**A. Scaled Cosine-similarity Attention**

Following the notations used in [3], the scaled dot-product attention can be formulated as Eqn. 1.

\[
\text{Att}_{sdp}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}
\]

where the input comprises queries \( Q \) and keys \( K \) of dimension \( d_k \), and values \( V \) of dimension \( d_v \). The scaled dot-product attention divides each dot product by \( \sqrt{d_k} \) and applies a softmax function to generate the attention weights. However, the variance of dot products grows large for large values of \( d_k \), pushing the softmax function into the regions of extremely small gradients, as shown in Eqn. 2.

\[
\begin{align*}
E(q_i) &= E(k_i) = 0, \\
\text{Var}(q_i) &= \text{Var}(k_i) = 1, \\
E(q \cdot k) &= E(\sum_{i=1}^{d_k} q_i k_i) = 0, \\
\text{Var}(q \cdot k) &= \text{Var}(\sum_{i=1}^{d_k} q_i k_i) = d_k,
\end{align*}
\tag{2}
\]

where the components of \( q \) and \( k \) are assumed as independent random variables with mean 0 and variance 1, and their dot product \( q \cdot k = \sum_{i=1}^{d_k} q_i k_i \) has mean 0 and variance \( d_k \). To counteract this effect discussed above, the scaled dot-product attention divides each dot product by \( \sqrt{d_k} \) to keep the variance being 1.

The dot product involves not only the direction but also the norm of vectors, as shown in Eqn. 3.

\[
q \cdot k = \|q\| \|k\| \cos \angle(q, k). \tag{3}
\]

If the norm \( \|k\| \) is large enough and \( \cos \angle(q, k) \neq 0 \), the dot product would be large too no matter what \( q \) is. After the softmax normalization, the weight for \( q \) on \( k \) would be much larger than other keys. Similarly, the scaled dot-product attention introduces both the content and intensity of feature points, and the strongly activated feature points would surpass others after the softmax normalization because of larger norm. Thus, the self-attention would focus more on the foreground and generates almost the same attention maps for different query positions.

Our proposed scaled cosine-similarity attention focuses on the content represented by the direction of feature points and avoids strongly activated keys surpassing other keys after the softmax normalization. To compute cosine similarity, we apply \( L_2 \) normalization to queries and keys before the dot product. The components of \( q \) and \( k \) are scaled down by the corresponding norm and have mean 0 and variance \( \frac{1}{d_k} \), as shown in Eqn 4.

\[
\begin{align*}
\sqrt{\sum_{i=1}^{d_k} q_i^2} &= 1, \\
E(\sum_{i=1}^{d_k} q_i^2) &= 1, \\
E(q_i^2) &= \frac{1}{d_k}, \\
\text{Var}(q_i) &= \frac{1}{d_k},
\end{align*}
\tag{4}
\]

where \( q' \) and \( k' \) are queries and keys after \( L_2 \) normalization. Consequently, the cosine similarity has mean 0 and variance \( \frac{1}{d_k} \) as presented in Eqn. 5.

\[
\begin{align*}
E(q'_i k'_i) &= 0, \\
\text{Var}(q'_i k'_i) &= \frac{1}{d_k^2}, \\
E(q' \cdot k') &= E(\sum_{i=1}^{d_k} q'_i k'_i) = 0, \tag{5}
\end{align*}
\]

\[
\begin{align*}
\text{Var}(q' \cdot k') &= \text{Var}(\sum_{i=1}^{d_k} q'_i k'_i) = \frac{1}{d_k}.
\end{align*}
\]

Similar to the scaled dot-product attention, to maintain the cosine similarity in magnitude at different values of \( d_k \), our proposed scaled cosine-similarity attention multiplies each cosine similarity by \( \sqrt{d_k} \), as shown in Eqn. 6.

\[
\text{Att}_{scs}(Q, K, V) = \text{softmax}(\sqrt{d_k} \cdot \text{Norm}(Q) \cdot \text{Norm}(K)^T)V, \tag{6}
\]
where \( \text{Norm} \) represents \( L_2 \) normalization along the channel dimension.

**B. Ablation study of self-attention mechanisms**


<table>
<thead>
<tr>
<th>Method</th>
<th>Atten.</th>
<th>( \text{AP}^b )</th>
<th>( \text{AP} )</th>
<th>( \text{AP}_{50} )</th>
<th>( \text{AP}_{75} )</th>
<th>( \text{AP}_S )</th>
<th>( \text{AP}_M )</th>
<th>( \text{AP}_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPN [1]</td>
<td></td>
<td>37.1</td>
<td>34.1</td>
<td>55.4</td>
<td>36.2</td>
<td>18.4</td>
<td>37.3</td>
<td>46.0</td>
</tr>
<tr>
<td>PAFPN [2]</td>
<td></td>
<td>37.6</td>
<td>34.4</td>
<td>55.9</td>
<td>36.4</td>
<td>18.7</td>
<td>37.5</td>
<td>47.2</td>
</tr>
<tr>
<td>w/ MGC D.P.</td>
<td></td>
<td>38.2</td>
<td>35.0</td>
<td>56.9</td>
<td>37.1</td>
<td>19.0</td>
<td>38.3</td>
<td>48.0</td>
</tr>
<tr>
<td>w/ MGC S.D.P.</td>
<td></td>
<td>37.8</td>
<td>34.6</td>
<td>56.2</td>
<td>36.7</td>
<td>18.4</td>
<td>38.0</td>
<td>47.8</td>
</tr>
<tr>
<td>w/ MGC S.C.S.</td>
<td></td>
<td><strong>38.6</strong></td>
<td><strong>35.4</strong></td>
<td><strong>57.4</strong></td>
<td><strong>37.5</strong></td>
<td><strong>19.5</strong></td>
<td><strong>38.6</strong></td>
<td><strong>48.3</strong></td>
</tr>
</tbody>
</table>

To verify the effectiveness of the proposed scaled cosine-similarity attention, we compare it with the dot-product attention in [4] and the scaled dot-product attention in [3]. Note that we remove the restriction on the norm in orthogonal regularization for the latter two, as shown in Eqn. 7.

\[
L_o = \lambda_o \| W \psi W^T \psi \odot (1 - I) \|^2_F. \tag{7}
\]

As shown in Table 1, our proposed method achieves 38.6% box AP and 35.4% mask AP, outperforming the dot-product attention by 0.4% box AP and 0.4% mask AP, and the scaled dot-product attention by 0.8% box AP and 0.8% mask AP.

**C. More Visual Comparisons**

As illustrated in Figure 1, we provide more instance segmentation result comparisons between Mask R-CNN w/ FPN and Mask R-CNN w/ \( A^2 \)-FPN on COCO val2017.

**D. More Visual Results**

As illustrated in Figure 2, we provide more instance segmentation results on COCO val2017, and these visualizations are based on \( A^2 \)-FPN equipped with HTC.

**References**


Figure 1. More instance segmentation result comparisons between FPN (odd rows) and $A^2$-FPN (even rows) on COCO val2017.
Figure 2. More instance segmentation results of $A^2$-FPN equipped with HTC on COCO val2017.