Paper: *Spatial Transformer Networks*

Nirajan Koirala

CSC 9010_001

Summary

In the given paper, authors introduce Spatial Transformer module to tackle the limitations of CNNs like having only a limited, pre-defined pooling mechanism for dealing with variations in the spatial arrangement of data. This module can be included into a standard NN architecture to provide spatial transformation capabilities. Unlike the pooling layers, where the receptive fields are fixed and local, the spatial transformer module is a dynamic mechanism that can actively spatially transform an image. This allows networks to transform the selected regions of an image that are most relevant to a canonical, expected pose to simplify inference in the subsequent layers. Also, these modules can be trained with standard back-propagation algorithm allowing an end-to-end training of the models. They can be incorporated into CNNs and can be used for various tasks like image classification, co-localization and spatial attention. They can also help in computational efficiency as the outputs of these modules have a relatively lower resolution than the input.

Section 2 of the paper discusses some related work and covers the central ideas of modelling transformation with NN, learning and analyzing transformation-invariant representations and feature selection using detection mechanisms. Authors discuss the works where networks with selective attention manipulate the data by taking corps, trained with reinforcement learning and use a differential attention mechanism by utilizing Gaussian kernels in a generative model. They mention that the framework presented in their paper can be seen as a generalization of differentiable attention to any spatial transformation. They describe the formulation and implementation the spatial transformer in section 3. The main mechanism of the transformer is split into three part: localization network, grid generator and sampler. A richer description of the components is provided in the following sub-sections. One of the important feature of these transformer is that they allow gradients to be backpropagated through from sample points to the localization network output. Also, spatial transformer is self-contained module which can be dropped into a CNN architecture at any point and in any number, giving rise to spatial transformer networks. Moreover, this module is computationally fast and does not impair the training speed. They allow the network to learn how to actively transform the feature maps to help minimize the overall cost function of the network during training. It is also possible to use them to downsample or oversample a feature map. Placing multiple transformers at increasing depths of a network also allows transformations of increasingly abstract representations and using multiple transformers in parallel can help focus on multiple objects or parts of interest in a feature map. However, one limitation of using parallel transformers is that they limit the number of objects that the network can model.

In section 4, authors explore the use of transformer networks on distorted versions of MNIST handwriting dataset, Street View House Numbers dataset and CUB-200-2011 birds dataset. They classify the MNIST data that has been distorted in various ways and find that spatial transformer enabled network outperforms its counterpart base networks. Moreover, ST-CNN models consistently

perform better than its fully connected counterpart due to max-pooling layers. In the Street Viiew House Numbers data, spatial transformer models obtain state-of-art results on 64 x 64 images but on 128 x 128 images, ST-CNN achieves a lower error than previous state of art error. As these transformers are able to crop and rescale the parts of feature maps that correspond to the digit, they achieve high accuracy at very low cost of computational speed when compared to CNN. In the bird classification task, 4 x ST-CNN outperforms the baseline model by 1.8%. During visualization, it is also found that they learn to focus on head and body parts of birds. Additionally, these transformers downsample the image's pixel counts before processing so they don't impact performance significantly.

In conclusion, these transformers help CNN models as they can be simply dropped into the network without any changes in the architectural design and loss function. Also, the regressed transformation parameters from the spatial transformer are available as an output and could be used for subsequent tasks.

Strengths

- The points made in the paper are clear, concise and to the point.
- Authors have tested their approach on 3 different type of datasets.
- The methods are easy to follow and the good organization of sections helps in readability.
- The animations provided for their experiments further helps in getting the intuition of the method.

Weaknesses

- I think some of the captions provided in the figures were a little lengthy.

Points of Confusion

- The workings of grid generator section of the transformer is confusing.
- I am not sure how one can use a spatial transformer to downsample or oversample a feature map as described in the paper.

Discussion Questions

- How/Can one use these type of transformers in RNNs and how would it affect the networks overall performance and accuracy?
- How/Can we use the parameters generated by these transformers to perform additional computer vision tasks?
- A discussion on Section 3.3 which provides the Math and working behind these transformers would be helpful.