Attention as Activation

Yimian Dai\textsuperscript{1}  Stefan Oehmcke\textsuperscript{2}  Fabian Gieseke\textsuperscript{2,3}
Yiquan Wu\textsuperscript{1}  Kobus Barnard\textsuperscript{4}

\textsuperscript{1} Nanjing University of Aeronautics and Astronautics
\textsuperscript{2} University of Copenhagen
\textsuperscript{3} University of Münster
\textsuperscript{4} University of Arizona

December 10, 2020
### Attention Models Achieve SOTA Performance in Many Tasks:

<table>
<thead>
<tr>
<th>Task</th>
<th>SOTA Attention Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Classification</td>
<td>SENet [1], ViT-H/14 [2]</td>
</tr>
<tr>
<td>Semantic Segmentation</td>
<td>ResNeSt [3]</td>
</tr>
<tr>
<td>Image Generation</td>
<td>Image Transformer [4]</td>
</tr>
<tr>
<td>Medical Image Segmentation</td>
<td>PraNet [5]</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>Transformer+BT [6]</td>
</tr>
<tr>
<td>Language Modelling</td>
<td>Transformer-XL [7]</td>
</tr>
<tr>
<td>Question Answering</td>
<td>LUKE [8]</td>
</tr>
</tbody>
</table>
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?
Disassemble a Residual Block:

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

1. Attention Mechanism Can Be Written As

\[ X' = G(X) \odot X, \]  

(1)

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

\[ X'_{[c,i,j]} = G(X)_{[c,i,j]} \cdot X_{[c,i,j]} = g_{c,i,j}(X) \cdot X_{[c,i,j]}. \]  

(2)

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

\[ X'_{[c,i,j]} = g'(X_{[c,i,j]}) \cdot X_{[c,i,j]}. \]  

(3)
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
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
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
**Table 1:** Difference between Attention Mechanism in SENet and ATAC

<table>
<thead>
<tr>
<th>Difference</th>
<th>SENet</th>
<th>ATAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>GAP + Fully Connected</td>
<td>Point-wise Conv</td>
</tr>
<tr>
<td>Attention Weight</td>
<td>Shared by a Feature Map</td>
<td>Element-wise</td>
</tr>
<tr>
<td>Context Scale</td>
<td>Global</td>
<td>Local / Point-wise</td>
</tr>
<tr>
<td>Usage</td>
<td>Block-wise Refinement</td>
<td>Layer-wise Activation</td>
</tr>
</tbody>
</table>
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:

\[ X^{(k)} \]

\[ X^{(k+1)} \]

Basic ATAC-ResNet Block

\[ C \times H \times W \]

ATAC Conv 2D

\[ C \times H \times W \]

\[ + \]

\[ \text{BN} \]

Conv 2D

\[ C \times H \times W \]

Conv 2D

\[ \text{BN} \]

BN

\[ C \times H \times W \]

Conv 2D

\[ 4C \times H \times W \]

BN

\[ \text{BN} \]

Conv 2D

\[ \text{BN} \]

Conv 2D

\[ \text{BN} \]

Conv 2D

\[ \text{ReLU} \]

Bottleneck ATAC-ResNet Block

\[ X^{(k)} \]

\[ X^{(k+1)} \]
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
Ablation Study – Importance of Locality

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

<table>
<thead>
<tr>
<th>Activation</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b = 1$</td>
<td>$b = 2$</td>
</tr>
<tr>
<td>ReLU</td>
<td>0.895</td>
<td>0.920</td>
</tr>
<tr>
<td>SEActivation</td>
<td>0.548</td>
<td>0.601</td>
</tr>
<tr>
<td>ATAC (ours)</td>
<td>0.906</td>
<td>0.927</td>
</tr>
</tbody>
</table>

**Locality** is Critical for Attentional Activation.
Ablation Study – Choice of Micro Structure

Architectures for Ablation Study on Choice of Micro Structure

1. Conv 2D
   - BN $C \times H \times W$
   - ReLU
   - Point-wise Conv
     - BN $C \times H \times W$
     - ReLU
     - Point-wise Conv
       - BN $C \times H \times W$
       - ReLU
       - NiN
   - $X^{(k+1)}$

2. Residual
   - BN $C \times H \times W$
   - Point-wise Conv
     - BN $C \times H \times W$
     - ReLU
     - Point-wise Conv
       - BN $C \times H \times W$
       - Sigmoid
       - LocalSENet Module
   - $X^{(k+1)}$

3. Conv 2D
   - BN $C \times H \times W$
   - Point-wise Conv
     - BN $C \times H \times W$
     - ReLU
     - Point-wise Conv
       - BN $C \times H \times W$
       - Sigmoid
       - ATAC
   - $X^{(k+1)}$
### Table 3: Validation on the Choice of Micro Structure

<table>
<thead>
<tr>
<th>Activation</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b = 1$</td>
<td>$b = 2$</td>
</tr>
<tr>
<td>NiN</td>
<td>0.893</td>
<td>0.917</td>
</tr>
<tr>
<td>LocalSENet</td>
<td>0.906</td>
<td>0.926</td>
</tr>
<tr>
<td>ATAC (ours)</td>
<td><strong>0.906</strong></td>
<td><strong>0.927</strong></td>
</tr>
</tbody>
</table>
The results suggest that given the same budget for parameters and computation costs, one should resort to the cost-effective ratio of the fully attentional network. The network depth, refining the feature maps is a more efficient way to increase the networks' performance. We also investigated the cost-effective ratio of the fully attentional network with all ReLUs being applied after every convolution (with each ATAC unit having only half the parameters compared to the LocalSENet module). The difference between LocalSENet and ATAC is that it is beneficial for activation functions. (c) By replacing ReLUs with the proposed ATAC unit—we conclude that channel-wise context is more effective than using individual ReLUs. It can be seen that: (a) The ATAC unit achieves a better accuracy compared to other activation units, namely ReLU [2], SELU [6], Swish [7], and xUnit [9]. These results reaffirm that one can obtain better activation functions and other state-of-the-art network competitors. We work's predictive performance on CIFAR-10 and CIFAR-100 with a gradual increase of network depth. It can be seen in Fig. 6(a) and Fig. 6(b), the performance tends to increase with more ATAC units. Therefore, a fully attentional network offers a way to obtain a performance increase with marginal additional costs. It can also be seen that the largest performance increase is obtained for the replacements made going towards a fully attentional network.

Fig. 6: Illustration of the performance gain tendency in the LocalSENet module is set to 1.0. The results suggest that we can obtain a better performance by gradually replacing ReLUs with ATAC units starting with the last layer and ending with the first layer. As it can be seen in Fig. 5, only 65% of the parameters are used, while the performance is still better than the ReLU-like activation units. These results ports our hypothesis that early attentional modulation enables by suppressing irrelevant low-level features and highlighting the network's predictive performance on CIFAR-10 and CIFAR-100 given a gradual increase of the depths of the networks. Finally, we address question Q4 raised at the beginning of this section by comparing our approach with several activation functions and other state-of-the-art network competitors. We start by comparing our approach with several activation functions. (b) By applying as many lightweight attention modules as possible, one can obtain a more efficient convolutional task. The results suggest that ATAC units being used (hence, no gain). The results suggest that instead of blindly increasing the number of parameters and computation costs, one should resort to a few times. instead of adopting the sophisticated attention modules only a few times.

Ablation Study – Efficiency of Fully Attentional Networks

![Graphs showing the contribution ratio of induced parameters by ATAC units for CIFAR-10 and CIFAR-100 datasets.](image)
Comparison to State-of-the-Art Activation Units

Fig. 6: Illustration of the performance gain tendency in CIFAR-10 and CIFAR-100. Here, a contribution percentage by gradually replacing ReLUs with the proposed ATAC unit. We analyze the network's performance on CIFAR-10 and CIFAR-100 with a gradual increase of network depth. As it can be seen, the ATAC-ResNet is not affected by the marginal additional costs. It can also be seen that the largest performance increase is obtained for the replacements made from the last layer and ending with the first layer. As it can be observed, the ATAC-ResNet offers a way to obtain a performance increase with even fewer layers or parameters per network by replacing ReLUs with ATAC units.

Fig. 7(b) provides the comparison on CIFAR-10 and CIFAR-100 given a gradual increase of the depths of the networks. The results reaffirm that one can obtain better activation functions and other state-of-the-art network competitors. We also considered GELU and PReLU, but their performance was not as good as the aforementioned baselines and were, hence, excluded from the overall comparison.

Table III provides the results, from which it can be seen that: (a) The ATAC unit achieves a better performance for all experimental settings, which demonstrates its superiority compared to other activation units, namely ReLU, SELU, Swish, and xUnit. First, we compared our proposed methods with the baseline Swish, which is also a non-linear gating function, ranked second in the list of activation functions and other state-of-the-art network competitors. We also showed that the ATAC-ResNet is not affected by the marginal additional costs.

Second, we compared Swish with the state-of-the-art networks. Fig. 8(a) and Fig. 8(b) illustrate the results given a gradual increase in the network depths for all the networks to encode higher-level semantics more efficiently by suppressing irrelevant low-level features and highlighting relevant features in the early layers of the networks.

The results suggest that given the same budget for parameters and computation costs, one should resort to ATAC, which suggests that having a small additional budget can replace the attention mechanism instead of a NiN-style block. This is beneficial for activation functions. (c) By replacing ReLUs with ATAC units, one can obtain a more efficient convolutional layer. Hence, to obtain the same number of parameters and the same computational costs, the channel-wise context is applied after every convolution (with each ATAC unit having the additional parameters). In contrast, the ATAC units are applied only half the parameters compared to the LocalSENet module.

The LocalSENet uses the attention mechanism only once with all ReLUs in the network. 2) The difference between LocalSENet and ATAC is that the network can obtain a consistent performance gain by replacing ReLUs with ATAC units and a ratio of 0.0 corresponds to no gain. The results suggest that the network can obtain a consistent performance gain by replacing ReLUs with ATAC units and a ratio of 0.0 corresponds to no gain. The results suggest that the network can obtain a consistent performance gain by replacing ReLUs with ATAC units and a ratio of 0.0 corresponds to no gain.
Comparison to State-of-the-Art Networks

![Graphs showing accuracy vs network parameters for CIFAR-10 and CIFAR-100 datasets.](image)

- **CIFAR-10**

- **CIFAR-100**

---

**Table IV: Classification comparison on ImageNet with other state-of-the-art networks.**

<table>
<thead>
<tr>
<th>Architecture</th>
<th>GFlops</th>
<th>Params</th>
<th>top-1 err.</th>
<th>top-5 err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA-ResNet-50</td>
<td>7.2</td>
<td>18.0M</td>
<td>22.40</td>
<td>6.20</td>
</tr>
<tr>
<td>AA-ResNet-50</td>
<td>8.3</td>
<td>25.8M</td>
<td>22.30</td>
<td>6.20</td>
</tr>
<tr>
<td>SE-ResNet-50</td>
<td>3.87</td>
<td>28.1M</td>
<td>22.12</td>
<td>5.99</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>3.86</td>
<td>25.6M</td>
<td>23.30</td>
<td>6.55</td>
</tr>
<tr>
<td>ResNet-32</td>
<td>4.4</td>
<td>28.0M</td>
<td>21.41</td>
<td>6.02</td>
</tr>
</tbody>
</table>

---

**References**


---

**Conclusion**

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61573183, the Nanjing Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant No. 201900029, the Association for Science and Technology, China Scholarship Council under Grant No. 201806830039.
**Comparison to State-of-the-Art Networks**

**Table 4**: Comparison on ImageNet

<table>
<thead>
<tr>
<th>Architecture</th>
<th>GFlops</th>
<th>Params</th>
<th>top-1 err.</th>
<th>top-5 err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA-ResNet-50 [14]</td>
<td>7.2</td>
<td>18.0M</td>
<td>22.40</td>
<td>/</td>
</tr>
<tr>
<td>GE-$\theta^+$-ResNet-50 [15]</td>
<td>3.87</td>
<td>33.7M</td>
<td>21.88</td>
<td>5.80</td>
</tr>
<tr>
<td>ATAC-ResNet-50 (ours)</td>
<td>4.4</td>
<td>28.0M</td>
<td><strong>21.41</strong></td>
<td>6.02</td>
</tr>
</tbody>
</table>
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!


