EECS 6322 Week 5, Paper 2
Attention Augmented Convolutional Networks

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a novel two-dimensional relative self-attention mechanism that combines with a convolution layer to achieve both parallel processing and long term memory. The model augments the convolutional operations with a self-attention mechanism and concatenates the feature maps produced by the convolutional layers with the feature maps produced by the self-attention layers. The architecture has then been tested on various datasets such as ImageNet for image classification and COCO for object detection, and has proved to outperform the current state of the art techniques in terms of accuracy.

Convolutional Neural Networks are an integral part of deep learning and are used extensively for applications such as image processing and object detection. It has several crucial properties that form an inductive bias while designing the models, such as locality and translational equivarience. However, since the CNNs operate over a limited receptive field of an image, it is unable to capture the global information from an entire image. This global information can help uncover new relations between the features which might not be dependent on the local features. Self-attention models, on the other hand, are known to uncover features through a long memory range and hence, are widely used in sequence models. The self-attention mechanism calculates the weights of the hidden layers and computes the key, value, and query pairs which capture the dependencies of inputs over a large range. This method does not increase the parameters of the model and is hence quite desirable.

The self-attention model is formed in the architecture using the following steps. First, a multihead attention is performed on a matrix obtained from the flattened input layer. The output from each head is then normalized using a softmax activation, and the output of all heads is then concatenated into a single matrix given by: $MHA(X) = Concat[O_1, \cdots, O_{Nh}]W^O$ *The self-attention mechanism is also modified to a relative self-attention model by independently adding relative width and relative height information. This makes the model translation equivariant, which proves to be efficient while dealing with images. These relative positional embeddings are shared across the heads, but not the layers of the model. The feature maps thus obtained from the convolutional and self-attention layers are then concatenated which can be given as:* $AAConv(X) = Concat[Conv(X), MHA(X)]$

The model formed using the above setup is thus translation equivalent and is able to operate on inputs from different spatial dimensions. It also decreases the number of parameters by replacing the $(3x3)$ *convolutions with the* $(1x1)$ *convolutions as compared to the fully convolutional models. Batch normalization is also applied to the augmented convolutionals to scale the contribution of the feature maps generated from the convolutional and self-attention layers. These augmented convolutions are then applied once per each residual block according to the memory constraints.*

The architecture was applied on baseline architectures such as ResNet and MnasNet, and was tested on several datasets such as CIFAR-100, ImageNet and COCO datasets. In all experiments, the convolutional feature maps were replaced with attentional feature maps for better comparison. On the CIFAR-100 dataset, a Wide-ResNet-28-10 architecture was used consisting of 3 sets of 4 residual blocks each with $(3x3)$ *convolutional layers, where a relative attention was added to the first convolution of all residual blocks. The attention-enhanced architecture outperformed all the other variations of the baseline model and achieved a better accuracy at a similar complexity and parameter cost.*

On the ImageNet dataset, attention augmentation was applied on the $(3x3)$ *convolutional layers on different ResNet models. Again, the architecture performed much better in terms of accuracy against the spatial and channel attention mechanisms. On the same ImageNet dataset, when the baseline architecture was replaced with the MnasNet architecture, the attention-augmented architectures performed better than their original baseline architectures with a small increase in the parameter sizes. Similarly, on the COCO dataset, a RetinaNet architecture was combined with the ResNet architecture and modified with the attention augmentations. This architecture also performed better as compared to its baselines and had a higher accuracy score.*

Thus, the experiments prove that the attention augmented models are more efficient than their original baselines on both image classification and object detection applications. Despite having fewer parameters in some applications, these models have a high accuracy and hence a better performance.

II. Strengths

The attention-augmented models combine the desirable factors of both the convolutional networks and the attention models. The proposed architecture conserves the benefits of parallel computations from the convolutional networks and long-range memory from the attention model. The model also *has fewer parameters as compared to a few of its baseline models, and is still able to outperform its original baseline counterparts.*

III. Weaknesses

The attention augmented models require large memory resources. The memory cost of the model is given by: $O(N_h(HW)^2)$ which reaches a very high value for large spatial dimensions. Hence, the models are required to restrict the augmented convolutions with self-attention according to the memory constraints of the system.

IV. Correctness

The claims and empirical methodology of the architecture proposed in the paper are correct based on both their theoretical grounding and comparative results with the original baselines architectures. The architecture has proven to have a better efficiency on several datasets of image classification and object detection when compared with their original baselines models. Hence, the attention-augmented models are seen to have a better generalizing power.

V. Clarity

The paper is well written in general with a clear formatting and adequate supplementary information such as graphs and comparative tables for the results. A step-wise explanation of the model is also given with mathematical proofs at each step to make the paper easy to follow. Overall, the paper has good clarity and is easy to read.

VI. Relation to prior work

Convolution neural networks have previously been used in the designs of many new architectures, which consists of convolutional operations over skip connections and spatial scales. They have been tested on image classification and segmentation tasks, along with object detection tasks on datasets such as CIFAR-10 and ImageNet. Attention models have been popularly used in sequence models, especially for applications such as language translations. This is because of their ability to capture long-range dependencies which helps the model generate better results. The combination of convolutional networks and self-attention models have mainly been used in Natural Language Processing applications, such as Machine Translations and Question Answering. A few models have also worked on video classification and object detection that consist of non-local residual blocks that use the self-attention mechanism in their convolutional structures.

In contrast to the previous works, the model proposed in the paper uses the self-attention mechanism along the entire architecture. The multi-head attention system jointly attends the feature and spatial subspaces, and the model achieves translational equivariance by extending the relative self-attention to two-dimensional inputs. In contrast to the previous models, the architecture also concatenates the feature maps from both the self-attention and convolutional mechanisms, instead of updating or enhancing the features maps obtained from the convolutional layers. This helps the architecture to control the ratio of the outputs from each layer, hence enabling the model to convert from a fully convolutional model to a fully attentional model, or a combination of both the models.

VII. Reproducibility

The implementation of the proposed architecture on baselines models such as RetinaNet has been provided by the authors on GitHub. The hyperparameters and other initializations used by the architectures are also given in the paper, along with the constraints for each architecture and dataset. Hence, the major results of the proposed work can be reproduced by the information and resources provided by the authors.

VIII. Additional comments

A comparative study of models including different ratios of attentional feature maps and convolutional feature maps in the architecture and their effect on the performance of the model, along with their pros and cons, would be interesting if included in the paper.