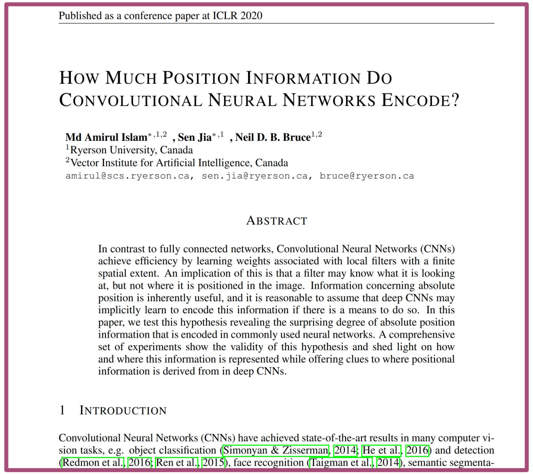
Global Pooling, More than Meets the Eye: Position Information is Encoded Channel-Wise in CNNs

Md Amirul Islam[1,6] Matthew Kowal[2,6] Sen Jia[4] Kosta Derpanis[2,5,6] Neil Bruce[3,6]

Ryerson University YORK UNIVERSITÉ UNIVERSITY UNIVERSITY of GUELPH LG SAMSUNG VECTOR INSTITUTE

Motivation


Published as a conference paper at ICLR 2020

HOW MUCH POSITION INFORMATION DO CONVOLUTIONAL NEURAL NETWORKS ENCODE?

Md Amirul Islam*,[1,2], Sen Jia*,[1], Neil D. B. Bruce[1,2]
[1]Ryerson University, Canada
[2]Vector Institute for Artificial Intelligence, Canada
amirul@scs.ryerson.ca, sen.jia@ryerson.ca, bruce@ryerson.ca

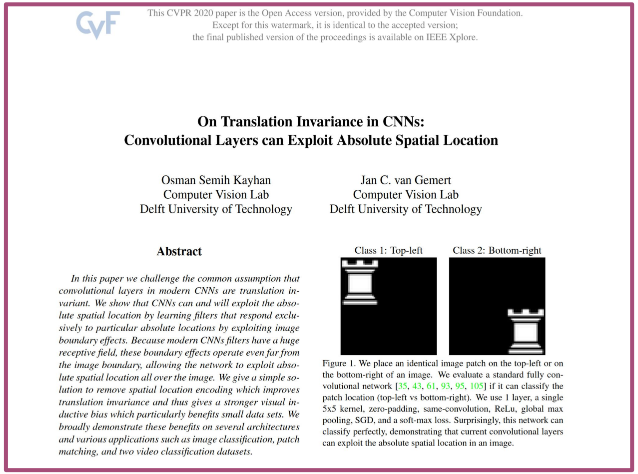
ABSTRACT

In contrast to fully connected networks, Convolutional Neural Networks (CNNs) achieve efficiency by learning weights associated with local filters with a finite spatial extent. An implication of this is that a filter may know what it is looking at, but not where it is positioned in the image. Information concerning absolute position is inherently useful, and it is reasonable to assume that deep CNNs may implicitly learn to encode this information if there is a means to do so. In this paper, we test this hypothesis revealing the surprising degree of absolute position information that is encoded in commonly used neural networks. A comprehensive set of experiments show the validity of this hypothesis and shed light on how and where this information is represented while offering clues to where positional information is derived from in deep CNNs.

1 INTRODUCTION

Convolutional Neural Networks (CNNs) have achieved state-of-the-art results in many computer vision tasks, e.g. object classification (Simonyan & Zisserman, 2014; He et al., 2016) and detection (Redmon et al., 2016; Ren et al., 2015), face recognition (Taigman et al., 2014), semantic segmenta-

Islam et. al (ICLR 2020)


This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

On Translation Invariance in CNNs: Convolutional Layers can Exploit Absolute Spatial Location

Osman Semih Kayhan, Computer Vision Lab, Delft University of Technology
Jan C. van Gemert, Computer Vision Lab, Delft University of Technology

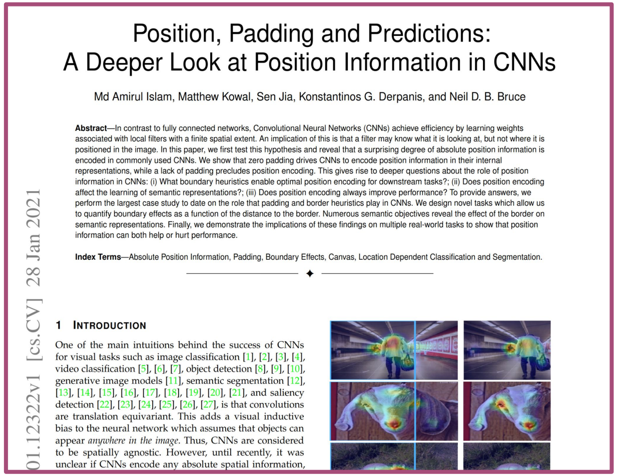
Abstract

In this paper we challenge the common assumption that convolutional layers in modern CNNs are translation invariant. We show that CNNs can and will exploit the absolute spatial location by learning filters that respond exclusively to particular absolute locations by exploiting image boundary effects. Because modern CNNs filters have a huge receptive field, these boundary effects operate even far from the image boundary, allowing the network to exploit absolute spatial location all over the image. We give a simple solution to remove spatial location encoding which improves translation invariance and thus gives a stronger visual inductive bias which particularly benefits small data sets. We broadly demonstrate these benefits on several architectures and various applications such as image classification, patch matching, and two video classification datasets.

Class 1: Top-left Class 2: Bottom-right

Figure 1. We place an identical image patch on the top-left or on the bottom-right of an image. We evaluate a standard fully convolutional network [35, 43, 61, 93, 95, 105] if it can classify the patch location (top-left vs bottom-right). We use 1 layer, a single 5x5 kernel, zero-padding, same-convolution, ReLu, global max pooling, SGD, and a soft-max loss. Surprisingly, this network can classify perfectly, demonstrating that current convolutional layers can exploit the absolute spatial location in an image.

Kayhan et. al (CVPR 2020)


Position, Padding and Predictions: A Deeper Look at Position Information in CNNs

Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G. Derpanis, and Neil D. B. Bruce

Abstract—In contrast to fully connected networks, Convolutional Neural Networks (CNNs) achieve efficiency by learning weights associated with local filters with a finite spatial extent. An implication of this is that a filter may know what it is looking at, but not where it is positioned in the image. In this paper, we first test this hypothesis and reveal that a surprising degree of absolute position information is encoded in commonly used CNNs. We show that zero padding drives CNNs to encode position information in their internal representations, while a lack of padding precludes position encoding. This gives rise to deeper questions about the role of position information in CNNs: (i) What boundary heuristics enable optimal position encoding for downstream tasks?; (ii) Does position encoding affect the learning of semantic representations?; (iii) Does position encoding always improve performance? To provide answers, we perform the largest case study to date on the role that padding and border heuristics play in CNNs. We design novel tasks which allow us to quantify boundary effects as a function of the distance to the border. Numerous semantic objectives reveal the effect of the border on semantic representations. Finally, we demonstrate the implications of these findings on multiple real-world tasks to show that position information can both help or hurt performance.

Index Terms—Absolute Position Information, Padding, Boundary Effects, Canvas, Location Dependent Classification and Segmentation.

01.12322v1 [cs.CV] 28 Jan 2021

1 INTRODUCTION

One of the main intuitions behind the success of CNNs for visual tasks such as image classification [1], [2], [3], [4], video classification [5], [6], [7], object detection [8], [9], [10], generative image models [11], semantic segmentation [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], and saliency detection [22], [23], [24], [25], [26], [27], is that convolutions are translation equivariant. This adds a visual inductive bias to the neural network which assumes that objects can appear *anywhere in the image*. Thus, CNNs are considered to be spatially agnostic. However, until recently, it was unclear if CNNs encode any absolute spatial information,

Islam et. al (arXiv 2021)

CNNs encode absolute position information

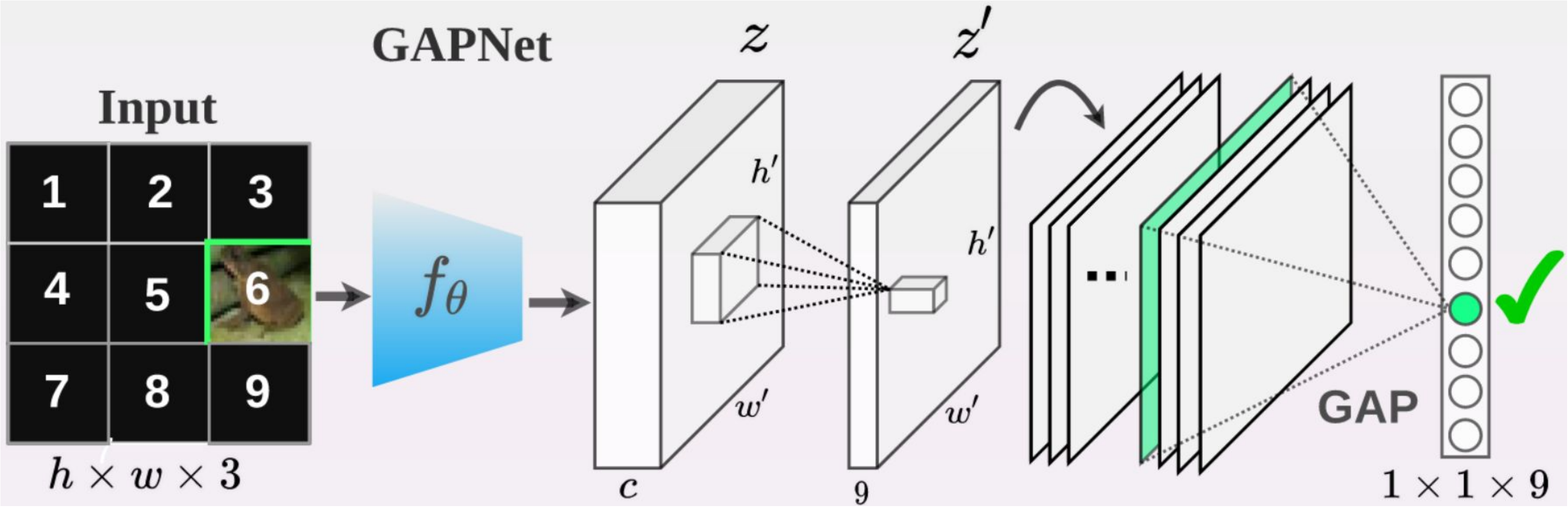
Motivation

How does a CNN contain positional information in the representations after a Global Average Pooling (GAP) layer?

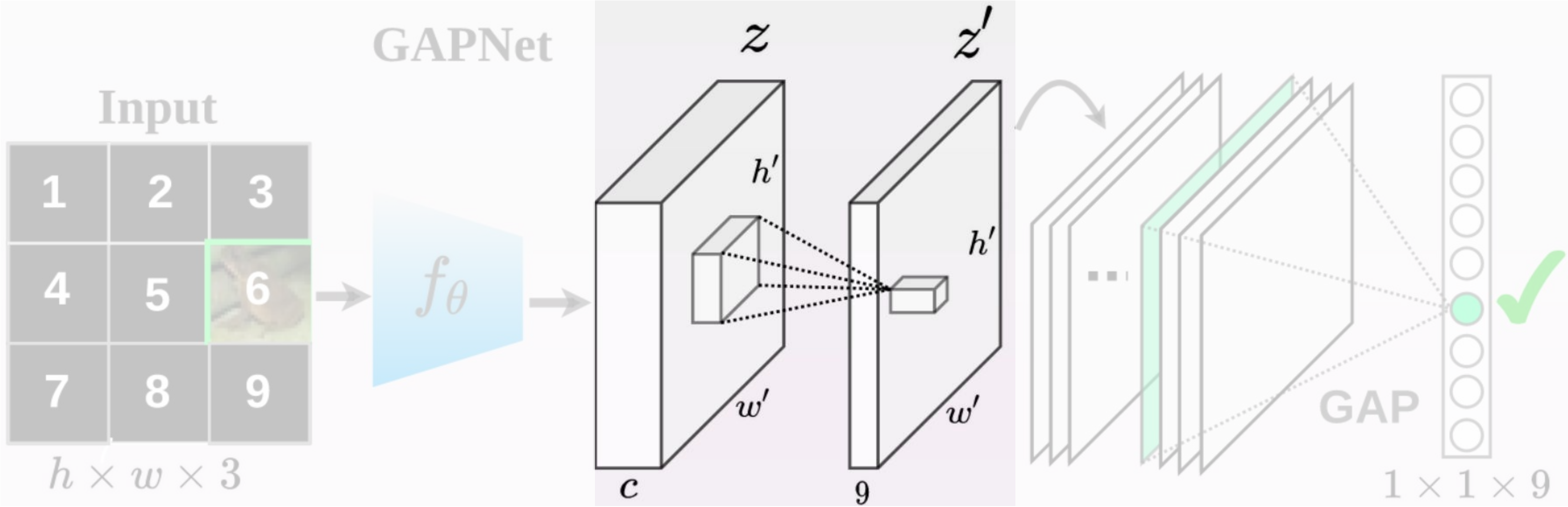
Hypothesis

CNNs encode absolute position information along the ordering of the channel dimension

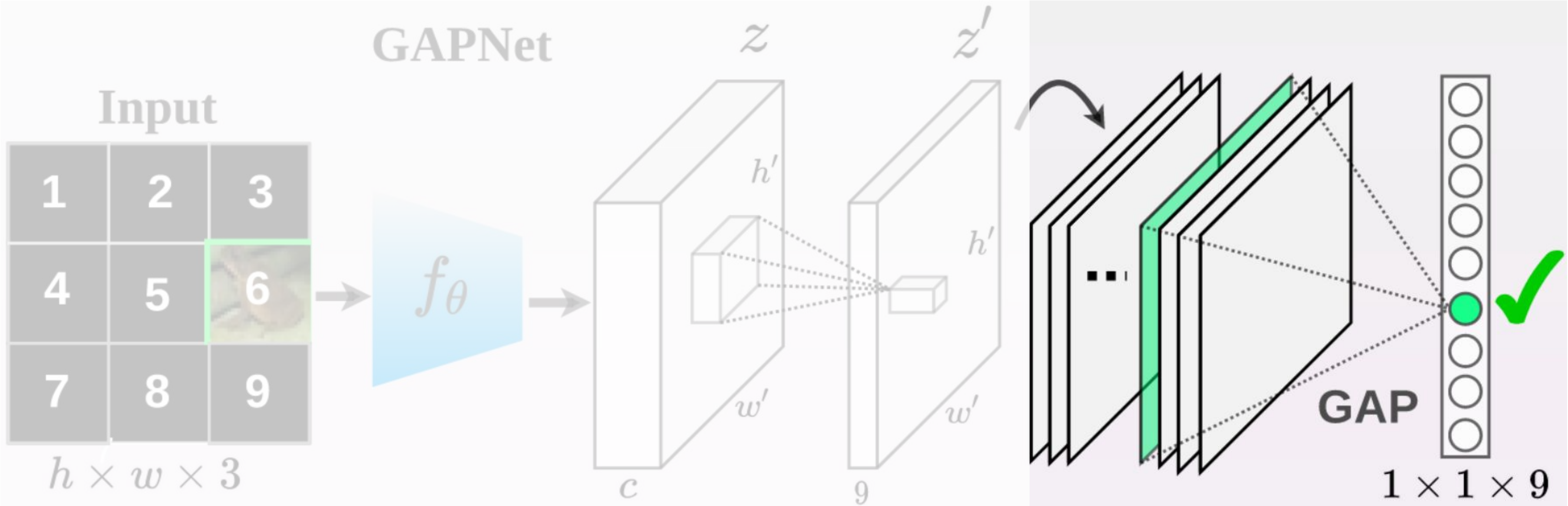
Learning Position with a GAPNet



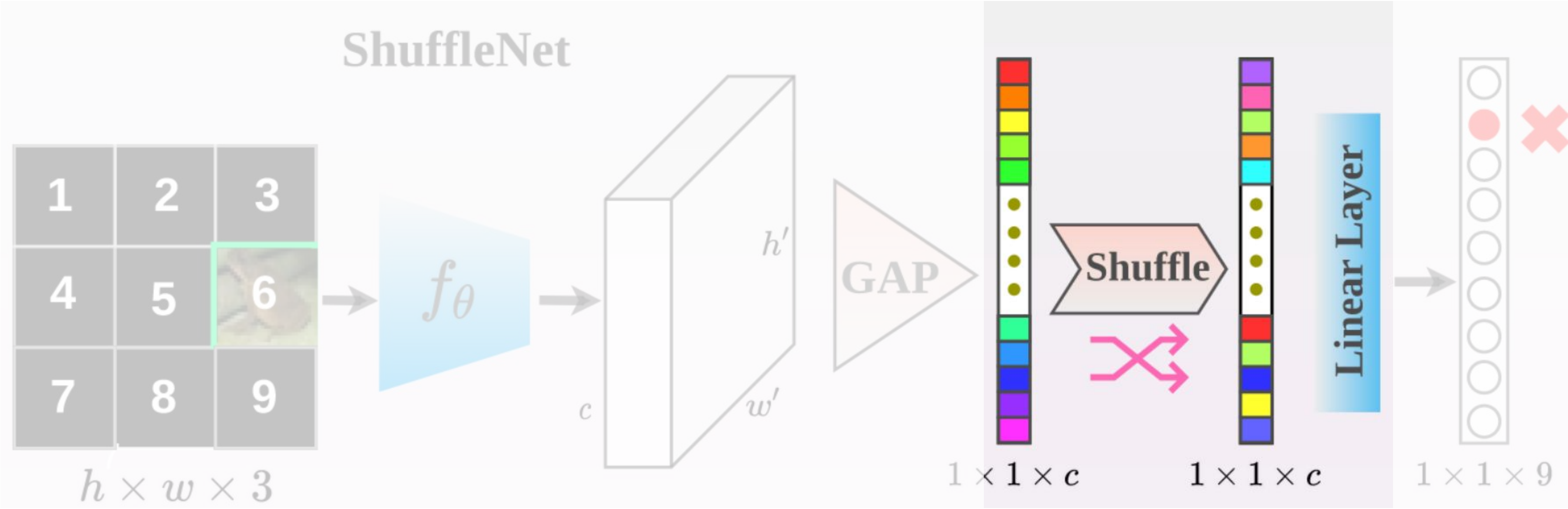
Learning Position with a GAPNet



Learning Position with a GAPNet



Learning Position with ShuffleNet



Evaluation of Channel-wise Position Encoding

Network		Loc. Classification		Image Classification	
		3x3	7x7	3x3	7x7
Res18	GAPNet	100	100	82.6	82.1
	PermuteNet	78.8	21.4	82.1	69.9

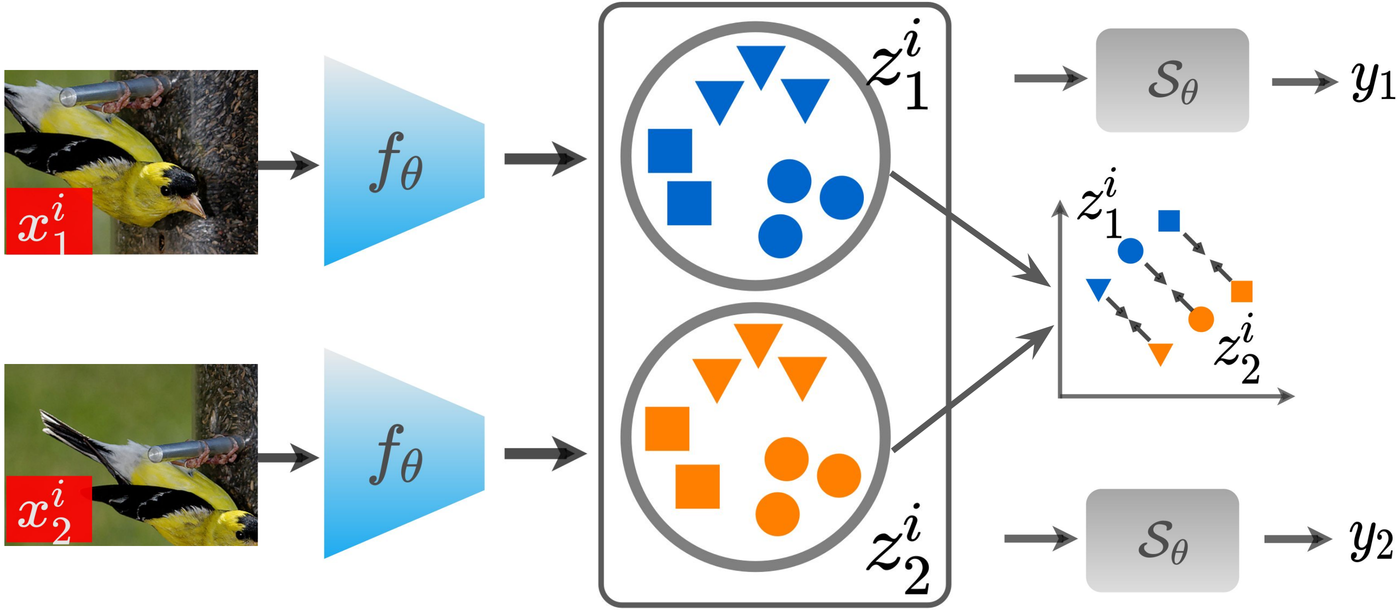
Results are on CIFAR-10 dataset

Applicability of Channel-wise Positional Encoding

Learning Translation Invariant Representation

Attacking Position Encoding Channels

Learning Translation Invariant Representations



Results: Translation Invariance

Network	CIFAR-10		CIFAR-100	
	Top-1 Acc.	Consistency	Top-1 Acc.	Consistency
ResNet-18	93.1	90.8	72.6	70.1
Blurpool	92.5	92.5	72.4	78.2
AugShift (Ours)	92.1	94.8	72.6	85.6

Results: Translation Invariance

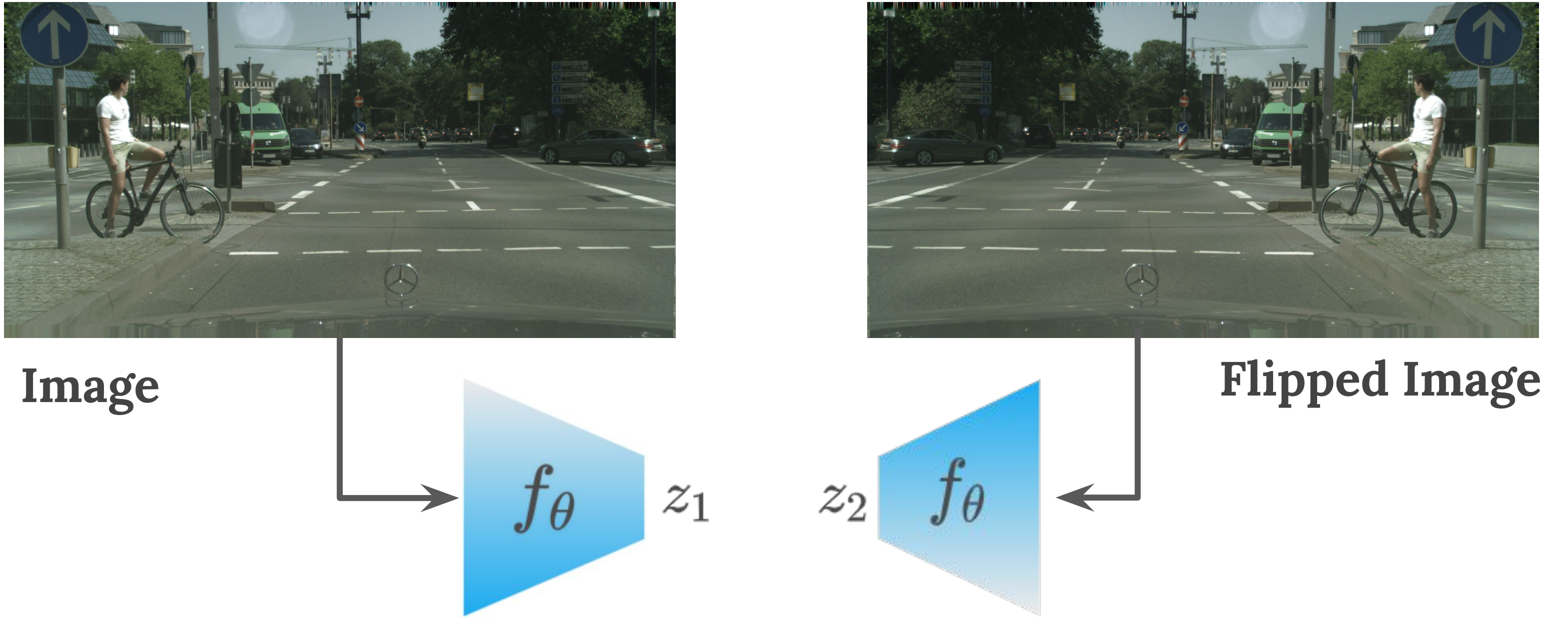
Network	CIFAR-10		CIFAR-100	
	Top-1 Acc.	Consistency	Top-1 Acc.	Consistency
ResNet-18	93.1	90.8	72.6	70.1
Blurpool	92.5	92.5	72.4	78.2
AugShift (Ours)	92.1	94.8	72.6	85.6

Attacking the Position Encoding Channels

1) Identify the position-specific channels

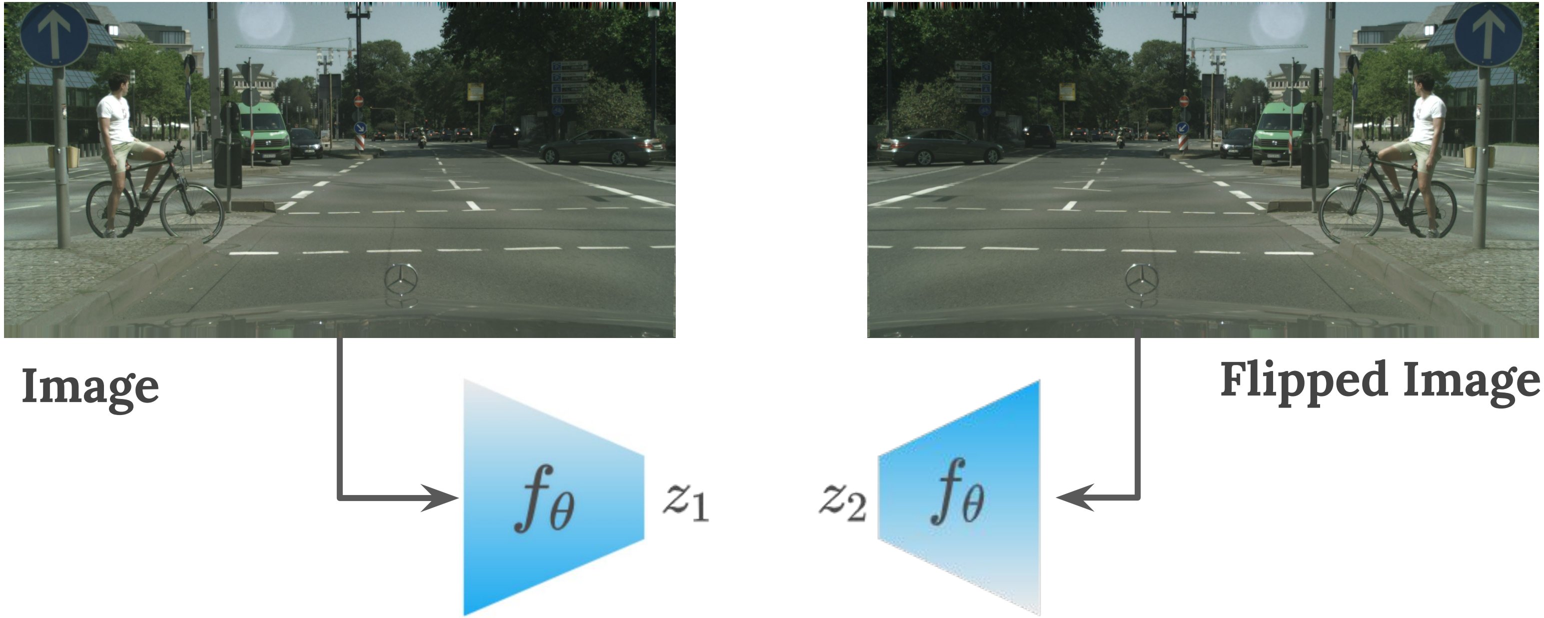
2) Target the position-specific channels

Identifying the Overall Position Encoding Channels



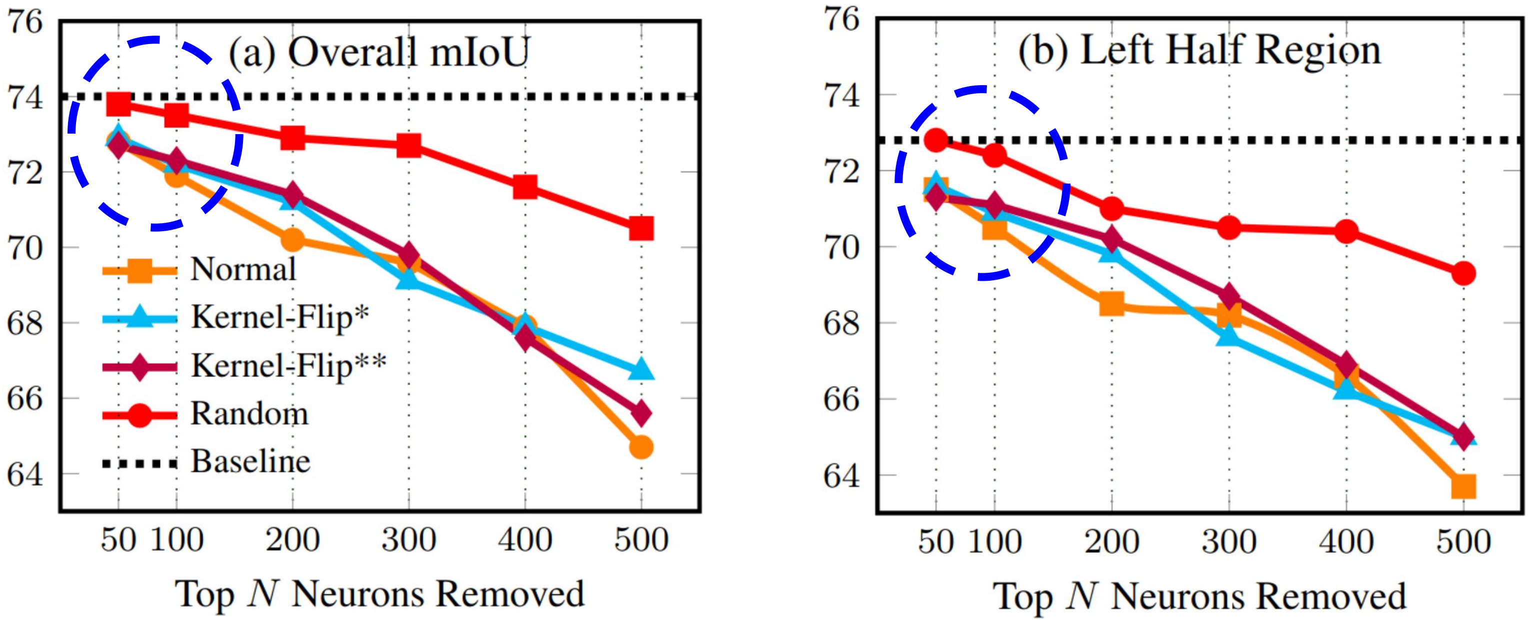
$$\hat{z} = \mathtt{argsort}_{j \in C}\left[\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} |\Delta z_i|\right]$$

Region-Specific Position Encoding Channels



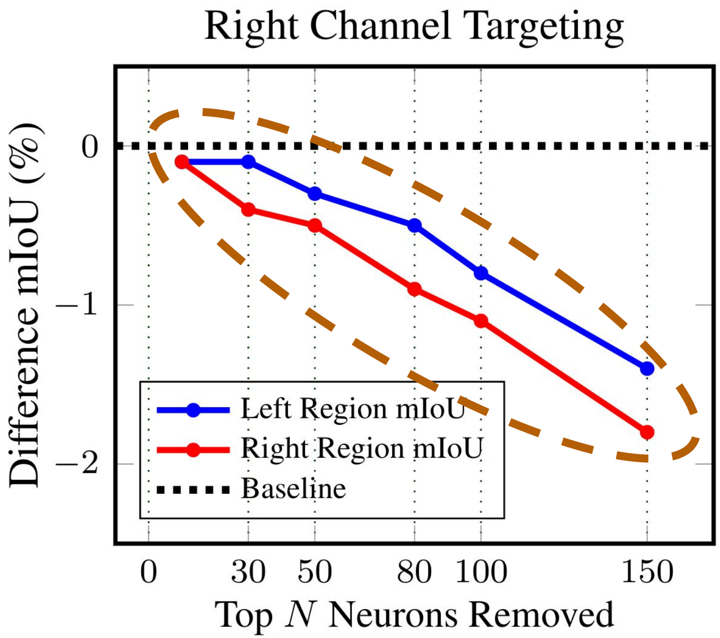
$$\hat{z}^l = \mathtt{argsort}_{j\in C}\left[\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\Delta z_i\right]$$

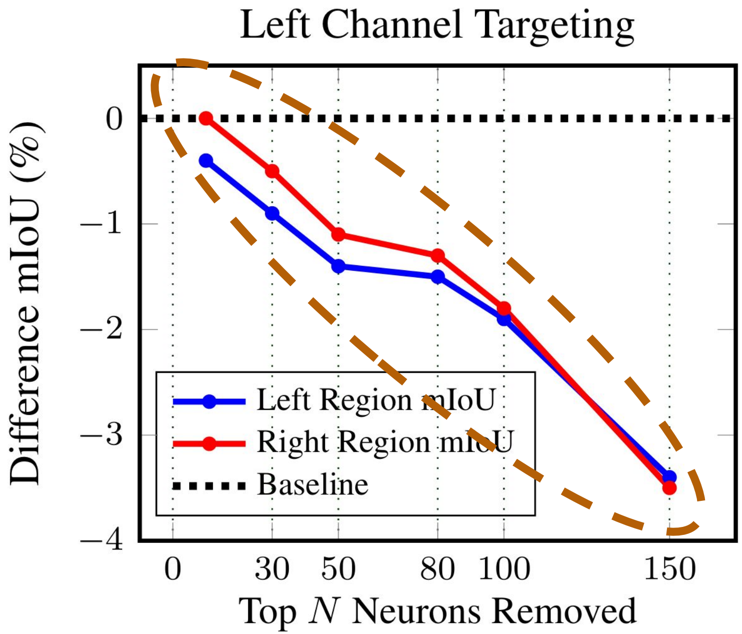
Targeting Position-specific Channels



Evaluated on Cityscapes pre-trained DeepLabv3-ResNet50 model

Targeting Region-specific Channels





Evaluated on Cityscapes pre-trained DeepLabv3-ResNet50 model

Take Away

- Position information is encoded based on the *ordering* of the channels while semantic information is largely not.

- Introduced a simple data augmentation strategy to improve translation invariance of CNNs.

- Introduced an intuitive technique to identify the position-specific neurons in a network’s latent representation.

Thanks for Listening

Code is available at: https://github.com/islamamirul/PermuteNet