Truly shift-invariant convolutional neural networks

Anadi Chaman*, Ivan Dokmanić†

*Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign
†Department of Mathematics and Computer Science, University of Basel

Introduction

Convolutional neural networks lose shift invariance due to downsampling (stride) layers.

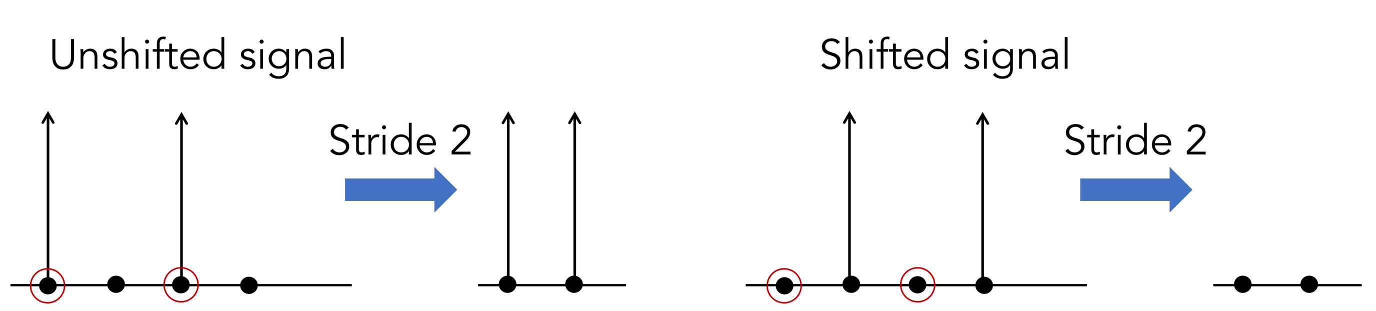
Existing solutions do not enable perfect shift invariance.

- **Data augmentation.** The obtained gains in shift invariance do not extend well to images outside the training distribution.
- **Anti-aliasing.** Improves robustness to shifts but gains limited by the action of non-linear activation functions like ReLU.

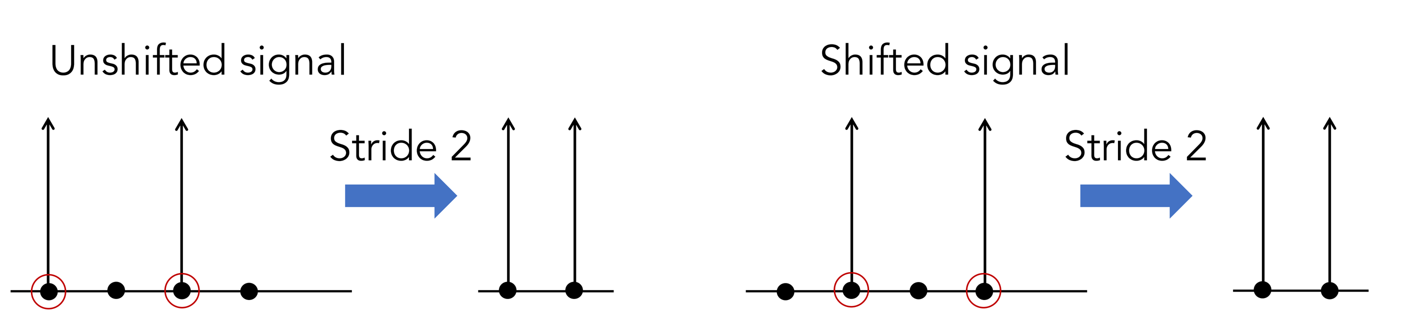
Our goal. To design a CNN architecture that exhibits perfect shift invariance without the above limitations.

The case for adaptive sampling

Conventional downsampling is not robust to shifts because it samples pixels along a fixed sampling grid.

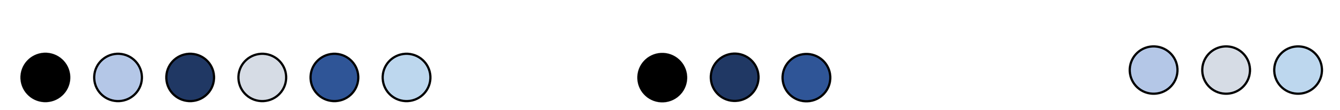


Key idea: Choose the sampling grid adaptively to ensure that same pixels are selected irrespective of shift.



Polyphase components

A 1-D signal can be uniformly downsampled with stride-2 in two ways.



$x(n)$ $\quad y_0 = x(2n)$ $\quad y_1 = x(2n+1)$

y_0 and y_1 are called polyphase components of x and are both equally *valid* results of downsampling.

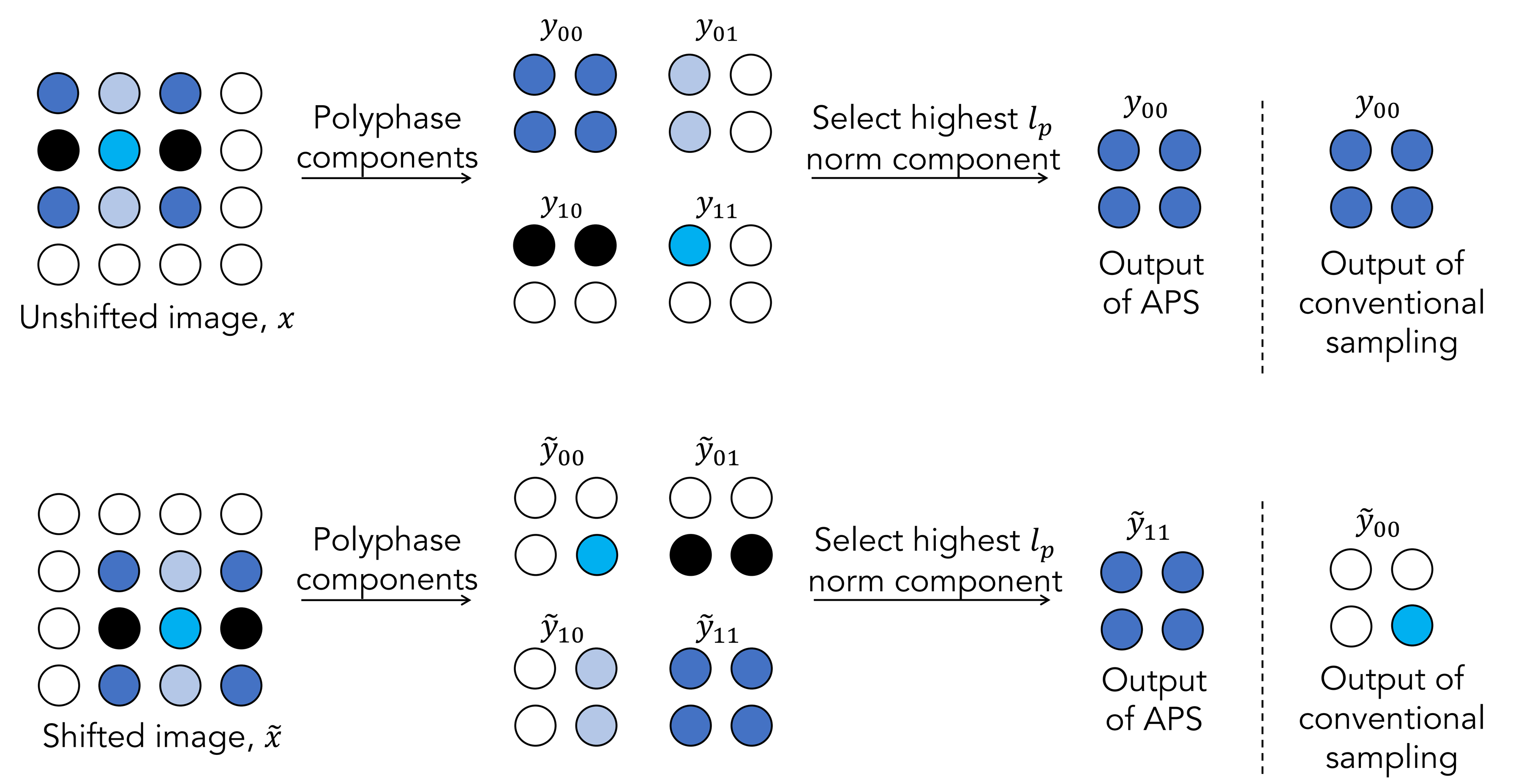
Images can similarly be downsampled to yield 4 polyphase components.

Our approach: Adaptive Polyphase Sampling (APS)

How to select polyphase component to result in shift invariance?

An image and its shift share the same set of polyphase components.

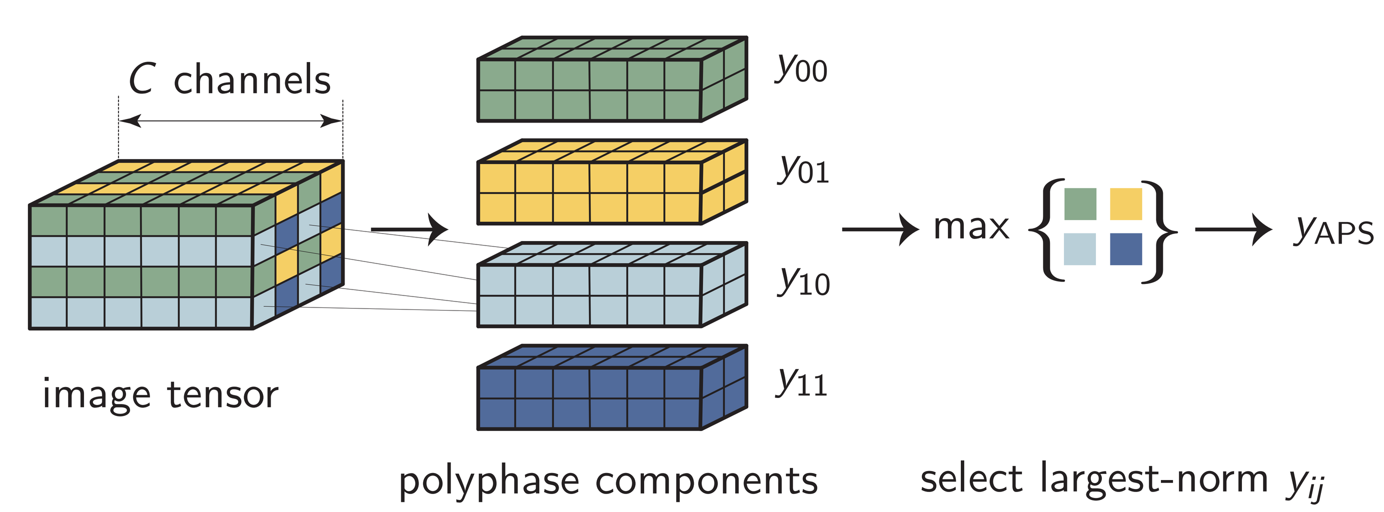
Choose the component with the highest norm as the subsampled output.



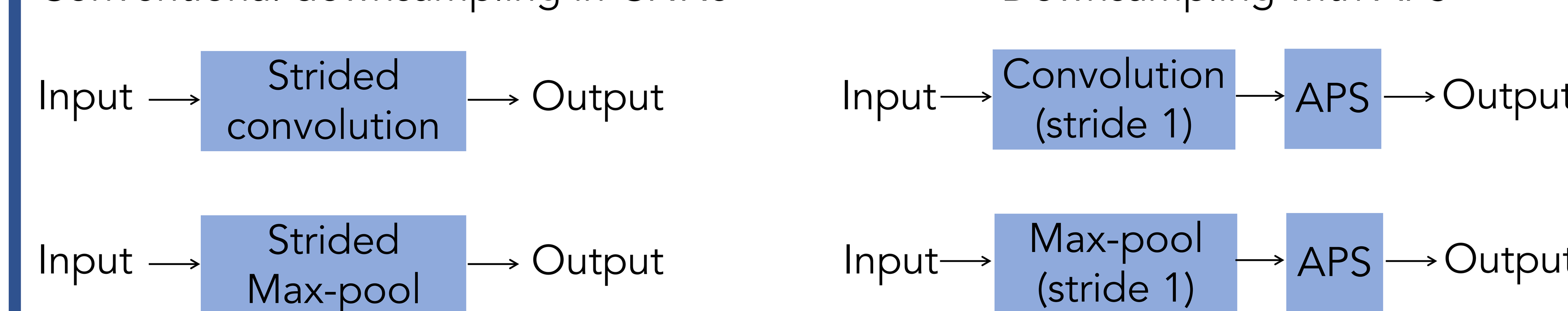
$$y_{APS} = y_{i_1, j_1},$$
$$s.t.\ i_1, j_1 = argmax \left\{ \|y_{i,j}\|_p \right\}_{i,j=0}^{1}$$

Shifting the input of APS by k pixels shifts its output by $\sim k/2$ pixels.

APS on multi-channel image tensor



Integrating APS in CNN architectures



APS does not require any additional learnable parameters.

Results

Classification consistency (shift invariance measure) & accuracy.

Networks containing APS are 100% consistent to shifts and also exhibit improved accuracy as a side benefit.

CIFAR-10 classification

Models	Baseline	APS	LPF-2	APS-2	LPF-3	APS-3	LPF-5	APS-5
Consistency	90.88%	**100%**	95.06%	**100%**	97.19%	**100%**	98.19%	**100%**
Accuracy (unshifted)	91.96%	93.97%	93.47%	94.38%	94.01%	**94.53%**	94.28%	94.48%

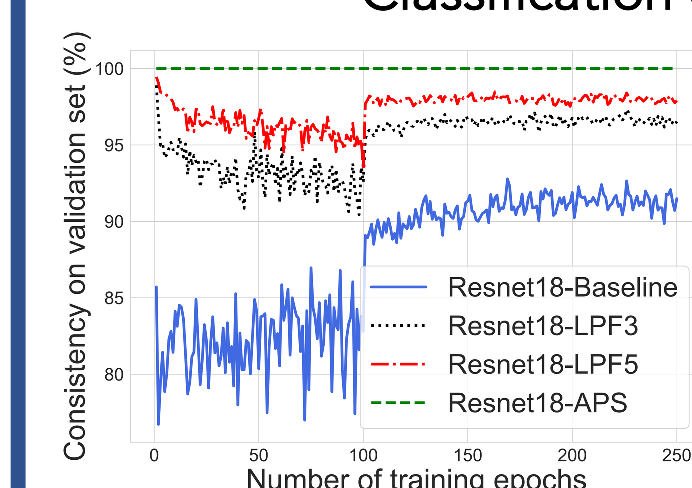
Tab 1. Consistency and accuracy evaluated for ResNet-18 on CIFAR-10 test set. Random circular shifts between -3 to 3 pixels used for evaluation. Models trained without random shifts.

ImageNet classification

Models	Baseline	APS	LPF-2	APS-2	LPF-3	APS-3	LPF-5	APS-5
Consistency	80.39%	**100%**	84.35%	**100%**	86.54%	**99.996%**	87.88%	**99.98%**
Accuracy (unshifted)	64.88%	67.05%	67.03%	**67.60%**	66.96%	67.43%	66.85%	67.52%

Tab 2. Consistency and accuracy evaluated for ResNet-18 on ImageNet validation set. Random circular shifts between -32 to 32 pixels used for evaluation. Models trained without random shifts.

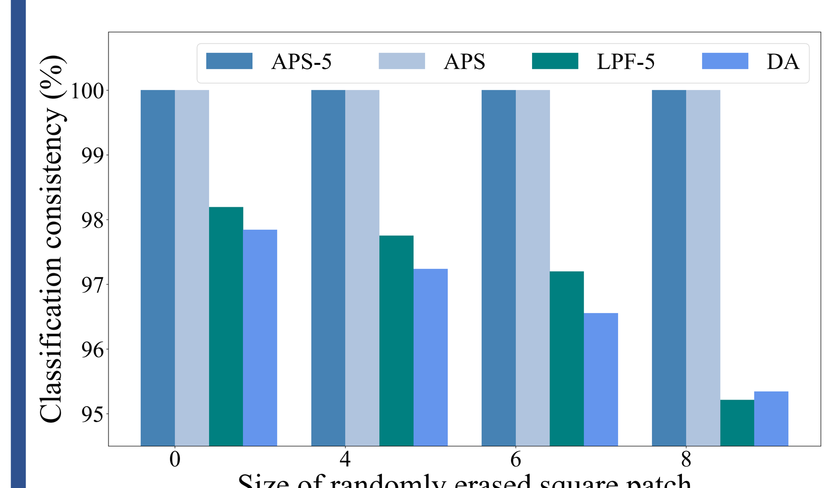
Classification consistency during training



Unlike prior methods, networks with APS are always 100% consistent, even before training.

The shift invariance prior is, therefore, truly embedded in the CNN architecture.

Shift invariance on out-of-distribution images



APS continues to enable perfect shift invariance even as images move away from the training distribution.

This is not true for methods like anti-aliasing (LPF) and data augmentation (DA).

Conclusion

- CNNs lose shift invariance due to downsampling layers.
- Prior methods like data augmentation and anti-aliasing have limitations and do not result in perfect shift invariance.
- We propose APS, a simple non-linear sampling scheme that enables perfect shift invariance without any loss in accuracy.
- APS does not need any additional learnable parameters and can be easily integrated into existing architectures.

Code available at https://github.com/achaman2/truly_shift_invariant_cnns.