Jenish Maharjan **Week 7**

Paper: *Spatial Transformer Networks*

Summary:

Jaderberg et al, in their paper “Spatial Transformer Networks” address the limitations of convolutional neural networks (CNNs) in efficiently learning the invariances in the input feature maps. Although CNNs do learn invariances, they require a lot of training data and a high number of network parameters to do so. This limitation of CNNs is credited to the limited and pre-defined pooling mechanism that CNNs implement to deal with spatial variations. In this paper, the authors introduce a module that can be incorporated into a standard neural network architecture to deal with spatial variations. The module allows for a dynamic mechanism to transform the input feature map by producing an appropriate transformation for each input sample which can then be performed on the entire feature map. The module can be used for different transformations including scaling, cropping, rotations and non-rigid deformations. Another key feature of the module is that it can be trained with standard back-propagation which makes it easy to integrate with any standard neural network without affecting the end-to-end training.

The authors describe how they could benefit different tasks such as image classification, co-localization and spatial attention. Image classification could be simplified by using spatial transformers to crop out and scale-normalize appropriate regions of the image. Spatial transformers can be used to localize different instances of the same class in an image. In tasks that require an attention mechanism, spatial transformers can be used for spatial attention in a flexible way while using the standard backpropagation without reinforcement learning. While there have been previous works to deal with spatial variations like creating a generative model composed of transformed parts or the implementing the idea of symmetry groups, this work is different in that they aim to achieve spatial invariance by manipulating the data rather than the feature extractors.

The spatial transformer is described as “a differentiable module which applies a spatial transformation to a feature map during a single forward pass, where the transformation is conditioned on the particular input, producing a single output feature map.” The transformer module consists of three parts – a localization network that takes the input feature map which produces transformation parameters to be applied to the feature map, a grid generator that generates a sampling grid to which the input map should be sampled to produce the transformed output and a sampler that produces the output map sampled from the input at the grid points. The localization network function should include a regression layer to produce the transformation parameters. The sampling grid generated is based on the task at hand. An example of the grid presented in the paper with only 6 parameters allows for cropping, translation, rotation, scaling and skewing to be applied to the input. A grid used for attention would be different and more constrained. The combination of these three components is a self-contained module that can be injected into a CNN architecture at any point and in any number, which makes a spatial transformer network. This module is computationally fast and does not affect the training speed of the network. However, using multiple modules in parallel in a network can limit the objects that the network can model.

The authors have explored the use of spatial transformer networks on MNIST handwriting recognition, Street view house numbers recognition and fine-grained classification. In each of these experiments, they start with a baseline fully connected networks and CNNs and then extend those models with spatial transformer modules. For each of these tasks, the performance of the spatial transformer networks comparatively better and more efficient than the baseline models. The authors also point out that these modules could be powerful in recurrent models and useful for tasks that require disentangling object reference frames.

Strengths:

- The content of the paper is very well organized and the flow of the paper makes it easy to read.
- The diagrams help understand the functioning of the spatial transformer module.
- Each of the three experiments and the implementations are explained in detail.

Weaknesses:

- The last few lines in the introduction and related works section that explain what the following sections talk about seem unnecessary. The only line that seems helpful is about the supplementary material.
- I think they could have talked about parameterized the three components a little more in detail.

Confusions:

- Would we need to implement different modules for different tasks?
- What is a bilinear sampling kernel?

Discussion Questions:

- How would spatial transformer modules be useful with recurrent networks, and with non-image data?
- A quick discussion on the previous work by Hinton and the others mentioned in the related works sections would be helpful.
- Could spatial transformer modules be considered a replacement for pooling layers?