Truly shift-invariant convolutional neural networks

Anadi Chaman
University of Illinois at Urbana-Champaign
achaman2@illinois.edu

Ivan Dokmanić
University of Basel
ivan.dokmanic@unibas.ch

Abstract

Thanks to the use of convolution and pooling layers, convolutional neural networks were for a long time thought to be shift-invariant. However, recent works have shown that the output of a CNN can change significantly with small shifts in input—a problem caused by the presence of downsampling (stride) layers. The existing solutions rely either on data augmentation or on anti-aliasing, both of which have limitations and neither of which enables perfect shift invariance. Additionally, the gains obtained from these methods do not extend to image patterns not seen during training. To address these challenges, we propose adaptive polyphase sampling (APS), a simple sub-sampling scheme that allows convolutional neural networks to achieve 100% consistency in classification performance under shifts, without any loss in accuracy. With APS the networks exhibit perfect consistency to shifts even before training, making it the first approach that makes convolutional neural networks truly shift invariant.

1. Introduction

The output of an image classifier should be invariant to small shifts in the image. For a long time, convolutional neural networks (CNNs) were simply assumed to exhibit this desirable property [35, 36, 37, 38]. This was thanks to the use of convolutional layers which are shift equivariant, and non-linearities and pooling layers which progressively build stability to deformations [6, 41]. However, recent works have shown that CNNs are in fact not shift-invariant. [2, 57, 17, 33, 30]. Azulay and Weiss [2] show that the output of a CNN trained for classification can change with a probability of 30% with merely a one-pixel shift in input images. Related works [30, 33] have also revealed that CNNs can encode absolute spatial location in images: a consequence of a lack of shift invariance.

One of the key reasons for why CNNs are not shift invariant is downsampling\(^1\) [2, 57], or stride, which is a linear operation that samples evenly spaced image pixels located at fixed positions on the grid and discards the rest. As shown in Fig. 1(a), the results of downsampling an image and its shifted version can be significantly different. This is because shifting an image can change the pixel intensities located over the sampling grid. Various measures have been proposed in the literature to counter this problem. With data augmentation [47], the output of a CNN can be made more robust to shifts by training it on randomly shifted versions of input images [2, 57]. This, however, improves the network’s invariance only for image patterns seen during training [2]. Anti-aliasing or blurring spreads sharp image features across their neighbouring pixels which improves structural similarity between subsampled outputs of an image and its shifted version (Fig. 2).

Figure 1. (a) Conventional downsampling is not shift invariant. It samples image pixels at fixed locations on the grid (shown with small squares). Shifting the image changes its pixel intensities located on the fixed grid, resulting in a different subsampled output. (b) By choosing the grid that supports pixels with highest energy, our approach results in shift invariance.

\(^1\)Layers like strided max-pooling in CNNs can be regarded as a combination of a dense max-pooling operation followed by downsampling.
Lack of shift invariance due to downsampling was noted early on by Simoncelli et al. [46]. They showed that shift
invariance in multi-scale convolutional transforms is not possible, and instead proposed the notion of shiftability, a weaker form of invariance associated with anti-aliasing. Azulay and Weiss [2] and Zhang [57] studied the lack of shift invariance due to sub-sampling in modern convolutional neural networks. Azulay and Weiss [2] showed that anti-aliasing can completely restore shift invariance in linear convolutional networks ending with global average pooling. They provided empirical evidence in favour of anti-aliasing for improving robustness to shifts in non-linear networks. Zhang [57] showed that by combining a dense max pooling layer with blurring before subsampling, both consistency and accuracy of classification can be improved. Zou et al. [58] used content-aware anti-aliasing to prevent signal loss from over-blurring. Instead of explicitly anti-aliasing the feature maps, Sundaramoorthi and Wang [48] showed that by parameterizing convolutional filters with smooth Gauss–Hermite basis functions, CNN classifiers can attain translation insensitivity, a weak form of shift invariance. While useful in practice, anti-aliasing offers only a partial solution. This is because the improved robustness to shifts is limited by the action of non-linear activation functions like ReLU, and does not generalize well to image distributions not seen in the training set [2]. Additionally, attempts to improve classification consistency beyond a point using anti-aliasing alone can adversely impact classification accuracy due to the risk of over-blurring [57].

Removing sub-sampling layers from CNNs can completely restore shift invariance [2]. This is indeed the case with networks based on à trous convolutions [56, 9, 8]. Alas, this leads to a dramatic increase in memory and computation requirements, rendering it an impractical strategy for large networks.

Shift invariance in convolutional neural networks can be lost due to boundary and padding effects that arise due to the finite support of input images [33, 30, 1]. This allows CNNs to encode absolute spatial locations in images.

3. Our proposed approach

3.1. Preliminaries

Shift invariance: An operation \( G \) is said to be shift invariant if for a signal \( x \) and its shifted version \( x_s \), \( G(x) = G(x_s) \). Similarly, it is termed as shift equivariant if \( G(x_s) = (G(x))_s \). Convolution is an example of a shift equivariant operator. We define \( G \) as sum-shift-invariant if \( \sum G(x_s) = \sum G(x) \), where the summation is over the pixels of \( G(x) \) and \( G(x_s) \) respectively.

Convolutional neural networks for classification end in fully connected layers at the end which are not shift invariant. As a result, any shifts in convolutional feature maps of the final layer can impact the classifier’s final output. Global average pooling, popularly used in CNN architectures like ResNet [24] and MobileNet [28] can solve this problem. These layers reduce the feature map of each channel in the final convolutional layer to a scalar by averaging. Thus, if the convolutional part of the network is sum-shift-invariant, the overall classifier architecture can be made shift invariant. Our analysis in subsequent sections assumes the use of global average pooling layers.

Polyphase components. For simplicity, we will consider downsampling of 1-D signals with stride 2. The analysis easily generalizes to images and volumes. Consider a 1-D signal \( x_0(n) \) with discrete-time Fourier transform (DTFT) \( X_0(e^{j\omega}) \), which we will denote by \( X_0(\omega) \) from hereon. The DTFT of a one-pixel-shifted version \( x_1(n) = x_0(n-1) \) is given by \( X_1(\omega) = X_0(\omega)e^{-j\omega} \). Given \( x_0(n) \), there are two ways to uniformly sample it with stride 2: we can choose to retain samples at either even or odd locations on the grid. These two possible downsampled outputs denoted by \( y_0 \) and \( y_1 \) are called the even and odd polyphase components of \( x_0 \), and can be expressed as \( y_0(n) = x_0(2n) \) and \( y_1(n) = x_0(2n-1) \). Notice that the even polyphase component of \( x_1 \) is the same as the odd counterpart of \( x_0 \) and vice versa. The downsampled outputs \( y_0(n) \) and \( y_1(n) \) have DTFTs given by

\[
Y_0(\omega) = \frac{X_0(\omega/2) + X_0(\omega/2 + \pi)}{2},
\]

\[
Y_1(\omega) = \frac{(X_0(\omega/2) - X_0(\omega/2 + \pi))e^{-j\omega/2}}{2}.
\]

The terms in (1) and (2) that correspond to \( (\omega/2 + \pi) \) are called aliased components. They arise when \( x_0 \) contains high frequencies, and can cause significant degradation of the sub-sampled outputs. This is traditionally countered by anti-aliasing [43], a signal processing technique which removes high frequencies in \( x_0 \) by blurring before downsampling.

Global average pooling operation on a signal \( x_0(n) \), results in its mean and, ignoring a normalizing constant, is equal to \( X_0(\omega = 0) \).

3.2. Key problem with downsampling

Downsampling is used in CNNs to increase the receptive field of convolutions, and to reduce the amount of memory and computation needed for training. With these goals, either of the two polyphase components of a 1-D signal is a ‘valid’ result of downsampling. However, when using conventional linear sampling, current neural network architec-
tures always select the even component, rejecting the odd one. As a result, downsampling $x_0$ and its shifted version $x_1$ always results in different signals $y_0$ and $y_1$ which are highly unlikely to be equal \[46\] or invariant under sum. Indeed, we can see from (1) and (2) that $Y_0(0) \neq Y_1(0)$. Blurring-based methods \[57\] attempt to improve invariance by promoting similarity between $y_0$ and $y_1$. For example, anti-aliasing based approaches improve structural similarity between the 2 signals and restore invariance under sum by promoting similarity between $y_0$ and $y_1$. For any integer $m > 1$, if $z_0$ and $z_1$ are obtained by applying a non-linear activation function $g(y) = y^m$ on $y_0$ and $y_1$ respectively, i.e. $z_0 = (y_0)^m$ and $z_1 = (y_1)^m$, then,

$$\sum_{n \in \mathbb{Z}} z_0(n) = \sum_{n \in \mathbb{Z}} z_1(n).$$

The above discussion on the impact of downsampling on sum-shift invariance applies to 2-D images as well, with the difference that instead of 2, there exist 4 polyphase components to choose from.

3.3. Adaptive polyphase sampling

Consider stride-2 subsampling of a single channel image $x$. As shown in Fig. 3(a)-(b), the image can be downsampled along 4 possible grids, resulting in the set of 4 potential candidates for sub-sampling. We refer to these candidate results of downsampling as polyphase components and denote them by \{y_{ij}\}_{i,j=0}^1. Similarly, the polyphase components of a 1-pixel shifted version of $x$, namely $\tilde{x} = x(m-1,n-1)$, are denoted by \{y_{ij}\}_{i,j=0}^1. Notice from Fig. 3(b) that $\{y_{ij}\}$ is just a re-ordered and potentially shifted version of the set \{y_{ij}\}. More formally,

$$\tilde{y}00 = y_{11}(n_1-1, n_2-1), \quad \tilde{y}10 = y_{01}(n_1, n_2-1), \quad (5)$$

$$\tilde{y}01 = y_{10}(n_1-1, n_2), \quad \tilde{y}11 = y_{00}(n_1, n_2). \quad (6)$$

As we saw in Section 3.2, the key reason why conventional sampling is not shift invariant is that it always returns the first polyphase component of an image as output. This results in $y_{00}$ and $\tilde{y}00$ as subsampled outputs, which from (5)-(6) are not equal. We propose adaptive polyphase sampling (APS) to address this challenge. The key idea that APS exploits is that \{y_{ij}\} and \{\tilde{y}_{ij}\} are sets of identical\footnote{The images in the two sets could have some shifts between them as well. However, this does not impact shift invariance for networks ending with global average pooling.} but re-ordered images. Therefore, the same subsampled output for $x$ and $\tilde{x}$ can be obtained by selecting the polyphase component with the highest $l_p$ norm from \{y_{ij}\} and \{\tilde{y}_{ij}\}. This is illustrated in Fig. 3(c). APS obtains its output $y_{APS}$ by using the following criterion with $p = 2$.

$$y_{APS} = y_{i_1,j_1}, \quad \text{s.t. } i_1,j_1 = \arg\max_{i,j} ||y_{ij}||_{l_p}^1 \quad (7)$$

where for reference, conventional sampling returns $y_c = y_{00}$ as the output for $x$. It is interesting to note that since we did not use blurring in (8), $y_{APS}$ will contain...
activations and APS layers results in identical feature maps

\[\tilde{y}_{ij} = y_{ij} + \text{aps}\]

As a result, passing an image \(x\) and its shift \(\tilde{x}\) through a cascade of convolutions, non-linear activations and APS layers results in identical feature maps

When \(x\) is an image with \(C\) channels given by \(x = \{x_k\}_{k=1}^C\), we construct its polyphase components \(\{y_{ij}\}_{i,j=0}^s\) by gathering the respective polyphase components for all channels, as illustrated in Fig. 4. In particular, if we assume each channel \(x_k\), to have components \(\{x_{k,ij}\}_{i,j=0}^s\), then for \(i, j \in \{0, 1\}\),

\[y_{ij} = (x_{k,ij})_{k=1}^C. \quad (9)\]

The output of subsampling \(x\) using APS, denoted by \(y_{\text{aps}}\), can then be obtained similar to (8). The above method can be extended to a general stride \(s\), in a straightforward manner by performing norm maximization over \(s^2\) polyphase components. The overall approach is summarized in Algorithm 1.

### Algorithm 1: Adaptive Polyphase Sampling with stride s

1. **Input:** An image \(x = \{x_k\}_{k=1}^C\) with \(C\) channels.
2. For \(i, j \in \{0, 1, \ldots, s-1\}\), polyphase components:

\[\{y_{ij}\} = x(sn_1 + i, sn_2 + j) = \{x_{k,ij}\}_{k=1}^C.\]

3. **Output:** \(y_{\text{aps}} = y_{i,j}\)

\[s.t. \quad \max_{i,j} \|||y_{ij}||_{ps}\}_{i,j=0}^{s-1}\]

at the final layer. Therefore, a convolutional neural network with APS stride layers and a global average pooling layer at the end, yields 100% classification consistency, thus exhibiting perfect shift invariance.

### 3.4. Impact of boundary effects on shift invariance

While training CNNs to be shift invariant via data augmentation, a standard practice is to show the networks randomly shifted crops of images in the training set. The shifted images obtained this way have minor differences near their boundaries. After each layer these differences are amplified and propagated across the whole image to the point that shift invariance is lost even in the absence of downsampling. One way to mitigate this is to pad images with enough zeros. This, however, leads to computation and memory overhead.

To separate the two sources of loss in shift invariance—downsampling and boundary effects—we use circular padded convolutions and shifts in our experiments [57]. With circular padding, CNNs with APS yield 100% classification consistency to shifts. We then train and evaluate the networks with standard padding and random crop based shifts as well, to still observe superior performance of APS over other approaches.

### 3.5. Combining APS with anti-aliasing

We saw in Section 3.3 that APS can achieve perfect shift invariance without blurring the feature maps. While anti-aliasing is not strictly needed for shift invariance, it is still a useful tool to use before sampling. This is because, as discussed in Section 3.1, it reduces information loss caused by aliased components during sampling. Hence, combining APS with anti-aliasing can help us in reaping the advantage of additional improvements in classification accuracy. This can be done by slightly blurring the feature maps before downsampling them with APS.

### 4. Experiments

We evaluate the performance of APS on CIFAR-10 [34] dataset. For CIFAR-10, a 0.9/0.1 training/validation fractional split is used over the 50k training set with the final results reported over the test set of size 10k. APS is
Table 1. Classification consistency and accuracy evaluated on CIFAR-10 test set with ResNet models using blurring (LPF) and APS based downsampling. Circular padding was used in convolutional layers and circular shifts were used for consistency evaluation. The models were trained without being shown shifted images during training.

| Model     | Accuracy | Consistency |
|-----------|----------|-------------|
|           | ResNet-20 | ResNet-36 | ResNet-18 | ResNet-50 | ResNet-20 | ResNet-36 | ResNet-18 | ResNet-50 |
| Baseline  | 89.76%   | 91.40%   | 91.96%   | 90.05%   | 90.83%   | 91.89%   | 90.88%   | 88.96%   |
| APS       | 90.88%   | 92.66%   | 93.97%   | 94.05%   | 100%     | 100%     | 100%     | 100%     |
| LPF-2     | 90.99%   | 92.07%   | 93.47%   | 91.61%   | 94.68%   | 94.44%   | 95.06%   | 92.47%   |
| APS-2     | 91.69%   | 92.28%   | 94.38%   | 94.27%   | 100%     | 100%     | 100%     | 100%     |
| LPF-3     | 91.01%   | 92.24%   | 94.01%   | 93.65%   | 95.23%   | 95.07%   | 97.19%   | 95.63%   |
| APS-3     | 91.78%   | 92.72%   | 94.53%   | 93.80%   | 100%     | 100%     | 100%     | 100%     |
| LPF-5     | 91.56%   | 92.98%   | 94.28%   | 94.12%   | 96.53%   | 96.90%   | 98.19%   | 97.38%   |
| APS-5     | 91.75%   | 92.93%   | 94.48%   | 94.07%   | 100%     | 100%     | 100%     | 100%     |

To separate the impact of boundary artifacts and downsampling on shift invariance, we first implement the CNNs with circular padding and evaluate consistency over random circular shifts by up to 3 pixels in each dimension.

We additionally experiment with 3 anti-aliasing filters of size $2 \times 2$, $3 \times 3$ and $5 \times 5$, similar to the ones used in [57]. ResNet model embedded with filter size $j \times j$ is denoted by ResNet-LPF$_j$. Similarly, we use ResNet-APS$_j$ to denote the models which combine $j \times j$ blur filter and APS. All networks have been trained with random horizontal flips, and no random shifts have been used during training unless mentioned otherwise. We label the models trained with random shifts as DA. Further details on how we train and embed the subsampling modules into the networks are provided in the supplementary material.

4.1. Classification consistency and accuracy on test set

We first train and evaluate networks with different subsampling modules on CIFAR-10 dataset. We use ResNet-20, 56, 18 and 50 for these experiments. Originally used in [24] for CIFAR-10 classification, ResNet-20 and 56 are small models that use downsampling twice with stride 2 and contain $\{16, 32, 64\}$ filters in different layers. ResNet-18 and 50 on the other hand, downsample thrice with a stride 2 and contain $\{64, 128, 256, 512\}$ filters. Table 1 shows consistency and accuracy of the models trained with circular padding. As expected, all networks containing APS modules exhibit perfect robustness to shifts evident from 100% classification consistency. Note that this is despite training the networks without showing any shifted versions of images. In contrast, the baseline ResNet-18 model is consistent 90.87% times, whereas its anti-aliased variants LPF-2, 3 and 5 show consistencies of 95.06%, 97.19%, 98.19% respectively.
Table 2. Test accuracy and consistency evaluated on CIFAR-10 dataset with ResNet models containing APS and blur based subsampling modules. Standard zero padded convolutions were used in the networks, and random crop based shifts were used for consistency evaluation. Shifted images were not shown to the models during training.

Similar to Zhang [57], we also observe increase in classification accuracy with improving shift invariance. For instance, APS increases the accuracy of baseline ResNet-18 from 91.96% to 93.97%. We observe that combining APS with anti-aliasing further improves accuracy. As seen in Table 1, for a given filter size, accuracy obtained with APS+Blur is typically higher than the case which only uses blurring. As stated in Section 3.5, we believe this is because of the combined benefits of perfect shift invariance prior from APS, and anti-aliasing’s ability to reduce information loss during sampling.

To understand the role of learned model weights on robustness to shifts, we compare how classification consistency on validation set varies while training ResNet-18 with the different sub-sampling modules. Fig. 5 shows that unlike the baseline and anti-aliasing based approaches, the validation consistency for APS is 100% throughout training. In fact, we observe perfect consistency in models with APS even before training, implying that APS truly embeds shift invariance into the architecture of CNNs. Additionally, notice that anti-aliasing based models have higher classification consistency in the beginning compared to the final converged values. We believe this is due to over-blurring caused by low-pass filters in the initial phase which is then countered by learning filters which emphasize high frequencies.

**Boundary effects.** We saw that APS completely addresses the loss in shift invariance caused by downsampling. However, while using zero padded same-convolutions and evaluating on random-crop based shifts, which is the general practice in modern CNNs, boundary effects can cause a loss in robustness to shifts. We investigate how these factors impact classification consistency and accuracy of the models trained with APS in comparison with blurring and baseline counterparts. ResNet models with standard zero padding are trained and evaluated on CIFAR-10. For consistency evaluation, images are padded with zeros of size 3 on all sides, and a crop of size $32 \times 32$ is randomly chosen. Results in Table 2 reveal that for a given blur filter size, combining APS with anti-aliasing consistently provides better robustness and accuracy compared to blurring alone. In fact, in most cases, consistency boost provided by APS with no anti-aliasing is still higher than the models that only use blurring.

**4.2. Shift invariance on out-of-distribution images**

Azulay and Weiss [2] showed that the robustness to shifts achieved via data augmentation and anti-aliasing gets poorer when the trained models are evaluated on images that differ substantially from the training distribution. In our experiments, we make a similar observation. On clean CIFAR-10 images, we train 4 variants of ResNet-18:
model with (i) APS, (ii) LPF-5, (iii) APS combined with blur (APS-5), and (iv) another with vanilla sub-sampling but trained with random circular shifts of training set (model referred to as DA). The trained models are then evaluated for consistency on test images with small patches of pixels randomly erased from different locations. Fig. 6 shows that for models trained with data augmentation and anti-aliasing, classification consistency continues to fall with increasing size of erased patches. On the other hand, it remains 100% for ResNet-18 with APS and APS-5. In addition, the model with APS and blurring combined exhibits highest accuracy for all sizes of erasures.

We also evaluate these models on a vertically flipped version of CIFAR-10 test set and observe similar results. As shown in Table 3, the consistency of both data augmentation and anti-aliased model suffers on the flipped set, while APS continues to remain 100% consistent. Since vertically flipping the test images semantically pushing them further away from the training set, it is expected for the classification accuracy of all the models to take a hit. However, in this case as well, we observe APS-5 to have higher accuracy on the flipped set than all the other models.

| Models  | Unflipped | Flipped | Unflipped | Flipped |
|---------|-----------|---------|-----------|---------|
| APS-5   | 94.48%    | 47.55%  | 100%      | 100%    |
| APS     | 93.97%    | 44.79%  | 100%      | 100%    |
| LPF-5   | 94.28%    | 46.21%  | 98.19%    | 89.21%  |
| DA      | 94.22%    | 44.97%  | 97.84%    | 84.94%  |

Table 3. Classification accuracy and consistency evaluated on vertically flipped CIFAR-10 test dataset. While models trained with APS continue to remain shift invariant on the vertically flipped set, whereas those trained with data augmentation and blurring alone lose consistency.

4.3. Stability of internal convolutional feature maps to small shifts

We compare the impact of shifting an input image by 1-pixel on the feature maps of ResNet-18 models using APS and blur based subsampling modules. A filter size of $5 \times 5$ has been used for the anti-aliased model. For a test image from CIFAR-10 set and its diagonally shifted version, we compute feature maps with the APS and LPF-5 model. Shift compensated error $\delta(\cdot)$ from (10) is used to compare the feature maps.

Fig. 7 shows the errors for feature maps from the last 3 residual layers of the models (stride-2 sampling used in each layer). In particular, for each layer, we plot the errors of the channels with the highest energy. The results indicate that while the feature maps with LPF-5 develop minor differences due to shift in input, the output of APS is completely stable.
5. Conclusion

Convolutional neural networks lose shift invariance due to sub-sampling (stride). We address this challenge by replacing the conventional linear sampling layers in CNNs with our proposed adaptive polyphase sampling (APS). A simple non-linear scheme, APS is the first approach that allows CNNs to be truly shift invariant. We show that with APS, the networks exhibit 100% consistency to shifts even before training. It also leads to better generalization performance, as evident from improved classification accuracy.

Acknowledgement

We thank Konik Kothari for useful discussions on the work presented in this paper. This research was supported by the European Research Council Starting Grant SWING, number 852821. Numerical experiments were partly performed at sciCORE (http://scicore.unibas.ch/), scientific computing center at University of Basel. We also utilized computational resources supported by the National Science Foundation’s Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign.

References

[1] Bilal Alsallakh, Narine Kokhlikyan, Vivek Miglani, Jun Yuan, and Orion Reblitz-Richardson. Mind the pad—cnn can develop blind spots. arXiv preprint arXiv:2010.02178, 2020.
[2] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? Journal of Machine Learning Research, 20(184):1–25, 2019.
[3] Étienne Barnard and David Casasent. Shift invariance and the neocognitron. Neural Networks, 3(4):403–410, 1990.
[4] Yoshua Bengio, Frédéric Bastien, Arnaud Audebert, Nicolas Boulanger-Lewandowski, Thomas Breuel, Yousouf Chherawala, Moustapha Cisse, Myriam Côté, Dumitru Erhan, Jeremy Eustache, Xavier Glorot, Xavier Muller, Sylvain Pannetier Lebeuf, Razvan Pascanu, Salah Rifai, François Savard, and Guillaume Sordat. Deep learning benefits from out-of-distribution examples. volume 15 of Proceedings of Machine Learning Research, pages 164–172, Fort Lauderdale, FL, USA, 11–13 Apr 2011. JMLR Workshop and Conference Proceedings.
[5] Alberto Bietti and Julien Mairal. Invariance and stability of deep convolutional representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30, pages 6210–6220. Curran Associates, Inc., 2017.
[6] Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. J. Mach. Learn. Res., 20(1):876–924, Jan. 2019.
[7] J. Bruna and S. Mallat. Invariant scattering convolution networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8):1872–1886, 2013.
[8] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4):834–848, 2018.
[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062, 2014.
[10] G. Cheng, P. Zhou, and J. Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 54(12):7405–7415, 2016.
[11] Taco Cohen and Max Welling. Group equivariant convolutional networks. In International conference on machine learning, pages 2990–2999, 2016.
[12] 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 (CVPR), June 2019.
[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.
[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.
[15] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, page 1889–1898. JMLR.org, 2016.
[16] S. Dodge and L. Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In 2017 26th International Conference on Computer Communication and Networks (ICCCN), pages 1–7, 2017.
[17] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludvig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. volume 97 of Proceedings of Machine Learning Research, pages 1802–1811, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
[18] Alhussein Fawzi and Pascal Frossard. Manifest: Are classifiers really invariant? 2015.
[19] Kunihiko Fukushima. Neocognitron for handwritten digit recognition. Neurocomputing, 51:161 – 180, 2003.
[20] Robert Geirhos, David HJ Janssen, Heiko H Schütt, Jonas Rauber, Matthias Bethge, and Felix A Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969, 2017.
[21] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-
[51] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016. 2

[52] Maurice Weiler, Fred A. Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[53] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[54] Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. Deep image: Scaling up image recognition. *arXiv preprint arXiv:1501.02876*, 7(8), 2015. 2

[55] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369*, 2014. 2

[56] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[57] Richard Zhang. Making convolutional networks shift-invariant again. volume 97 of *Proceedings of Machine Learning Research*, pages 7324–7334, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1, 2, 3, 4, 5, 6, 7

[58] Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee. Delving deeper into anti-aliasing in convnets. *arXiv preprint arXiv:2008.09604*, 2020. 3
A. Non-linear activation functions and shift invariance

We saw in Section 3.2 of the paper that anti-aliasing a signal before downsampling restores sum-shift invariance. In particular, consider a 1-D signal \( x_0(n) \) and its 1-pixel shift \( x_1(n) = x_0(n - 1) \). Blurring the two signals (with an ideal low pass filter) followed by downsampling with stride 2 results in \( y_0^0(n) \) and \( y_1^0(n) \) with DTFTs

\[
Y_0^0(\omega) = \frac{X_0(\omega/2)}{2}, \quad Y_1^0(\omega) = \frac{X_0(\omega/2)e^{-j\omega/2}}{2}, \quad (12)
\]

that satisfy \( Y_0^0(0) = Y_1^0(0) \). Azulay and Weiss pointed out in [2] that the sum-shift invariance obtained via anti-aliasing is lost due to the action of ReLU non-linear activation functions in convolutional neural networks. They postulated that this happens through the generation of high-frequency content after applying ReLU. We elaborate on this phenomenon here and also show that high frequencies alone do not provide a full picture.

Let \( g(\cdot) \) be a generic pointwise non-linear activation function applied to the outputs of anti-aliased downsampling. Owing to the pointwise nature of \( g \), the stride operation and the non-linearity can be interchanged. As shown in Fig. A.1(b), despite anti-aliasing \( x_0 \) with an ideal low pass filter LPF, \( g \) generates additional high frequencies which result in aliasing during downsampling. Note that one can not simply add another LPF block after \( g \) as shown in Fig. A.2(a) and hope to get rid of the aliasing caused by the non-linearity. This is because the new LPF block when interchanged with stride results in a dilated filter which is not low pass anymore (Fig. A.2(b)).

While non-linear activations do generate high frequencies, this does not necessarily lead to invariance loss (as we show in A.1). Therefore, in Section A.2 we take a closer look at how the ReLU affects sum-shift invariance in terms of its thresholding behavior.

A.1. Action of polynomial non-linearities on sum-shift invariance

In Theorem 1 from Section 3.3.1, we stated that for any integer \( m > 1 \), non-linear activation functions of the form \( g(y) = y^m \) do not impact sum-shift invariance. We provide the proof below.

**Proof.** Let the DTFTs of \( z_0 \) and \( z_1 \) be \( Z_0(\omega) \) and \( Z_1(\omega) \). Then by definition of the DTFT,

\[
Z_0(0) = \sum_{n\in\mathbb{Z}} z_0(n), \quad \text{and} \quad Z_1(0) = \sum_{n\in\mathbb{Z}} z_1(n). \quad (13)
\]

Since \( z_0 = (y_0)^m \), and \( z_1 = (y_1)^m \), we have

Figure A.1. Pointwise non-linearity can be interchanged with the stride operation. Despite anti-aliasing \( x_0 \) with LPF block, \( g(\cdot) \) generates high frequencies which can lead to additional aliasing during downsampling.

Figure A.2. Additional low pass filtering after \( g(\cdot) \) in (a) does not eliminate the impact of aliasing. This is because, as shown in (b), interchanging the final LPF block with stride operation, results in a dilated version of the filter which is not low-pass any more.

\[
Z_0(\omega) = \left( Y_0^0(\omega) \circledast Y_0^0(\omega) \circledast \cdots \circledast Y_0^0(\omega) \right)_{\text{m times}}, \quad (14)
\]

\[
Z_1(\omega) = \left( Y_1^0(\omega) \circledast Y_1^0(\omega) \circledast \cdots \circledast Y_1^0(\omega) \right)_{\text{m times}}, \quad (15)
\]

where \( \circledast \) represents circular convolution. We can write \( Z_0(\omega) \) as

\[
Z_0(\omega) = \left( \frac{1}{2\pi} \right)^{m-1} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} Y_0^0(\alpha_1) \cdots Y_0^0(\omega - \sum_{i=1}^{m-1} \alpha_i) d\alpha, \quad (16)
\]

\[
= \frac{1}{2} \left( \frac{1}{4\pi} \right)^{m-1} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} X_0(\frac{\alpha_1}{2}) \cdots X_0(\frac{\omega - \sum_{i=1}^{m-1} \alpha_i}{2}) d\alpha, \quad (17)
\]

where \( \bar{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{m-1}) \). Similarly, \( Z_1(\omega) \) is given by

\[
\]
illustrate how sum-shift invariance is lost due to ReLUs. We therefore consider not straightforward when using the ReLU non-linearity. Z expression of functions of the form

From (13) and (21), we have which results in

The result in Theorem 1 can be extended to arbitrary polynomial activation functions of the form \( g(y) = \sum_{i=0}^{m} \alpha_i y^i \) with \( m > 1 \).

### A.2. ReLU spoils sum-shift invariance

Unlike the case with polynomials, deriving a closed form expression of \( Z_0(\omega) \) and \( Z_1(\omega) \) for arbitrary \( x_0 \) and \( x_1 \) is not straightforward when using the ReLU non-linearity. We therefore consider \( x_0 \) to be a narrow-band cosine, and illustrate how sum-shift invariance is lost due to ReLUs.

Let \( x_0 \) be an \( N \) length 1-D cosine signal and \( x_1 = x_0(n-1) \) be its 1-pixel shift. We define the two signals as

\[
x_0 = \cos \left( \frac{2\pi n}{N} \right), \quad x_1 = \cos \left( \frac{2\pi (n-1)}{N} \right)
\]

\( n \in \{0, 1, \ldots, N-1\} \).

For any \( N > 4 \), \( x_0 \) satisfies the Nyquist criterion and is anti-aliased by default. The downsampled outputs \( y_0^\alpha \) and \( y_1^\alpha \) are then defined as follows.

\[
y_0^\alpha = x_0(2n) = \cos \left( \frac{2\pi n}{N'} \right), \quad n \in \{0, 1, \ldots, N'-1\},
\]

\[
y_1^\alpha = x_1(2n) = \cos \left( \frac{2\pi (n-1/2)}{N'} \right), \quad n \in \{0, 1, \ldots, N'-1\},
\]

where \( N' = N/2 \). Note that \( y_0^\alpha \) and \( y_1^\alpha \) are structurally similar signals, and can be interpreted as half-pixel shifted versions of each other. The action of \( g(x) = \text{relu}(x) \) on \( y_0^\alpha \) can be regarded as multiplication by a window function which is zero for any \( n \), s.t. \( y_0^\alpha(n) < 0 \). We construct sets \( \{S_i^+\}_{i=0}^{1} \) which define \( n \) where \( y_0^\alpha(n) > 0 \). For simplicity in constructing the sets, we assume \( N' > 6 \) and divisible by 4 (similar conclusions from below can be reached without these simplifying assumptions as well). Then we have

\[
S_0^+ = \{n : n \in \mathbb{Z}, n \in [0, N'/4 - 1] \cup [\frac{3N'}{4} + 1, N' - 1]\}
\]

\[
S_1^+ = \{n : n \in \mathbb{Z}, n \in [0, \frac{N'}{4}] \cup [\frac{3N'}{4} + 1, N' - 1]\}
\]

Notice that the supports \( S_0^+ \) and \( S_1^+ \) differ slightly by 1 pixel near \( n = \frac{N'}{4} \). This is because despite being structurally similar, \( y_0^\alpha \) and \( y_1^\alpha \) have slightly different zero crossings, which results in some differences in the support of thresholded outputs. We can now compute the sums \( \sum_{n \in \mathbb{Z}} g(y_0^\alpha) \) and \( \sum_{n \in \mathbb{Z}} g(y_1^\alpha) \).
Equation (33) illustrates the loss in sum shift invariance caused by the action of ReLU. Notice that the differences in sum arise due to minor differences in the signal content in \( y_0^a \) and \( y_1^a \), which are magnified by ReLU. The term \( \sin(\cdot) \) arises due to a 1-pixel difference in the supports of \( g(y_0^a) \) and \( g(y_1^a) \), whereas the \( \cos(\cdot) \) term is associated with \( e^{-j\omega/2} \) from (12), again depicting the impact of small differences in \( y_0^a \) and \( y_1^a \).

B. Implementation details

We trained ResNet models with APS, anti-aliasing based and baseline conventional downsampling approaches on CIFAR-10 dataset, and compared their achieved classification consistency and accuracy. Four variants of the architecture were used: ResNet-20, 56, 18 and 50. ResNet 20 and 56 were originally introduced in [24] for CIFAR-10 classification and are smaller models with number of channels: \{16, 32, 64\} in different layers, and use stride 2 twice, which results in a resolution of \( 8 \times 8 \) in the final convolutional feature maps. On the other hand, ResNet-18 and 50 contain \{64, 128, 256, 512\} number of channels, and downsample three times with a stride 2, resulting in final feature map resolution of \( 4 \times 4 \). Similar to the experiments with CIFAR-10 in [24], we use a convolution with stride 1 and kernel size of \( 3 \times 3 \) in the first convolutional layer. In all architectures, global average pooling layers are used at the end of the convolutional part of the networks. Fig. A.3 illustrates the architectures of ResNet-18 and 20.

The original training set of the CIFAR-10 dataset was split into training and validation subsets of size 45k and 5k. All models were trained with batch size of 256 for 250 epochs using stochastic gradient descent (SGD) with momentum 0.9 and weight decay \( 5e^{-4} \). The initial learning rate was chosen to be 0.1 and was decayed by a factor of 0.1 every 100 epochs. Training was performed on a single NVIDIA V-100 GPU. All the models were randomly initialized with a fixed seed before training. We evaluated on the models which reported the highest validation accuracy during training.

We were able to show substantial improvements in accuracy and consistency with APS over baseline and anti-aliased downsampling without substantial hyper-parameter tuning. Further improvements in the results with better hyper-parameter search is therefore possible.

B.1. Embedding APS in ResNet architecture

We replace the baseline stride layers in the ResNet architectures with APS modules. To ensure shift invariance, a consistent choice of polyphase components in the main and residual branch stride layer is needed. APS uses a shift invariant criterion (like \( \arg\max \)) to choose the polyphase component to be sampled in the main branch. The index of the chosen component is passed to the residual branch where the polyphase component with the same index is sampled. An illustration is provided in Fig. B.4.
C. Impact of the APS criterion on accuracy

In the paper, we saw that APS achieves perfect shift invariance by selecting the polyphase component with the highest $l_2$ norm, i.e.

$$y_{APS} = y_{i_1,j_1}, \quad \text{s.t. } i_1,j_1 = \argmax_{i,j} \{\|y_{i,j}\|_2\}_{i,j=0}$$

This can also be achieved, however, with other choices of shift invariant criteria. Here, we study the impact of different such criteria on classification accuracy. In particular, we explore maximization of $l_p$ norms with $p = 1, \infty$ in addition to $p = 2$. We also consider minimization of $l_1$ and $l_2$ norms. We run the experiments on ResNet-18 architecture with 9 different initial random seeds and report the mean and standard deviation of achieved accuracy on the test set.

Table A1 shows that choosing the polyphase component with the largest $l_\infty$ norm provides the highest classification accuracy which is then followed by choosing the one with the highest $l_2$ norm and $l_1$ norm. Additionally, the accuracy obtained when choosing polyphase component with minimum $l_2$ norm is somewhat lower than the case which chooses maximum $l_2$ norm. We believe this could be due to the polyphase components with higher energy containing more discriminative features.

Note that for all cases in Table A1, the achieved classification accuracy is $\sim 2\%$ higher than that of baseline ResNet-18 (reported in the paper). This is because in each case, APS enables stronger generalization via perfect shift invariance prior.

D. Experiments with data augmentation

As shown in Section 4.1 of the main paper, for ResNet models trained without seeing random shifts during training, APS can improve classification consistency by more than $10\%$ and accuracy by more than $2\%$ over the baseline on CIFAR-10 dataset. Here, we assess the impact of APS on the performance of models trained with random shifts in data augmentation (labeled as DA). The results are reported in Table A2.

While data augmentation does substantially improve classification consistency, it is still lower than APS which yields perfect shift invariance. However, we observe higher classification accuracy for models trained with randomly shifted images in comparison to the ones trained without it. This is not surprising because data augmentation explicitly trains the models to attain higher accuracy on random shifts of images similar to the ones seen in training set. On the other hand, as reported in the paper, the accuracy for networks trained with APS is more robust to image corruptions, and the models continue to yield $100\%$ classification consistency on all image distributions.

| APS criterion | Accuracy     | Consistency |
|---------------|--------------|-------------|
| argmax ($l_1$) | 93.89±0.27\% | 100\%       |
| argmax ($l_2$) | 94.03±0.26\% | 100\%       |
| argmax ($l_\infty$) | 94.14±0.25\% | 100\%       |
| argmin ($l_1$) | 93.92±0.12\% | 100\%       |
| argmin ($l_2$) | 93.90±0.16\% | 100\%       |

Table A1. Impact of APS criterion on CIFAR-10 classification accuracy.
Table A2. Impact of APS on the performance of models trained with random shifts in data augmentation. Models trained without data augmentation are also shown as reference.

| Model          | ResNet-18 | ResNet-50 | ResNet-18 | ResNet-50 |
|----------------|-----------|-----------|-----------|-----------|
| Baseline       | 91.96%    | 90.05%    | 90.88%    | 88.96%    |
| APS-3          | 94.53%    | 93.80%    | 100%      | 100%      |
| Baseline + DA  | 94.33%    | **94.77%**| 97.84%    | 97.64%    |
| APS-3 + DA     | **94.61%**| 94.39%    | **100%**  | **100%**  |

Accuracy and Consistency.