

Attention Augmented Convolution

Shishir Roy¹ Meghal²

¹M.Tech A.I.

Indian Institute of Science, Bangalore

²M.Tech A.I.

Indian Institute of Science, Bangalore

Advanced Image Processing, May 2024



Motivation

- Convolutional Networks has been the go-to model architecture for image classification tasks, but it has local receptive field as it only operates on the local neighbourhood.
- To deal with this problem one has to either use a larger kernel size or make the network deeper or both which increases the model complexity.
- Self-Attention, on the other hand, has emerged as the choice of model architecture to capture long range interactions.
- Bello et al. [2020] has implemented Attention Augmented Convolutional Networks which combines both Convolutional Networks and Self-Attention.

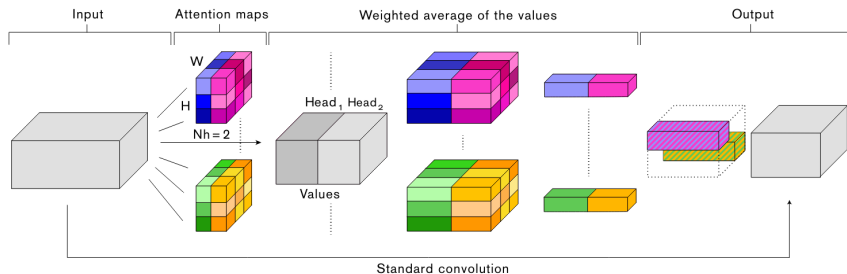


Objective

- Bello et al. [2020] has used Vaswani et al. [2023]'s implementation of self attention (sdpa) which is quadratic in time and space complexities.
- Over the years many sub-quadratic attention mechanisms have been proposed, we took Shen et al. [2024]'s implementation which is linear in time and space complexities.
- We intended to see how it affects the time and accuracy of the network.



What is Attention Augmented Convolutional Network



Attention Augmented Convolutional Network (AA-Conv)

- For Vaswani et al. [2023]'s implementation given input $X \in \mathbb{R}^{n \times m}$, we have 3 weight matrices $W_q \in \mathbb{R}^{m \times d_k}$, $W_k \in \mathbb{R}^{m \times d_k}$, $W_v \in \mathbb{R}^{m \times d_v}$ that transforms the input as $Q = XW_q$, $K = XW_k$, $V = XW_v$ and the attention output as $O_h = \rho(\frac{QK^T}{\sqrt{d_k}})V$.
- We have input image to the AA-Conv with dimensions $F_{in} \times h \times w$, this is reshaped to dimensions $(hw) \times F_{in}$ and this is feed to the Attention Mechanism as input. Thus making the time and space complexities to be $O((hw)^2(d_k + d_v))$.



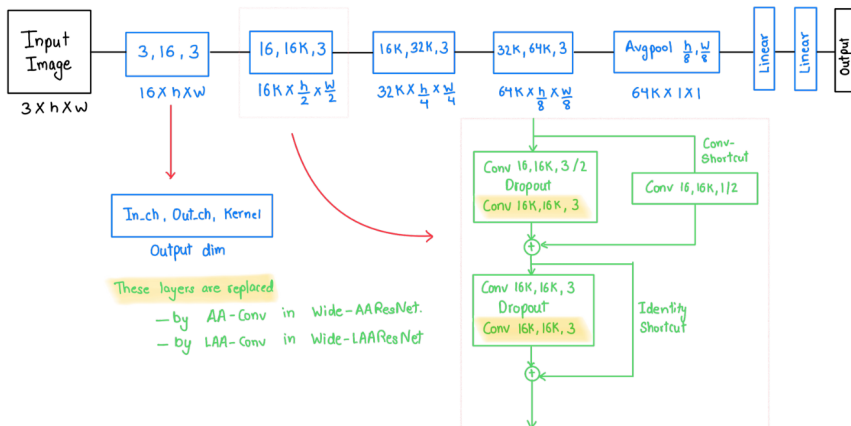
Linear Attention Augmented Convolutional Network (LAA-Conv)

- For Shen et al. [2024]'s implementation given input $X \in \mathbb{R}^{n \times m}$, we have 3 weight matrices $W_q \in \mathbb{R}^{m \times d_k}$, $W_k \in \mathbb{R}^{m \times d_k}$, $W_v \in \mathbb{R}^{m \times d_v}$ that transforms the input as $Q = XW_q$, $K = XW_k$, $V = XW_v$ and the attention output as $O_h = \rho_{col}(Q)(\rho_{row}(K))^T V$.
- We have input image to the LAA-Conv with dimensions $F_{in} \times h \times w$, this is reshaped to dimensions $(hw) \times F_{in}$ and this is feed to the Attention Mechanism as input. Thus making the time and space complexities to be $O(d_k d_v(hw))$.



Experiments

- We implemented 3 models viz Wide-ResNet, Wide-AAResNet and Wide-LAAResNet each for the two datasets viz CIFAR-100 and Tiny-Imagenet.

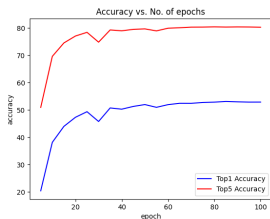
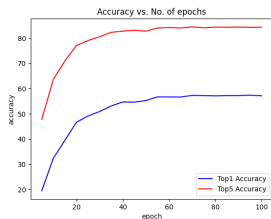
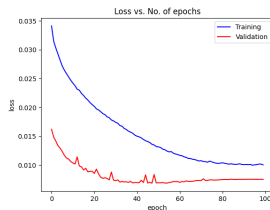
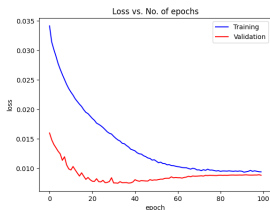
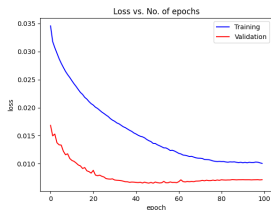


Experiments

- In each block of Wide-ResNet, we changed one conv layer to AA-conv layer (LAA-conv layer) to get Wide-AAResNet (Wide-LAAResNet).
- Bello et al. [2020] used learnt positional embeddings stating that they got bad results using sine-cosine positional embeddings used by Vaswani et al. [2023]. But in our experiments the later gave better results so we stuck with that.



Experimental Results - CIFAR-100



Wide-ResNet

57.11, 84.33

Wide-AAResNet

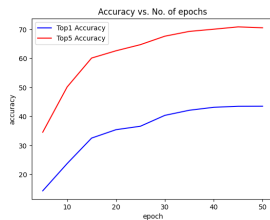
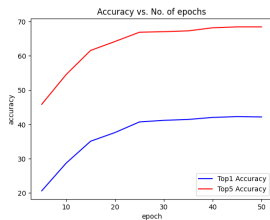
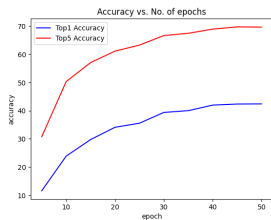
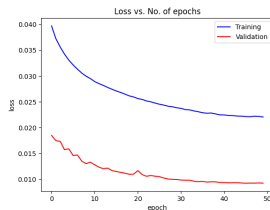
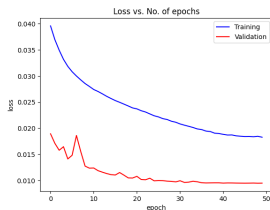
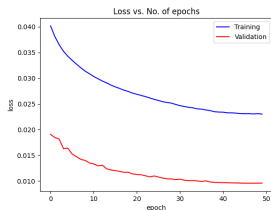
52.91, 80.31

Wide-LAAResNet

55.32, 82.95



Experimental Results - Tiny-Imagenet



Wide-ResNet

42.39, 69.61

Wide-AAResNet

42.18, 68.44

Wide-LAAResNet

43.51, 70.52



Conclusion

- Thank You!
- You can find all the code and results of the experiments
<https://github.com/yoR-rihsihS/aip-andromeda>.



References

- I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le. Attention augmented convolutional networks, 2020.
- Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li. Efficient attention: Attention with linear complexities, 2024.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

