Generalization to translation shifts: a study in architectures and augmentations

Anonymous Author(s)
Affiliation
Address
email

Abstract

We provide a detailed evaluation of various image classification architectures (convolutional, vision transformer, and fully connected MLP networks) and data augmentation techniques towards generalization to large translation shifts. We make the following observations: (a) In the absence of data augmentation, all architectures, including convolutional networks suffer degradation in performance when evaluated on translated test distributions. Understandably, both the in-distribution accuracy and degradation to shifts is significantly worse for non-convolutional architectures. (b) Across all architectures, even a minimal augmentation of 4 pixel random crop improves the robustness of performance to much larger magnitude shifts of up to $1/4$ of image size (8-16 pixels) in the test data – suggesting a form of meta generalization from augmentation. For non-convolutional architectures, while the absolute accuracy is still low, we see dramatic improvements in robustness to large translation shifts. (c) With sufficiently advanced augmentation (4 pixel crop+RandAugmentation+Erasing+MixUp) pipeline all architectures can be trained to have competitive performance, in terms of in-distribution accuracy as well as generalization to large translation shifts.

1 Introduction

Convolutional neural networks (ConvNets) are a natural architectural choice for a variety of computer vision tasks. The built-in structure from localization and translation equivariance of the convolutional layers is intrinsically useful in many image processing scenarios [Krizhevsky et al., 2012, LeCun et al., 1989, Fukushima and Miyake, 1982]. For over a decade ConvNets were the backbone of computer vision and continue to be one of the most important class of models. At the same time, advances in large datasets and data augmentation techniques have made it possible to train general purpose architectures to be competitive with ConvNets despite lacking any image specific priors. Most popular among these are the Vision Transformers (ViTs) and their variants [Vaswani et al., 2017, Dosovitskiy et al., 2020, Touvron et al., 2021c,b]. When pretrained on ultra-large datasets like ImageNet-21k (14 million images) or JFT-300/3B (300 million/3 billion weakly labeled images, respectively), ViTs can outperform similarly pretrained ConvNets on diverse vision tasks. While the scale of the data was original thought to be crucial, follow up work show that competitive accuracies can also be achieved in small-to-medium data regimes using advanced data augmentation [Touvron et al., 2021b] or optimization techniques [Chen et al., 2021]. Detailed experimentation on competitive benchmarks by Steiner et al. [2021] showed that ViTs trained with extensive data augmentation can recover the performance gains from ~ 10x larger independently annotated dataset. In such large data or extensive augmentation regimes, even simpler fully connected multi-layer perceptron (MLP) architectures [Tolstikhin et al., 2021, Touvron et al., 2021a] can achieve competitive performance. In this work we focus on the role of data augmentation in learning from general purpose architectures.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Can data augmentation capture the inductive biases of carefully designed architectures? Beyond accuracy, the design of architectures is often motivated by domain knowledge of desired invariances. One of the fundamental image priors is the invariance of object labels to spatial shift or translation of its position. Indeed, the success and motivation of ConvNets is often attributed to their component convolutional operators being definitionally translation equivariant. In this work, we study to what extent the shift-equivariant convolutional layers make ConvNets robust to translation shifts compared to ViTs and MLPs? How effective are data augmentation techniques in encouraging similar behaviors?

There is a rich literature on understanding the translation invariance properties of ConvNets (*cf.* Section 1.1). It has been shown that despite the equivariance of the convolutional operator, other architectural components like non-linearities and strides can cause the networks lose their invariance to translations [Zhang, 2019, Azulay and Weiss, 2019, Chaman and Dokmanic, 2021, Engstrom et al., 2019, Xiao et al., 2018, Alsallakh et al., 2020]. Despite these findings, it is reasonable to believe that ConvNets will still have substantial edge in robustness to translation when compared to ViT or MLP models. At the same time, the competitive performance of latter architectures suggests that good image priors can also be learned from rich training data. The goal of this work is to quantify the relative effectiveness of architectures design and data augmentation towards robustness to translation shifts. Algorithmically, these are complementary tools for incorporating domain knowledge. On one hand, it is conceptually simpler to generate data with desired invariant transformation rather than hard code it in the architecture. On the other hand, data augmentation only provides a weak supervision about the invariant properties and is limited by the biases in the training samples.

Generalization vs invariance. In our experiments, rather than chase the gold standard of *strict* invariance on all inputs, we work with a data-centric measure of *generalization to structured translation shifts* in the test distribution. We evaluate our models for accuracy on specific out-of-distribution test datasets where the object locations are systematically shifted without creating additional distortions or domain gaps from training distribution. See Section 2 for full setup.

In the strict sense, a classifier $f : X \to Y$ on input space X is invariant to set of spatial translations T if for all $t \in T$ and all inputs $x \in X$, $f(t(x)) \approx f(x)$. Often we are not concerned with performance on all inputs but rather on typical samples from a task distribution, say $(x, y) \sim D$. We thus evaluate our models a structured out-of-distribution accuracy, wherein we train on samples from D, but test on translation shifted inputs $t(x)$, *i.e.,* generalized accuracy to shift t is $\mathbb{E}_{(x,y)\sim D}\mathbf{1}[f(t(x)) = y]$. This evaluation also differs from adversarial robustness to translation shifts as studied in Engstrom et al. [2019], Xiao et al. [2018]. In our notation, the adversarial accuracy metric would be $\mathbb{E}_{(x,y)\sim D}\mathbf{1}[f(t_x(x)) \neq y]$, where $t_x \in T$ is adversarially chosen per-sample (x, y). Finally, we clarify that data augmentation does change the training distribution D. If inputs x were to be augmented with their respective transformations $\{t(x) : t \in T\}$ (and no other augmentation), then there is no distribution shift. However, our experiments are never in the no-distribution shift regime. Importantly, all our augmentation pipelines uses random crop of at most 4 pixels (some are further restricted to 1 or 2 pixels), but we evaluate our models on much larger translation shifts of up to $1/4$th of image dimension (8 pixels on CIFAR and 16 on TinyImageNet). Thus, any generalization to larger translation shifts from limited augmentation can be thought of as a form of meta-generalization.

1.1 Related work

Translation invariance in ConvNets. ConvNets have been extensively studied on various approximate measures of invariance to translation shifts including, but not limited to, Goodfellow et al. [2009], Zeiler and Fergus [2014], Fawzi and Frossard [2015], Kanbak et al. [2018], Azulay and Weiss [2019], Zhang [2019], Engstrom et al. [2019], Kayhan and Gemert [2020], Chaman and Dokmanic [2021]. For example, Zeiler and Fergus [2014] provide visualizations of the hidden layer filters in early ConvNets that show their sensitivity to small changes in translation, scale, and rotation. Similar visual evaluation was also more recently used in Alsallakh et al. [2020]. Other work Fawzi and Frossard [2015], Kanbak et al. [2018], Azulay and Weiss [2019], Zhang [2019], Engstrom et al. [2019] focus on more quantitative measures of invariance such as mean change in top-1 prediction or class probabilities between images within a distortion range. These works collectively establish that despite the built-in inductive bias, ConvNets are not “truly" invariant even to small translation. Some recent work, importantly [Zhang, 2019, Chaman and Dokmanic, 2021], address this shortcoming of ConvNets by designing new architectural modifications to ConvNets. These prior works show that strict spatial invariance is a strong measure that even ConvNets with their built-in priors do

not satisfy. We choose a data-centric measure in our evaluation that is more practical for general architectures. Our experimental and evaluation protocol has some key differences from these prior works studying invariance: (a) Our evaluation is designed to introduce only translations shifts in the test data without confounding with any other domain gap between train and test distributions. For example, in the experiments of Azulay and Weiss [2019], Zhang [2019] the test samples are padded and resized during evaluation, while the models were trained on regular dataset without such padding or resizing. This leads the test dataset to differ from the training data in ways that are not just translation of the objects. (b) Our evaluation is non-adversarial in that it measures degradation in average test loss and not in the worst-case drop in performance on any single image. Further, by comparing the drop in performance from translation shift to the unperturbed in-distribution accuracy, we inherently down-weight the non-robustness on hard-to-learn inputs on which the classifier had inaccurate prediction to begin with. (c) Finally, our evaluation does not penalize models from learning position dependent features as observed by Kayhan and Gemert [2020]. For example, a network has the flexibility to use its large representation power to create a separate model for an object (say a cat) at each pixel location — while such a model will be inefficient, it would still do a good job at detecting cats in translated test distribution. These differences are nuanced but significant, which makes our results complementary to prior work in this space. Such differences may or may not be important depending on the application.

Robustness under adversarial perturbations and other OOD benchmarks. In a work closest to ours, Engstrom et al. [2019] study test accuracy degradation from adversarially chosen translations and rotations on test images. While much of the work in adversarial robustness focus on ℓ_∞ or ℓ_2 norm bounded perturbations to the inputs, Engstrom et al. [2019] show that ConvNets can be effectively "attacked" even when the perturbed inputs are merely small rotations and/or translations of the input. Similar adversarial attack on ConvNets based on more complex spatial transformations was previously studied in [Xiao et al., 2018]. Our evaluation is closer to "random perturbation" evaluation, which is only briefly explored in Engstrom et al. [2019]. In comparison to these studies, our distribution shift is picked non-adversarially and independent of input samples.

Complementing the work on adversarial robustness, there has also been lot of interest in evaluating models on other out-of-distribution robustness. Generalization to out-of-distribution test datasets is a broad umbrella topic, and aside from adversarial robustness, many work also focus on robustness performance on benchmarks for benign "natural" perturbations Recht et al. [2019], Hendrycks and Dietterich [2018], Koh et al. [2021], Djolonga et al. [2021]. For ConvNets, inspired by the neural networks *scaling laws* line of work, Djolonga et al. [2021] probe for relation between OOD robustness and learning choices like data size, model size, optimization algorithm, and model choices like model sizes, and normalization (they do not consider data augmentation in detail though). They also propose a synthetic benchmark SI-SCORE for controlled image invariance evaluation. Expanding on this line of work, Yung et al. [2021], Bhojanapalli et al. [2021], Bai et al. [2021], Mahmood et al. [2021], Shao et al. [2021], Paul and Chen [2021], Pinto et al. [2022], compare ConvNets and ViTs on adversarial robustness and/or out-of-distribution robustness benchmarks.

Aside from the methodological differences, much of the work in this space has focused on the relative merits, demerits, and robustness of different model choices and the role of training data size. The effects of data augmentation is only minimally considered, if at all. In contrast, our goal in this work is to specifically quantify how much different data augmentation pipelines can capture the inductive biases in a carefully designed architecture.

2 Experimental setup

All our experiments are conducted on the CIFAR-10, CIFAR-100, and TinyImageNet datasets. Since our study involves training from scratch and testing on large models in numerous configurations, it is beyond the scope of the paper to extend such a detailed study to larger benchmarks like full ImageNet. Moreover, we emphasize that our goal here is not to get the state-of-the-art accuracy/robustness on benchmarks, but rather to understand how much data augmentation captures the benefits of the convolutional architecture. Arguably, it is also the small data regime where the inductive biases from architecture and/or augmentations play more important roles. In ultra-large scale datasets, accuracy/robustness might naturally come from dataset size itself rather than model priors.

In the main paper we focus on results from CIFAR-10 and CIFAR-100 datasets which consists of 32×32 pixel RGB images balanced across 10 and 100 classes, respectively. We defer the discussion on TinyImageNet, which a subset of the more diverse ImageNet benchmark, to Appendix C. To study large translation shifts without introducing domain gaps, we modify the dataset by symmetrically padding all the CIFAR images with 8 pixels ($1/4$ of image size) on each side. The padded pixels contain the mean channel values of the entire training dataset, which ensures that the channel-wise means and standard deviations across training dataset remains the same as the original un-padded dataset (see illustration in Figure 1). This padded dataset allows us to evaluate large translation shifts of up to 16 pixels (Hamming distance) in the test dataset without creating additional confounding factors. *Importantly, in all the shifted test sets, there is no cropping or loss of the image content and the entire image is available to the network at the same scale as seen during training.* After padding with a mean-valued canvas, we resize the resulting $48 \times 48 \times 3$ images to $224 \times 224 \times 3$ (the standard input size for ImageNet) using bilinear interpolation. This up-sampling step helps avoid extensive hyperparameter tuning of the models, especially, the ViT and MLP models.

We briefly discuss the alternative evaluation methodologies. The more natural random cropping of images to evaluate robustness to translation is inherently limited by the number of pixels we can faithfully forgo without losing information and hence cannot capture large translation shifts. In prior work, Azulay and Weiss [2019] also used similar padded images to investigate translation shifts. A key difference in our methodology is that we have our entire training and testing pipeline on the preprocessed images (with padding), while the latter paper evaluated models pretrained on standard ImageNet without any padding – this creates an uncontrolled distribution shifts. Another technical difference is that Azulay and Weiss [2019] downsampled the images, which leads to loss in resolution, while our preprocessing is non-lossy. Finally, synthetic benchmarks such as SI-Score proposed in Djolonga et al. [2021] are a good alternative to our setup. However, for our simple controlled setting, the conceptual advantage of padding is that it does not change the natural distribution of foreground and background which would be lost in the cut-and-paste protocol of Djolonga et al. [2021]. Furthermore, the segmentation process in Djolonga et al. [2021] appears to be not perfect which might create additional confounders. To further ensure that our "synthetic" padding is benign, we verified that the test accuracy of our models trained on the padded dataset is comparable to the standard train-test pipeline on $32 \times 32 \times 3$ CIFAR inputs without any padding (see Table 1).

(a) Original (b) Mean padded (c) Resize to 224×224 (d) Translation shift

Figure 1: Sample images from the prepossessing steps

Training All the models are trained on the mean-padded datasets for 1600 epochs on $8 \times V100$ GPUs. We use implementations (with suitable modifications) of the models from various open source repositories, most notably Wightman [2019] and Liu [2017]. We performed basic hyperparameter tuning in a small grid around the parameters reported in the respective papers. The exact value of hyperparameters used in experiments along with code are provided in the supplementary material. All the evaluation metrics reported in this paper are median performance over 3 runs.

2.1 Architectures

Our goal is to compare fundamentally different architectures on generalization to translation shifts. After initial experiments with different variants, we choose the following models in our evaluation.

- **resnet18_bn (11M parameters)**: We use ResNets He et al. [2016] as our representative ConvNet. In our initial experiments, larger ResNets and the other ConvNets like RegNet Radosavovic et al. [2020] did not yield qualitative difference in performance on our datasets.
- **resnet18_gn (11M parameters)**: It has been observed that batch normalization often leads to poor performance in transfer learning as the batch statistics from source task could be widely off for target task (see *e.g.,* Kolesnikov et al. [2020], Wu and Johnson [2021]). The same reasoning also applies when dealing with distribution shifts, where batch statistics could become irrelevant

even when the test distribution shifts systematically Djolonga et al. [2021]. To overcome this, we consider a variant of *resnet18* with group normalization and weight standardization Wu and He [2018], Qiao et al. [2019] in place of batch normalization. This modification indeed leads to more stable performance in our experiments, specially in the absence of data augmentation.

- **antialiased_resnet18 (11M parameters)**: Azulay and Weiss [2019], Zhang [2019] show that the downsampling layers (pooling/strides) make ConvNets non-shift invariant. To remedy this, Zhang [2019] proposed a a specialized modification to ConvNets to improve their invariance to spacial shifts by introducing a *BlurPool* layer as an antialiasing-filter before downsampling. This constitute a model with more specialized priors about translation invariance built into its design.
- **cait_xxs36 (17M parameters)**: CaiT architecture Touvron et al. [2021c] is a variant of the basic vision transformer (ViT) Dosovitskiy et al. [2020] that leads to more efficient training of deeper models. We use CaiT as our representative transformer model as it had the best performance in initial experiments. Other ViT variants, including larger models and the distilled variant DeiT Touvron et al. [2021b] did not provide significant performance boost on our small scale datasets.
- **resmlp_12 (18M parameter)**: Among the MLP models for image classification, we tried MLP-mixer Tolstikhin et al. [2021] and ResMLP Touvron et al. [2021a] in our initial experiments. We stick with ResMLP for detailed experimentation as it had slightly better performance.

Although the above model configurations are not the state-of-the-art on larger benchmarks, on smaller scale CIFAR and TinyImageNet datasets, they have competitive performance as their larger or more complex counterparts. Since our goal is to evaluate the relative degradation in performance with translation shifts, we do not overly optimize for top-accuracy.

2.2 Augmentations

We first consider four data augmentation pipelines while training the models described above. In the appendix, we look at more minimal augmentations to elaborate on our findings.

- **No Augmentation *(NoAug)***: We use this setting as a baseline for purely evaluating the merits of an architecture in generalization to translation shifts.
- **Basic augmentation *(BA)***: The basic augmentation consists of a random flip and a random crop with up to 4×4 pixel padding. This minimal augmentation has been a de-facto standard in many vision tasks, and it already gives over 5% boost in accuracy even without considering its effects on distribution shifts. Note that unlike in standard training pipeline, with our padding of training images, the random crop does not lose any original image pixels.
- **Advanced augmentation *(AA)***: The current slate of image data augmentation techniques are more varied and less intuitive compared to the simple transformations described above. In our version of advanced augmentation (AA), we use the following pipeline: first we apply (a) the basic augmentation (BA) described above, then (b) RandAugment Cubuk et al. [2019], then (c) random erasing Zhong et al. [2020], and finally use (d) MixUp Inoue [2018]. RandAugment uses a randomly chosen composition of transformations from a predefined list. We use the standard RandAugment list from Wightman [2019] but without the `TranslationX` and `TranslationY` as these are covered with more control within basic augmentation.
- **AA without translation *(AA(no tr))***: We also consider a variant of advanced augmentation (AA) where we remove any augmentations that are explicitly related translation shifts. Specifically, we remove random crop from basic augmentation (BA) and in the RandAugment transformations list, along with previously removed `TranslationX` and `TranslationY`, we also remove `ShearX` and `ShearY`. In this case, any improvements over NoAug arise only indirectly.

2.3 In-distribution test accuracy

Table 1 gives the test accuracies on the in-distribution test dataset, which is preprocessed with the same padding configuration as the train dataset (*i.e.,* with 8 pixels padded symmetrically on all sides). We use this as a reference performance without any distribution shifts (a.k.a. the *in-distribution* accuracy). Ideally, we would expect a classifier that that learns good image priors to maintain their reference performance even when the test dataset is shifted by object invariant properties.

		resent18_bn	resnet18_gn	antialiased _resnet18	cait_xxs36	resmlp_12
CIFAR-10	NoAug	$90.85_{\pm 0.19}$	$91.48_{\pm 0.09}$	$92.55_{\pm 0.21}$	$77.58_{\pm 0.11}$	$79.99_{\pm 0.15}$
	BA	$96.10_{\pm 0.05}$	$95.96_{\pm 0.06}$	$95.63_{\pm 0.15}$	$87.69_{\pm 0.50}$	$87.73_{\pm 0.07}$
	AA(no-tr)	$96.35_{\pm 0.06}$	$96.06_{\pm 0.07}$	$96.98_{\pm 0.07}$	$95.09_{\pm 0.19}$	$91.90_{\pm 0.10}$
	AA	$97.74_{\pm 0.03}$	$98.03_{\pm 0.06}$	$97.77_{\pm 0.10}$	$97.25_{\pm 0.01}$	$96.03_{\pm 0.09}$
CIFAR-100	NoAug	$67.62_{\pm 0.65}$	$64.62_{\pm 0.29}$	$70.37_{\pm 0.09}$	$43.43_{\pm 0.12}$	$52.79_{\pm 0.21}$
	BA	$78.68_{\pm 0.18}$	$78.42_{\pm 0.08}$	$78.20_{\pm 0.16}$	$57.62_{\pm 0.78}$	$60.52_{\pm 0.60}$
	AA(no-tr)	$74.56_{\pm 0.46}$	$74.09_{\pm 0.22}$	$77.62_{\pm 0.25}$	$77.74_{\pm 1.40}$	$65.43_{\pm 0.30}$
	AA	$82.98_{\pm 0.14}$	$82.09_{\pm 0.27}$	$81.54_{\pm 0.17}$	$82.46_{\pm 0.27}$	$78.63_{\pm 0.28}$

Table 1: Accuracy on in-distribution test dataset (*i.e.,* test and train images have same padding).

3 Generalization to translation shifts

With the flexibility of padded pixels in our preprocessed training data, we can now create test datasets with up to 16 pixel translations (in Hamming distance) from the training image distribution by moving the test images anywhere within the 48×48 frame. This allows us to evaluate large translation while not distorting the contents of the image itself. In the extreme locations (see, *e.g.,* corners of grid in Figure 2) there only 25% overlap with the training distribution.

For each trained model, we can look at a grid of 17×17 test evaluations on modified test datasets. Each cell in the evaluation grid corresponds to the position of the 32×32 test images within the 48×48 frame (see illustration in Figure 2 for a *resnet18_bn* network). The center cell of the grid acts as the reference performance and corresponds to the no distribution shift, *i.e.,* the test images are centered on the frame, same as the train images. As we move away from the center, we analogously translate the position of the object image in the test dataset. The model is then evaluated for classification accuracy on the shifted test dataset. Thus, the generalization or robustness of trained models to translation shifts can be comprehensively summarized by such a grid.

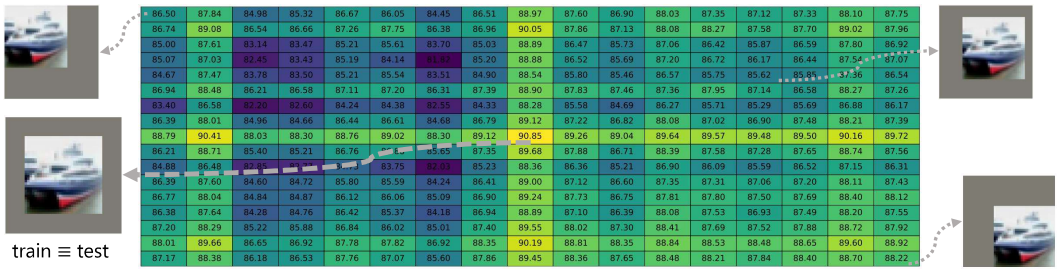
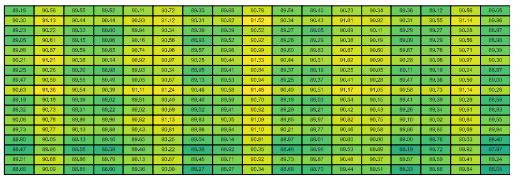


Figure 2: *Generalization to translation shifts of a resnet18_bn trained without data augmentation (NoAug) for* 1600 *epochs on CIFAR-10 dataset:* Each cell in the grid corresponds to the model performance on a test dataset with specific positioning of the object image. The center cell corresponds to no translation shift from training; and the distance of the cells from the center corresponds to respective position shift between train and test dataset. The values in each cell is the accuracy on the shifted test dataset. The color corresponds to "relative" drop in performance from the in-distribution performance: yellow the maximum accuracy in the grid (typically the center cell), while the dark blue is saturated at 90% of the max-value in the grid, *i.e.,* 10% drop in accuracy. For *resnet18_bn* +NoAug on CIFAR-10, max accuracy on the grid is 90.85% while min accuracy is 81.82%.

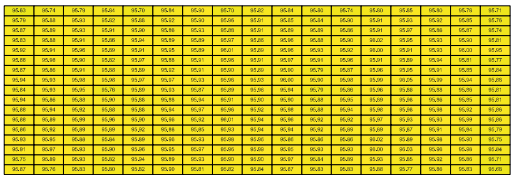
3.1 Case study: Convolutional networks

We first describe some observations from our initial experimentation exclusively on ConvNets trained on CIFAR-10. Our evaluations on other architectures are provided in Section 4. In this subsection, we mainly compare variations of *resnet18* trained in different configurations. First, in Figure 2 along with the illustration of translation shift grid, we show the evaluation of the basic *resnet18_bn* (with batch normalization) on generalization to translation shifts in CIFAR-10. This model was trained without any augmentation to demonstrate the baseline performance of convolutional architecture. We observe that, *despite the built-in image priors in ConvNets and despite using a weaker notion*

than translation invariance, there are significant drops in performance on even small translations shifts. For example, in the non-yellow cells close to center in Figure 2, we notice that a two pixel hamming distance translation can lead to $> 5\%$ *drop in performance.* It is worth forward referencing at this point that in spite of these drops in performance, when compared to other architectures (see Section 4), we find that ConvNets are relatively more resilient to translation shifts in the absence of data augmentation – the worst case drop in performance is $\sim 10\%$ for ConvNets, while for other architectures, the drop in performance could be as large as $30 - 50\%$ (see Figure 4).



(a) *resnet18_gn*+NoAug: max = 91.48, min = 87.97.



(b) *resnet18_gn*+BA: max = 95.96, min = 95.63

Figure 3: Generalization to translation shifts of *resnet* variants using the same evaluation as in Figure 2. (a) *resnet18_gh* network group norm and weight standardization again trained without any augmentation. (b) *resnet18_gn* trained with basic augmentation (BA) consisting of random horizontal flip and random crop at most 4 pixels. For quick reference, the sub-captions mention the max and min accuracy of the models over the grid.

Batchnorm vs groupnorm+weight standardization. Prior work most notably Azulay and Weiss [2019], Zhang [2019], Chaman and Dokmanic [2021] attribute the lack of invariance to strides and ReLU non-linearities in the standard networks. We believe these factors also affect the weaker notion of generalization to translation shifts that we study. Indeed, the "near" periodic locations of high-performing (yellow) cells is in line with prior observations that strides in ResNets would tend to have periodic translation invariance (Azulay and Weiss [2019], Fukushima and Miyake [1982]). Additionally, our experiments suggest that batch normalization is yet another factor that contributes to lack of translation invariance. In Figure 3(a), we show the performance of *resnet18_gn* model trained with the same configuration as *resnet18_bn* in Figure 2, but with a modification that all the batch normalization layers replaced by group normalization and weight standardization Wu and He [2018], Qiao et al. [2019], as was also done in Kolesnikov et al. [2020]. We see that this simple modification already improves the generalization of the ResNet to translation shifts. In hindsight, it is understandable that batch normalization would have detrimental effects when the test distribution shifts from the training distribution as the batch statistics obtained from exponential moving average of training statistics no longer remains accurate Wu and Johnson [2021].

Training with basic augmentation (BA). Experiments in Azulay and Weiss [2019] show that even after using data augmentation, ConvNets are not translation invariant when tested on inputs that are not from the same manifold as training images. On the other hand, in our evaluation of a weaker notion of generalization to translation shifts, we observe a different conclusion. In Figure 3(b), we show that even a small nudge using basic augmentation (BA) can make the models remarkably robust. A simple augmentation using random crops of up to 4 pixels and horizontal flips, not only improve in-distribution accuracy by over 5%, but also make the networks near-perfectly robust to up to 8 pixel translations in test distribution – indicating a form of meta-generalization from augmentations. In the appendix we provide further evidence of such meta-generalization: (a) In (Appendix B) ConvNets trained with a more minimal *BA-lite* and *BA-liter* augmentation with smaller range of random crops of at most 2 and 1 pixels, respectively already help robustness to translation shifts of up to 8 pixels on CIFAR-10, and (b) In Appendix C our results on TinyImageNet dataset show robustness to a larger range of translation shifts of up to 16 pixel shifts, even though BA still uses only 4 pixel shifts.

4 Architectures and augmentations for generalization to translation shifts

The summary grid view of evaluations on translation shifts (as in Figure 2-3) is more comprehensive and we will revisit them in the appendix. However, it is not ideal for comparing different configurations of architectures and augmentations. In this section, we use an alternative visualization and plot the test accuracies as a function of Hamming distance between the position of images in the test and training datasets. The performance of all our models and augmentations are summarized in Figure 4.

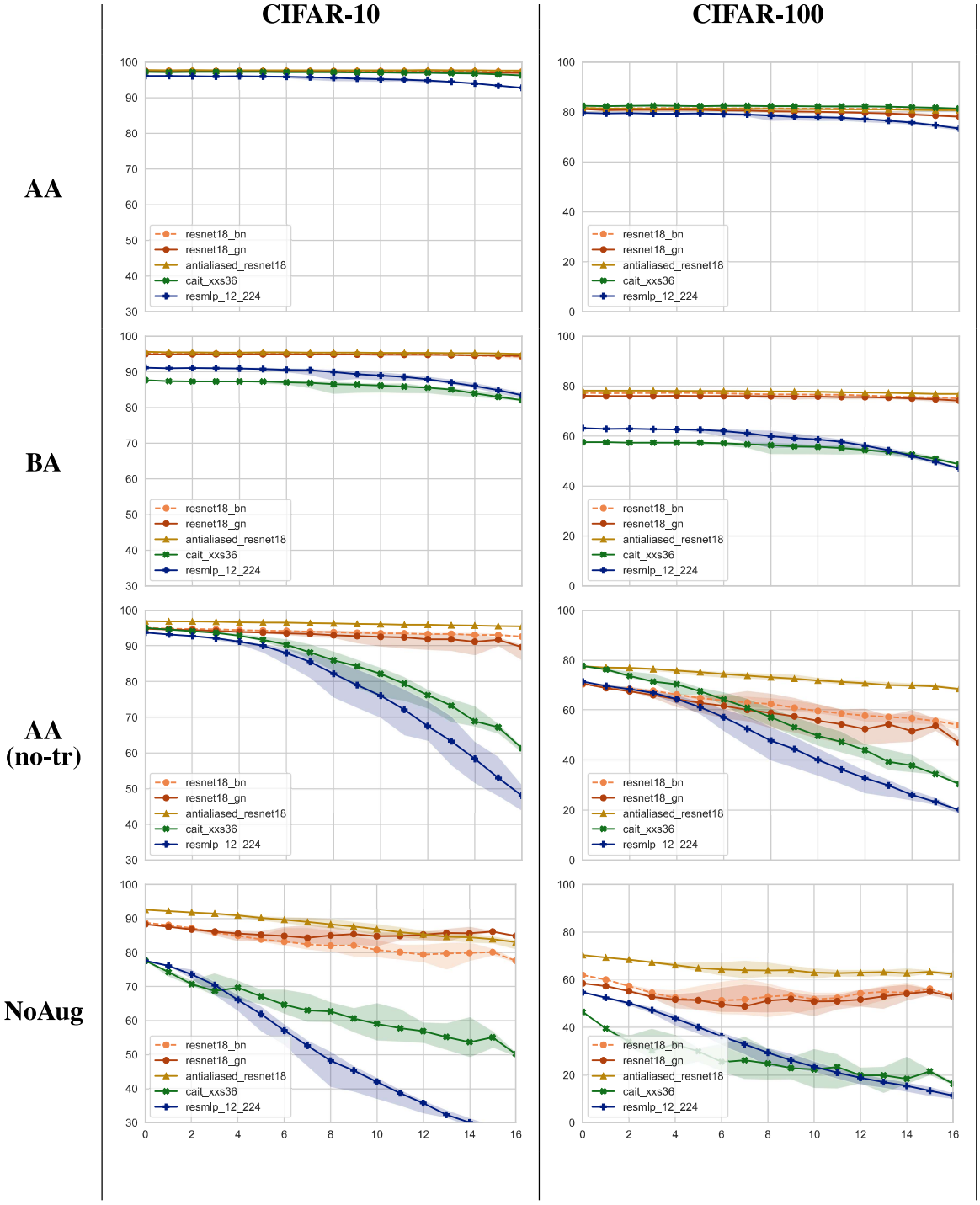


Figure 4: Generalization to translation shifts. The x-axis in each plot is the Hamming distance of the location of test images (within the 48×48 frame) relative to the location in the training images (center of the frame). The larger the x-axis value, the larger is the shift from train distribution. For a given value of x, there might be multiple test configurations that are x-Hamming distance away (like shift of 1 pixel to top, bottom, right, or left when $x = 1$). The median of these values is plotted as line plot, while the shaded region covers the min and max values of the list. The left and right plots correspond to accuracies on the CIFAR-10 and the more challenging CIFAR-100 dataset, respectively. The y-axis for each dataset (column) is normalized to be on the same scale.

4.1 Key takeaways

We make the following observations that are supported by our experiments. We provide additional experiments and discussions from our study including evaluation on TinyImageNet in the appendix.

1. Without data augmentation, even ConvNets with designed architecture has a noticeable drop in performance on spatially shifted test dataset. From Figures 2-3, one can see that merely switching batchnorm with groupnorm appears to mitigate some performance gap. Using a specialized antialiased modification by Zhang [2019] further improves the performance, albeit not to perfect invariance as the non-linearities could be a source of breaking spacial invariance Chaman and Dokmanic [2021]. Nonetheless, the architectural inductive biases in these ConvNets are still useful, if not perfect, as we can predictably see that the drop in performance is much more dramatic for the non-convolutional architectures without any image priors. In CaiT and ResMLP, without augmentation, even $1-2$ pixel translation shift can lead to dramatic drop in performance.
2. In the other extreme, with an Advanced Augmentation (AA) pipeline, all the architectures are remarkably robust even to large translation shifts in test distribution. Note that even with AA, the maximum translation augmentation we provide (in the form of random crop) is at most 4 pixels (8 pixels in Hamming distance), but we see robustness to up to 8 pixel shifts (16 pixels in Hamming distance). This supports a notion of meta generalization in robustness performance. In Appendix C, we see that ResNets continue to be robust to even larger translation shifts of up to 16 pixels on each direction on TinyImageNet dataset. To further support the idea of meta-generalization, we also show in Appendix B, even more minimal 1 or 2 pixel random crop augmentation already boosts robustness to translation shifts.
3. Furthermore, with AA, the performance on in-distribution test error becomes significantly closer for all architectures. Specially, the performance of resnet18 and cait_xxs36 are statistically identical in this setting even though we trained on the small-medium scale CIFAR datasets. The performance of resmlp_12 is however relatively suboptimal even with AA pipeline. First, there is non-trivial gap in the in-distribution accuracies. Secondly, the robustness to translation shifts is not nearly as good as with ResNet or CaiT. Despite these differences, even for ResMLP, the augmentations dramatically boost the generalization to translation shifts and the differences in relative drop start to appear only after 10-12 pixel hamming distance shifts in test distribution. These experiments suggest that with sufficient augmentation, the relative benefits or shortcomings of the architectures are effectively diminished.
4. Even with a minimal Basic Augmentation (BA), we see significant improvement in robustness to translations. In fact, for ConvNets, BA is sufficient to achieve the near perfect generalization on our canvas. This further highlights the benefit of the built-in inductive biases in ConvNet. For non-convolutional architectures, this simple augmentation is not sufficient to achieve optimal absolute test performance, but the relative robustness is still uniformly improved.
5. Finally, an intriguing phenomenon is observed when training with advance augmentation but without translation related augmentations (AA(no-tr)). Here the absolute test accuracy of all the models improves (presumably from learning some useful priors). For ConvNets on CIFAR-10, even such indirect augmentation is effective in making the models robust to translation shifts – but this does not appear to uniformly hold across datasets, so the conclusion might be spurious. For transformer and MLP architectures, the robustness does not improve significantly, even though the in-distribution accuracies are significantly higher.
6. Somewhat tangentially, our experiment slightly supports the position that data augmentation can recover some of the benefits of large datasets even when learning with general architectures like ViT and MLP are beneficial. In ImageNet scale datasets this was previously observed in Touvron et al. [2021c], Steiner et al. [2021] for ViTs and Touvron et al. [2021a] for MLPs. Our experiments show similar validation even on small CIFAR-10 datasets.

In summary, even though convolutional networks are not invariant or robust to translations in absolute sense, they clearly fare much better compared to other general-purpose architectures. Specially, ConvNets learn robust models with minimal augmentation, while it appears the transformer and MLP architectures require more sophisticated augmentations. At the same time, our experiments suggest that data augmentation can enforce learning of the right inductive bias with comparable or more effectiveness than the network architecture.

References

Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. Mind the pad–cnns can develop blind spots. *arXiv preprint arXiv:2010.02178*, 2020.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20:1–25, 2019.

Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34, 2021.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10231–10241, October 2021.

Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3783, 2021.

Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16458–16468, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019.

Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? In *British Machine Vision Conference (BMVC)*, 2015.

Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. *Advances in neural information processing systems*, 22, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020.

Hiroshi Inoue. Data augmentation by pairing samples for images classification. *arXiv preprint arXiv:1801.02929*, 2018.

Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4441–4449, 2018.

Osman Semih Kayhan and Jan C van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14274–14285, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Kuang Liu. Train cifar10 with pytorch. `https://github.com/kuangliu/pytorch-cifar`, 2017.

Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2(3), 2021.

Francesco Pinto, Philip HS Torr, and Puneet K Dokania. An impartial take to the cnn vs transformer robustness contest. *arXiv preprint arXiv:2207.11347*, 2022.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Weight standardization. 2019.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *arXiv e-prints*, pages arXiv–2103, 2021.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021a.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021b.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, October 2021c.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Yuxin Wu and Justin Johnson. Rethinking" batch" in batchnorm. *arXiv preprint arXiv:2105.07576*, 2021.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

Jessica Yung, Rob Romijnders, Alexander Kolesnikov, Lucas Beyer, Josip Djolonga, Neil Houlsby, Sylvain Gelly, Mario Lucic, and Xiaohua Zhai. Si-score: An image dataset for fine-grained analysis of robustness to object location, rotation and size. *arXiv preprint arXiv:2104.04191*, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.