(Supplementary Material)

In-sample Contrastive Learning and Consistent Attention for Weakly Supervised Object Localization

Anonymous ACCV 2020 submission

Paper ID 51

This document provides supplementary material mentioned in the main paper.

It contains three parts:

- Sec. A provides the location of our non-local attention blocks.
- Sec. B compares our contrastive attention loss ($\mathcal{L}_{\text{ca}}$) and its variant: $\mathcal{L}_{\text{ca}}$ with normalized temperature-scaled cross-entropy (*NT-Xent*).
- Sec. C shows more visual results of our method.

A Location of our non-local attention block

We build the proposed method upon three CNN backbones: VGG16 [1], InceptionV3 [2], ResNet50 [3]. We insert three non-local attention blocks at different locations for each backbone. For VGG16 [1], our non-local attention blocks are inserted after `conv_5_3`, `pool_4` and `pool_3` layers. For InceptionV3 [2], `mixed_5d`, `mixed_6e` and `mixed_A3_2b` are chosen. For ResNet50 [3], first residual block of `layer3`, first block and last block of `layer4` are chosen.

B *NT-Xent* for contrastive attention loss

NT-Xent for contrastive attention loss compares our contrastive attention loss ($\mathcal{L}_{\text{ca}}$) and its variant: $\mathcal{L}_{\text{ca}}$ with normalized temperature-scaled cross-entropy (*NT-Xent*) [4, 5]. We simply replace Eq. 3 of the main paper with *NT-Xent* loss. We extract three in-sample masked features ($z_{\text{dfg}}, z_{\text{fg}}$ and z_{bg}) and calculate *NT-Xent* as

$$\tilde{\mathcal{L}_{\text{ca}}} = -\log\left(\frac{\exp(sim(z_{\text{dfg}}, z_{\text{fg}})/\tau)}{\exp(sim(z_{\text{dfg}}, z_{\text{fg}})/\tau)) + \exp(sim(z_{\text{dfg}}, z_{\text{bg}})/\tau)}\right) \tag{1}$$

where *sim* is the cosine similarity between two features and τ denotes a temperature parameter, which is set to 0.07.

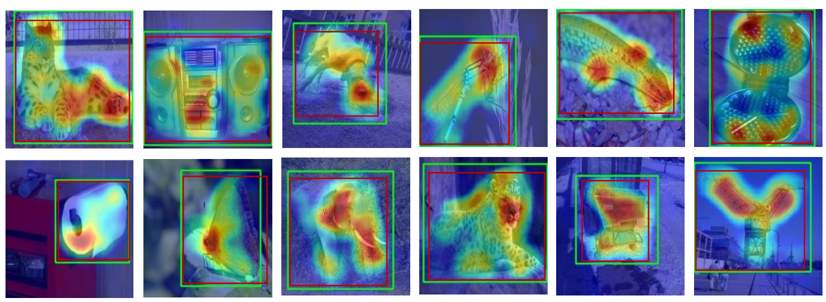
Table 1 compares `MaxBoxAccV2` [6] of the baseline [7], ours, and the *NT-Xent* variant. The *NT-Xent* variant improves the performance in all extents compared to baseline [7]. However, it is slightly inferior to ours in IoU 0.7.

Table 1. Effect of *NT-Xent* [4, 5] on our contrastive attention loss upon ResNet50 [3] on CUB-200-2011 [8]. The mean column is for the average of the three IoU thresholds 0.3, 0.5, and 0.7.

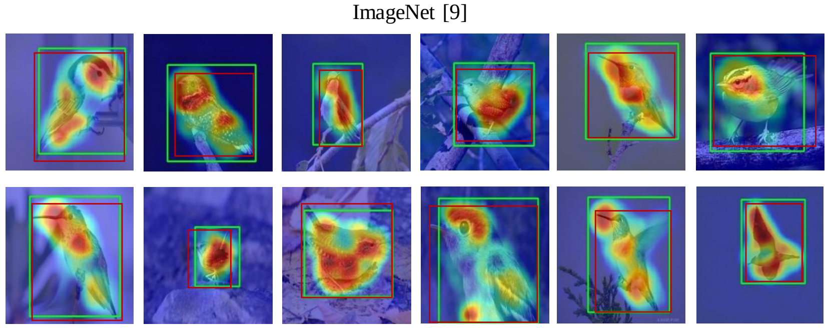
Methods	`MaxBoxAccV2`@IoU (%)			
	0.3	0.5	0.7	Mean
Baseline [7]	91.82	64.78	18.43	58.34
Ours (full) w *NT-Xent*	96.70	73.55	20.03	63.43
Ours (full) w triplet	96.18	72.79	20.64	63.20

C Additional qualitative results

We show additional qualitative results of our method on CUB-200-2011 [8] and ImageNet [9]. Fig. 1 illustrates activation maps and estimated bounding boxes at the optimal activation threshold. Our method not only spreads out to the less discriminative parts but also restrains the activations in the object regions.



ImageNet [9]



CUB-200-2011 [8]

Fig. 1. Qualitative examples of activation map and localization produced by our model on the ImageNet [9] and CUB-200-2011 [8] `test` split. The red boxes are the ground-truth and the green boxes are the predicted ones. The activation map is colored in heatmap scale (red: high, blue: low). Best viewed in color.

References

1. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR). (2015)
2. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 2818–2826
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 770–778
4. He, K., Fan, H., Wu, Y., Xe, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 9729–9738
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). (2020)
6. Choe, J., Oh, S., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
7. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 2219–2228
8. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Technical report cns-tr-2011-001, California Institute of Technology (2011)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., et al., S.S.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115** (2015) 211–252