On genuine invariance learning without weight-tying

Artem Moskalev [1] **Anna Sepliarskaia** [2] **Erik J. Bekkers** [3] **Arnold Smeulders** [3]

Abstract

In this paper, we investigate properties and limitations of invariance learned by neural networks from the data compared to the genuine invariance achieved through invariant weight-tying. To do so, we adopt a group theoretical perspective and analyze invariance learning in neural networks without weight-tying constraints. We demonstrate that even when a network learns to correctly classify samples on a group orbit, the underlying decision-making in such a model does not attain genuine invariance. Instead, learned invariance is strongly conditioned on the input data, rendering it unreliable if the input distribution shifts. We next demonstrate how to guide invariance learning toward genuine invariance by regularizing the invariance of a model at the training. To this end, we propose several metrics to quantify learned invariance: *(i)* predictive distribution invariance, *(ii)* logit invariance, and *(iii)* saliency invariance similarity. We show that the invariance learned with the invariance error regularization closely reassembles the genuine invariance of weight-tying models and reliably holds even under a severe input distribution shift. Closer analysis of the learned invariance also reveals the spectral decay phenomenon, when a network chooses to achieve the invariance to a specific transformation group by reducing the sensitivity to *any* input perturbation.

1. Introduction

The ability to abstract from irrelevant details and focus on core aspects is a foundational property of intelligent systems. Invariance, a crucial step of this abstraction process, enables neural networks to recognize patterns regardless of their transformations. Achieving effective invariance is vital for the robust performance of deep learning models.

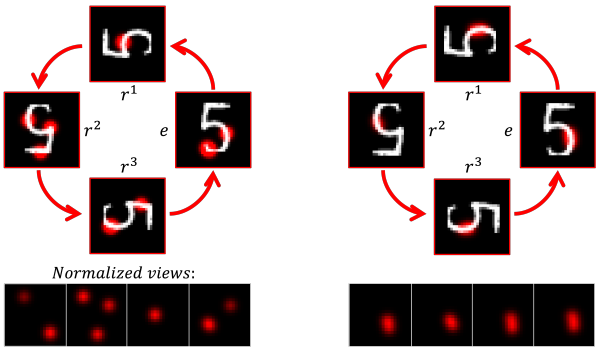

Figure 1. **Left:** The network learns a separate set of features for each of the orientations as indicated by divergent saliency maps. **Right:** saliency maps of the networks with group-invariant weight-tying. *In both cases, predicted class distributions are identical for all of the orientations.* **Bottom row:** saliency maps normalized to a common orientation.

There exist two approaches for invariance in neural networks: invariant weight-tying and learning invariance from data. Networks with built-in invariant weight-tying (Cohen and Welling, 2016; Worrall et al., 2017; Weiler and Cesa, 2019; Worrall and Welling, 2019; Sosnovik et al., 2020; Bekkers, 2020) offer *genuine* invariance, but require knowledge of geometrical priors and incurs high computational and memory costs (Sosnovik et al., 2021a;b). Alternatively, neural networks can learn invariance directly from data. Recent works (Olah et al., 2020; Benton et al., 2020; Moskalev et al., 2022a) demonstrate that neural networks successfully learn invariant priors without any architectural modifications. However, the nature of learned invariance remains largely unexplored, particularly regarding whether it resembles the genuine invariance of weight-tying methods at any level. Consequently, this raises concerns about how much we can rely on the learned invariance when operating conditions evolve. In this work, we investigate the properties of learned invariance to better understand its potential and limitations.

To investigate properties of the learned invariance, we adopt the group theoretical perspective and analyze invariance learning without weight-tying constraints. Firstly, we an-

[1]UvA-Bosch Delta Lab, University of Amsterdam, The Netherlands [2]University of Twente & Booking.com, The Netherlands [3]University of Amsterdam, The Netherlands. Correspondence to: Artem Moskalev <ammoskalevartem@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

alyze the saliency maps of no weight-tying networks with learned invariance. We demonstrate that even when such networks learn to correctly classify samples on a group orbit, the underlying decision-making process does not attain genuine invariance, see Figure 1. Instead of learning genuinely invariant weight-tying, unconstrained networks choose to learn a separate set of features for each of the transformations from a group orbit, even when invariance is enforced by strong data augmentation. This results in learned invariance being strongly conditioned on the input data. Consequently, the effectiveness of learned invariance degrades rapidly when operating conditions evolve, e.g. under input distribution shift. This renders neural networks with learned invariance less reliable.

Secondly, we tackle the problem of aligning learned invariance with the genuine invariance of weight-tying networks. To do so, we propose several measures to quantify the invariance error; we next use those measures to regularize the task loss to promote genuine invariance learning. We conduct experiments with rotation and translation groups, and we show that the proposed regularization significantly aligns learned invariance with the genuine invariance achieved through the weight-tying. However, the alignment also induces performance decay on a downstream task. This presents a new challenging problem of achieving genuine invariance by learning through data augmentation and specialized losses, while also maintaining the downstream task performance.

Thirdly, we investigate the performance decay under the learned invariance. To this end, we analyze the training dynamics of the invariance error minimization from the perspective of the gradient flow. We show that minimizing the invariance error without weight-tying implicitly promotes attaining the invariance to a certain group of transformations by reducing the sensitivity to *any* input perturbation. This has an effect similar to training a network with a large weight decay, which motivates the performance drop. We conduct experiments and demonstrate that this phenomenon holds for various transformations and various forms of invariance error minimization.

To sum up, we make the following contributions:

- We demonstrate that data-driven invariance learning fails to learn genuine invariance as in weight-tying networks.
- We show that it is possible to attain genuine invariance through invariance regularization, but at the cost of the downstream task performance.
- We attribute the performance decay under learned invariance to the training dynamics of the invariance error minimization, which constrains the sensitivity of a network to input perturbations in general.

2. Related work

Weight-tying invariance Weight-tying is the approach for invariance in neural networks that is based on the concept of group equivariant networks (Cohen and Welling, 2016). Group equivariant networks explicitly embed equivariance, or invariance as a special case, for specific transformation groups into a network architecture. The principle traces back to convolutional networks (LeCun et al., 1999) which incorporate translation symmetry. The scope of equivariant networks has since expanded to include other transformations such as rotations (Cohen and Welling, 2016; Worrall et al., 2017; Weiler and Cesa, 2019; Jenner and Weiler, 2022), permutations (Zaheer et al., 2017), and scaling (Worrall and Welling, 2019; Sosnovik et al., 2020; Bekkers, 2020; Sosnovik et al., 2021c;a;b). Another line of work focuses on advancing group equivariant networks by enabling them to learn symmetries directly from the data (Anselmi et al., 2019; Zhou et al., 2020; Dehmamy et al., 2021; Sanborn et al., 2023). This allows the model to adjust to specific symmetries present in the training dataset, eliminating the need for prior knowledge of geometrical priors. Yet, these methods still require modifying the architecture to train invariance.

In this work, we treat the weight-tying methods as oracle invariance learners and investigate whether networks without specific architectural modifications can learn the degree and quality of invariance comparable to the weight-tying approaches.

Data-driven invariance learning Another approach for achieving invariance is to learn it directly from the data. Recent and earlier works (Goodfellow et al., 2009; Lenc and Vedaldi, 2014; Benton et al., 2020; Moskalev et al., 2022a; Kvinge et al., 2022) demonstrate that neural networks can learn invariance without relying on specialized architectural modifications. Additionally, training with data augmentation has long been seen as a method to increase invariance of a model for input transformations (Perez and Wang, 2017; Shorten and Khoshgoftaar, 2019; Cubuk et al., 2018). Invariance learning that does not require specialized architectural modification is advantageous as it does not incur additional memory or computational costs. However, the nature of the learned invariance and its comparability to the genuine invariance obtained through the weight-tying remains an open question. The properties and reliability of such learned invariance are not well understood, which motivates the study in this paper.

3. Learning invariances from data

We take a group-symmetry perspective on data-driven invariance learning when the downstream task is classification. That is to say, we define a set of transformations to be a sym-

metry group a network needs to learn to be invariant to when classifying input signals. We start by briefly introducing group symmetry and invariance.

3.1. Group symmetry

Group A group $\langle \mathcal{G}, \circ \rangle$ is a set $\mathcal{G}$ with a group binary operation $\circ$ called the group product. For convenience, it is common to simplify the notation $a \circ b$ to ab. The group product combines two elements from $\mathcal{G}$ to a new element so that the following group axioms are satisfied. *Closure:* for all $a, b \in \mathcal{G}$, the element $ab \in \mathcal{G}$. *Associativity:* for all $a, b, c \in \mathcal{G}$, $(ab)c = a(bc)$. *Identity:* there is an element $e \in \mathcal{G}$ such that $ea = ae = a$ for every element $a \in \mathcal{G}$. *Inverse:* for each $a \in \mathcal{G}$ there exist $a^{-1} \in \mathcal{G}$ such that $a^{-1}a = aa^{-1} = e$.

Group actions & Symmetry Group actions are a way of describing symmetries of objects using groups. A group action of a group $\mathcal{G}$ on a set $\mathcal{X}$ maps each element $g \in \mathcal{G}$ and each element $x \in \mathcal{X}$ to an element of $\mathcal{X}$ in a way that is compatible with the group structure. In other words, $ex = x$ and $(g_1 g_2)x = g_1(g_2 x)$ for any $x \in \mathcal{X}$ and $g_1, g_2 \in \mathcal{G}$.

Group-invariance Group-invariance is a property of a function $f : \mathcal{X} \to \mathcal{Y}$ under a group action from a group $\mathcal{G}$. A function f is said to be group-invariant if $f(gx) = f(x)$ for $g \in \mathcal{G}$. This means that the value of f at x is unchanged by the action of any group element.

Group orbit The group orbit of an element $x \in \mathcal{X}$ under a group action from a group $\mathcal{G}$ is the set of all points in $\mathcal{X}$ that can be reached by applying the group action on x. More formally, the orbit of x is defined as the set $\mathcal{O}_x = \{gx | g \in \mathcal{G}\}$. This concept encapsulates the idea that the group action can move the element x around within the set, and the orbit describes all the possible positions x can be moved to by the group action.

3.2. Measuring learned invariance

Next, we explain how to measure group-invariance learned by a neural networks from the data. We assume we are given a neural network $f : \mathcal{X} \to \mathcal{Y}$ that maps inputs to logits, a group $\mathcal{G}$ and the dataset $\mathcal{D}$. We define three types of measures: *(i) predictive distribution invariance* to measure the average change of a network's output distribution when a symmetry transformation is applied, *(ii) logit invariance* to measure the change of raw network's logits and *(iii) saliency invariance similarity* to evaluate the consistency of network's decisions under group transformations.

Predictive distribution invariance Since the downstream task of interest is classification, it is natural to measure the invariance by evaluating the shift of the predictive distribution when transformations from a group orbit are applied. Practically, we can utilize Kullback–Leibler divergence between output softmax-distributions of $f(x)$ and $f(gx)$. With this, we can write the predictive distribution invariance error DI_f:

$$DI_f(\mathcal{D}, \mathcal{G}) = \sum_{x \sim \mathcal{D}} \sum_{g \sim \mathcal{G}} D_{KL}(u_x \parallel q_{gx}) \tag{1}$$

where u_x and q_{gx} denote the *softmax* applied to the logits $f(x)$ and $f(gx)$ respectively.

Since DI_f operates directly on the level of predictive distributions, it is the most useful to evaluate the invariance tackled to the downstream classification task.

Logit invariance Next, we define the logit invariance error to measure the shift of raw logits under group actions. Practically, we utilize average squared L_2 distance between the logits $f(x)$ and $f(gx)$:

$$LI_f(\mathcal{D}, \mathcal{G}) = \sum_{x \sim \mathcal{D}} \sum_{g \sim \mathcal{G}} \frac{1}{2} \big\| f(x) - f(gx) \big\|_2^2 \tag{2}$$

Note that the logit invariance error is a more strict invariance measure compared to $DI_f(\mathcal{D}, \mathcal{G})$. This is due to a scalar addition invariance of the predictive softmax-distribution. That means $LI_f(\mathcal{D}, \mathcal{G}) = 0$ implies $DI_f(\mathcal{D}, \mathcal{G}) = 0$, but not vice versa. With this, the logit invariance error is the most useful to characterize the absolute invariance of a function to group transformations regardless of a particular downstream task.

Saliency invariance similarity Lastly, we propose saliency invariance similarity SI_f to measure the consistency of the decision-making process of a neural network under input transformations. Let $m_f : \mathcal{X} \times \mathcal{Y} \to \mathcal{S}$ be a saliency map function for the network f (Sundararajan et al., 2017; Mundhenk et al., 2019). We then compute the similarity between $m_f(x)$ and $g^{-1} m_f(gx)$, where g^{-1} is needed to ensure a common orientation of the saliency maps. Practically, we adopt the cosine similarity and compute the average saliency similarity as:

$$SI_f(\mathcal{D}, \mathcal{G}) = \sum_{x \sim \mathcal{D}} \sum_{g \sim \mathcal{G}} \frac{m_f(x) \cdot g^{-1} m_f(gx)}{\|m_f(x)\|_2 \|g^{-1} m_f(gx)\|_2} \tag{3}$$

The saliency invariance similarity $SI_f(\mathcal{D}, \mathcal{G})$ reflects how much the direction of the most important features, that a network bases its decisions on, change under transformations from a group orbit. Saliency invariance similarity differs

from the previous two metrics as it considers the structure of the input data and not just the output of the network. This makes SI_f particularly useful to understand how group transformations alter a network's internal decision-making process.

3.3. Invariance regularization

Constrained invariance learning We next consider the task of facilitating learning invariances from the data. The natural way to do so is to optimize the task performance subject to a low invariance error. Practically, with the dataset $\mathcal{D}$ and a group of interest $\mathcal{G}$, invariance learning boils down to the constrained optimization approach $\min_\theta \mathcal{L}_f(\mathcal{D})$ *s.t.* $I_f(\mathcal{D}, \mathcal{G}) = 0$; then, to train a neural network, we can simply optimize the relaxation:

$$\min_\theta \mathcal{L}_f(\mathcal{D}) + \nu I_f(\mathcal{D}, \mathcal{G}) \tag{4}$$

where θ denotes the parameters of f, $\mathcal{L}_f$ is a downstream task loss functions, $I_f(\mathcal{D}, \mathcal{G})$ is an invariance regularizer with respect to the group $\mathcal{G}$ and ν regulates how much invariance we want to achieve at the training. Practically, we observed that using the logit invariance error as a regularizer provides better overall invariance and accuracy than other forms of the invariance error, see Section 4.4.

Adding an invariance-regularizer to the original loss yields a simple approach to facilitate data-driven invariance learning. We experimentally demonstrate that invariance regularization significantly improves the quality of learned invariance, closing the gap with the genuine invariance of weight-tying methods. However, we also observe that the improvement in the quality of invariance comes at the cost of downstream task performance, as we demonstrate in Section 4.4.

Invariance-induced spectral decay In order to analyze the causes of performance decay under the invariance error minimization, we analyze the training dynamics of the learned invariance through the lens of its gradient flow. We show that a neural network opts for achieving the invariance to a particular transform group by reducing the sensitivity to any input variations. We use a maximum singular value $\sigma_{\max}$ of network's weights as a sensitivity measure (Yoshida and Miyato, 2017; Khrulkov and Oseledets, 2018); and we analyze the gradient flow for the logit invariance error with a class of linear neural networks. We firstly show that the logit invariance error minimization implicitly constrains the maximum singular value of network's weights, thereby reducing its input sensitivity. Then, we experimentally demonstrate that this result also holds for more complex neural networks and the various forms of invariance errors.

Consider a linear neural network $h(x) = Wx$. Without loss of generality, we analyze the sensitivity to the action of a single group element g, instead of the full orbit of the group $\mathcal{G}$. Let G be a linear representation of the group acting on x and consider invariance error minimization over t steps.

Proposition 3.1 (*Invariance-induced spectral decay*)**.** *Logit invariance error minimization implies* $\sigma_{max}(W(t)) \leq \sigma_{max}(W(0))$ *when* $t \to \infty$.

Proof. The optimization of the parameter matrix W takes the form of $W^{t+1} = W^t - \alpha \nabla LI^t$, where $\nabla LI^t = W^t(x - Gx)(x - Gx)^T$ is a gradient of the logit invariance error (Equation 2) at the time step t.

Let $\epsilon = x - Gx$ and $\Sigma = \epsilon\epsilon^T$. With the infinitesimally small learning rate α, we can write the gradient flow of W as:

$$\frac{d}{dt}W = -W\Sigma \tag{5}$$

For a fixed Σ we can solve the gradient flow above analytically as:

$$W(t) = W(0)\exp(-\Sigma t) \tag{6}$$

Next, we consider a maximum singular value $\sigma_{\max}(W) = \|W\|_2$, when the model is trained, i.e. $W(t)$ with $t \to \infty$. Applying Cauchy–Schwarz we can write:

$$\|W(t)\|_2 \leq \|W(0)\|_2 \| \exp(-\Sigma t)\|_2 \tag{7}$$

With a spectral decomposition $\Sigma = U\Lambda U^T$, we can write $\| \exp(-\Sigma t)\|_2 = \| \exp(-\Lambda t)\|_2$. Since Σ is a rank-one matrix, it contains all zero eigenvalues except of the one, which equates to $\lambda_{\max}(\Sigma) = \epsilon^T\epsilon$. Thus, eigenvalues of the matrix $\exp(-\Lambda t)$ are all ones except of the eigenvalue, which equates to $\lambda_\epsilon(t) = \exp(-t \cdot \epsilon^T\epsilon)$. Note that $\lambda_\epsilon(t) \leq 1$, hence $\| \exp(-\Lambda t)\|_2 = 1$. Plugging into Equation 7 gives $\|W(t)\|_2 \leq \|W(0)\|_2$ with $t \to \infty$.

□

This reveals the non-increasing spectral norm constraint that invariance error minimization induces. Also, initialization routines for W, e.g. (Glorot and Bengio, 2010; He et al., 2015), yield small $\|W(0)\|_2$ at the beginning of the training, further restricting the sensitivity of a network when optimizing for the low invariance.

$\mathcal{G}$	Model	Acc. (%)	$LI \Downarrow$	$DI \Downarrow$	$SI \Uparrow$
$\mathbb{R}_4^2$	WT	94.6 ± 0.1	0.00 ± 0.0	0.0 ± 0.0	1.00 ± 0.00
	DA	94.0 ± 0.5	98.6 ± 5.4	0.3 ± 0.1	0.17 ± 0.04
	IR	87.9 ± 0.8	0.02 ± 0.0	0.0 ± 0.0	0.95 ± 0.03
$\mathbb{T}_3^2$	WT	96.6 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	1.00 ± 0.0
	DA	96.2 ± 0.2	50.8 ± 6.8	0.1 ± 0.0	0.57 ± 0.08
	IR	93.1 ± 0.1	0.01 ± 0.0	0.0 ± 0.0	0.95 ± 0.07

Table 1. Classification accuracy and invariance for the data augmentation [DA], weight-tying [WT], and the model trained with the logit invariance error as the regularizer [IR] on the Transformed-MNIST dataset. *LI* - logit invariance; *DI* - predictive distribution invariance; *SI* - saliency invariance similarity.

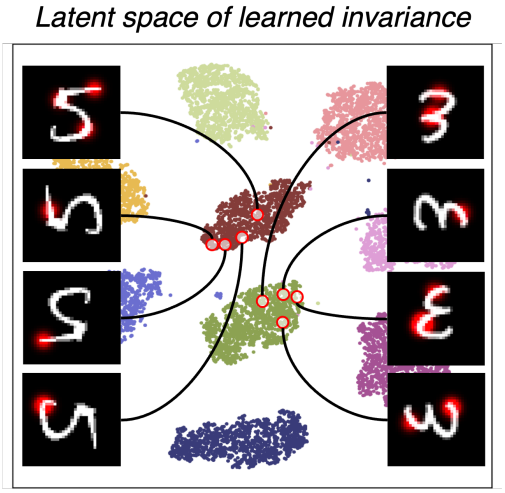

Figure 2. T-SNE of the representations in the pen-ultimate layer of the model with $\mathbb{R}_4^2$ learned invariance. Different orientations of a single sample are mapped to distant points in the latent space. The saliency of the network is highlighted by the red regions in the images.

4. Experiments

In this section, we experimentally investigate the properties of learned group-invariance. As groups of interest we choose the $\mathbb{R}_4^2$ group of 4-fold rotations and the $\mathbb{T}_3^2$ group of 3-fold cyclic translations along the x-axis. We examine how well the learned invariance is aligned with the downstream task performance and the genuine invariance of weight-tying methods. Then, we analyze the reliability of the learned invariance under the data distribution drift. Lastly, we investigate the invariance-induced spectral decay phenomenon for various forms of the invariance error.

4.1. Implementation details

Datasets We construct the Transforming-MNIST and Transforming-FMNIST datasets. Both dataset consist of MNIST and F-MNIST (Xiao et al., 2017) (28×28 black an white images of clothing categories) with $\mathbb{R}_4^2$ or $\mathbb{T}_3^2$ group transformations applied. We additionally leave out the digit 9 from the Transforming-MNIST dataset to avoid the confusion with 6, when studying the rotation invariance. We also leave out the last class of the Transforming-FMNIST to make number of classes equal to the Transforming-MNIST. We extend the resolution from 28×28 to 36×36 by zero-padding data samples. We use $10k/50k/2k$ splits for for *train* / *test* / *validation*. The datasets are normalized to zero mean and unit standard deviation.

Models We employ 5-layer perceptron with ReLU non-linearities and the hidden dimension of 128, resulting in total of $230k$ parameters. For the group-invariant model, we utilize group weight-tying with a pooling over a group to achieve the invariance. We only utilize group-invariant weight-tying for the first layer of a network.

Training details We train all models for 300 epochs using Adam optimizer with the batch size of 512 and the learning rate of 0.0008. For all models, we use data augmentation with transformations randomly sampled from a group of interest. For invariance regularization, we employ logit invariance error as a regularizer $\mathcal{I}_f$. We tune the weighting of the regularizer, such that the resulting saliency invariance similarity $SI_f \geq 0.95$. Final models are selected based on the best validation accuracy.

Saliency map function m_f The choice of the saliency map function to compute saliency invariance score is a hyper-parameter. We employ the following procedure to generate saliency maps of a network. First, we accumulate absolute values of integrated gradients (Sundararajan et al., 2017) with respect to the target class prediction. Second, we threshold the values that are less than 0.9 of a maximum value of the absolute integrated gradient. Finally, we apply a Gaussian filter with a kernel size of 3 and a standard deviation of 1 to smooth the resulting saliency map.

4.2. Learning invariance from the data

In this experiment, we compare models that learn invariance with data augmentation [DA] and invariance regularization [IR], and models that have the invariant weight-tying built-in [WT]. We evaluate the models by the classification accuracy, the logit and distribution invariance errors, and also by the saliency invariance similarity. The results are reported in Table 1. All results are averaged over 4 common random seeds.

WT The networks with group-invariant weight tying deliver the highest classification accuracy in both $\mathbb{R}_4^2$ and $\mathbb{T}_3^2$

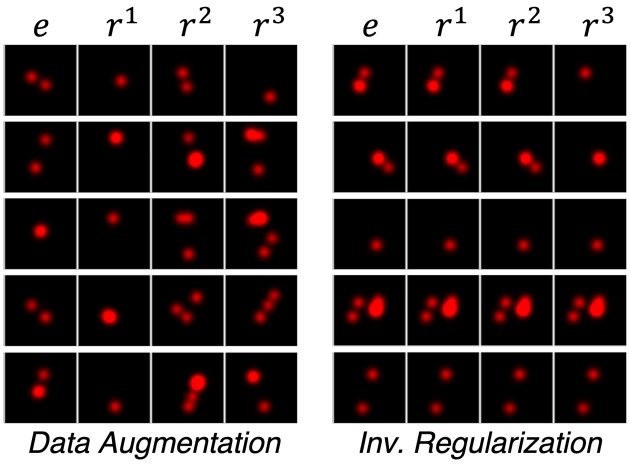

Figure 3. Saliency maps for the models with learned $\mathbb{R}_4^2$ invariance. Rows correspond to data samples and columns correspond to transformations from the group orbit applied to the sample. All saliency maps are realigned to a common orientation.

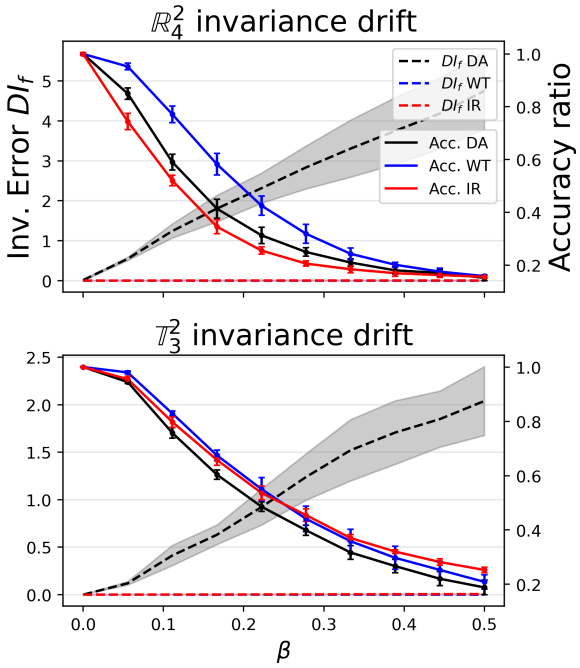

Figure 4. Predictive distribution invariance error DI_f (left y-axis) and classification accuracy drop ratio (right y-axis) for the data augmentation, weight-tying and invariance regularization models over increasing degree of the data distribution drift (x-axis).

scenarios. We thus treat the weight-tying models as a performance upper-bound and an oracle for the genuine invariance when further analyzing models with invariance learned.

DA We observe that the models trained with data augmentation fail to learn genuine group invariance as indicated by high logit invariance error LI_f and lower saliency similarity score SI_f. Interestingly, these models still provide moderately low predictive distribution invariance error DI_f and high classification accuracy under group transformations. This implies that *neural networks can learn to solve an invariant task without learning a genuinely invariant decision making-process*.

To visualize this phenomenon, we depict the T-SNE of the latent space of the model with learned $\mathbb{R}_r^2$ invariance (Figure 2), and we trace different orientations of a sample in the latent space. We observe that different orientations of one sample can land far away from each other in the representation space, but still within the boundaries of its class. When such configuration fully satisfies the downstream task objective, there is apparently no reason for a network to learn genuine invariance.

IR The models trained with invariance regularization achieve low logit and distribution invariance errors on par with the weight-tying models. Also, high saliency invariance similarity indicates that invariance regularization guides a model towards learning genuinely invariance decision-making process. In Figure 3, we visualize examples of saliency maps of the network with $\mathbb{R}_4^2$ invariance learned by data augmentation and invariance regularization. In contrast to the saliency maps of the model trained solely with data augmentation, saliency maps of the model with invariance regularization are well-aligned over the group orbit.

4.3. Reliability of learned invariance

We next investigate the reliability of the models with the learned invariance when operating conditions evolve. We simulate changing operating conditions as the data distribution drift from Transforming-MNIST to Transforming-FMNIST datasets. Practically, we linearly interpolate between those two dataset as $\mathcal{D}_{1\to2}(\beta) = (1-\beta)\mathcal{D}_1 + \beta\mathcal{D}_2$ to obtain the dataset with the drift degree of β. We compare the models by measuring the invariance error and the accuracy drop ratio on the drifted dataset. The accuracy drop ratio on the drifted dataset $\mathcal{D}_{1\to2}(\beta)$ is computed as $\text{Acc}(\mathcal{D}_{1\to2}(\beta))/\text{Acc}(\mathcal{D}_1)$, where $\text{Acc}(\mathcal{D})$ is the model's accuracy on the dataset $\mathcal{D}$. The accuracy ratio indicates how much of the original accuracy is preserved when a model is tested on the drifted dataset. The results are presented in Figure 4.

We observe that the invariance learned by data augmentation deteriorates rapidly, even under a slight degree of data drift. This turns into a major flaw if a user anticipates a certain level of invariance from the model, but then invariance instantly fails upon encountering unseen data. This

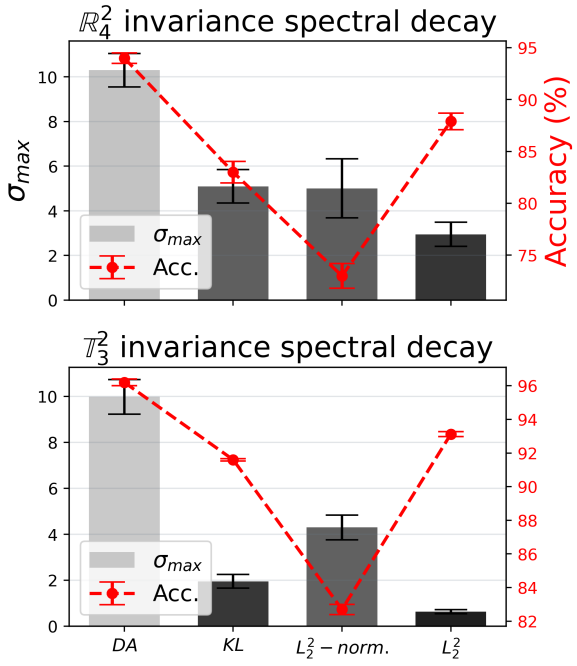

Figure 5. Sensitivity of a network to input perturbations measured by the maximum singular value of its Jacobian (left y-axis) for various forms of invariance regularization (x-axis). Models trained with invariance regularization come with overall reduced sensitivity to input perturbations.

also obscures the interpretability of predictions, thereby complicating the explainability of model decisions, even if accuracy is sustained. This yields invariance learned by data augmentation unreliable. *Conversely, models with weight-tying and invariance regularization maintain low invariance error even under substantial distribution drift.*

Also, we observe that networks with invariant weight-tying sustain higher classification accuracy under the distribution shifts. This also holds for the models trained with invariance regularization for $\mathbb{T}^2_3$, but interestingly, not for the $\mathbb{R}^2_4$ invariance. We hypothesize this can be attributed to the *accuracy-on-the-line* effect (Miller et al., 2021), where models with higher in-domain accuracy tend to also deliver higher accuracy on out-of-distribution data.

4.4. Invariance-induced spectral decay

Lastly, we take a closer look at the invariance-induced spectral decay and we verify if it holds for different forms of invariance regularization. We investigate invariance regularization with *(i)* distribution invariance error with KL divergence, *(ii)* logit invariance error with squared L_2 distance, and *(iii)* logit invariance error with the squared L_2 distance normalized by the magnitude of the logits, i.e. $\|f(x) - f(gx)\|_2^2 / \|f(x)\|_2^2$. We tune the weighting of the regularizer for all of the models such that saliency invariance similarity $SI_f \geq 0.95$. We then evaluate the sensitivity of a network to input perturbations as a maximum singular value of its Jacobian $\sigma_{\max}(J)$; and we compare it to the sensitivity of the networks trained solely with data augmentation. The results are presented in Figure 5.

We observe that models trained with invariance regularization come with overall reduced sensitivity to input perturbations as indicated by considerably smaller $\sigma_{\max}(J)$. This phenomenon holds for both $\mathbb{R}^2_4$ and $\mathbb{T}^2_3$ groups and various forms of invariance regularization. Note that all forms of the invariance regularization we examine also induce an accuracy drop for the model.

5. Discussion

Summary Our study sheds light on the properties and limitations of data-driven invariance learning within neural networks. First, we proposed several measures to evaluate learned invariance: predictive distribution invariance and logit invariance errors, and saliency invariance similarity. With this, we study networks with learned group invariance and demonstrate that high performance and low invariance error do not guarantee a genuine invariant decision-making process. This leads to a notable risk, when learned invariance immediately fails beyond the training data distribution, making neural networks with learned invariance less reliable. Then, we showed that it is possible to promote genuine invariance learning by regularizing invariance during the training. Yet, such an approach leads to a spectral-decay phenomenon, when a network opts for reducing input sensitivity to all perturbations to achieve invariance to a specific group of transformations. These findings bring us a step closer to deciphering the intricate dynamics of learning inductive biases from the data.

Broader Impact Our work, while primarily considering invariance to group symmetries, has potential implications for a much broader class of invariances. The increasing reliance on data-driven models, particularly in the era of large-scale machine learning, highlights the critical need to comprehend the properties of inductive biases that these models learn. Thus, understanding learned invariance, as one of the key inductive biases, becomes paramount for ensuring the fairness and interpretability of network's decisions.

Limitations and Future Work While our study provides several key insights, some limitations remain. Firstly, the generalizability of our findings to other types of neural networks, other data modalities and other training regimes, e.g. self-supervised learning (Chen et al., 2020; He et al., 2021; Moskalev et al., 2022b), remains an interesting future direction to explore. Secondly, the way we design invariance

regularization assumes a known group of transformations, which may not always be accessible in practice. Future work could look into methods for learning genuine invariance to unknown transformations without architectural modification. Lastly, our results also highlight a trade-off between learning the genuine invariance and the downstream task performance, opening a direction for future research into strategies for mitigating this trade-off.

References

Anselmi, F., Evangelopoulos, G., Rosasco, L., and Poggio, T. (2019). Symmetry-adapted representation learning. *Pattern Recognition*, 86:201–208.

Bekkers, E. J. (2020). B-spline {cnn}s on lie groups. In *International Conference on Learning Representations*.

Benton, G., Finzi, M., Izmailov, P., and Wilson, A. G. (2020). Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33:17605–17616.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Cohen, T. and Welling, M. (2016). Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA. PMLR.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2018). Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Dehmamy, N., Walters, R., Liu, Y., Wang, D., and Yu, R. (2021). Automatic symmetry discovery with lie algebra convolutional network. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Goodfellow, I., Lee, H., Le, Q., Saxe, A., and Ng, A. (2009). Measuring invariances in deep networks. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2021). Masked autoencoders are scalable vision learners. *arXiv:2111.06377*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Jenner, E. and Weiler, M. (2022). Steerable partial differential operators for equivariant neural networks. In *International Conference on Learning Representations*.

Khrulkov, V. and Oseledets, I. (2018). Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kvinge, H., Emerson, T., Jorgenson, G., Vasquez, S., Doster, T., and Lew, J. (2022). In what ways are deep neural networks invariant and how should we measure this? In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, pages 319–345. Springer Verlag. International Workshop on Shape, Contour and Grouping in Computer Vision ; Conference date: 26-05-1998 Through 29-05-1998.

Lenc, K. and Vedaldi, A. (2014). Understanding image representations by measuring their equivariance and equivalence.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. (2021). Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR.

Moskalev, A., Sepliarskaia, A., Sosnovik, I., and Smeulders, A. (2022a). Liegg: Studying learned lie group generators. In *Advances in Neural Information Processing Systems*.

Moskalev, A., Sosnovik, I., Volker, F., and Smeulders, A. (2022b). Contrasting quadratic assignments for set-based representation learning. In *European Conference on Computer Vision*.

Mundhenk, T. N., Chen, B. Y., and Friedland, G. (2019). Efficient saliency maps for explainable ai. *arXiv preprint arXiv:1911.11293*.

Olah, C., Cammarata, N., Voss, C., Schubert, L., and Goh, G. (2020). Naturally occurring equivariance in neural networks. *Distill*. https://distill.pub/2020/circuits/equivariance.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Sanborn, S., Shewmake, C. A., Olshausen, B., and Hillar, C. J. (2023). Bispectral neural networks. In *The Eleventh International Conference on Learning Representations*.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Sosnovik, I., Moskalev, A., and Smeulders, A. (2021a). Disco: accurate discrete scale convolutions. *arXiv preprint arXiv:2106.02733*.

Sosnovik, I., Moskalev, A., and Smeulders, A. (2021b). How to transform kernels for scale-convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1092–1097.

Sosnovik, I., Moskalev, A., and Smeulders, A. W. (2021c). Scale equivariance improves siamese tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2765–2774.

Sosnovik, I., Szmaja, M., and Smeulders, A. (2020). Scale-equivariant steerable networks. In *International Conference on Learning Representations*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Weiler, M. and Cesa, G. (2019). General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Worrall, D. and Welling, M. (2019). Deep scale-spaces: Equivariance over scale. *Advances in Neural Information Processing Systems*, 32.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., and Brostow, G. J. (2017). Harmonic networks: Deep translation and rotation equivariance. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7168–7177.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

Yoshida, Y. and Miyato, T. (2017). Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., and Smola, A. (2017). Deep sets.

Zhou, A., Knowles, T., and Finn, C. (2020). Meta-learning symmetries by reparameterization.