l} \sum_{\beta=-k}^k x_{\alpha+\beta,i}^l \omega_{ij,\beta}^{l+1}. \end{cases} \quad (2)$$

Here $\alpha \in [d]$ denote the spatial location, $i/j \in [n]$ denotes the fan-in/fan-out channel indices. For notational convenience, we assume circular padding and stride equal to 1 for all layers. The features of the penultimate layer are 2D tensors and there are two commonly used approaches to map them to the logit layer: (a) VEC_n , which vectorizes the 2D tensor to a 1D vector, yielding a translation-equivariant inductive bias, or (b) GAP_n , which applies a global average pooling layer to each channel, yielding a translation-invariant inductive bias. The readout layers for these models

can be written as,

$$x_j^{L+1} = \frac{1}{\sqrt{dn}} \sum_{i \in [n]} \sum_{\alpha \in [d]} x_{\alpha,i}^L w_{\alpha,ij}^{L+1} \quad (\text{VEC}_n), \quad (3)$$

$$x_j^{L+1} = \frac{1}{\sqrt{n}} \sum_{i \in [n]} \left(\frac{1}{d} \sum_{\alpha \in [d]} x_{\alpha,i}^L \right) w_{ij}^{L+1} \quad (\text{GAP}_n). \quad (4)$$

The key difference between the two architectures is that, in VEC_n , each pixel in the penultimate layer has its own readout variable, whereas in GAP_n the pixels within the same channel share the same readout variable. It is clear that the function space of VEC_n contains that of GAP_n .

Locally Connected Networks (LCN_n) (Fukushima, 1975; Lecun, 1989) are convolutional networks *without* weight sharing between spatial locations. They share the connectivity pattern and topology of a standard convolutional network, but the weights are not shared across spatial patches. Mathematically, the network is defined as in Equation 2, but with all the *shared* parameters $\omega_{ij,\beta}^l$ replaced by *unshared* $\omega_{ij,\alpha,\beta}^l \sim \mathcal{N}(0, 1)$. In this work, we assume that the LCN_n are always associated with a vectorization readout layer and it is therefore clear that the function space of LCN_n is a super set of VEC_n . Interestingly, LCN_n is also a subset of VEC_{dn} when the network is d times wider.

Theorem 3.1 (Sec. C). *Let $\text{VEC}_n/\text{LCN}_n/\text{GAP}_n/\text{FCN}_n$ denote the set of functions that can be represented by a L -hidden layer $\text{VEC}_n/\text{LCN}_n/\text{GAP}_n/\text{FCN}_n$ network with hidden-layer width n . Then*

$$\text{GAP}_n \subseteq \text{VEC}_n \subseteq \text{LCN}_n \subseteq \text{VEC}_{dn} \subseteq \text{FCN}_{d^2n}. \quad (5)$$

As remarked above, the random initialization of parameters endows GAP_n with a translation-invariant prior, which may be well-suited to many image-classification tasks. This observation, combined with the result from Theorem 3.1 that the function space of GAP_n is smallest, suggests that networks with this architecture may enjoy favorable generalization properties. Indeed, prior work (Novak et al., 2019b; Lee et al., 2020; Neyshabur, 2020) has found that GAP_n can significantly outperform VEC_n , LCN_n , and FCN_n , a conclusion we also find in Sec. 5.

We emphasize that the above observation relies on a notion of the prior induced by initialization and says little about the effect of optimization. It is possible that gradient descent could update the readout weights of a network from the VEC_n class toward a configuration that approximately implements average pooling, thereby pushing the model closer to a member of the GAP_n class. Alternatively, if the weights remain close to their initial random values, the function might more closely resemble a member of the LCN_n class. This perspective gives some intuition for how the inference algorithm \mathcal{I} can interact with the model \mathcal{M}

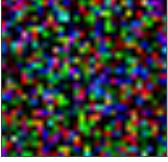
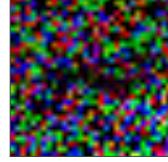
Models	FCN _{n/∞} (iid Gaussian)	FCN _n (iid Non-Gaussian)	VEC _∞ / LCN _{n/∞} (iid Gaussian)	VEC _n / GAP _{n/∞} (iid Gaussian)	
Symmetry Groups	O(3d)	P(3d)	O(3)^d	O(3) ⊗ I_d	I_{3d}
Rotated Images					

Figure 1: **Models and associated symmetry groups.** Top: architectures (initialization scheme). Middle: implied symmetry groups. Bottom: rotated images, where the rotation is randomly drawn from the corresponding symmetry group. The largest symmetry group for which a random transformation does not obfuscate the underlying image is evidently $\text{O}(3) \otimes \mathbf{I}_d$.

to produce predictive functions with significantly different generalization properties. We return to this theme in Sec. 6.

3.2. Infinite Network: Gaussian Processes and the Neural Tangent Kernels

In order to better disambiguate the effects of the model \mathcal{M} from the inference algorithm \mathcal{I} , it is useful to examine our model families in the limit of infinite width. This limit facilitates simpler theoretical analysis while simultaneously preserving most of the salient ingredients of the finite-width models. Below, we briefly review several useful aspects of infinitely-wide networks and refer interested readers to the references for a more comprehensive introduction.

Neural Networks as Gaussian Processes (NNGP). As the width $n \rightarrow \infty$, at initialization the output $f_0(\mathcal{X})$ forms a Gaussian Process $f_0(\mathcal{X}) \sim \mathcal{GP}(0, \mathcal{K}(\mathcal{X}, \mathcal{X}))$, known as the NNGP (Neal, 1994; Lee et al., 2018; Matthews et al., 2018). Here \mathcal{K} is the GP kernel and can be computed in closed form for a variety of architectures (Novak et al., 2019a). By treating this infinite width network as a Bayesian model (aka Bayesian Neural Networks) and marginalizing over the training set $(\mathcal{X}_T, \mathcal{Y}_T)$, the posterior is also a GP whose mean is given by $\mathcal{K}(\mathcal{X}_*, \mathcal{X}_T)\mathcal{K}^{-1}(\mathcal{X}_T, \mathcal{X}_T)\mathcal{Y}_T$.

Neural Tangent Kernels (NTK). Recent advances in global convergence theory of large width networks (Jacot et al., 2018a; Du et al., 2018; Allen-Zhu et al., 2018; Zou et al., 2020; Lee et al., 2019) have shown that under certain assumptions, the gradient descent dynamics of a network converge to that of kernel gradient descent as the widths approach infinity, where the kernel is the NTK (Jacot et al., 2018a), denoted by Θ . As such, in the infinite width limit, when the loss is the mean squared error (MSE), the mean prediction (marginalized over random initialization) can be solved analytically. In particular, when the training time

$t = \infty$, the prediction is given by

$$f(\mathcal{X}_*) = \Theta(\mathcal{X}_*, \mathcal{X}_T)\Theta^{-1}(\mathcal{X}_T, \mathcal{X}_T)\mathcal{Y}_T, \quad (6)$$

For convenience, we use $\text{FCN}_\infty(x)$, $\text{LCN}_\infty(x)$, $\text{VEC}_\infty(x)$ and $\text{GAP}_\infty(x)$ to denote the infinite width solutions (either via NNGP inference or NTK regression) for the corresponding architectures.

4. Symmetries of Machine Learning Systems

In this section, we study the symmetry properties of various machine learning systems $(\mathcal{D}, \mathcal{M}, \mathcal{I})$, focusing on how invariances implied by the learning algorithm $(\mathcal{M}, \mathcal{I})$ interact with the data distribution $\mu_{\mathcal{D}}$.

To provide concrete context, let us consider solving image classification using kernel regression with an inner-product kernel K (i.e. $K(x, \bar{x}) = k(\langle x, \bar{x} \rangle)$ for some function k). Since K is invariant to rotations of its inputs (i.e. for any rotation operator τ , $K(\tau x, \tau \bar{x}) = K(x, \bar{x})$), if we rotate all (both train and test) images by any fixed rotation τ , the kernel is unchanged and so are the predictions. Because natural images surely exhibit spatial structure that is destroyed by rotations (see Fig. 1), we might expect such kernels to perform poorly on image classification tasks, and indeed we find this to be the case (see Sec. 5). In this sense, the symmetry properties implied by the learning algorithm are incompatible with the data distribution $\mu_{\mathcal{D}}$, and we regard them as *spurious*. Below we describe some notation and results that will help us analyze this type of interaction more systematically.

For a deterministic (stochastic) learning algorithm $\mathcal{A} = (\mathcal{M}, \mathcal{I})$, we use $\mathcal{A}(\mathcal{D}_T)$ to denote the learned function (distribution of the learned functions) using training set \mathcal{D}_T . We use $\mathcal{A}^\tau(\mathcal{D}_T)$ to denote the learned function(s) using $\tau(\mathcal{D}_T)$, which make predictions on the transformed test point $\tau(\mathcal{X}_*)$. In other words, all inputs, including training and test inputs, are pre-processed by a common transformation τ before

feeding to the learning algorithm \mathcal{A} . For convenience, for random variables A and B , we use $A =^d B$ to indicate that they have the same distribution.

Definition 1. Let \mathcal{G} be a group of transformations $\mathbb{R}^{3d} \rightarrow \mathbb{R}^{3d}$. A deterministic (stochastic) learning algorithm $\mathcal{A} = (\mathcal{M}, \mathcal{I})$ is g -invariant if $\mathcal{A} = \mathcal{A}^g$ ($\mathcal{A} =^d \mathcal{A}^g$). In this case, we say the system $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ is g -invariant and use the notation $(\mathcal{D}, \mathcal{M}, \mathcal{I}) = (g\mathcal{D}, \mathcal{M}, \mathcal{I})$. If this holds for all $g \in \mathcal{G}$, then we say the algorithm and the system are \mathcal{G} -invariant.

Comparing with Functional Invariance. The flavor of invariance studied this paper is *algorithmic* invariance, as it concerns a *system* or a *learning algorithm*, and is qualitatively different from the *functional* invariance studied elsewhere (Cohen & Welling, 2016). Recall that a function f is (functionally) invariant to a group \mathcal{G} if $f(\tau x) = f(x)$ for all $\tau \in \mathcal{G}$. Natural images are often considered to be translationally invariant, which is a key motivation for the usage of *convolutional* networks in computer vision. These symmetries are *hard-coded* into the architectures¹ as a kind of *inductive bias* and the (post-pooling) representations are invariant to this group by design. As a consequence, the corresponding hypothesis class, defined by these networks, is more restricted, which could lead to better generalization performance if the symmetry were exact (Shalev-Shwartz & Ben-David, 2014). Algorithmic invariance is weaker than functional invariance. For example, in the discussion above, the learning system defined by kernel regression with an inner product kernel is algorithmically invariant to the rotational group \mathcal{G} since $K(\tau x, \tau \bar{x}) = K(x, \bar{x})$ for all $\tau \in \mathcal{G}$, but the learned function itself is not *functionally* invariant, which would require the kernel to be \mathcal{G} -invariant $K(\tau x, \bar{x}) \neq K(x, \bar{x})$. Nevertheless, we could *augment*² this learning system to be *functionally* invariant by hard-coding the symmetries into the system by defining an *invariant* kernel $K^{\text{inv}}(x, \bar{x}) = \int_{\tau \in \mathcal{G}} K(x, \tau \bar{x}) d\tau$. The regressor is invariant to \mathcal{G} as the kernel itself is \mathcal{G} -invariant: $K^{\text{inv}}(\tau x, \bar{x}) = K^{\text{inv}}(x, \bar{x})$, which is stronger than $K(\tau x, \tau \bar{x}) = K(x, \bar{x})$.

To present the implied symmetry properties of the main architectures under study, we need to introduce some notation. Let $O(3d)$ denote the orthogonal group on the flattened input space \mathbb{R}^{3d} . The subgroup $O(3)^d \leq O(3d)$ operates on the un-flattened input $(\mathbb{R}^3)^d$, whose elements rotate each pixel $x_\alpha \in \mathbb{R}^3$ by an independent element $\tau_\alpha \in O(3)$. The smaller subgroup $O(3) \otimes \mathbf{I}_d \leq O(3)^d$ applies the *shared* rotation, i.e. $\tau_\alpha = \tau$ to all x_α for $\alpha \in [d]$. Similarly, we use $P(3d)$ to denote the permutation group on \mathbb{R}^{3d} and $P(3)^d$

¹In the idealized setting when circular padding is applied and the readout layer is a global average pooling.

²This is similar to augment the training set by group actions (Chen et al., 2020).

and $P(3) \otimes \mathbf{I}_d$ are defined similarly. Note that rotating \mathcal{X} by τ is equivalent to transforming the underlining coordinate systems of the input by the adjoint $\tau^* = \tau^{-1}$. Fig 1 displays an image from CIFAR-10 under five families of rotations. We use FCN_n , VEC_n , etc. to denote the output function (distribution) of a finite-width network obtained by SGD, in which the random initialization is the only source of randomness.

Theorem 4.1 (Sec.D). *If the initial parameters of the networks defined in Sec. 3.1 are iid samples from $\mathcal{N}(0, 1)$, then the predictions from finite-width networks trained by SGD or infinite-width networks trained by kernel regression enjoy the following symmetries:*

- $\text{FCN}_{n/\infty}$ are $O(3d)$ -invariant
- $\text{LCN}_{n/\infty}$ and VEC_∞ are $O(3)^d$ -invariant
- VEC_n and $\text{GAP}_{n/\infty}$ are $O(3) \otimes \mathbf{I}_d$ -invariant.

The $O(3d)$ -invariance of $\text{FCN}_{n/\infty}$ follows from the rotational invariance of the Gaussian measure, and has been observed in many prior works, including (Wadia et al., 2020; Li et al., 2020). Rotating the input by $\tau \in O(3d)$ is equivalent to rotating the weight matrix ω of the first layer by τ^* , and since $\tau^* \omega =^d \omega$ for $\omega \sim \mathcal{N}(0, 1)^{3d}$, the distribution of the output functions at step 0 (aka initialization) is invariant. This observation implies that the first gradient is also $O(3d)$ -invariant, which further implies the $O(3d)$ -invariance of the output function after the first gradient update. By induction, this invariance property holds throughout the course of gradient descent training, even with L^2 -regularization as the L^2 -norm is rotationally invariant. Such invariant property also holds for (finite-width) Bayesian posterior inference thanks to the Bayes theorem: $P(\tau \mathcal{X}^* | \tau \mathcal{X}_T) = P(\mathcal{X}^* | \mathcal{X}_T)$ because $P(\tau \mathcal{X}^*, \tau \mathcal{X}_T) / P(\tau \mathcal{X}_T) = P(\mathcal{X}^*, \mathcal{X}_T) / P(\mathcal{X}_T)$.

For the same reason, LCN_n is $O(3)^d$ -invariant because each patch of the image uses independent Gaussian random variables. In addition, weight-sharing in VEC_n and GAP_n breaks the $O(3)^d$ symmetry, reducing it to $O(3) \otimes \mathbf{I}_d$.

For infinite networks, $\text{LCN}_\infty = \text{VEC}_\infty$ (Xiao et al., 2018b; Novak et al., 2019b; Garriga-Alonso et al., 2019). The kernels of VEC_∞ and GAP_∞ are of the forms

$$\Theta_{\text{VEC}}(x, x') = k(\{\langle x_\alpha, x'_\alpha \rangle\}_{\alpha \in [d]}) \quad (7)$$

$$\Theta_{\text{GAP}}(x, x') = k(\{\langle x_\alpha, x'_{\alpha'} \rangle\}_{\alpha, \alpha' \in [d]}). \quad (8)$$

The former depends only on the inner product between pixels in the *same* spatial location, breaking the $O(3d)$ symmetry of FCN_∞ and reducing it to $O(3)^d$. In addition, the latter depends also on the inner products of pixels across different spatial locations due to pooling, which breaks the $O(3)^d$ symmetry and reduces it to $O(3) \otimes \mathbf{I}_d$. Noting that

$\dim(\mathrm{O}(3d)) = 3d(3d - 1)/2 > \dim(\mathrm{O}(3)^d) = 3d > \dim(\mathrm{O}(3) \otimes \mathbf{I}_d) = 3$, we see that $\mathrm{LCN}_n/\mathrm{VEC}_\infty$ dramatically reduces the dimensionality of the symmetry group. As we will observe in Secs. 5 and 6, if a symmetry group is inconsistent with the data distribution, the performance of the associated learning algorithm tends to diminish in proportion to the dimension of the spurious symmetry group; see Fig. 2.

The results of the paper are presented in the most *vanilla* setting. Our methods can easily extend to more complicated architectures like ResNet(He et al., 2016), MLP-Mixer(Tolstikhin et al., 2021), etc. The symmetry groups of such systems need to be computed in a case-by-case manner by identifying the invariant group of the random initialization and training procedures. For example, the orthogonal group type of symmetries needed to be replaced by the permutation-type of symmetries if non-Gaussian i.i.d. initialization or/and L^p ($p \neq 2$) regularization. However, we empirically observe that swapping the Gaussian initialization by the uniform initialization in the first layer does not essentially change the performance of the network; see Sec. F.2. This observation indicates that the permutation group may exhibit a similar degree of spuriousness as the rotation group; however, more rigorous and thorough experiments are needed to confidently confirm this claim, which is left for future work. Moreover, the invariance property studied here is mainly coming from the first layer and it is possible that later layers could contribute new invariances to the system. For example, owing to the non-overlapping between patches in ViT (Dosovitskiy et al., 2020), there could be permutation symmetries between the patches in the subsequent self-attention layer (assuming no positional encoding). Finally, for the sake of simplicity, we use NTK-parameterization (Jacot et al., 2018a) but our results apply to other network parameterizations, including standard- (Sohl-Dickstein et al., 2020), meanfield- (Mei et al., 2018), μ -parameterizations (Yang & Hu, 2020). In particular, both finite- and infinite-width FCNs still suffer from the most spurious symmetries $\mathrm{O}(3d)$ for all such parameterization scheme, which may explain the poor performance of FCN in the “feature learning” regime (e.g. 61.5% accuracy on CIFAR-10, Table 1. in Yang et al. (2022).)

5. Empirical Analysis

This section focuses on empirical analysis. Details regarding the experimental setup of this and next section can be found in Sec. B. The goal is to (1) verify Theorem 4.1, (2) study the consistency between \mathcal{D} and $\mathcal{A} = (\mathcal{M}, \mathcal{I})$ via the lens of symmetries, and (3) study the effect of symmetry breaking by comparing VEC_n to VEC_∞ . In light of the visualization of the images in Fig. 1, as well as the numerical performance of the various methods mentioned below, $\mathrm{O}(3) \otimes \mathbf{I}_d$ is the

largest symmetry group that is compatible with μ_D . In what follows, we regard all larger symmetries $(\mathrm{O}(3)^d, \mathrm{P}(3d)$ and $\mathrm{O}(3d))$ as spurious.

5.1. Experimental Setup.

We conduct experiments on $\mathcal{D} = \text{CIFAR-10}$ (Krizhevsky et al., 2009), which is a standard image dataset that consists of 50,000/10,000 training/test images. We vary each member of $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ as follows. **Five Datasets.** We create 5 families of new datasets $\tau\mathcal{D}$ by rotating all input images in \mathcal{D} by τ , a fixed random element of one of the five groups: identity operator \mathbf{I}_{3d} , pixel-wise shared-rotations, $\mathrm{O}(3) \otimes \mathbf{I}_d$, pixel-wise (unshared-)rotations $\mathrm{O}(3)^d$, permutations $\mathrm{P}(3d)$ and global rotations $\mathrm{O}(3d)$. **Six Models.** In addition to FCN, LCN, VEC and GAP, we add $\mathrm{LAP}^{4/8}$ which are the same as GAP except the readout layer is replaced by the Local Average Pooling with window size $4 \times 4/8 \times 8$. All networks have 8 hidden layers. **Three Inference Algorithms.** (1) NTK regression³ (aka infinite-width networks), (2) NN, our baseline for finite-width networks which is trained with momentum using a *small learning rate* without L^2 regularization; (3) NN+ := NN+LR+L2, i.e. using a larger learning rate (+LR) and adding L^2 regularization (+L2). We plot the test accuracy for each $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ (a total of $90 = 3 \times 6 \times 5$) in Fig. 2. The accuracy for finite-width networks is averaged over 5 runs⁴ for each $(\mathcal{D}, \mathcal{M}, \mathcal{I})$, and in each run the rotation τ is resampled. Note that the total variance across runs is small, indicating that the particular choice of τ has negligible effect on the results.

5.2. Verifying Theorem 4.1

As expected from Theorem 4.1, Fig. 2 shows that across NTK/NN/NN+, the performance of $\mathrm{FCN}_{n/\infty}$ is invariant to all symmetry transformations, the performance of LCN_n and $\mathrm{VEC}_\infty = \mathrm{LCN}_\infty$ are invariant to $\mathrm{O}(3)^d$ (and its subgroups), and the performance of VEC_n , $\mathrm{LAP}^{4/8}$ and $\mathrm{GAP}_{n/\infty}$ are invariant to $\mathrm{O}(3) \otimes \mathbf{I}_d$.

5.3. Effect of spurious symmetries

In order to analyze the consistency of the data \mathcal{D} and algorithm \mathcal{A} , we examine performance in the presence of various spurious symmetries, which we introduce through the five rotated datasets and through the six different model families. We focus on two main findings: (1) the strength of

³ When investigating the impact of inference algorithms, it would be preferable to compare finite-width Bayesian inference to SGD. Unfortunately, Bayesian inference is too expensive to perform exactly and approximations may induce unwanted biases. As such, we instead use infinite-width Bayesian inference (i.e. NNGP regression (Hron et al., 2020a)), whose performance is usually very similar to NTK regression.

⁴ A few $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ have only 1 or 2 successful runs.

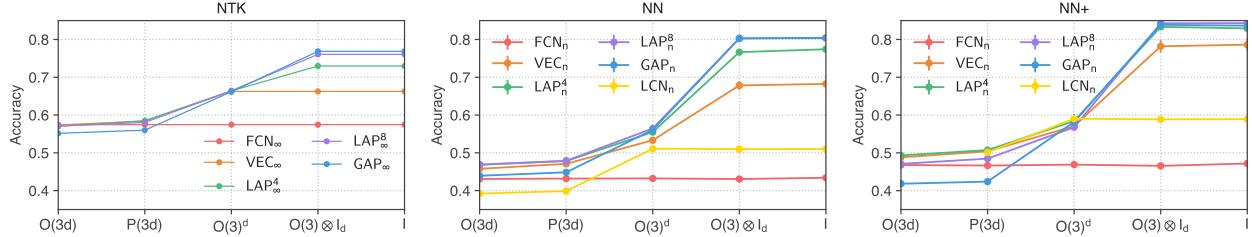


Figure 2: **Performance of various architectures under different data transformations.** Left: network trained using NTK. Middle: finite-width network trained with a small learning rate and no L2-regularization. Right: larger learning rate and L2-regularization. We transform all images in CIFAR-10 by a random element from one of the five groups (x -axis) and plot the accuracy (y -axis) for 6 architectures. Performance degrades in the presence of spurious symmetry, and the decrease is similar regardless of whether that symmetry arises from the data or from the model (see Thm. 4.1).

the spurious symmetry controls performance, regardless of how it is introduced; and (2) SGD confers its main benefits in the absence of spurious symmetries.

Performance dictated by spurious symmetries. For each fixed \mathcal{I} , we choose the highest-performing triplet $(\mathcal{D}, \text{GAP}, \mathcal{I})$ as a baseline that exhibits the strongest consistency between \mathcal{A} and $\mu_{\mathcal{D}}$. We then progressively break this consistency by injecting spurious symmetries in two ways: (1) fixing $(\text{GAP}, \mathcal{I})$ and changing the dataset to $\tau\mathcal{D}$; and (2) fixing $(\mathcal{D}, \mathcal{I})$ and changing GAP to \mathcal{M}_{τ} , where \mathcal{M}_{τ} represents an architecture that is τ -invariant (c.f.Theorem 4.1).

From Fig. 2, we see that for each fixed $\mathcal{I} \in \{\text{NTK}, \text{NN}, \text{NN+}\}$, test performance monotonically decreases as the symmetries become more “spurious”. We also observe that performance is to a good approximation determined by the spurious symmetry itself, independent of the way it was introduced. In particular, across all settings we observe that the performance of $(\tau\mathcal{D}, \mathcal{M}, \mathcal{I})$ is close to that of $(\mathcal{D}, \mathcal{M}_{\tau}, \mathcal{I})$ which is itself nearly the same as $(\tau\mathcal{D}, \mathcal{M}_{\tau}, \mathcal{I})$. As a concrete example of this relationship, in Fig. 2 for $\mathcal{I} = \text{NTK}$, the performance of GAP_{∞} under an $\text{O}(3)^d$ data transformation equals the performance of VEC_{∞} with no data transformation, which is expected since Theorem 4.1 implies VEC_{∞} is $\text{O}(3)^d$ -invariant.

Spurious symmetries eliminate the benefit of SGD. We examine the benefits of SGD by varying \mathcal{I} from NTK (no SGD-impact, left panel of Fig. 2) to NN (weak SGD-impact, middle panel) to NN+ (strong SGD-impact, right panel). We find that the behavior depends strongly on whether or not $(\mathcal{D}, \mathcal{M})$ has spurious symmetries. In the presence of spurious symmetries, i.e. when either \mathcal{M} is invariant to a symmetry group larger than $\text{O}(3) \otimes \mathbf{I}_d$ or when \mathcal{D} is rotated by an element from such a group, there is no benefit from SGD, as we observe that NTK outperforms both NN and NN+. In the absence of spurious symmetries, i.e. when $\mathcal{M} \in \{\text{GAP}, \text{VEC}, \text{LAP}^{4/8}\}$ and when the transformation

applied to \mathcal{D} is from a group no larger than $\text{O}(3) \otimes \mathbf{I}_d$, we observe a significant boost in performance when changing the inference algorithm from $\mathcal{I} = \text{NTK}$ to $\mathcal{I} = \text{NN+}$, e.g. $77\% \rightarrow 84\%$ when $\text{GAP}_{\infty} \rightarrow \text{GAP}_n$. The performance gain of VEC is most significant ($67\% \rightarrow 78\%$ when $\text{VEC}_{\infty} \rightarrow \text{VEC}_n$), which will be discussed in detail in the following section. Overall, our empirical results suggest that SGD does not improve the performance when $(\mathcal{D}, \mathcal{M})$ has spurious symmetries, at least for CIFAR-10 without data-augmentation. We argue that, when studying the benefits of feature learning, it is essential to take into account both the data distribution and the architecture choices.

5.4. Symmetry breaking at finite width in VEC_n .

Weight-sharing implies that VEC_n is $\text{O}(3) \otimes \mathbf{I}_d$ -invariant; however, when $n \rightarrow \infty$, this symmetry group is enlarged to $\text{O}(3)^d$ (see Theorem 4.1). This observation highlights a novel and previously underappreciated aspect of convolutional models with vectorization: symmetry breaking at finite width. As we will see, this broken symmetry can have a significant impact on performance.

In Fig.3, we plot the performance on CIFAR-10 of VEC_n as a function of width under the interventions: (1) rotating the data by $\text{O}(3) \otimes \mathbf{I}_d$ (which is actually a null operation with no impact) or by $\text{O}(3)^d$, and (2) training under NN or NN+ (i.e. +LR+L2). We summarize our findings below.

Strong benefits of finite width (Blue Lines). Both with and without the intervention of +LR+L2, the performance decreases monotonically and dramatically⁵ towards or across the NTK performance (**Faded Orange Lines**) when n passes a threshold ($n \approx 2^6/2^7$ for NN/NN+).

Spurious symmetry eliminates benefits of finite width. When changing $\mathcal{D} \rightarrow \tau\mathcal{D}$ where $\tau \in \text{O}(3)^d$, a spurious symmetry incompatible with weight-sharing convolution, modern wisdom in machine learning is restored: perfor-

⁵This phenomenon was first observed in (Lee et al., 2020).

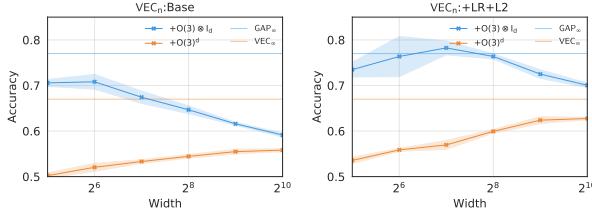


Figure 3: Finite width Effect for VEC_n . Left: networks trained with a small learning rate and no L2-regularization. Right: with a larger learning rate and L2-regularization. With $O(3)^d$ imposed on the data, performance of VEC_n is far below the performance of VEC_∞ (67%, orange horizontal line), but improves monotonically as a function of n . However, with smaller rotation group $O(3) \otimes \mathbf{I}_d$ (has no impact on VEC_n), performance degrades substantially after the peak (blue lines). This suggests a strong finite-width effect of VEC_n due to symmetry breaking.

mance improves with overparameterization, gradually approaching NTK performance (Orange Lines).

Large fluctuations at small n (Blue shaded area in Fig. 3.) The intervention +LR+L2 not only improves the performance of VEC_n but also increases the variance of the performance substantially⁶ for small width ($n = 2^5, 2^6$). For example, when $n = 2^6$ the validation accuracy are 68%, 78%, 78%, 79%, 79%. Such variability does not occur in the presence of spurious symmetries ($O(3)^d$).

One interpretation of Thm. 3.1 and the above observations is that strong algorithmic forces are needed to help VEC_n escape from the undesirable function class LCN and move towards the more desirable properties of GAP. As discussed earlier, one mechanism behind such movement could be learning an approximation of pooling in the readout layer.

6. Data Improves Data Efficiency

In the previous section, we investigated the consistency of various machine learning systems through the lens of symmetries. In this section, we further investigate the interplay of the components of the $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ triplet by conducting a fine-grained analysis of learning curves on various SoTA vision models.

Recently, Kaplan et al. (2020); Bahri et al. (2021) and other authors suggest that for real-world problems the learning objective often has a power-law $\sim m^{-\alpha}$ dependence of training set size m , where the exponent α is a constant that usually independent from m . A surprising finding here

⁶For each configuration, the standard deviation is computed over 5 different random initialization. Some of the runs fail and we only plot the configurations with at least 4 successful runs (best training accuracy is at least 95%).

is that, for certain triplets $(\mathcal{D}, \mathcal{M}, \mathcal{I})$, α can grow as m becomes larger, i.e. data improves data efficiency (DIDE); see Fig. 4. In what follows, we first examine the learning curve of VEC_n to better understand the huge performance gap between VEC_n and VEC_∞ . We then move to SoTA models, in which we observe a “cusp” in the learning curve. Finally, we provide possible explanations about the observed phenomena.

6.1. DIDE for VEC_n

We vary the training set size of CIFAR-10 from 320 to 45k (the whole un-augmented training set) and then to 90k (adding left-right flip augmentation) and plot the learning curves in Fig. 4 for various $(\mathcal{D}, \mathcal{M}, \mathcal{I})$. We observe a dramatic speedup of learning in our baseline setting $(\mathcal{D}, \mathcal{M}, \mathcal{I}) = (\text{CIFAR-10}, \text{VEC}_n, \text{SGD})$ (Blue Lines). Pictorially, the slope of the learning curve is steepened substantially in the log-log plot. We then did an ablation study by changing one member in $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ at a time: (1) **Inference Algorithm \mathcal{I}** (Orange Lines): SGD to NTK⁷ (2) **Model \mathcal{M}** (Red Lines): $\text{VEC}_n \rightarrow \text{LCN}_n$, and (3) **Data \mathcal{D}** (Green Lines): $\mathcal{D} \rightarrow \tau \mathcal{D}$, where $\tau \in O(3)^d$ randomly selected. In all cases above, this phenomenon disappeared.

Possible Explanation. Recall from Theorem that 3.1, the function class of VEC_n is sandwiched in-between GAP_n and LCN_{dn} . In the small dataset regime, the algorithm is unable to move VEC_n far away from the LCN_{dn} -like regime ($O(3)^d$ -invariance). This behavior is reflected from Fig. 4: for $m < 10^4$ the Blue Lines (VEC_n) is very close to the Green Lines (VEC_n with $O(3)^d$ spurious symmetry applied to the data), and the slopes of all learning curves are comparable. With more data, VEC_n is able to break the $O(3)^d$ symmetry and being to perform feature learning, which moves the model away from LCN-like regime and towards GAP_n . We provide further empirical support of this hypothesis in Sec. E by showing that better performing VEC_n -learners are closer to GAP_n -learners and further from the VEC_∞ -learners, and vice versa.

6.2. DIDE for ImageNet

Larger deep learning systems can exhibit qualitative differences from smaller ones. As such, we examine the DIDE phenomenon for SoTA models on ImageNet (Deng et al., 2009), which has $m_{\text{ImageNet}} \sim 10^6$ training samples with size $224 \times 224 \times 3 (= 3d)$.

We subsample $m_i = 2^{-i/2} m_{\text{ImageNet}}$ images as our training set, where $i \in [12]$. In Fig. 5, we plot the learning curves for four $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ triplets. We set \mathcal{M} to be ResNet101 or a (small MLP-)Mixer (Tolstikhin et al., 2021), \mathcal{D} to be

⁷This requires changing VEC_n to VEC_∞ .

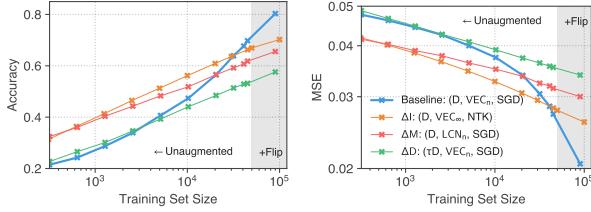


Figure 4: More Data Improves Learning Curve of VEC_n . Left: test accuracy. Right: MSE loss. The slope of the learning curve of the optimal baseline $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ configuration (blue) increases significantly in the large-data regime. This data-improved efficiency disappears under interventions that suboptimally deform the data \mathcal{D} (green), the model \mathcal{M} (red), or the inference algorithm \mathcal{I} (orange).

the original (unrotated) ImageNet \mathcal{D}_{IN} or a rotated version $\tau\mathcal{D}_{\text{IN}}$, for some $\tau \in O(3)^d$. We keep \mathcal{I} (SGD, see Sec. B.2) fixed. For each configuration, we interpolate the first/last six points (i.e. $i \geq 6/i < 6$) with straight lines (in the log-log plot) and compute the slope α (see legends in Fig. 5). We treat ResNet101 trained on clean images as our baseline, which is the most efficient and consistent $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ system among the four. We observe the following. (1) Almost perfect power-law scaling for the baseline (**Blue Dashed Line**). (2) A cusp around $m_{i=6}$ for the remaining learning curves, which have two phases: relatively flat in the first phase and steepened in the second one. (3) Surprisingly, the slopes ($\alpha = 0.49, 0.38$) of $(\text{ResNet101}, \tau\mathcal{D}_{\text{IN}}, \mathcal{I})$ and $(\text{Mixer}, \mathcal{D}_{\text{IN}}, \mathcal{I})$ in the second phase essentially catch up with that ($\alpha = 0.41$) of the best one $(\text{ResNet101}, \mathcal{D}_{\text{IN}}, \mathcal{I})$. These observations suggest that with more data the system can overcome the spurious symmetries $O(3)^d$.

Finally, to test the limit of deep learning systems in overcoming spurious $O(3)^d$ symmetries, we further scale up \mathcal{M} in the ResNet family (He et al., 2016) and in the EfficientNet family (Tan & Le, 2019). In the right panel of Fig. 4, we make a scatter plot showing the accuracy of the $O(3)^d$ -rotated vs original datasets. Each datapoint corresponds to one model. For the ResNet family, the top-1 accuracy gap between the rotated and the unrotated dataset drops from about 10% (ResNet-18) to about 6% (ResNet-200) and for the EfficientNet family, this gap drops from about 4% (EfficientNet B0) to *only* about 1% (EfficientNet B7⁸), which is quite remarkable.

Discussion of DIDE. The change of the slopes of the learning curves suggests that the function classes on the left/right of the cusp might be qualitatively quite different. The cusp happens around $m_{i=6} \sim 2 \times 10^5$, which is of the same order of $\dim(O(3)^d) = 3d = 150528$ and

⁸Trained by about 180 epochs.

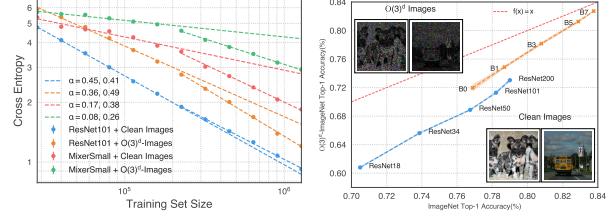


Figure 5: SoTA models overcome $O(3)^d$ spurious symmetries. Left: learning curves for four $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ triplets. There is a cusp in all learning curves except the optimal configuration (Blue curve.) Learning efficiency significantly improves for those curves after the cusp. Right: Top-1 accuracy for the ResNet and the EfficientNet families when the dataset is $O(3)^d$ -rotated (x -axis) and unrotated (y -axis.) SoTA models can overcome the $O(3)^d$ spurious symmetries.

$\dim(O(16^2 \times 3)) = 294528$ (the size of the patches in the Mixer is $(16, 16, 3)$). This agreement suggests that to overcome spurious symmetry \mathcal{G} (either from the models or data), at least $\sim \dim(\mathcal{G})$ extra training points are needed. We also test the capability of the ResNet family in overcoming the $P(3d)$ ($< O(3d)$) symmetry (Sec.F.2), but the test and training accuracy remain below 35% for all ResNet models. If we extrapolate the dimension counting argument, $\dim(O(3d)) \sim 10^{10}$ many training points may be needed to overcome the $O(3d)$ symmetries.

7. Conclusion

We consider machine learning methods as an integrated system of data, models and inference algorithms and study the basic symmetries of various machine learning systems $(\mathcal{D}, \mathcal{M}, \mathcal{I})$. We examine the relation between the performance and the consistency of the triplet $(\mathcal{D}, \mathcal{M}, \mathcal{I})$ through the lens of symmetries. We find that learning is most efficient when the learning algorithm is consistent with the data distribution. Finally, we observe that, for many triplets $(\mathcal{D}, \mathcal{M}, \mathcal{I})$, the slopes of the scaling law curves can improve with more data, suggesting the function class is transitioning to a new one that is dramatically more data-efficient than the one obtained from initialization. Theoretical characterization of how does this transition occur might be a crucial step to the understanding of feature learning in neural networks.

Acknowledgment.

We thank Jascha Sohl-dickstein, Sam Schoenholz, Jaehoon Lee, Roman Novak, and Yasaman Bahri for insightful discussion, and Sam Schoenholz, Jaehoon Lee, Roman Novak, and Mingxing Tang for engineering support. We are also grateful to Roman Novak and the anonymous reviewers for feedback and suggestions on an earlier draft of this work.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016.
- Adlam, B. and Pennington, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization, 2020.
- Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, 2018.
- Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, 2019.
- Bach, F. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.
- Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.
- Bietti, A. Approximation and learning with deep convolutional models: a kernel perspective, 2021.
- Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks, 2021.
- Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/google/jax>.
- Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.
- Chen, S., Dobriban, E., and Lee, J. A group-theoretic framework for data augmentation. *Advances in Neural Information Processing Systems*, 33:21321–21333, 2020.
- Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1305–1338. PMLR, 09–12 Jul 2020. URL <http://proceedings.mlr.press/v125/chizat20a.html>.
- Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.
- Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.
- Daniely, A. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2017.
- Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pp. 2253–2261, 2016.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.
- Favero, A., Cagnetta, F., and Wyart, M. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios, 2021.
- Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3-4):121–136, 1975.
- Garriga-Alonso, A., Aitchison, L., and Rasmussen, C. E. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019.
- Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.
- Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.
- Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., and van Zee, M. Flax: A neural network library and ecosystem for JAX, 2020. URL <http://github.com/google/flax>.
- Hron, J., Bahri, Y., Novak, R., Pennington, J., and Sohl-dickstein, J. Exact posterior distributions of wide bayesian neural networks. *arXiv preprint arXiv:2006.10541*, 2020a.
- Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. Infinite attention: Nngp and ntk for deep attention networks, 2020b.
- Hu, W., Xiao, L., Adlam, B., and Pennington, J. The surprising simplicity of the early-time learning dynamics of neural networks. *arXiv preprint arXiv:2006.14599*, 2020.
- Ingrosso, A. and Goldt, S. Data-driven emergence of convolutional structure in neural networks. *arXiv preprint arXiv:2202.00565*, 2022.
- Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018a.
- Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018b.
- Ji, Z. and Telgarsky, M. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798, 2019a.
- Ji, Z. and Telgarsky, M. J. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019b.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- Lecun, Y. Generalization and network design strategies. In *Connectionism in perspective*. Elsevier, 1989.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=B1EA-M-0Z>.
- Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.
- Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: an empirical study. *arXiv preprint arXiv:2007.15801*, 2020.
- Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.
- Li, Z., Zhang, Y., and Arora, S. Why are convolutional nets more sample-efficient than fully-connected nets? *arXiv preprint arXiv:2010.08515*, 2020.
- Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SJeLIGBKPS>.
- Matthews, A., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.
- Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
- Mei, S., Misiakiewicz, T., and Montanari, A. Learning with invariances in random features and kernel models, 2021.
- Nakkiran, P., Kaplun, G., Kalimeris, D., Yang, T., Edelman, B. L., Zhang, F., and Barak, B. Sgd on neural networks learns functions of increasing complexity. *arXiv preprint arXiv:1905.11604*, 2019.
- Neal, R. M. Priors for infinite networks (tech. rep. no. crg-tr-94-1). *University of Toronto*, 1994.

- Neyshabur, B. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.
- Neyshabur, B. Towards learning convolutions from scratch. *arXiv preprint arXiv:2007.13657*, 2020.
- Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural tangents: Fast and easy infinite neural networks in python. *arXiv preprint arXiv:1912.02803*, 2019a.
- Novak, R., Xiao, L., Lee, J., Bahri, Y., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019b.
- Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2020. URL <https://github.com/google/neural-tangents>.
- Petrini, L., Favero, A., Geiger, M., and Wyart, M. Relative stability toward diffeomorphisms indicates performance in deep nets. *Advances in Neural Information Processing Systems*, 34, 2021.
- Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, 2016.
- Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.
- Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.
- Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. *International Conference on Learning Representations*, 2017.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792): 706–710, 2020.
- Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- Sohl-Dickstein, J., Novak, R., Schoenholz, S. S., and Lee, J. On the infinite width limit of neural networks with a standard parameterization. *arXiv preprint arXiv:2001.07301*, 2020.
- Song, M., Montanari, A., and Nguyen, P. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.
- Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70), 2018.
- Su, L. and Yang, P. On learning over-parameterized neural networks: A functional approximation perspective. *arXiv preprint arXiv:1905.10826*, 2019.
- Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.
- Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- von Luxburg, U. and Bousquet, O. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5: 669–695, 2004.
- Wadia, N. S., Duckworth, D., Schoenholz, S. S., Dyer, E., and Sohl-Dickstein, J. Whitening and second order optimization both destroy information about the dataset, and can make generalization impossible. *arxiv preprint arXiv:2008.07545*, 2020.
- Xiao, L. Eigenspace restructuring: a principle of space and frequency in neural networks, 2021.
- Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. 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*, 2018a.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. 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*, pp. 5393–5402, 2018b.

Xu, Z. J. Understanding training and generalization in deep learning by fourier analysis. *arXiv preprint arXiv:1808.04295*, 2018.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

Yang, G. and Hu, E. J. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Yang, G. and Salman, H. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

Yang, G., Santacroce, M., and Hu, E. J. Efficient computation of deep nonlinear infinite-width neural networks that learn features. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=tUMr0Iox8XW>.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

Supplementary Material

A. Glossary

We use the following abbreviations in this work:

- +L2: Adding L2 regularization.
- +LR: Using a large learning rate.
- FCN_n : Fully-connected networks with width n .
- FCN_∞ : Infinite width FCN_n .
- VEC_n : Convnet with width n and a flattening readout layer.
- VEC_∞ : Infinite width VEC_n .
- LCN_n : Locally-connected network with width n .
- LCN_∞ : Infinite width LCN_n , which is the same as VEC_∞ .
- GAP_n : Convnet with width n and a global average readout layer.
- GAP_∞ : Infinite width GAP_n .
- LAP_n^k : Similar to GAP_n , except the readout layer is a (k, k) average pooling.
- LAP_∞^k : Infinite width LAP_n^k .

B. Experimental details

We use Neural Tangents (NT) library (Novak et al., 2020) built on top of JAX (Bradbury et al., 2018) for all Cifar10 experiments, and the ImageNet codebase from FLAX⁹(Heek et al., 2020) for ResNets and Mixer experiments on ImageNets.

B.1. Cifar10 Experiments

The experimental setup is almost the same as in (Lee et al., 2020).

Architectures. For FCN_n , LCN_n , VEC_n and GAP_n , the number of hidden layers are 8 and the widths (number of channels) are $n = 1024, 32, 128$ and 128 , resp. For all CIFAR10 experiments, we only use Relu as the activation function. We use NTK parameterization and the variances of initialization are chosen to be $\sigma_\omega^2 = 2$ and $\sigma_b^2 = 0.01$ for the weights and biases, resp.

Training Details. We use MSE as our loss function, which is defined to be

$$L(\theta; D_{\text{mini}}) = \frac{1}{2 \times K |D_{\text{mini}}|} \sum_{(x_i, y_i) \in D_{\text{mini}}} |f_\theta(x_i) - y_i|^2 + \frac{\lambda}{2} \|\theta\|_2^2 \quad (\text{S1})$$

where D_{mini} is a mini-batch with $|D_{\text{mini}}| = 40$ and $K = 10$ is the number of classes. The regularization is set to be $\lambda = 10^{-7}$ if L^2 regularization (+L2) is applied otherwise 0. SGD + Momentum (mass = 0.9) is used for all experiments. The learning rate is set to be $\eta = c\eta_0$, where $c = 8$ if using a larger learning rate (+LR) and 1 otherwise. Here η_0 ¹⁰ is estimated by $\frac{2}{\lambda_{\max}}$, where λ_{\max} is the largest eigenvalue of the finite-width NTK (estimated by MC sampling). With $c = 8$, we are about a factor of $2 \sim 4$ smaller than the maximal feasible learning rate observed empirically.

We use 45k images as training set and reserve the remaining 5k as validation set. All finite width experiments are initially set to be trained for at least 10^6 steps (about 900 epochs), but will be early-stopped if the training accuracy reaches 100% with fewer steps. Among the successful runs (with training accuracy $\geq 95\%$), we pick the highest test accuracy along training and average them over 5 random runs (if all runs succeed.)

⁹<https://github.com/google/flax/blob/main/examples/imagenet/README.md>

¹⁰We use `max_learning_rate` function from Neural Tangents Library to estimate η_0 .

B.2. ImageNet Experiments

We use the ImageNet codebase from FLAX¹¹(Heek et al., 2020) for our ResNets experiments. We adopt most of the training configurations except change the number of epochs to 150. Note that we also rotate (see Sec. B.3) and/or subset the dataset when needed.

The model for the Mixer experiments on ImageNets is adopted from Sec.E in (Tolstikhin et al., 2021). The training configurations are identical to that of the ResNets above. The EfficientNet family models are trained using Tensorflow (Abadi et al., 2016), which are adopted from <https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet>.

B.3. Data Transformation.

Let $\mathcal{D} = \{(x_i, y_i), i \in [m_{\text{train}} + m_{\text{test}}]\}$ be the data set (e.g. ImageNet, Cifar10.) Here $x_i \in \mathbb{R}^{3d}$ is a flattened input image, and m_{train} and m_{test} are the numbers of images in the training set (including data-augmentation) and in the test set, resp.

O(3d) Transformation. Randomly sample $Q \in \text{O}(3d)$, a $3d \times 3d$ orthogonal matrix. The transformed dataset $\tau\mathcal{D}$ is

$$\tau\mathcal{D} := \{(Qx_i, y_i) : i \in [m_{\text{train}} + m_{\text{test}}]\}. \quad (\text{S2})$$

The $P(3d)$ (permutation) transformation is defined similarly.

O(3)^d Transformation. We reshape each $x_i \in \mathbb{R}^{3d}$ into $x_i \in \mathbb{R}^{H \times W \times 3}$ where H and W are the height and width of the images (e.g. $H = W = 32$, $d = 32^2$ for Cifar10 and $H = W = 224$, $d = 224^2$ for ImageNet). Independently sample HW many 3×3 random orthogonal matrix $(Q_{h,w})_{h \in [H], w \in [W]}$. The transformed dataset $\tau\mathcal{D}$ is

$$\tau\mathcal{D} := \left\{ \left((Q_{h,w}x_{i;h,w})_{h \in [H], w \in [W]}, y_i \right) : i \in [m_{\text{train}} + m_{\text{test}}] \right\}. \quad (\text{S3})$$

O(3) \otimes I_d Transformation. In this case, we sample only *one* 3×3 orthogonal matrix and the transformed dataset is defined to be

$$\tau\mathcal{D} := \left\{ \left((Qx_{i;h,w})_{h \in [H], w \in [W]}, y_i \right) : i \in [m_{\text{train}} + m_{\text{test}}] \right\}. \quad (\text{S4})$$

Note that the same rotation matrix Q is applied to all pixels regardless of their spatial locations.

C. Proof of Theorem 3.1

We use FCN_n to denote the class of functions that can be expressed by L -hidden layer fully-connected networks whose widths are equal to n . Similar notation applies to other architectures.

Corollary 1. *We have the following*

$$\text{GAP}_n \subseteq \text{VEC}_n \subseteq \text{LCN}_n \subseteq \text{VEC}_{dn}, \quad \text{LCN}_n \subseteq \text{FCN}_{dn} \quad (\text{S5})$$

Proof. We only need to prove $\text{LCN}_n \subseteq \text{VEC}_{dn}$ because the others are obvious. Let $\text{LCN}_n(x)_{\alpha,i}^l$ denote the post-activation at layer l , spatial location α and channel index i of a LCN_n with input x and $\text{VEC}_n(x)_{\alpha,i}^l$ is defined similarly. It suffices to prove that for any LCN with width n there is a VEC with width dn such that for any $l \geq 1$ (i.e. not the input layer)

$$\text{VEC}_{dn}(x)_{\alpha,\alpha n+i}^l = \text{LCN}_n(x)_{\alpha,i}^l \quad (\text{S6})$$

since we could choose the readout weights of VEC_{dn} at locations $(\alpha, \alpha n + i)$ to match the one of LCN_n at location (α, i) and zero out the remaining entries. We prove this by induction and assume it holds for l (the base case $l = 1$ is obvious). Then the LCN_n and VEC_n at layer $l + 1$ can be written as

$$\text{LCN}_n(x)_{\alpha,j}^{l+1} = \phi \left(\frac{1}{\sqrt{n(2k+1)}} \sum_{i \in [n], \beta \in [-k,k]} \text{LCN}_n(x)_{\alpha+\beta,i}^l \omega_{\beta,ij}^{l+1}(\alpha) \right)$$

¹¹<https://github.com/google/flax/blob/main/examples/imagenet/README.md>

and

$$\text{VEC}_{dn}(x)_{\alpha,j}^{l+1} = \phi \left(\frac{1}{\sqrt{dn(2k+1)}} \sum_{i \in [dn], \beta \in [-k,k]} \text{VEC}_{dn}(x)_{\alpha+\beta,i}^l \tilde{\omega}_{\beta,ij}^{l+1} \right)$$

One can show that Equation S6 holds for $(l+1)$ by choosing the parameters of VEC_{dn} as follows

$$\tilde{\omega}_{\beta,ij}^{l+1} = \sqrt{d} \omega_{\beta,i-(\alpha+\beta)n,j-\alpha n}^{l+1} \quad \text{if } \alpha n \leq j < \alpha(n+1) \quad \text{and} \quad (\alpha+\beta)n \leq i < (\alpha+\beta)(n+1)$$

and 0 otherwise.

□

D. Proof of Symmetries

Proof. For simplicity, we present the proof for full-batch training. The proof can be applied to mini-batch training as long as order of the mini-batch is fixed. Let τ be a rotation in $O(3d)$ or $O(3)^d$ or $O(3) \otimes \mathbf{I}_d$, depending on the architectures ($\text{FCN}_n, \text{LCN}_n, \text{VEC}_n, \text{GAP}_n$) and the tuple θ and γ denote the parameters of the first and remaining layers of the network, respectively. Let $h(\tau x, \theta) = \langle \tau x, \theta \rangle$ denote the pre-activations of the first-hidden layer in the rotated coordinate. Here $\langle \cdot, \cdot \rangle$ is the bilinear map (a dense layer or a convolutional layer with or without weight-sharing, etc.), not the inner product. The loss with L^2 -regularization is

$$R_\lambda(\theta, \gamma) = L(h(\tau \mathcal{X}, \theta), \gamma) + \frac{1}{2} \lambda (\|\theta\|_2^2 + \|\gamma\|_2^2) \quad (\text{S7})$$

where $L(h(\tau \mathcal{X}, \theta), \gamma)$ is the raw loss of the network. For each random instantiation $\theta = \theta_0$ with θ_0 drawn from standard Gaussian iid, we instantiate a coupled network from the un-rotated coordinates but with a different instantiation in the first layer $\theta^\tau = \tau^* \theta_0$ and keep the remaining layers unchanged, i.e. $\gamma^\tau = \gamma_0$. Here τ^* is the adjoint of τ and note that $\tau^* \theta_0$ and θ_0 have the same distribution by the Gaussian initialization of θ_0 and the definition of τ . The regularized loss associated to this instantiation is

$$R_\lambda(\theta^\tau, \gamma^\tau) = L(h(\mathcal{X}, \theta^\tau), \gamma^\tau) + \frac{1}{2} \lambda (\|\theta^\tau\|_2^2 + \|\gamma^\tau\|_2^2) \quad (\text{S8})$$

It suffices to prove that for each instantiation $\theta = \theta_0$ drawn from Gaussian, the following holds for all gradient steps t

$$(\theta_t^\tau, \gamma_t^\tau) = (\tau^* \theta_t, \gamma_t). \quad (\text{S9})$$

We prove this by induction on t and $t = 0$ is true by definition. Assume it holds when $t = t$. Now the update in γ and γ^τ with learning rate η are

$$\gamma_{t+1} = \gamma_t - \eta \left(\frac{\partial L}{\partial \gamma} \Big|_{(h(\tau \mathcal{X}, \theta_t), \gamma_t)} \right)^T - \eta \lambda \gamma_t \quad (\text{S10})$$

$$\gamma_{t+1}^\tau = \gamma_t^\tau - \eta \left(\frac{\partial L}{\partial \gamma} \Big|_{(h(\mathcal{X}, \theta_t^\tau), \gamma_t^\tau)} \right)^T - \eta \lambda \gamma_t^\tau \quad (\text{S11})$$

It is clear $\gamma_{t+1} = \gamma_{t+1}^\tau$ by induction since $h(\tau \mathcal{X}, \theta_t) = h(\mathcal{X}, \theta_t^\tau)$. Similarly,

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\partial L}{\partial h} \frac{\partial h}{\partial \theta} \Big|_{(\tau \mathcal{X}, \theta_t)} \right)^T - \lambda \theta_t \quad (\text{S12})$$

$$\theta_{t+1}^\tau = \theta_t^\tau - \eta \left(\frac{\partial L}{\partial h} \frac{\partial h}{\partial \theta^\tau} \Big|_{(\mathcal{X}, \theta_t^\tau)} \right)^T - \lambda \theta^\tau \quad (\text{S13})$$

Note that by the chain rule and induction assumption

$$\frac{\partial h}{\partial \theta^\tau} \Big|_{(\mathcal{X}, \theta_t^\tau)} = \frac{\partial h}{\partial \theta} \Big|_{(\mathcal{X}, \theta_t^\tau)} \frac{\partial \theta^\tau}{\partial \theta} = \frac{\partial h}{\partial \theta} \Big|_{(\mathcal{X}, \theta_t^\tau)} \tau \quad (\text{S14})$$

This implies $\theta_{t+1}^\tau = \tau^* \theta_{t+1}$.

□

Remark S1. It is not difficult to see the apply proof apply to Non-Gaussian i.i.d. initialization (e.g. uniform distribution) and/or adding L^p -regularization when the rotation groups are replaced by the corresponding permutation groups. Empirically, we observe that replacing the first layer Gaussian initialization by uniform distribution does not change the performance of the network much. See Fig.S5.

Remark S2. The proof works for other parameterization methods, including NTK-parameterization(Jacot et al., 2018b), standard parameterization (Sohl-Dickstein et al., 2020), mean-field parameterization(Song et al., 2018) and ABC-parameterization (Yang & Hu, 2020)

E. Measuring the Effect of Symmetry Breaking of VEC_n .

The discussion in the main text suggests that breaking the $O(3)^d$ symmetry, making the network to exploit the smaller symmetry group $O(3) \otimes \mathbf{I}_d$ might be important to good performance of VEC_n . To measure the effect of symmetry breaking and the reliance of VEC_n on the $O(3) \otimes \mathbf{I}_d$ symmetry, we compare the distance of VEC_n to VEC_∞ ($O(3)^d$ invariant) and to GAP_n ($O(3) \otimes \mathbf{I}_d$ invariant). More precisely, for two learning algorithm \mathcal{A}_1 and \mathcal{A}_2 trained on \mathcal{D}_T , we defined the square distance between them to be

$$\text{S-Dist}(\mathcal{A}_1, \mathcal{A}_2) = \mathbb{E}_{x \sim \mathcal{X}} |\mathcal{A}_1(\mathcal{D}_T)(x) - \mathcal{A}_2(\mathcal{D}_T)(x)|^2. \quad (\text{S15})$$

If the learning algorithm \mathcal{A}_i is stochastic, then we use the mean prediction in the above definition. E.g, if \mathcal{A}_1 depends on the initialization θ_0 which is a random variable and \mathcal{A}_2 is deterministic, then we define the squared distance to be

$$\text{S-Dist}(\mathcal{A}_1, \mathcal{A}_2) = \mathbb{E}_{x \sim \mathcal{X}} |\mathbb{E}_{\theta_0} \mathcal{A}_1(\mathcal{D}_T; \theta_0)(x) - \mathcal{A}_2(\mathcal{D}_T)(x)|^2. \quad (\text{S16})$$

Using this definition, we can measure the discrepancy between two rotated systems $(\tau_1 \mathcal{D}, \mathcal{M}, \mathcal{I})$ and $(\tau_2 \mathcal{D}, \mathcal{M}, \mathcal{I})$ by computing $\text{S-Dist}(\mathcal{A}^{\tau_1}, \mathcal{A}^{\tau_2})$, where $\mathcal{A} = (\mathcal{M}, \mathcal{I})$ and $\tau_{1/2}$ are coordinate transformations. Note that if the system is strictly \mathcal{G} invariant, then $\text{S-Dist}(\mathcal{A}^{\tau_1}, \mathcal{A}^{\tau_2}) = 0$ for all $\tau_{1/2} \in \mathcal{G}$.

We use the exponential map to construct a continuous path¹² from the identity operator \mathbf{Id} to a random element in $O(3)^d$. More precisely, we randomly sample d 3×3 skew-symmetric matrices $A = (A_0, \dots, A_{d-1}) \subseteq (\mathbb{R}^{3 \times 3})^d$ and define $\tau = \exp(-A)$ and $\tau_t = \exp(-tA)$ for $t \in [0, 1]$. Then $(\tau_t)_{t \in [0, 1]} \subseteq O(3)^d$ is a continuous path from \mathbf{Id} to τ . We then construct new datasets $\tau_t \mathcal{D}$ (see Fig. S2 for a sample of the continuously rotated images) and study the behavior of the corresponding systems

$$\{(\tau_t \mathcal{D}, \mathcal{M}, \mathcal{I}) : t \in [0, 1]\} \quad (\text{S17})$$

We vary the width $n = 64$ to $n = 512$ dyadically and t from $[0, 1]$ with equal distance and train the networks on CIFAR10 as in the NN+ setting (+LR+L2). Finally, we average the predictions of the learned network over 10 random initialization as an approximation of $\mathbb{E}\text{VEC}_n^{\tau_t}(x)$ and etc. We summarize the observation below.

- As n and/or t increases, the (ensemble) test performance decays monotonically (left panel in Fig.S1). This is because increasing n and/or t discourages VEC_n to utilize the smaller symmetry group $O(3) \otimes \mathbf{I}_d$.
- As a function of n or t , $\text{S-Dist}(\text{VEC}_n^{\tau_t}, \text{GAP}_n)$ increases monotonically (middle panel) while $\text{S-Dist}(\text{VEC}_n^{\tau_t}, \text{VEC}_\infty)$ decreases monotonically (right panel). Thus, small n moves the VEC_n learner towards the GAP_n learner while increasing n and/or the strength of rotation moves it away from the GAP_n learner and towards the VEC_∞ learner.

¹²More precisely, the path lies in $\text{SO}(3)^d \subseteq O(3)^d$.

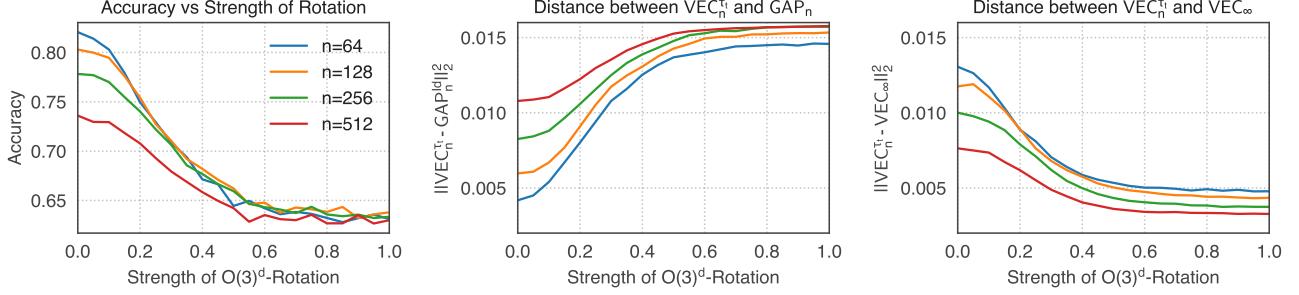


Figure S1: High performing VEC_n -learner is closer to GAP_n -learner and far away from VEC_{∞} -learner, and vice versa. Even in the NN+ setting, VEC_n is closer to GAP_n for small n and moves towards VEC_{∞} with stronger symmetry and/or larger n and accuracy drops.

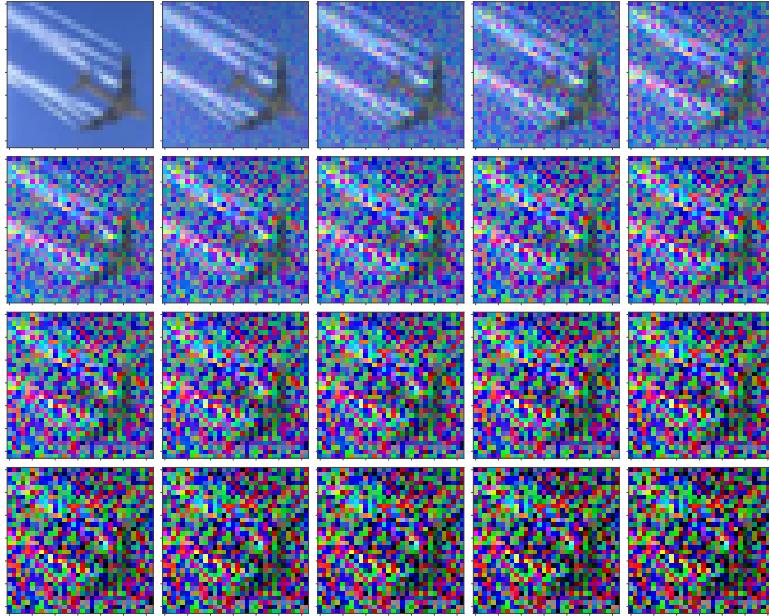


Figure S2: Continuous deformation from a clean image to the $O(3)^d$ -rotated image.

F. Plots Dump

F.1. Scaling Plots for ResNet34 and ResNet101

F.2. Learning dynamics of ImageNet for various Rotations

F.3. Gaussian vs Uniform Initialization

We compare initializing the first layer of the networks using iid Gaussian vs iid Uniform distribution. We observe that the difference is very small; see Fig. S5

F.4. Scaling Law for Infinite Networks

F.5. ImageNet Samples

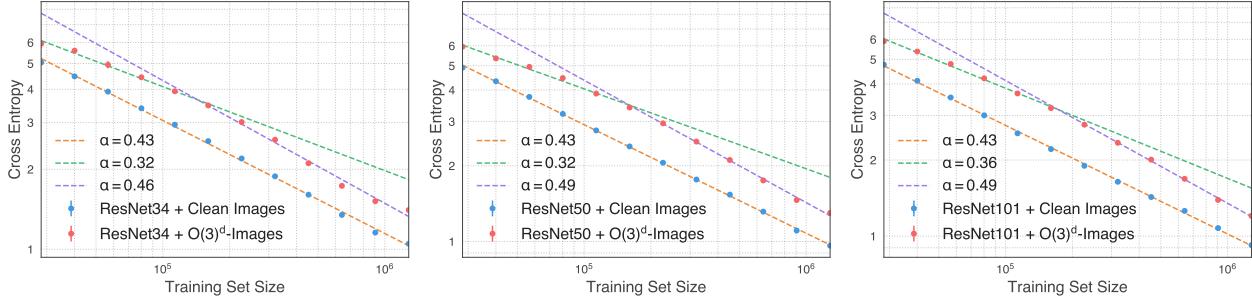


Figure S3: Scaling vs Rotation

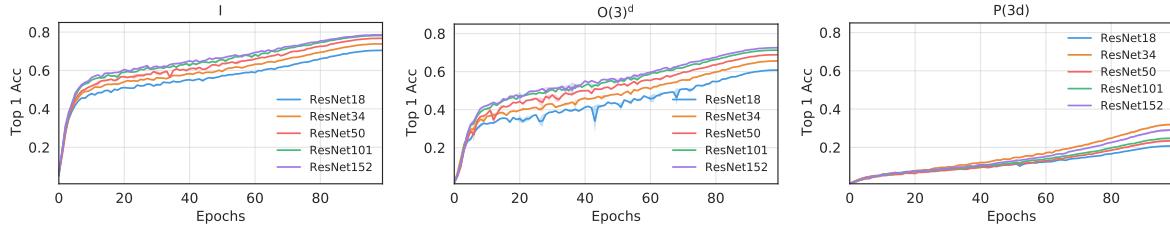


Figure S4: ResNet can overcome the spurious symmetries $O(3)^d$ (middle) but not $P(3d)$ symmetries (right) Learning dynamics (test accuracy vs epochs) of rotated ImageNets. Averaged over 3 runs. Left: no rotation. Middle: $O(3)^d$ rotation. Right: $P(3d)$ rotation. When the dataset is $O(3)^d$ -rotated, the models are still able to obtain decent performance, which monotonically improves as the model becomes larger. However, when the dataset is $P(3d)$ -rotated, test accuracy is below 35%.

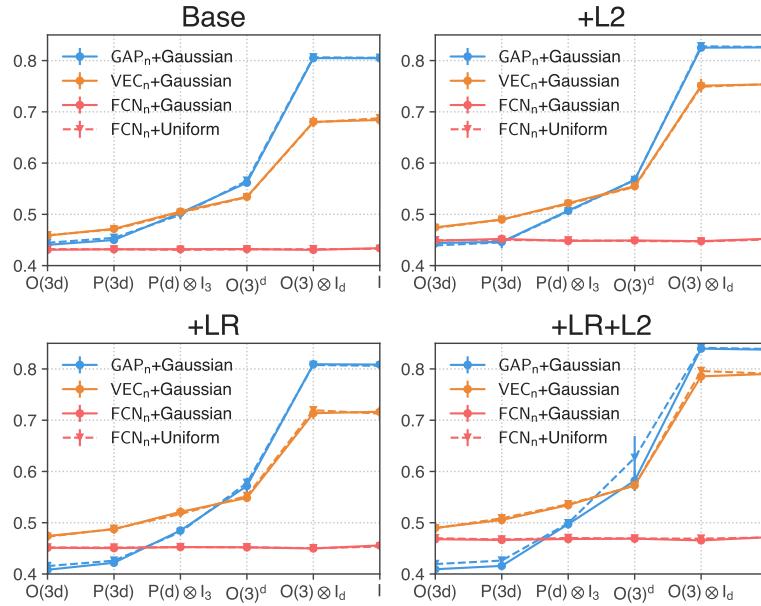


Figure S5: Replacing the Gaussian initialization by uniform distribution does not change the performance much.

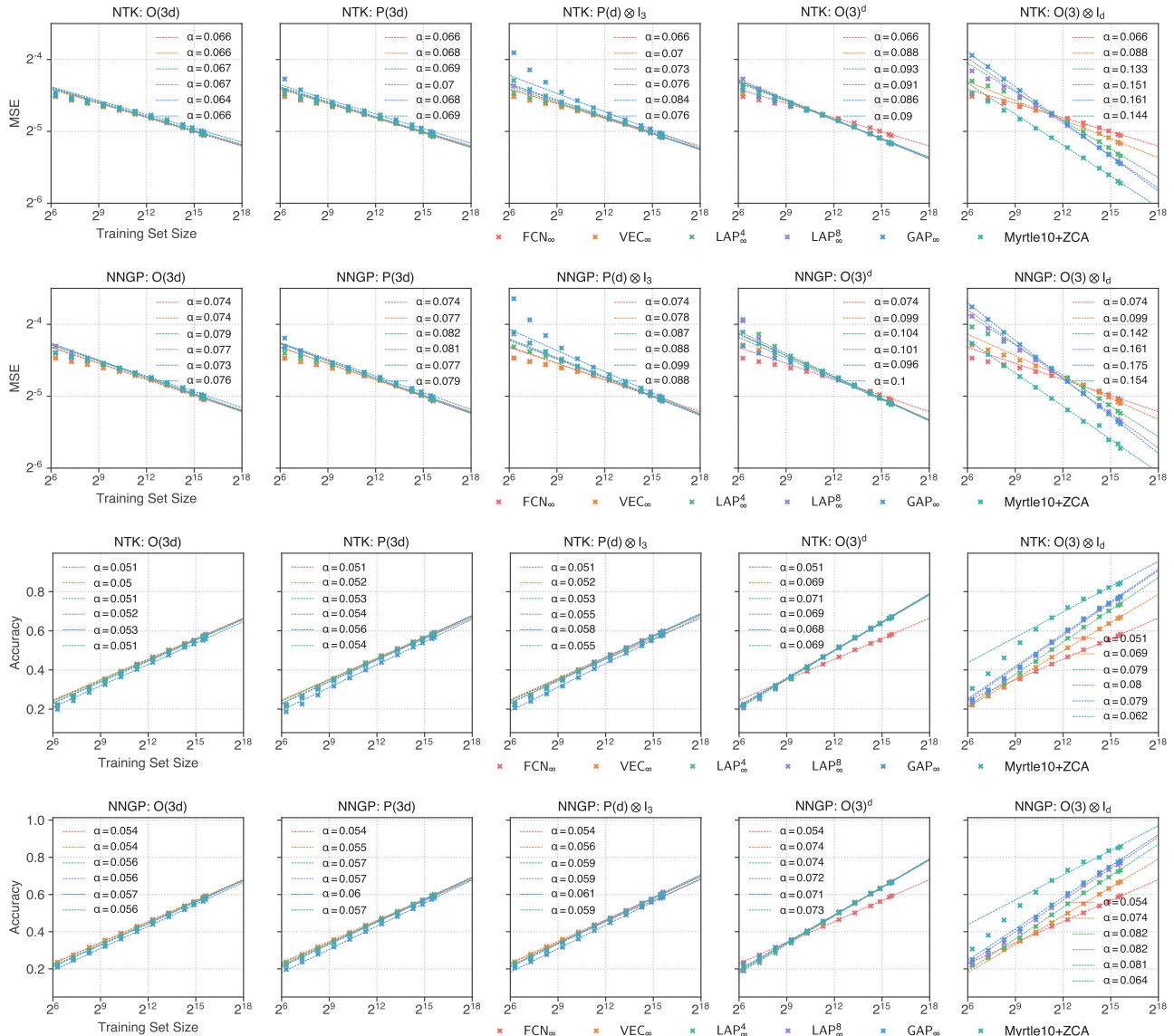


Figure S6: Scaling Law of Infinite Network vs Different Symmetries vs Architectures. We see clean power-law lines for most of the learning curves in the MSE plots.



Figure S7: $O(3)^d$ —Rotated ImageNet Samples. Seed=1

Figure S8: $O(3)^d$ —Rotated ImageNet Samples. Seed=2

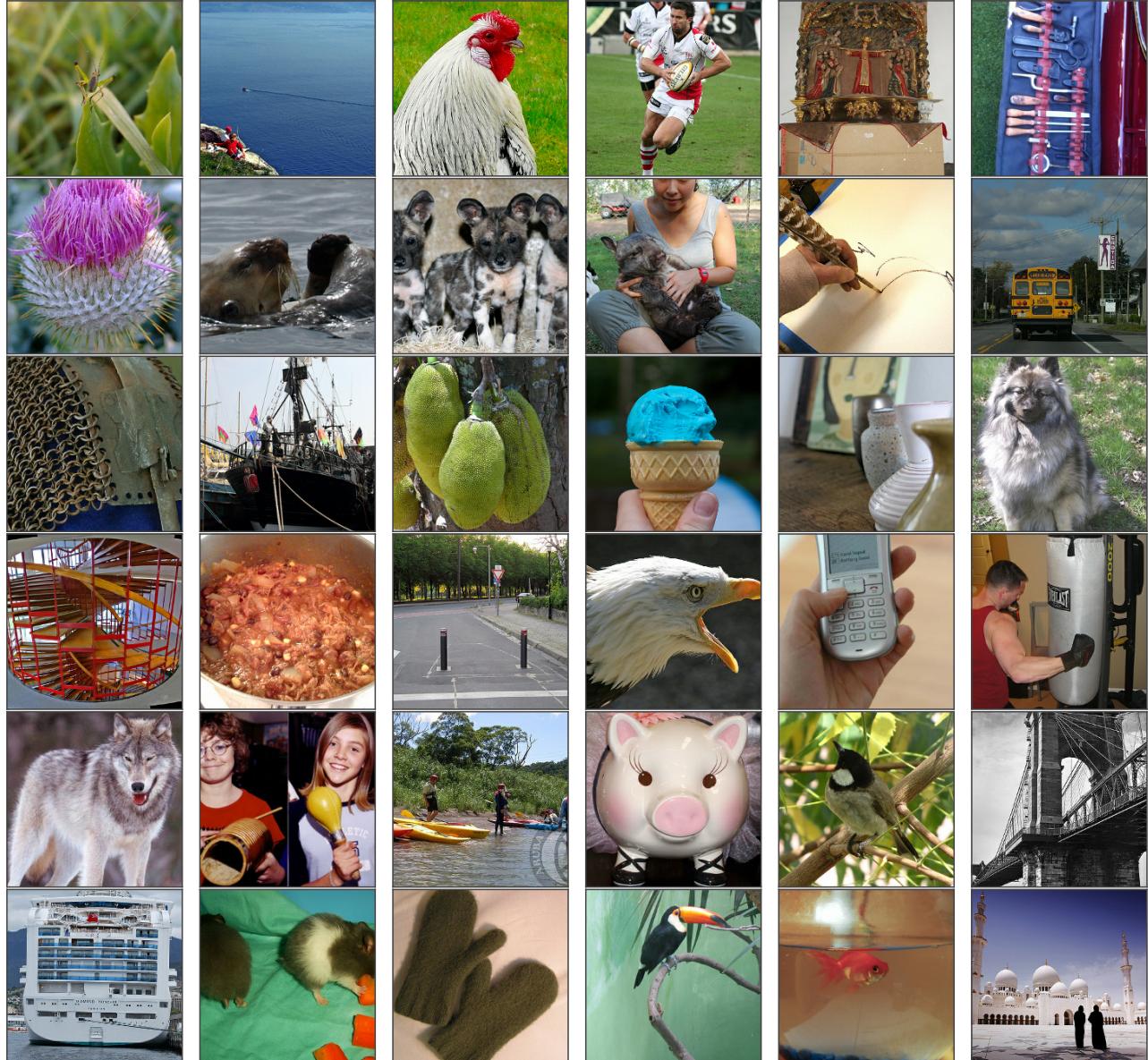


Figure S9: Clean ImageNet Samples