Attend To Convolution

2023 DATA302 Final Project : Team 6

Joungbin Lee
2021320334

Seungwon Ha
2020320161

Chaehyeon Kim
2020250028

Abstract—**While ViTs have shown promising results in many tasks, they require large amounts of training data and computation resources to achieve high accuracy. This is due to lack of inductive bias and many research have been conducted to introduce CNN, as the architecture forces to capture local spatial structure, and thus achieves the properties of shift, scale and distortion invariance. We have researched on a few convolution and attention combined models and found some interesting hypothesis that could enhance the performance and may become the new state of the art.**

1. Introduction

In the field of computer vision and deep learning, Convolutional Neural Networks(CNNs) have been very successful, and Transformer has recently performed well. Motivated by several studies that combine these two methods, we note the efficient combination of convolution and transformer. We conduct this study based on four assumptions:(1) We expect that convolution after attention has a receptive field for the entire image. (2) Convolution token embedding with stride will adapt the low-level feature information of ResNet and the high-level feature information of Transformer. (3) Convolution is more better for learning images than self-attention because of inductive bias. (4) The model we propose, ATC, will be a generalization of CvT and ResNet. We propose a model called Attendant to Convolution (ATC), which improves the baseline model called CvT based on four assumptions. In addition, we identify assumptions through various experiments and explore the need for components.

2. Related Works

2.1. Backbone Architectures

Residual Network ResNet[6] is a deep residual learning framework to address the degradation problem of substantially deeper networks than those used previously. Instead of hoping each few stacked layers directly fit a desired underlying mapping $\mathcal{H}(\mathbf{x})$, they explicitly let stacked nonlinear layers fit a residual mapping $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$, hypothesizing that it is easier to optimize the residual mapping with reference to the layer inputs than to optimize the original unreferenced mapping. The formulation of $\mathcal{F}(\mathbf{x}) + \mathbf{x}$ can be realized by FFN with "shortcut connections", skipping one or more layers. In ResNet, they simply perform identity mapping, and their outputs are added to the outputs of the stacked layers, while adding neither extra parameter nor computational complexity. ResNet is easier to optimize, and can gain accuracy from considerably increased depth.

Vision Transformer ViT[4] is simple yet scalable strategy that directly applies standard Transformers to image recognition, with the fewest possible modifications. In vision, attention is either applied in conjuction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. ViT show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. They split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer, which means they interpret an image as a sequence of patches. They train the model on image classification in supervised fashion. ViT attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

2.2. Transformers with Convolutions

Rank with Attention In [3], they interpreted the role of skip connections and MLPs in Transformer from a rank perspective. They proposes an new way to understand self-attention networks, showing that their output can be decomposed into a sum of smaller terms or paths each involving the operation of a sequence of attention heads across layers. Using path decomposition, they prove that self-attention possess a strong inductive bias towards token uniformity. Specifically, without skip connections or multi-layer perceptrons (MLPs), the output of pure attention converges doubly exponentially to a rank-1 matrix, in other words, loses rank doubly exponentially with depth. On the other hand, skip connections and MLPs stop the output from degeneration.

Transformer-convolutinoal hybrids Hybrids of Transformer and Convolutional are a recent innovation and in the past multiple models have proposed different approaches to constructing such hybrids as said in [1]. In CoAtNet[2], they show that while Transformers tend to have larger model capacity, their generalization can be worse than CNNs due to the lack of the right inductive bias. To effectively combine the strengths from both architectures, they present CoAtNets, a family of hybrid models built from 2 key insights: (1) depthwise convolution and self-attention can be naturally unified via simple relative attention; (2) vertically stacking convolution layers and attention layers in a principled way is surprisingly effective in improving generalization, capacity and efficiency. In [9], they recommend to use a standard, lightweight convolutional stem for ViT models as a more robust architectural choices compared to the original ViT model design, which dramatically increases optimization stability and improves peak performance while maintaining flops and runtime. They conjectured that the substandard optimizability issue of ViT models compared to CNN models lies with the patchify stem, which is implemented by a stride-p $p{\times}p$ convolution applied to the input image. In other words, they conjectured that this large-kernel plus large-stride convolution runs counter to typical design choices of convolutional layers in neural networks. By replacing the ViT stem by a small number of stacked stride-two 3×3 convolutions, they find that this small change of atypical design in early visual processing results in markedly different training behavior in terms of the sensitivity to optimization settings as well as the final model accuracy, while the vast majority of computation in the two ViT designs is identical. In [1], they introduce Astroformer, a method to learn from less amount of data. They propose using a hybrid transformer-convolutional architecture drawing much inspiration from the success of models like CoAtNet. Concretely, they use the transformer-convolutional hybrid with a new stack design for the network, a different way of creating a relative self-attention layer, and pair it with a careful selection of data augmentation and regularization techniques. In experiments, they also showed that the greater the proportion of convolution than the proportion of Transformers, the better it works.

2.3. CvT: Introducing Convolutions to Vision Transformers

In [8], they proposed a new architecture names Convolutional vision Transformer (CvT), that improves Vision Transformer (ViT) in performance and efficiency by introducing convolutions into ViT to yield the best of both designs. This was accomplished through two primary modifications: a hierarchy of Transformers containing a new convolutional token embedding, and a convolutional Transformer block leveraging a convolutional projection. These changes intoruduced desirable properties of convolutional nerual networks to the ViT architecture (i.e. shift, scale and distortion invariance) while maintaining the merits of Transformers (i.e. dynamic attention, global context, and better generalization). Introduced convolutional token embedding and convolutional projection, along with the multi-stage design of the network enabled by convolutions, make CvT architecture achieved good performance while maintaining computational efficiency. Also, due to the built-in local context structure introduced by convolutions, CvT no longer requires a positional embedding, giving it a potential advantage for adaption to a wide range of tasks requiring variable input resolution. Figure 1 of [8] shows the pipeline of the CvT architecture.

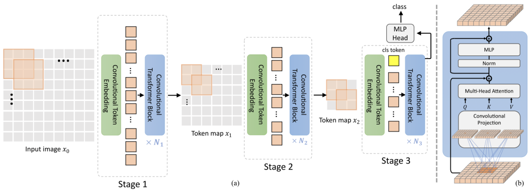

Figure 1. Figure from [7]. The pipeline of the CvT architecture. (a) Overall architecture showing the hierarchical multi-stage structure facilitated by the Convolutional Token Embedding layer. (b) Details of the Convolutional Transformer Block, which contains the convolution projection as the first layer.

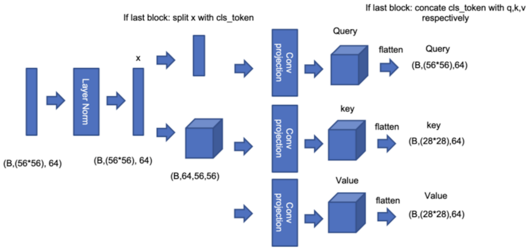

Figure 2. Overall method of ATC.

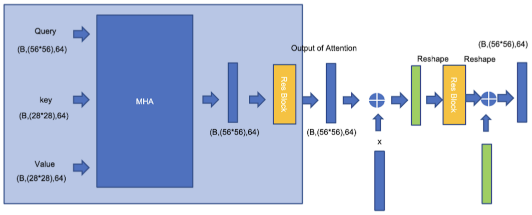

Figure 3. Overall architecture of ATC.

3. Method with Hypothesis

In this section, we would like to explain the method implemented based on the hypotheses we established. In section 3.1., we propose “Receptive Field of Entire Image” method, thinking about receptive field in perspective of input resolution, motivated by [2]. In section 3.2., we propose “Strided Convolutional Token Embedding” method that tried to adapt the low-level feature information of ResNet and the high-level feature information of Transformer. In section 3.3., we argue “Image Learning with Inductive Bias of Convolution” hypothesis. Lastly, in section 3.4., we propose “Generalization of CvT and ResNet”, saying that our model, ATC can be viewed as a generalized version of CvT[8] and ResNet[6].

3.1. Receptive Field of Entire Image

We assume that convolution after attention has receptive field of entire image resolution. In our model, ATC, multi-head self-attention and depth-wise separable convolution are consisting one block. However, in existing studies that combine convolution to the front and transformer to the back, the receptive field of former convolution is limited around convolution. We assume that convolution after attention has receptive field of entire image resolution. Accordingly, by connecting the convolution for the information obtained after applying attention, the convolution can have a receptive field corresponding to the entire image, thereby having the advantage of proceeding with convolution that can encompass both local and global level features. In our model, ATC, multi-head self-attention and depth-wise separable convolution are consisting one block.

3.2. Strided Convolutional Token Embedding

We assume that strided convolutional token embedding can match low-level information of ResNet and high-level information of Transformer. In other words, we expect that it can match “level of information”.

In CvT, they proposed a convolution operation, convolutional token embedding, that aims to model local spatial context, from low-level edges to higher order semantic primitives, over a multi-stage hierarchy approach, similar to CNNs. The convolutional token embedding layer allows us to adjust the token feature dimension and the number of tokens at each stage by adjusting paramters and stride of the convolution operation. In this manner, in each stage we progressively decrease the token sequence length, while increasing the token feature dimension. This gives the tokens the ability to represent increasingly complex visual patterns over increasingly larger spatial footprints, similar to feature layers of CNNs.

Based on these characteristics, we can further interpret that by applying convolutional token embedding with stride in the ATC structure that combines convolution and self-attention, it is possible to extract the high level feature values of key and value required for attachment from the low level feature.

3.3. Image Learning with Inductive Bias of Convolution

As stated in [5], the success of deep learning has largely been fueled by models with strong inductive biases, allowing efficient training across domains. The use of CNNs epitomizes this trend. Inductive biases are hard-coded into the architectural structure of CNNs in the form of two strong constraints on the weights: locality and weight sharing. By encouraing translation equivariance and translation invariance, the convolutional inductive bias makes models more sample-efficient and parameter-efficient. However, the rise of models based purely on attention in recent years calls into question the necessity of hard-coded inductive biases. Anyway, thanks to inductive bias, convolution is relatively easy to learn images compared to self-attention.

3.4. Generalization of CvT and ResNet

The Transformer block which consists of the convolutional projection layer, proposed by CvT, is a generalization of the original Transformer block design. By applying 1× 1 convolution to convolution embedding of CvT, ViT model can be simply implemented. In additionally, The Transformer block proposed by ATC removes all linear projections of CvT model and replace them with depth-wise separable residual blocks. Also, by applying 1× 1 convolution to all convolutional residual blocks, CvT model can be simply implemented. If the output of MHSA is zero, by making learned weights of MHSA into zero, it is equal to identity mapping by skip connection. Since only residual block is learned, it can be viewed as a generalization of ResNet. In conclusion, we can argue that ATC model is a generalization of CvT and ResNet.

model	Stage 1	Stage 2	Stage 3
ATC(Ours) 3-32x32	1	1	1
ATC(Ours) 5-32x32	1	1	3
ATC(Ours) 5-32x32	1	3	1
ATC(Ours) 5-32x32	3	1	1

TABLE 1. The number of blocks in ATC

method	Before	After
Conv Token Embed	2D Conv	ResBlock
Linear Porj in MHA	Linear	ResBlock
FFN	Linear	ResBlock

TABLE 2. Ablation Study on Methods

Model	Param(M)	top-1(%)	top-5(%)
CvT 13-32x32	20.0	73.934	92.179
ATC(Ours) 3-32x32	2.9	71.440	91.480
ATC(Ours) 9-32x32	12.0	77.440	93.920
ATC(Ours) 13-32x32	19.2	79.350	94.340
ATC(Ours) 15-32x32	19.7	80.320	94.500

TABLE 3. Results of Experiment 1

Model	Param(M)	top-1(%)	top-5(%)
ATC(Ours) 3-32x32	2.9	71.440	91.480
ATC(Ours) 5-32x32	6.4	75.360	93.400
ATC(Ours) 5-32x32	3.3	73.800	92.860
ATC(Ours) 5-32x32	2.9	72.780	92.470

TABLE 4. Results of Experiment 2

Model	Param(M)	top-1(%)	top-5(%)
ATC(Ours) 3-32x32	2.9	71.440	91.480
ATC(Res) 3-32x32	3.3	74.740	93.680
ATC(Ours) 15-32x32	19.7	80.320	94.500
ATC(Res) 15-32x32	20.1	80.370	94.840

TABLE 5. Results of Experiment 3

Model	Param(M)	top-1(%)	top-5(%)
CvT	5.8	61.2	85.24
ATC(Ours) : Res+MLP	3.9	65.850	88.160
ATC(Ours) : MLP+Res	2.9	67.510	88.720
ATC(Ours) : Res+Res	2.9	71.440	91.480
ATC(Ours) : Res*+Res*	3.3	74.740	93.680

TABLE 6. RESULTS OF EXPERIMENT 4

4. Experiments & Analysis

We conducted a total of four experiments for performance evaluation, and the results of the experiments are shown in Table 3 6. We used CIFAR100 as our dataset. The number of blocks used in ATC is shown in Table 1. The options related with ablation study are shown in Table 2.

Experiment 1 is Comparison with CvT on CIFAR100. We directly compare performance of our model, ATC, and our baseline model CvT on CIFAR100. Results are shown in Table 3. Experiment 2 is Study of number of blocks in ATC. We would like to know whether which stages are more important. So we differentiate the number of blocks in each stages and compare the performance. Experiment 3 is Replace Convolutional Token Embedding with Residual Block with Downsampling. Experiment 4 is Ablation Study. We explored which modules that we newly introduced in our model have higher importance.

We compared the number of parameters and top-1 and top-5 accuracy of CvT and a total of four ATC models that applied MLP and residual block to two places. The results are shown in Table 6. In Table 6, ”Res*” refers to the simple residual convolution, and ”Res” refers to the residual convolution that uses depth-wise separable convolution. Among the five models, CvT(corresponding to the first row) requires the most parameters. Compared to CvT, the ATC model, which requires about half of the parameters, is efficient in terms of parameters. In addition, when compared with only accuracies, the ATC model of Res*+Res* shows the best top-1 and top-5 accuracies. Overall, even though the ATC model requires fewer parameters than the CvT, the performance tends to be better. It is also possible to compare more specifically with numbers. Comparing the CvT of first row and ATC with Res+MLP of second row, the number of parameters was reduced from 5.8M to 3.9M, and the top-1 accuracy improved by 4.65%p, and the top-5 accuracy by 2.92%p. Also, even when comparing the CvT and ATC with MLP+Res of thired row, the number of parameters decreased from 5.8M to 2.9M, and the top-1 accuracy improved by 6.31%p, and the top-5 accuracy iby 3.48%p. In conclusion, CvT uses linear operation, while ATC uses convolution operation to improve performance. ATC replaces linear operation with depth-wise separable convolution, which can be interpreted as achieving good performance by efficiently reducing the number of parameters while recognizing the image more evolutionary.

5. Conclusion

We presented a model of a new structure called Attend to Convolution and analysis through various experiments. Our conclusions are as follows.

- It can be understood that the convolution after attention has a receptive field for the entire image.
- Convolution token embedding with stride can adapt the low-level feature information of ResNet and the high-level feature information of Transformer.
- Thanks to inductive bias, convolution is more advantageous for learning images than self-attention.
- Our model, ATC, can be regarded as a generalization of CvT and ResNet.

In the future, research that combines convolution with transformers should be conducted more actively.

References

[1] Rishit Dagli. Astroformer: More data might not be all you need for classification. *arXiv preprint arXiv:2304.05350*, 2023. 1, 2

[2] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 1, 2

[3] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[5] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[7] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[8] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 2

[9] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021. 2