Indian Institute of Technology Bombay

Machine Learning for Remote Sensing - II
Paper Review - 2

Dynamic Convolutions: Exploiting Spatial Sparsity for Faster Inference

Utkarsh Ranjan: 200050147
Mahesh Bhupati: 200260027

1 Problem Statement

In many images, every pixels or region are not equally important and thus don't require equal computation to extract important features from them. In contrast to other modern convolution neural networks this paper suggests a novel method to dynamically apply convolutions on the input image thus reducing computational complexity without compressing existing architecture or making a tradeoff between performance and accuracy.

2 Model Architecture

The structure of the paper was quite relevant as it displayed the very need of the reducing computations by executing conditionally in the spatial domain and how it was unique from already existing designs. The method proposed in the paper was fairly easy and straightforward to integrate with existing architectures like the Residual Network (ResNets). Some salient features of the model were the following:

- The model introduced a small gating network(mask unit) which choosed the location over the image where the dynamic convolutions had to be applied.
- Gating decisions were trained end-to-end using the Gumbel-Softmax trick with a focus on efficiency.
- Those decisions progress throughout the network: the first stages extract features from complex regions in the image, while the last layers use higher-level information to focus on the region of interest only.

Furthermore, for efficient spatial inference, non-spatial operations were implemented in the standard way while the 3x3 convolutions were modified to operate on the intermediate tensor. CUDA implementation was also used to speed up the model.

3 Loss Function

In order to achieve reduction in computation a sparsity criterion was added to the task loss. Thus the final loss was given by:

$$L = L_{task} + \alpha(L_{sp,net} + L_{sp,lower} + L_{sp,upper})$$

The last two components were proposed for proper initialization of each block as it kept the percentage of executed operation between an upper and lower bound.

4 Experiments

- **CIFAR-10 and ImageNet:** When experiment was performed with ResNet-32 on the standard train/validation split of CIFAR-10 and was evaluated on different budget targets $\theta \in \{0.1, 0.2, ..., 0.9\}$, rougly 1-2% increase in Top1 accuracy was observed conpared to ResNet, SACT and ConvNet-AIG. A similar result was obtained with the ImageNet dataset.

- **Human Pose Estimation:** In such a task which is inherently spatially sparse, DynConv achieved a 60% increase in processing speed without any loss in accuracy. On MPII dataset the model always outperformed baseline models with the same depth and amount of computations. Author depicted the results in graph using mean Percentage of Correct Keypoints as a metric.

5 Critical Analysis

The novelty and approach mentioned in the paper can be extended to approach many problems like processing higher resolution images within reasonable time. As the models detects critical regions in the image very efficiently it can be used in many devices which are not computationally powerfull like mobile phones, surviellence cameras, driverless cars etc.