Attention as Activation

Yimian Dai[1] **Stefan Oehmcke**[2] **Fabian Gieseke**[2,3]
Yiquan Wu[1] **Kobus Barnard**[4]

[1] Nanjing University of Aeronautics and Astronautics
[2] University of Copenhagen
[3] University of Münster
[4] University of Arizona

December 10, 2020

Background

Attention Models Achieve SOTA Performance in Many Tasks:

Task	SOTA Attention Model
Image Classification	SENet [1], ViT-H/14 [2]
Semantic Segmentation	ResNeSt [3]
Image Generation	Image Transformer [4]
Medical Image Segmentation	PraNet [5]
Machine Translation	Transformer+BT [6]
Language Modelling	Transformer-XL [7]
Question Answering	LUKE [8]

Background

It raises a natural question:

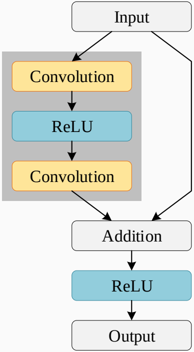
- The more attention modules, the better the performance?

If yes, then

- How to add more attention modules, after all SENet has already implemented attention modules in every block?

Background

Disassemble a Residual Block:



1. Conv => Deformable Kernels [9]
2. ReLU => **Attentional Activation**
    - Motivated by the Similarity between Activation and Attention

Motivation

Observation: Unification of Attention and Activation

1. Attention Mechanism Can Be Written As

$$\mathsf{X}' = \mathsf{G}(\mathsf{X}) \odot \mathsf{X}, \tag{1}$$

2. The Scalar Form of Eq. (1) Can Be Expressed As

$$\mathsf{X}'_{[c,i,j]} = \mathsf{G}(\mathsf{X})_{[c,i,j]} \cdot \mathsf{X}_{[c,i,j]} = g_{c,i,j}(\mathsf{X}) \cdot \mathsf{X}_{[c,i,j]}. \tag{2}$$

3. Activation Function Can Also Be Expressed in a Similar Form

$$\mathsf{X}'_{[c,i,j]} = g'\left(\mathsf{X}_{[c,i,j]}\right) \cdot \mathsf{X}_{[c,i,j]}. \tag{3}$$

Motivation

Observation: Unification of Attention and Activation

1. Both can be expressed as a nonlinear adaptive gating function
2. **Difference**: The gating function input in activation is a scalar, while in attention is the entire feature map
3. **A Unified Perspective**:
   - Attention Mechanism: A Context-Aware Activation Unit
   - Activation Unit: An Extremely Simplified Attention Module
   - Examples:
     - ReLU: Indicator Function
     - Swish [10]: Sigmoid Function
     - SIREN [11]: Sinc Function

Motivation

Using Lightweight Attention Modules as Activation Units:

1. The Basic Function of Introducing Nonlinearity into Networks
2. Dynamic, Context-Aware Feature Refinement Layer by Layer

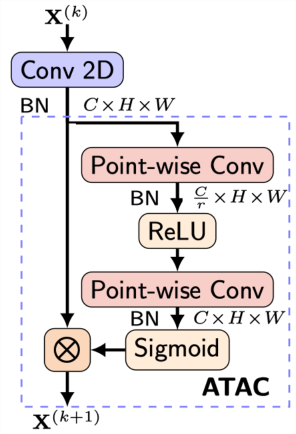
Formulation

A Bottleneck of Point-wise Conv:

$$X' = G(X) \odot X,$$

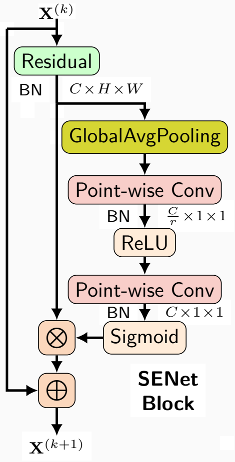
A Parameterless Version – Swish

$$x' = x \cdot \sigma(x)$$



Revisiting Channel Attention in SENet:

1. Question: Can Channel Attention Only Be Global?
2. Argument: Spatial Pooling Size Is the Scale of Channel Attention
3. Perspective: SENet Adopts an Extreme Coarse (**Global**) Scale Biased to Large Objects
4. Our Hypothesis: **Locality** Is Important for Activation Units



Discussion

Table 1: Difference between Attention Mechanism in SENet and ATAC

Difference	SENet	ATAC
Architecture	GAP + Fully Connected	Point-wise Conv
Attention Weight	Shared by a Feature Map	Element-wise
Context Scale	Global	Local / Point-wise
Usage	Block-wise Refinement	Layer-wise Activation

Fully Attentional Model

With ATAC Units, We Can Construct a Fully Attentional Model By

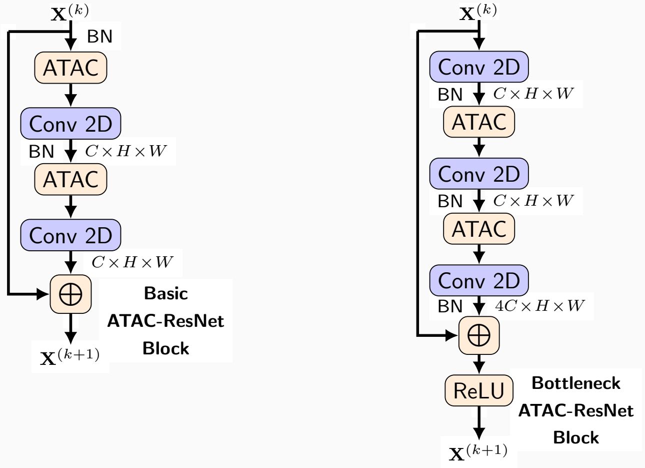
- Replacing ReLUs with ATAC Units

Hypothesis of a Fully Attentional Model:

1. Refining Features at Very Early Stages, Even after the First Convolutional Layer
2. Enable Networks to Encode Higher-Level Semantics More Efficiently.

Fully Attentional Model

Examples:

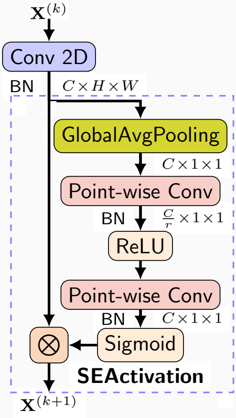


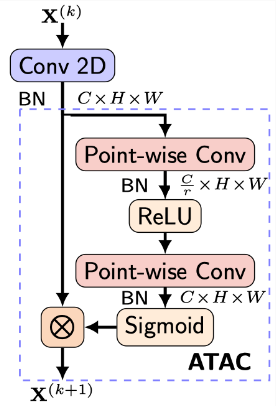
Experiments

Experiment outline

- Ablation Study
  1. Is **Locality** Critical for Attentional Activation?
  2. Choice of Micro Structure: NiN, SENet, or ATAC?
  3. Verification of the Efficiency of the Fully Attentional Network
- Comparison to State-of-the-Art

Architectures for Ablation Study on Importance of Locality





The Same #Params, Only Different in Context Aggregation **Scale**

Ablation Study – Importance of Locality

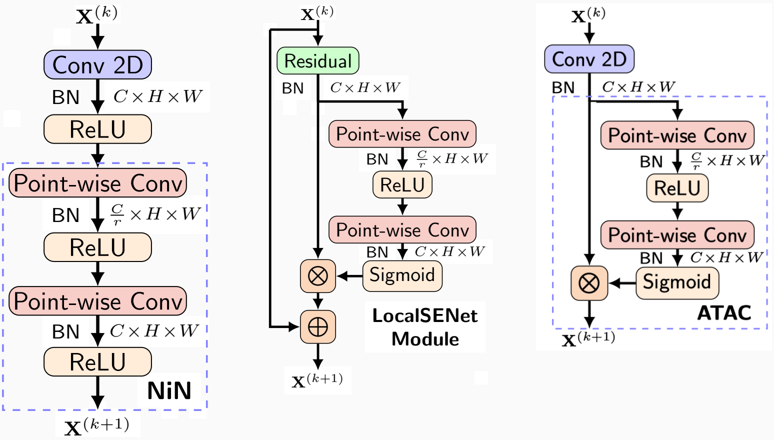
Table 2: Validation on the Importance of Contextual Aggregation Scale

Activation	CIFAR-10				CIFAR-100			
	$b = 1$	$b = 2$	$b = 3$	$b = 4$	$b = 1$	$b = 2$	$b = 3$	$b = 4$
ReLU	0.895	0.920	0.929	0.935	0.737	0.785	0.799	0.806
SEActivation	0.548	0.601	0.613	0.622	0.388	0.432	0.452	0.456
ATAC (***ours***)	**0.906**	**0.927**	**0.936**	**0.939**	**0.764**	**0.796**	**0.812**	**0.821**

Locality Is Critical for Attentional Activation.

Ablation Study – Choice of Micro Structure

Architectures for Ablation Study on Choice of Micro Structure

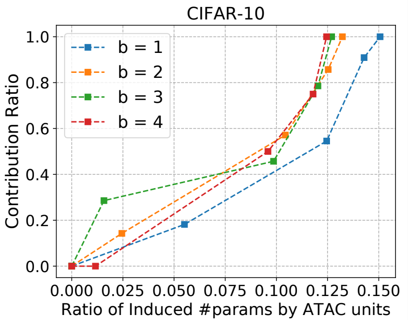


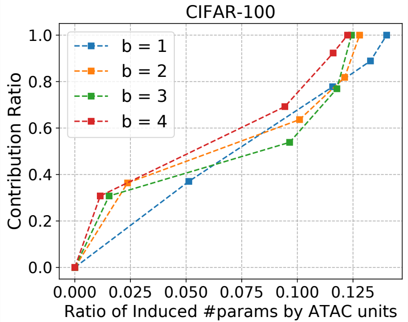
Ablation Study – Choice of Micro Structure

Table 3: Validation on the Choice of Micro Structure

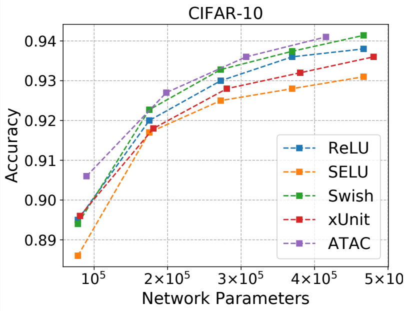
Activation	CIFAR-10				CIFAR-100			
	$b=1$	$b=2$	$b=3$	$b=4$	$b=1$	$b=2$	$b=3$	$b=4$
NiN	0.893	0.917	0.922	0.926	0.743	0.776	0.792	0.796
LocalSENet	0.906	0.926	0.931	0.937	0.762	0.794	0.805	0.811
ATAC (***ours***)	**0.906**	**0.927**	**0.936**	**0.939**	**0.764**	**0.796**	**0.812**	**0.821**

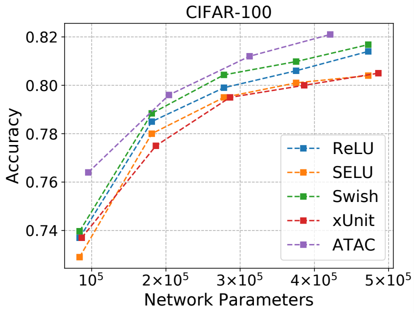
Ablation Study – Efficiency of Fully Attentional Networks



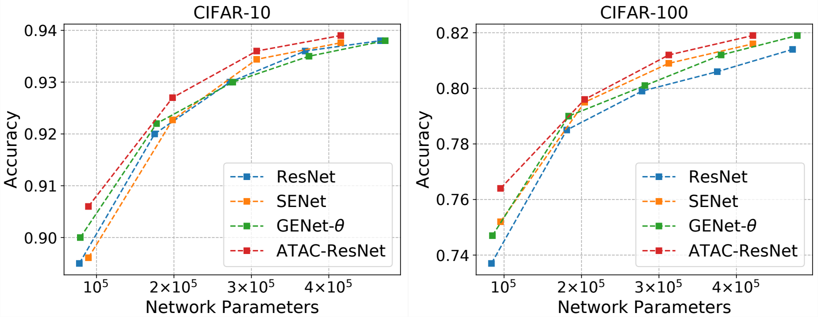


Comparison to State-of-the-Art Activation Units





Comparison to State-of-the-Art Networks



Comparison to State-of-the-Art Networks

Table 4: Comparison on ImageNet

Architecture	GFlops	Params	top-1 err.	top-5 err.
ResNet-50 [12]	3.86	25.6M	23.30	6.55
SE-ResNet-50 [1]	3.87	28.1M	22.12	5.99
AA-ResNet-50 [13]	8.3	25.8M	22.30	6.20
FA-ResNet-50 [14]	7.2	18.0M	22.40	/
GE-θ^+-ResNet-50 [15]	3.87	33.7M	21.88	5.80
ATAC-ResNet-50 (***ours***)	4.4	28.0M	**21.41**	6.02

Conclusion

1. A Unified Perspective for Attention and Activation
2. An Instance of Attentional Activation (ATAC) Unit
3. A Way to Fully Attentional Networks

Codes and Trained Models

https://github.com/YimianDai/open-atac

Thank you!

[1] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," CoRR, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[3] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-Attention Networks," arXiv e-prints, p. arXiv:2004.08955, Apr. 2020.

[4] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 4052–4061. [Online]. Available: http://proceedings.mlr.press/v80/parmar18a.html

[5] D. Fan, G. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI, ser. Lecture Notes in Computer Science, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., vol. 12266. Springer, 2020, pp. 263–273. [Online]. Available: https://doi.org/10.1007/978-3-030-59725-2_26

[6] X. Liu, K. Duh, L. Liu, and J. Gao, "Very deep transformers for neural machine translation," CoRR, vol. abs/2008.07772, 2020. [Online]. Available: https://arxiv.org/abs/2008.07772

[7] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, A. Korhonen, D. R. Traum, and L. Màrquez, Eds. Association for Computational Linguistics, 2019, pp. 2978–2988. [Online]. Available: https://doi.org/10.18653/v1/p19-1285

[8] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: deep contextualized entity representations with entity-aware self-attention," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 6442–6454. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.523/

[9] H. Gao, X. Zhu, S. Lin, and J. Dai, "Deformable kernels: Adapting effective receptive fields for object deformation," in 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020, pp. 1–12.

[10] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," in 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 2018, pp. 1–13.

[11] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in ArXiv Preprint, vol. abs/2006.09661, 2020, pp. 1–11.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[13] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in 2019 IEEE International Conference on Computer Vision (ICCV), October 2019, pp. 3286–3295.

[14] N. Parmar, P. Ramachandran, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in Annual Conference on Neural Information Processing Systems (NeurIPS) 2019, Vancouver, BC, Canada, 2019, pp. 68–80.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in Annual Conference on Neural Information Processing Systems (NeurIPS), 2018, pp. 9423–9433.