Supplementary Materials: Visual Attention Consistency under Image Transforms for Multi-label Image Classification

Hao Guo[‡], Kang Zheng[‡], Xiaochuan Fan[‡], Hongkai Yu[♯], Song Wang[†,‡,*]
[†]Tianjin University, [‡]University of South Carolina, [♯]University of Texas - Rio Grande Valley
{hguo, zheng37}@email.sc.edu, efan3000@gmail.com, hongkai.yu@utrgv.edu, songwang@cec.sc.edu

Average precision improvement for each label

To further verify that considering visual attention consistency under certain image transforms can benefit the multi-label image classification, we compare the average precisions (APs) achieved for each label by different models trained on WIDER Attribute dataset [6] in Table 1. The models are denoted in the same way with the Table 1 in the original paper: R50, R50+t, R50+r, R50+s, R50+f, R50+ACt, R50+ACr, R50+ACs, R50+ACf, R50+ACfs, R101, R101+ACt, R101+ACr, R101+ACs, R101+ACf, and R101+ACfs. Note that:

- models R50 and R101 are the **baseline models** with ResNet50 and ResNet101 as backbone, respectively.
- models R50+t, R50+r, R50+s and R50+f are the models using *translation*, *rotation*, *scaling* and *flipping* as **data augmentation**, **without** considering attention consistency under these image transforms, respectively. To be specific, these models are trained from the proposed two-branch network by removing the attention consistency loss. The backbone is ResNet50.
- models R50+ACt, R50+ACr, R50+ACs, R50+ACf and R101+ACt, R101+ACr, R101+ACs, R101+ACf are the models considering **attention consistency** under *translation*, *rotation*, *scaling* and *flipping*, respectively. The backbones of the above two sets of models are ResNet50 and ResNet101, respectively.
- models R50+ACfs and R101+ACfs are the models considering **attention consistency under both scaling and flipping** with backbones of ResNet50 and ResNet101, respectively.

We can notice that the average precision of every label is improved by considering attention consistency, especially that under rotation, scaling, flipping and both scaling and flipping.

*Corresponding author.

Selection of multi-label image classification loss

Though there exist various loss functions that can be used as multi-label image classification loss, we simply use the weighted sigmoid cross entropy loss [4, 5, 7] as classification loss in the proposed network. To verify that the classification loss of the proposed network is not limited to weighted sigmoid cross entropy loss, we further replace it with multi-label soft margin loss [1] (which is also modified from cross entropy loss) and train models R50 and R50+ACf denoted in Table 1. As shown in Table 2, the proposed network can also work well if the multi-label image classification loss is changed.

Impact of hyper-parameter λ in Eq. (5)

In the original paper, we simply set $\lambda = 1$ in Eq. (5), since we find that multi-label image classification loss and attention consistency loss are in the same magnitude at the start of the model training. We further try different values assigned to λ for the training of model R50+ACf denoted in Table 1 and show the achieved mAPs in Table 3. The mAP results show that our selection of hyper-parameter $\lambda = 1$ is reasonable.

Impact of transformed image size for attention consistency under image scaling

In the original paper, we mainly considering attention consistency under *translation*, *rotation*, *scaling* and *flipping*. There may be more applicable transforms that can be embedded into the proposed network.

Existing CNNs usually resize the input images to a fixed size for image classificaition task, e.g., 227×227 for AlexNet [3], 224×224 for VGG [8], ResNet [1], DenseNet [2], etc. We all know that the input size of a CNN can influence the performance of trained model, since resizing to a smaller size may result in more information loss. Therefore, for fair comparison with the existing works, we fix the original image size as 224×224 (default input size

[1]Provided by PyTorch

Table 1. Average precisions (APs) of each label achieved by baseline models and models trained from the proposed network.

model	*male*	*long hair*	*sunglasses*	*hat*	*t-shirt*	*long sleeves*	*formal*	*short*	*jeans*	*long pants*	*skirt*	*face mask*	*logo*	*stripe*	mAP
R50	94.3	84.1	70.8	93.8	76.8	95.0	80.7	90.3	77.0	94.6	81.4	75.3	88.4	64.7	83.4
R50+t	94.7	85.9	69.2	93.8	78.8	95.3	81.3	91.1	76.7	95.6	82.0	75.4	89.0	62.0	83.7
R50+r	94.3	85.0	68.7	93.8	77.7	95.0	81.4	90.8	76.0	95.2	80.4	73.9	88.7	64.2	83.2
R50+s	94.6	85.5	71.0	94.5	77.2	95.2	81.5	91.1	77.0	94.9	81.6	76.2	88.7	64.3	83.9
R50+f	95.0	85.6	73.8	94.4	76.8	95.1	81.3	91.2	77.0	95.1	81.8	77.2	88.7	64.8	84.2
R50+ACt	95.0	86.0	71.3	93.9	77.4	95.4	81.8	90.6	76.4	95.2	82.2	76.3	88.3	63.8	83.9
R50+ACr	95.1	86.7	71.1	94.7	79.6	95.8	82.6	91.5	79.6	**96.0**	81.9	76.5	90.3	68.5	85.0
R50+ACs	95.4	87.4	73.2	95.0	80.3	96.0	84.2	92.2	79.9	95.9	84.2	77.2	90.3	66.6	85.6
R50+ACf	95.7	87.8	75.4	95.1	81.0	96.1	84.2	92.7	80.3	95.9	84.8	**80.5**	90.2	68.0	86.3
R50+ACfs	**96.3**	**88.7**	**76.9**	**95.3**	**82.6**	**96.2**	**85.2**	**93.1**	**80.7**	**96.0**	**85.8**	78.5	**90.9**	**68.7**	**86.8**
R101	95.1	85.8	72.7	94.4	79.0	95.6	82.4	91.6	78.7	95.2	83.0	76.3	89.8	66.6	84.8
R101+ACt	95.0	86.3	71.2	94.3	78.3	95.6	83.1	91.1	78.7	95.3	83.4	75.7	89.3	65.5	84.6
R101+ACr	95.5	88.5	72.8	95.3	81.7	96.2	83.2	92.8	80.9	95.9	84.2	77.4	90.9	68.5	86.0
R101+ACs	95.6	88.4	73.7	95.1	81.8	96.3	84.8	93.0	81.1	95.9	86.1	78.1	90.8	69.1	86.5
R101+ACf	96.0	88.9	**76.4**	95.7	81.8	96.5	84.9	93.8	82.0	**96.5**	86.7	**80.2**	90.5	69.2	87.1
R101+ACfs	**96.2**	**89.4**	75.7	**96.0**	**83.4**	**97.0**	**85.4**	**94.0**	**82.6**	96.4	**87.4**	79.4	**91.5**	**69.4**	**87.5**

Table 2. Performance of baseline model R50 and model R50+ACf (attention consistency under flipping) using different classification loss functions: ℓ_{c1} – weighted sigmoid cross entropy loss; ℓ_{c2} – multi-label soft margin loss. (on WIDER dataset)

model	mAP	mA	F1-C	P-C	R-C	F1-O	P-O	R-O
R50+ℓ_{c1}	83.4	82.0	73.9	**79.5**	69.4	79.4	82.3	76.6
R50+ACf+ℓ_{c1}	**86.3**	**84.5**	**76.4**	78.9	**74.3**	**81.2**	**82.6**	**79.8**
R50+ℓ_{c2}	83.5	81.8	73.8	80.3	68.7	79.4	83.3	75.8
R50+ACf+ℓ_{c2}	**86.2**	**82.6**	**75.5**	**83.3**	**69.7**	**81.1**	**85.3**	**77.3**

Table 3. Impact of different values of λ on mAP for model R50+ACf, considering attention consistency under flipping.

$\lambda =$	0.1	1	2	10	20
mAP (%)	85.3	86.3	86.3	85.6	85.0

of ResNet50 / 101, which is also used by other methods), when training the proposed network on WIDER dataset.

However, when the attention consistency under image scaling is considered by the proposed network, the transformed (scaled) images are resized to a different size. If the transformed images are upscaled to a size larger than 224×224, there may be performance improvement of the proposed network resulting from larger input size. To focus on the performance improvement from considering attention consistency, we form the scaled images by downscaling the original images to 192×192 in the original paper, when considering attention consistency under scaling. Note that comparing the performance of model R50+ACs and model R50+s (using multi-scale input as data augmentation) with model R50 has already verified that the performance improvement are mainly from considering attention consistency, not the multi-scale input.

To further verify the impact of different input sizes of the branch taking transformed images as input, we further conduct experiments of fixing the size of transformed images to 160×160 and 256×256, respectively, to train model R50+ACs with attention consistency under scaling. As shown in Table 4, when the input size of the branch taking transformed images increases, the performance of mean average precision (mAP, %) is improved ($85.2\% \rightarrow 85.6\% \rightarrow 86.1\%$). This result suggests that the performance of the proposed network may be further improved by upscaling the input.

Table 4. mAP (%) performance achieved on WIDER dataset by the proposed network considering attention consistency under scaling with fixed original image size 224×224 and different transformed image sizes, separately.

scaled image	160×160	192×192	256×256
original image 224×224	85.2	85.6	86.1

Impact of different usages of certain image transform

Besides the horizontal flipping embedded in the proposed network, we also conduct experiment of embedding vertical flipping in the proposed network. As shown in Table 5, even though the vertical flipping is not normal in practice, considering attention consistency under vertical flipping can also slightly improve the multi-label image classification performance. We can also notice that considering attention consistency under certian transform can perform much better than using the transform as data augmentation.

Table 5. Performance comparison of using flipping transform differently.

model	usage of flipping	mAP	mA	F1-C	P-C	R-C	F1-O	P-O	R-O
R50	without flipping	83.4	82.0	73.9	79.5	69.4	79.4	82.3	76.6
R50+f	horizontal flipping as data augmentation	84.2	82.8	74.6	79.5	70.7	80.0	82.9	76.9
R50+ACf	attention consistency under horizontal flipping	**86.3**	84.5	76.4	78.9	74.3	81.2	82.6	79.8
	attention consistency under vertical flipping	84.9	83.3	74.9	78.0	72.2	80.1	81.9	78.4

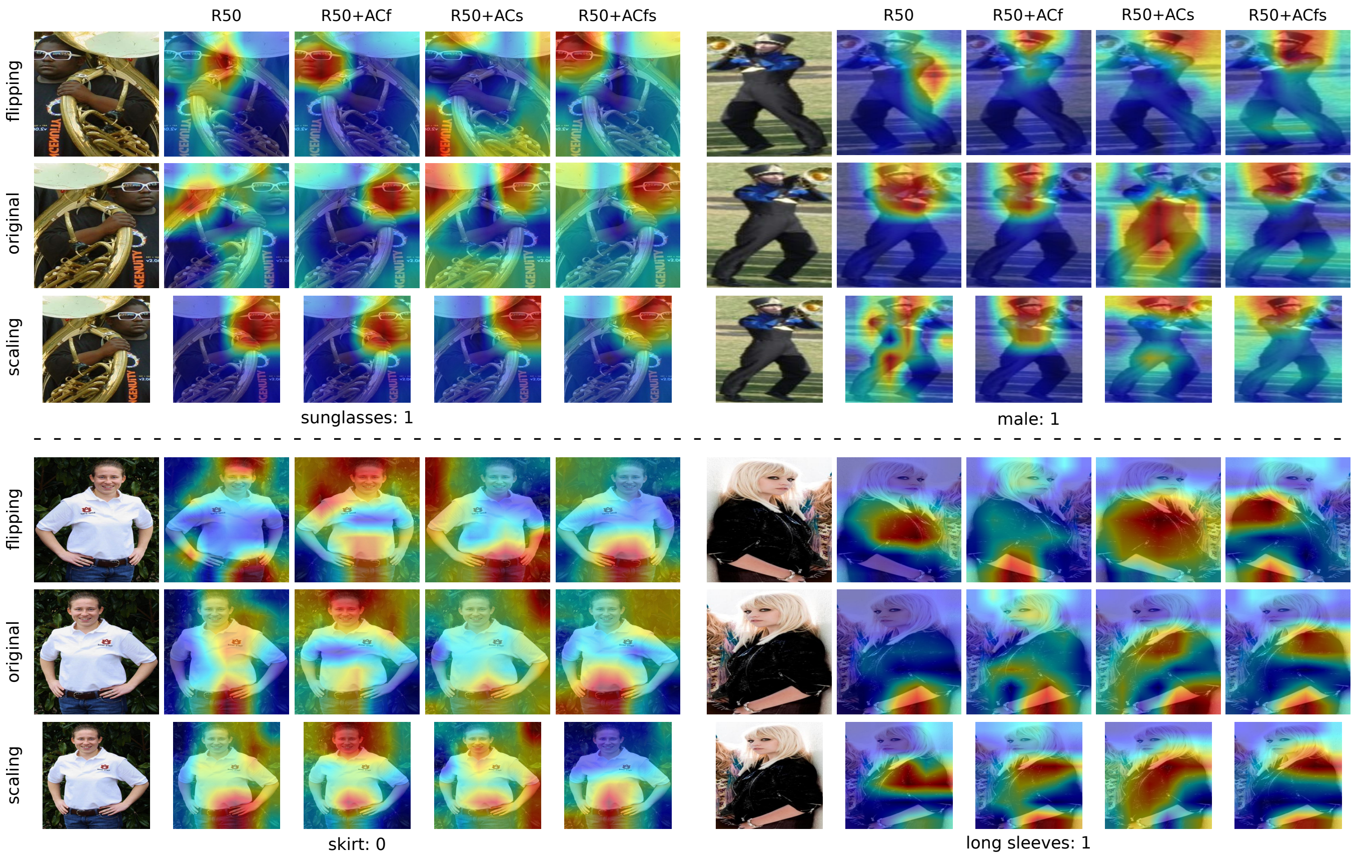

Figure 1. Attention heatmaps for classifying different labels from flipped, original and scaled images using different models. The red color indicates attention regions.

Supplementary qualitative analysis

To further verify that enforcing attention consistency under certain transforms can help CNNs focus attention on regions more relevant to each label, we show more qualitative results of attention heatmaps for classifying different labels from flipped, original and scaled images using different models, respectively, in Fig. 1 (similar to Fig. 5 in the original paper). From Fig. 1, we can notice that R50 usually focuses attention on inconsistent regions of the original and the transformed images. Even worse, the attention of current R50 may cover many regions irrelevant to the specific label. As the attention consistency under a transform (flipping / scaling / both) is enforced by the proposed network, the attention regions usually become consistent under this transform. Besides, as attention regions are forced to be consistent under certain transform, they may be focused on regions more relevant to the specific label.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[2] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 1

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[4] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance

scenarios. In *Asian Conference on Pattern Recognition*, pages 111–115. IEEE, 2015. 1

[5] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 1

[6] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, pages 684–700. Springer, 2016. 1

[7] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision*, pages 1–9, 2017. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1