A Generic Visualization Approach for Convolutional Neural Networks Supplementary Material

Ahmed Taha, Xitong Yang, Abhinav Shrivastava, and Larry Davis

University of Maryland, College Park

S1 Extended Related Work

Classification networks learn class-logits $\in R^{N_c}$. The number of logits is equal to the number of classes N_c. There is a clear *one-to-one mapping* between classes and logits. This mapping is vital for class-activation mapping (CAM) and Grad-CAM approaches because their visualizations rely on the weights or gradients of a particular logit. In contrast, retrieval networks learn a feature embedding $\in R^d$. The output dimensionality does not equal the number of classes. Thus, there is no one-to-one mapping between classes and output dimensions. This lack of mapping is why CAM and Grad-CAM suffer on retrieval networks. To highlight this limitation, we train a retrieval network with various ranking losses. The following paragraphs review the two ranking losses employed in the main paper.

Retrieval networks learn a feature embedding where objects within the same class are closer than objects from different classes. To learn this feature embedding, a retrieval network is trained with ranking losses such as contrastive, triplet, and N-pair losses.

In the main paper, we employ triplet loss [6] for its simplicity and efficiency. Equation S1 shows the triplet loss formulation

$$TL(a,p,n) = [(D(\lfloor a \rfloor, \lfloor p \rfloor) - D(\lfloor a \rfloor, \lfloor n \rfloor) + m)]_+, \tag{S1}$$

where $[\bullet]_+ = max(0, \bullet)$ is the hinge function and m is the margin between different classes in the feature embedding. $\lfloor \bullet \rfloor$ and $D(,)$ are the embedding and the Euclidean distance functions, respectively. This formulation attracts an anchor image a of a specific class closer to a positive image p from the same class than it is to a negative image n.

We leverage the semi-hard sampling [6] strategy. In semi-hard negative sampling, instead of picking the hardest positive-negative samples, all anchor-positive pairs and their corresponding semi-hard negatives are considered. Semi-hard negatives are further away from the anchor than the positive exemplar, yet within the banned margin m as shown in Figure S1.

The performance of triplet loss relies heavily on the sampling strategy because every anchor sample is paired with a single negative sample. N-pair loss

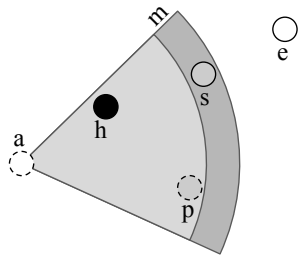

Fig. S1: Triplet loss tuple (anchor, positive, negative) and margin m. The (h)ard, (s)emi-hard and (e)asy negatives are highlighted in black, gray, and white, respectively.

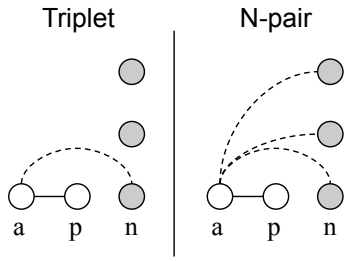

Fig. S2: The difference between triplet and N-pair losses using a single positive pair (a, p) and three negative (n) samples. The triplet loss pushes the anchor a away from a selected negative sample while N-pair pushes the anchor a away from all negative samples. The N-pair all-negatives approach relaxes the requirement for an efficient negative mining strategy.

mitigates this limitation by pairing every anchor with all negative samples within a mini-batch. Figure S2 depicts the difference between triplet and N-pair losses. Equation S2 shows the N-pair loss formulation

$$\text{NPL} = -\log \frac{exp(\lfloor a \rfloor \lfloor p \rfloor)}{exp(\lfloor a \rfloor \lfloor p \rfloor) + \sum_{n \in B} exp(\lfloor a \rfloor \lfloor n \rfloor)}, \tag{S2}$$

For N-pair loss, a training batch contains a single positive pair from each class. Thus, a mini-batch will have $b/2$ positive pairs and every anchor is paired with $b-2$ negatives, where b is the mini-batch size.

Weakly supervised object localization (WSOL) approaches localize objects inside images using the class label only. Attention visualization approaches (*e.g.*, CAM) generate class-specific attention heatmaps. A simple segmentation of the heatmap provides a localization bounding box. Attention-based approaches do not require bounding box annotations during training. Thus, these approaches

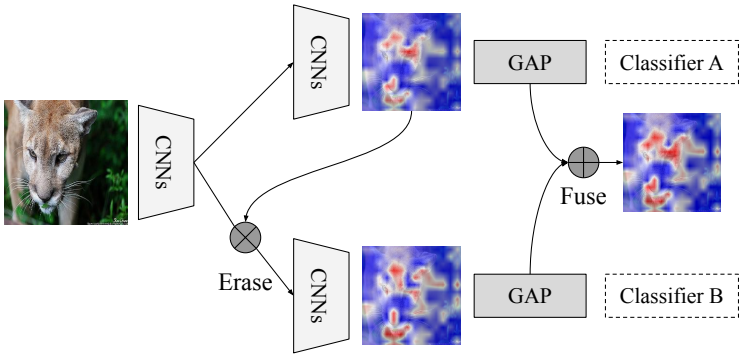

Fig. S3: An illustration of the ACoL method; A classification network is trained with two complementary classifier heads (A and B). Classifier A is presented with a localization map that highlights the most discriminative parts. The discriminative-parts' features are erased from the input features of classifier B. Accordingly, classifier B learns complementary parts of an object. GAP refers to global average pooling.

reduce the cost of data annotation; yet, they tend to localize the most discriminative part of an object, not the entire object. For instance, an attention-based approach would focus on the cat's head and ignore other parts such as legs. Thus, the result bounding box partially covers the object (*e.g.*, cat's head) while it should cover all its parts.

Attention-based approaches focus on the most discriminative part because classification CNNs focus on the most discriminative features to boost their classification performance. To mitigate this limitation, Choe and Shim [2] proposed an **a**ttention-based **d**ropout **l**ayer (ADL) while Zhang *et al.* [11] proposed **a**dversarial **co**mplementary **l**earning (ACoL). Both approaches have the same core objective, *i.e.*, hide the most discriminative feature (*e.g.*, cat's head feature) so the classifier identifies less discriminative parts. The following paragraphs review ACoL and ADL.

Zhang *et al.* [11] train a classification network with two classification heads (A and B). During training, the localization heatmap for classifier A is obtained. This localization heatmap identifies the most discriminative region. Zhang *et al.* [11] use this heatmap to guide an erasing operation on the intermediate feature maps of classifier B. This drives classifier B to discover complementary object-related regions. Thus, the two classifiers are trained to exploit complementary object regions and obtain integral object localization. Figure S3 depicts an illustration for this training strategy.

To eliminate the auxiliary classification head in ACoL, Choe and Shim [2] proposed an attention-based dropout layer (ADL). Similar to ACoL [11], ADL obtains a localization heatmap during training. From the heatmap, ADL produces both a drop-mask and an importance-map through simple-thresholding and sigmoid-activation, respectively. Applying the drop-mask drives the model

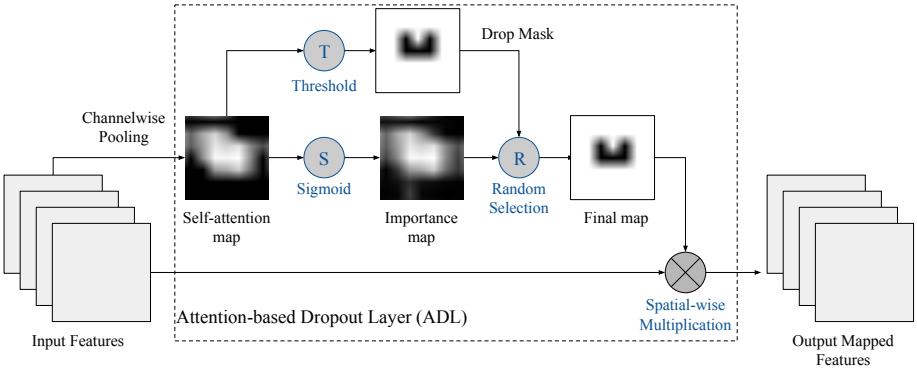

Fig. S4: ADL illustration diagram. The self-attention map is generated by channelwise average pooling of the input feature map. Based on the self-attention map, a drop-mask is produced by thresholding and an importance-map is produced by a sigmoid activation. At every training iteration, either the drop-mask or the importance-map is selected and applied to the input feature map.

to learn the less discriminative parts, which improves the localization performance. In contrast, applying the importance-map highlights the most discriminative region which improves the classification performance. During training, either the drop-mask or the importance-map is stochastically selected at each iteration, and then the selected one is applied to the input feature map through a spatialwise multiplication as shown in the next Figure S4

S2 Extended Experiments

Implementation Details For Retrieval Networks: To evaluate the localization performance quantitatively, we leverage both triplet [6] and N-pair [8] ranking losses. We use the default settings for each loss; the N-pair's embedding is unnormalized while the triplet loss's embedding is normalized to the unit-circle and a margin $m = 0.2$ is utilized. We employ ResNet-50 [3] and GoogLeNet [9] as backbones. These are standard architectures for evaluating ranking losses [8,10,5]. Both architectures are trained for 5K iterations. VGG architecture is omitted because it overfits on these datasets. Similar to Hermans *et al.* [4], the last convolution layer is followed by a global average pooling layer then a single fully connected layer, *i.e.*, a feature embedding $\in R^{128}$.
Evaluation metrics: For retrieval, we utilize both Recall@1 (R@1) and the Normalized Mutual Information (NMI) metrics. NMI score $\in [0, 1]$ measures the agreement between the true and predicted cluster assignments. NMI $= \frac{I(\Omega,C)}{\sqrt{H(\Omega)H(C)}}$, where $\Omega = \{\omega_1, .., \omega_n\}$ is the ground-truth clustering while $C = \{c_1, ...c_n\}$ is a clustering assignment for the learned embedding. $I(\bullet, \bullet)$ and $H(\bullet)$ denote mutual information and entropy, respectively. We use K-means to compute C. For localization, we follow the same evaluation procedure in [12,7] for

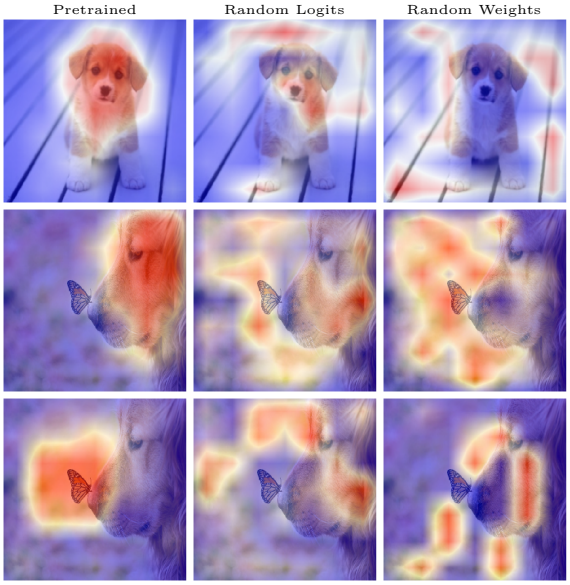

Fig. S5: First column depicts attention using a pretrained network–nothing random. Second and third columns depict attention when logits and weights (all-layers) are randomized.

classification networks. We replace the top-1 by R@1 metric to decide if the network's output is correct or not. The same IoU $> 50\%$ criterion is used to evaluate localization.

Figures S5 and S6 show how random initialization for the logit-layer, or the whole network, affects attention visualization. These sanity checks [1] emphasize a high dependency between the proposed L2-CAF and the weights of the network.

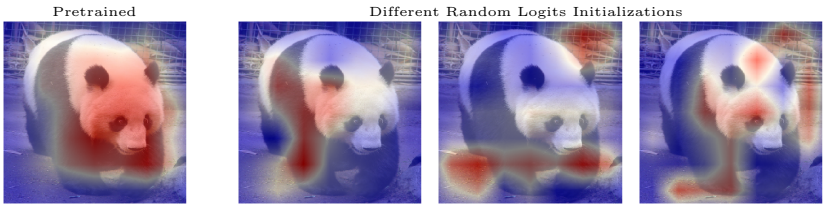

Fig. S6: Different random logit initializations, columns 2-4, generate different heatmaps.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NIPS (2018)
2. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: CVPR (2019)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
4. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
5. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR (2016)
6. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR (2015)
7. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
8. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: NIPS (2016)
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
10. Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: ICCV (2017)
11. Zhang, X., Wei, Y., Feng, J., Yang, Y., S. Huang, T.: Adversarial complementary learning for weakly supervised object localization. In: CVPR (2018)
12. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)