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

Md Amirul Islam¹,６ Matthew Kowal²,６ Sen Jia⁴ Kosta Derpanis²,⁵,６ Neil Bruce³,６
Motivation

Published as a conference paper at ICLR 2021

How Much Position Information Do Convolutional Neural Networks Encode?

Md Amirul Islam 1, 2, 3, Sen Jair 4, Neil D. B. Bruce 1, 2
1Kensura University, Canada
2Vector Institute for Artificial Intelligence, Canada
amirul.islam@kensura.com, sen.jair@kensura.com, bruce@kensura.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, 2015; He et al. 2016) and detection (Redmon et al., 2016; He et al., 2016), face recognition (Huang et al., 2015), semantic segmentation.

CNNs encode absolute position information

Islam et. al (ICLR 2020)

Kayhan et. al (CVPR 2020)

Islam et. al (arXiv 2021)

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

Omar Senthil Kayhan
Computer Vision Lab
Delft University of Technology

Jan C. van Gennip
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 absolute spatial location by learning filters that respond strongly to particular absolute locations by exploiting image boundaries that surround receptive fields, thereby effectively shaping the feature maps at the output of the convolutional layers. We demonstrate that this is true for a wide variety of architectures and datasets, including object detection, image classification, and unsupervised learning. Our results provide a fresh perspective on how CNNs perceive the world and can be extended to other tasks such as image completion and object localization.

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

Md Amirul Islam, Matthew Kowai, Sen Jair, 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 show that hypothesis and reveal a surprising degree of absolute position information in the weights of CNNs, which can be exploited in various applications, such as image classification, object detection, and semantic segmentation, and in many real-world tasks that require precise localization. We demonstrate that the predictions of these findings on multiple real-world tasks are that position information is encoded in a meaningful manner.
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

Results are on CIFAR-10 dataset.

<table>
<thead>
<tr>
<th>Network</th>
<th>Loc. Classification</th>
<th>Image Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3x3</td>
<td>7x7</td>
</tr>
<tr>
<td>GAPNet</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>PermuteNet</td>
<td>78.8</td>
<td>21.4</td>
</tr>
</tbody>
</table>
Applicability of Channel-wise Positional Encoding

Learning Translation Invariant Representation

Attacking Position Encoding Channels
Learning Translation Invariant Representations
## Results: Translation Invariance

<table>
<thead>
<tr>
<th>Network</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1 Acc.</td>
<td>Consistency</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>93.1</td>
<td>90.8</td>
</tr>
<tr>
<td>Blurpool</td>
<td>92.5</td>
<td>92.5</td>
</tr>
<tr>
<td>AugShift (Ours)</td>
<td>92.1</td>
<td>94.8</td>
</tr>
</tbody>
</table>
## Results: Translation Invariance

<table>
<thead>
<tr>
<th>Network</th>
<th>CIFAR-10</th>
<th></th>
<th>CIFAR-100</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-1 Acc.</td>
<td>Consistency</td>
<td>Top-1 Acc.</td>
<td>Consistency</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>93.1</td>
<td>90.8</td>
<td>72.6</td>
<td>70.1</td>
</tr>
<tr>
<td>Blurpool</td>
<td>92.5</td>
<td>92.5</td>
<td>72.4</td>
<td>78.2</td>
</tr>
<tr>
<td>AugShift (Ours)</td>
<td>92.1</td>
<td>94.8</td>
<td>72.6</td>
<td>85.6</td>
</tr>
</tbody>
</table>
Attacking the Position Encoding Channels

1) Identify the position-specific channels

2) Target the position-specific channels
Identifying the Overall Position Encoding Channels

\[ \hat{z} = \text{argsort}_{j \in C} \left[ \frac{1}{|D|} \sum_{i=1}^{|D|} |\Delta z_i| \right] \]
Region-Specific Position Encoding Channels

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