Paper: Spatial Transformer Networks

Summary

In this paper, the authors present a module that can be included in standard neural network architecture to provide spatial transformation capabilities. This paper aims at addressing the issue with CNN in being spatially invariant to the input data in a computationally and parameter efficient manner. Even though the introduction of max-pooling layers on CNN has helped the network to be somewhat spatially invariant to the position of features, the spatial support by the max-pooling is typically small. The action of the spatial transformer is conditioned on individual data samples, with the appropriate behavior learned during training. Unlike 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 by producing an appropriate transformation for each input image. The transformation is performed in the entire feature map and can include scaling, cropping, rotations, and non-rigid transformations. This allows the network to select regions of an image that are most relevant. It also helps transform those regions into a canonical, expected pose to simplify inference in the subsequent layers. One important feature of spatial transformers is they can be trained with standard backpropagation without any change in the loss function. This allows for end-to-end training of the neural network model. Other benefits of incorporating spatial transformers into CNN are image classification, co-localization, and spatial attention.

The spatial transformer has three parts: localization net, grid generator, and sampler. The localization network takes the input feature map U with width W, height H, and channels C and outputs Θ and the parameters of the transformation T_Θ to be applied to the feature map. Then the predicted transformation parameters are used to create a sampling grid, which is a set of points where the input map should be sampled to produce the transformed output. This is done by a grid generator. The authors use various transformation T_Θ in their experiment such as affine, projective, and thin-plate spline. They also mention that the transformation can have any parameterized form as long as it is differentiable with respect to the parameters which allow gradients to be backpropagated. Finally, to perform the spatial transformation of the input feature map, the sampler takes the set of sampling points $T_\Theta(G)$ along with the input feature map U and produce the sampled output feature map V.

Spatial Transformer is a self-contained module that can be dropped into a CNN architecture at any point and in any number. It is computationally very fast and does not impair training speed, causing very little time overhead. Placing spatial transformer within a CNN allows 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 a spatial transformer to downsample or oversample a feature map. The authors showed that the spatial transformer can also be used in parallel if there are multiple objects or parts of interest in a feature map that should be focused individually.

The first result they include is of the experiment they performed on the MNIST handwriting dataset as a testbed for exploring the range of transformation to which a network can learn invariance. They trained the FCN, CNN, as well as spatial transformer induced FCN and CNN. All spatial transformer used bilinear sampling but various transformation functions: affine (Aff), projective (Proj), thin-plate spline (TPS). They found out that spatial transformer enables network outperforms their counterpart base network. The thin-plate spline transformation is the

most powerful being able to reduce error on elastically deformed digits by reshaping the input into a prototype instance of the digit, reducing the complexity of the task for the classification network, and does not overfit on simpler data.

On the street view house number dataset (SVHN), the spatial transformer models obtain state-of-the-art results with only a single forward pass of a single model. This accuracy is achieved since the spatial transformers crop and rescale the parts of the feature maps that correspond to the digit, focusing resolution and network capacity only on these areas.

Finally, they tested the spatial transformer network with multiple transformers in parallel to perform fine-grained bird classification. They found out that the transforms predicted by a network have learned to detect the head and central part of the body of a bird. The resulting output is also somewhat posed normalized representation of a bird. Spatial transformers allow for the use of the high-resolution image without any impact on performance.

Strengths

- The visual representing of the network and the results are intuitive and hence, easy to understand.
- The authors have done a good job of having separate sections for each part of their network architecture and explaining them in detail.

Weaknesses

- They could have discussed a bit more about the different transformations they used in their module.

Confusions

- How does affine, projective, and thin-plate spline transformation work?
- How does a spatial transformer network with multiple transformers in parallel work?

Discussions

- How can one put a spatial transformer anywhere in the network and still be able to back-propagate without any change in the existing network?
- What exactly are Θ and T_Θ that is given by the localization net and how are they used by grid generator?
- What is a more sophisticated use of spatial transformer at present?