A^2-FPN: Attention Aggregation based Feature Pyramid Network for Instance Segmentation — Supplementary Material

Miao Hu Yali Li* Lu Fang Shengjin Wang
Department of Electronic Engineering, Tsinghua University
hum19@mails.tsinghua.edu.cn, {liyali13, fanglu, wgsgj}@tsinghua.edu.cn

A. Scaled Cosine-similarity Attention

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

$$\boldsymbol{Att}_{sdp}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

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.

$$\begin{aligned} E(q_i)&=E(k_i)=0,\\ Var(q_i)&=Var(k_i)=1,\\ E(q_ik_i)&=0,\ Var(q_ik_i)=1,\\ E(q\cdot k)&=E(\sum_{i=1}^{d_k}q_ik_i)=0,\\ Var(q\cdot k)&=Var(\sum_{i=1}^{d_k}q_ik_i)=d_k,\end{aligned} \tag{2}$$

where the components of q and k are assumed as independent random variables with mean 0 and variance 1, and their dot product $q\cdot k=\sum_{i=1}^{d_k}q_ik_i$ 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.

$$q\cdot k=\|q\|\|k\|\cos<q,k>. \tag{3}$$

If the norm $\|k\|$ is large enough and $cos<q,k>\neq 0$, 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.

$$\begin{aligned} &\sqrt{\sum_{i=1}^{d_k}{q'}_i^2}=1, && E(\sum_{i=1}^{d_k}{q'}_i^2)=1,\\ &E({q'}_i^2)=\frac{1}{d_k}, && Var({q'}_i)=\frac{1}{d_k},\end{aligned} \tag{4}$$

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.

$$\begin{aligned} E(q'_ik'_i)&=0, Var(q'_ik'_i)=\frac{1}{d_k^2},\\ E(q'\cdot k')&=E(\sum_{i=1}^{d_k}q'_ik'_i)=0,\\ Var(q'\cdot k')&=Var(\sum_{i=1}^{d_k}q'_ik'_i)=\frac{1}{d_k}.\end{aligned} \tag{5}$$

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.

$$\boldsymbol{Att}_{scs}(Q,K,V)=\mathrm{softmax}(\sqrt{d_k}\cdot\mathrm{Norm}(Q)\mathrm{Norm}(K)^T)V, \tag{6}$$

*Corresponding author.

where Norm represents L_2 normalization along the channel dimension.

B. Ablation study of self-attention mechanisms

Table 1. **Ablation study of different self-attention mechanisms** on COCO *val2017*. "Atten." means self-attention mechanism. D.P.: dot-product attention, S.D.P.: scaled dot-product attention, S.C.S.: scaled cosine-similarity attention.

Method	Atten.	AP^{bb}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FPN [1]	-	37.1	34.1	55.4	36.2	18.4	37.3	46.0
PAFPN [2]	-	37.6	34.4	55.9	36.4	18.7	37.5	47.2
w/ MGC	D.P.	38.2	35.0	56.9	37.1	19.0	38.3	48.0
w/ MGC	S.D.P.	37.8	34.6	56.2	36.7	18.4	38.0	47.8
w/ MGC	S.C.S.	**38.6**	**35.4**	**57.4**	**37.5**	**19.5**	**38.6**	**48.3**

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.

$$L_o = \lambda_o \| W_\psi W_\psi^T \odot (\mathbf{1} - I) \|_F^2. \tag{7}$$

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

[1] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[2] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 2

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2

[4] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

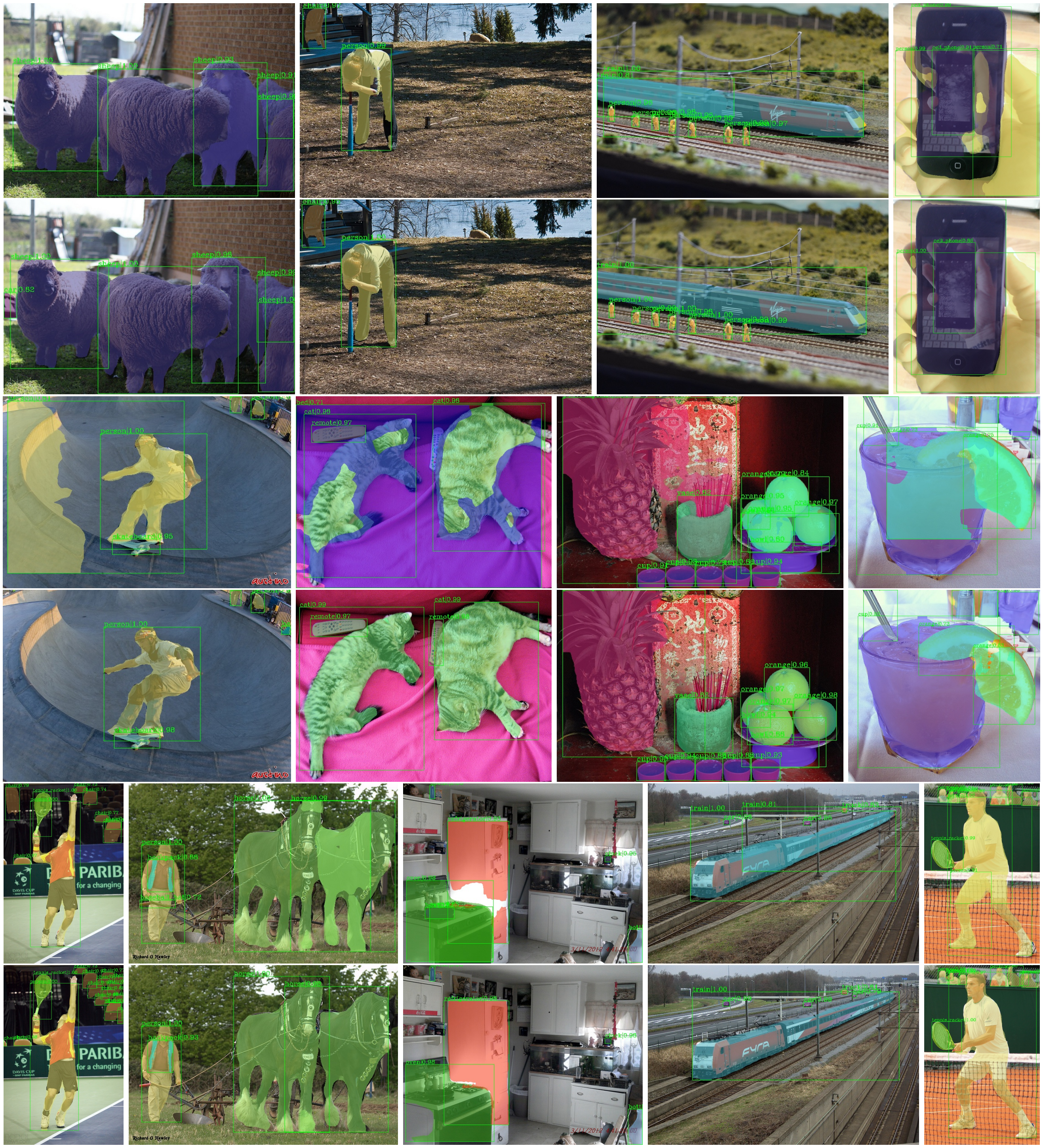

Figure 1. More instance segmentation result comparisons between **FPN** (odd rows) and A^2-**FPN** (even rows) on COCO *val2017*.

Figure 2. More instance segmentation results of A^2-**FPN** equipped with HTC on COCO *val2017*.