Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu

1. Experiment Setups and More Results

1.1. Class Activation Maps

Class Activation Maps. More examples of the class activation maps generated by GoogLeNet-GAP using the CAM technique are shown in Figure 1. The comparison results from different CNN-GAPs along with backpropagation results are shown in Figure 2.

Localization results. More localization results done by GoogLeNet-GAP using the CAM technique are shown in Figure 3

Comparison of CNN-GAP and CNN-GMP. The class activation maps generated from GoogLeNet-GAP and GoogLeNet-GMP are shown in Figure 4. The proposed Class Activation Mapping technique could be applied to CNNs both trained with global average pooling and CNNs trained with global maximum pooling. But from the comparison of the heatmaps from two networks, we can see that network with global average pooling could generate better heatmaps, highlighting larger object regions with less background noise.

1.2. Classification using deep features

In SUN397 experiment [8], the training size is 50 images per category. Experiments are ran on 10 Splits of train set and test set given in the dataset. In MIT Indoor67 experiment [6], the training size is 100 images per category. Experiment is ran on 1 split of train set and test set given in the dataset. In the Scene15 experiment [3], the training size is 50 images per category. Experiments are ran on 10 random splits of train set and test set. In the SUN Attribute experiment [5], the training size is 150 images per attribute. The report result is average precision. The splits of train set and test set are given in the paper. In Caltech101 and Caltech256 experiment [1, 2], the training size is 30 images per category. The experiments are ran on 10 random splits of train set and test set. In Stanford Action40 experiment [9], the training size is 100 images per category. Experiments are ran on 10 random splits of train set and test set. The reported result is classification accuracy. In UIUC Event8 experiment [4], training size is 70 per category and the testing size is 60 images per category. The experiments are ran on 10 random splits of train set and test set.

We plot more class activation maps from Stanford Action 40 dataset and SUN397 in Figure 6 and Figure 7.

1.3. Pattern Discovery

Discovering the informative objects in the scenes. The list of 10 scene categories with fully annotations from SUN397 are bathroom, bedroom, building facade, dining room, highway, kitchen, living room, mountain snowy, skyscraper, street, totally 4675 images. We train a one-vs-all linear SVM for each class.

Concept localization in weakly labeled images. The detail of the concept discovery algorithm is in [10]. Basically for each concept in the text which have enough images, we learn a binary classifier, then we use the weights of the classifier to generate the class activation maps for those top ranked images detected with this concept.

Weakly supervised text detector. There are 350 Google street images in the SVT dataset [7], which are taken as positive training set. We randomly select 1750 images from SUN attribute dataset [5] as negative training set, then train a linear SVM.

2. Visualizing Class-Specific Units

More class-specific units from the AlexNet*-GAP trained on ImageNet and Places are plotted in Figure 8.

References

[1] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 2007. 1

[2] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 1

[3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. CVPR*, 2006. 1

[4] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. *Proc. ICCV*, 2007. 1

Figure 1. Class activation maps generated by GoogLeNet+GAP using the CAM technique. The class categories selected from ILSVRC dataset are goldfish, ostrich, spotted salamander, Norwegian elkhound, Sealyham terrier, fireboat, and football helmet.

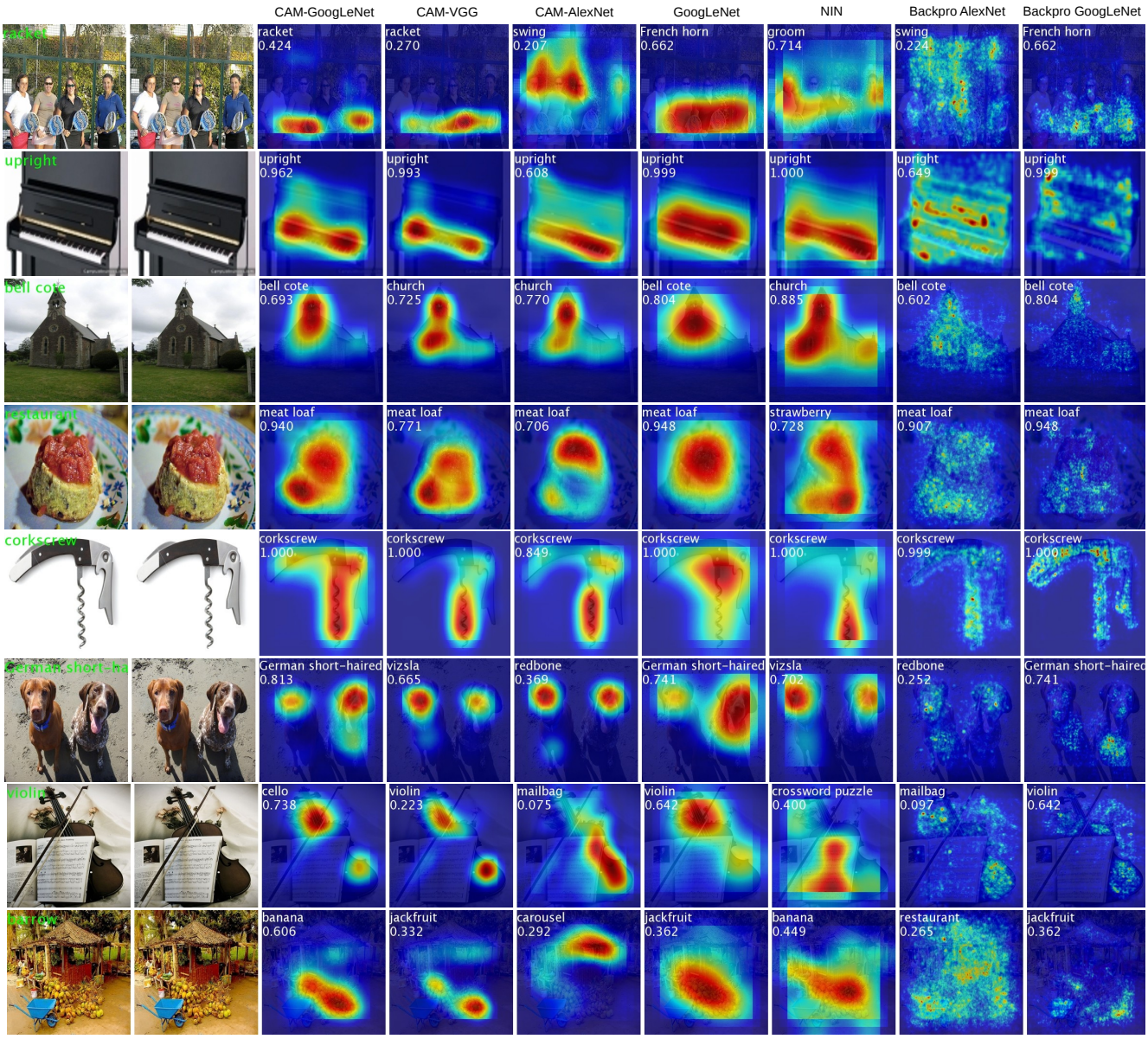

Figure 2. More examples of the class activation maps generated by several CNN+GAPs from CAM technique, along with the class-specific saliency maps generated by backpropagation on AlexNet and GoogLeNet.

[5] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. *Proc. CVPR*, 2012. 1

[6] A. Quattoni and A. Torralba. Recognizing indoor scenes. *Proc. CVPR*, 2009. 1

[7] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. *Proc. ICCV*, 2011. 1

[8] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *Proc. CVPR*, 2010. 1

[9] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. *Proc. ICCV*, 2011. 1

[10] B. Zhou, V. Jagadeesh, and R. Piramuthu. Conceptlearner: Discovering visual concepts from weakly labeled image collections. *Proc. CVPR*, 2015. 1

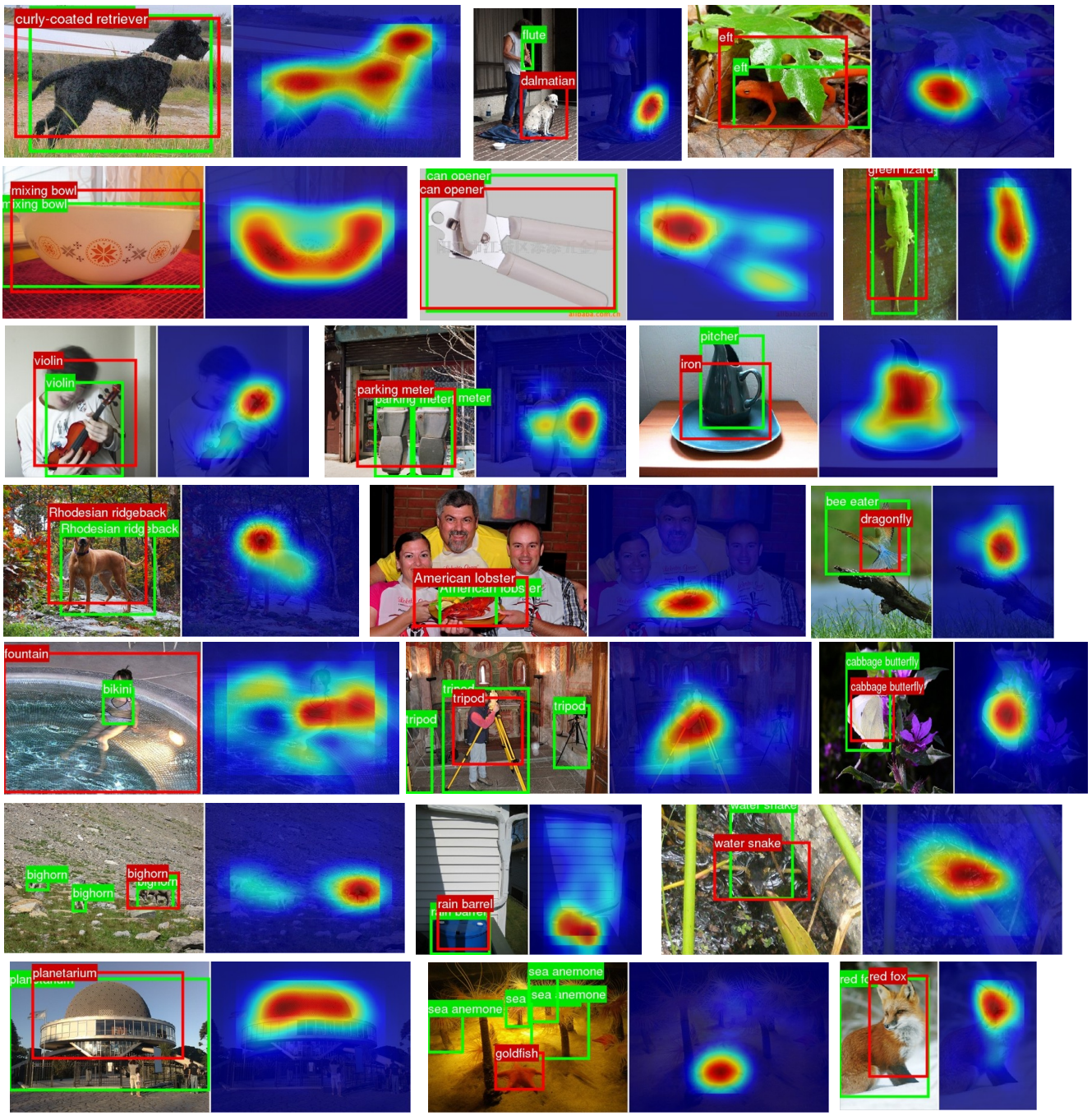

Figure 3. Localization results by GoogLeNet+GAP using the CAM technique. Groundtruth bbox and its class annotation is in green, and the predicted bbox and class label from our method is in red.

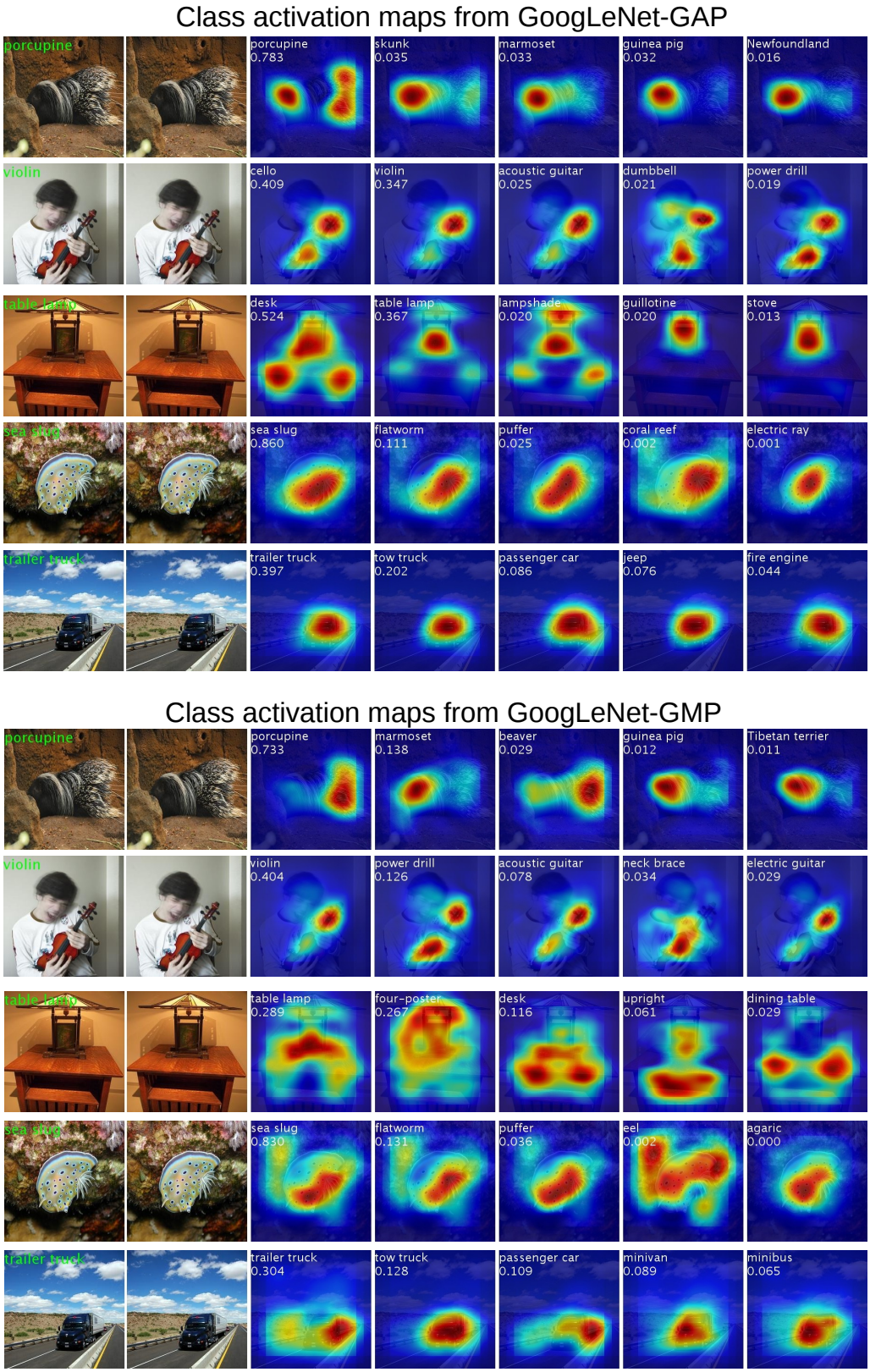

Figure 4. Class activation maps generated the GoogLeNet-GAP and GoogLeNet-GMP. Class activation map from GoogLeNet-GAP highlights more complete object regions and less background noise.

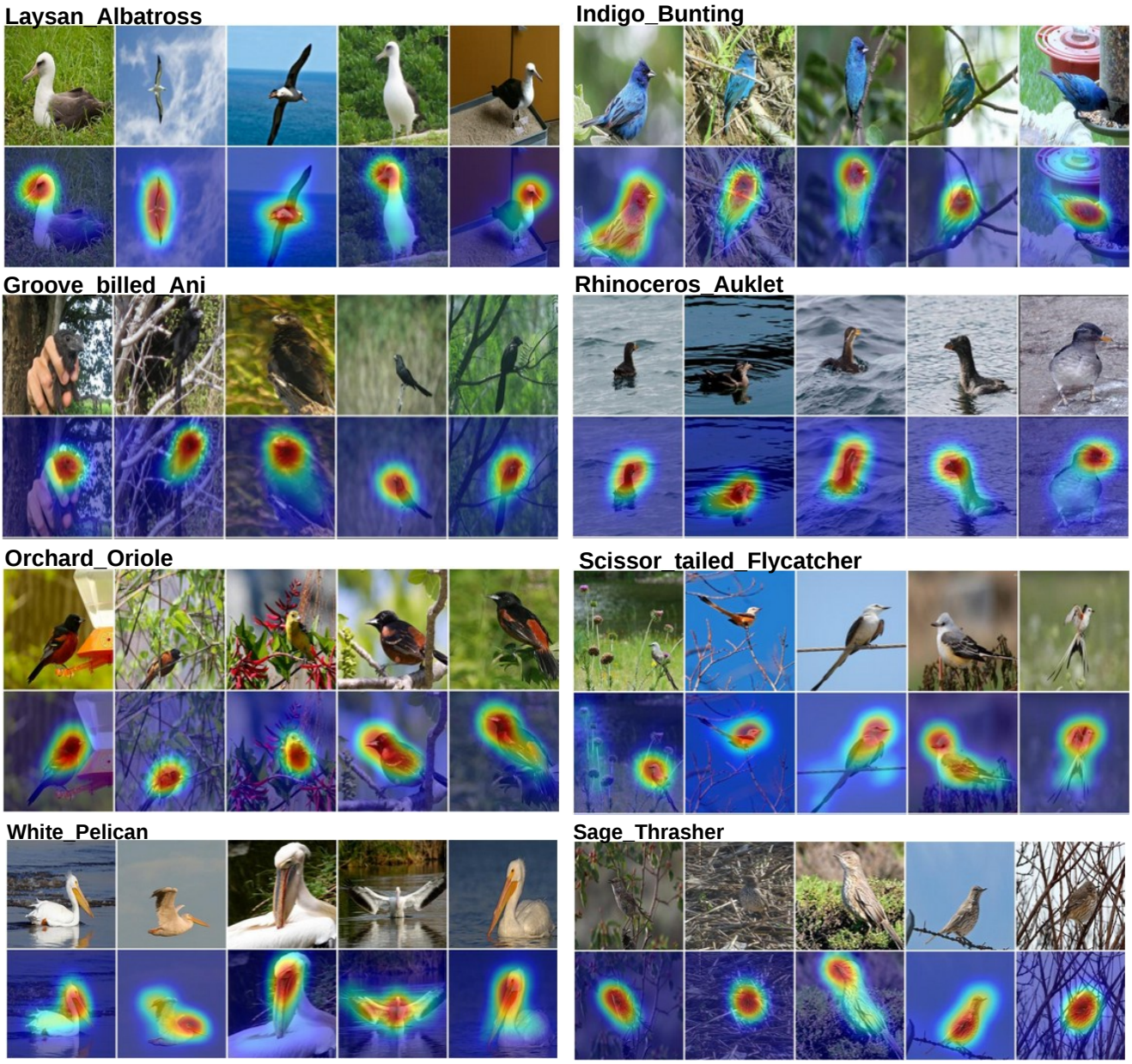

Figure 5. Class activatiom maps for 8 bird classes from CUB200.

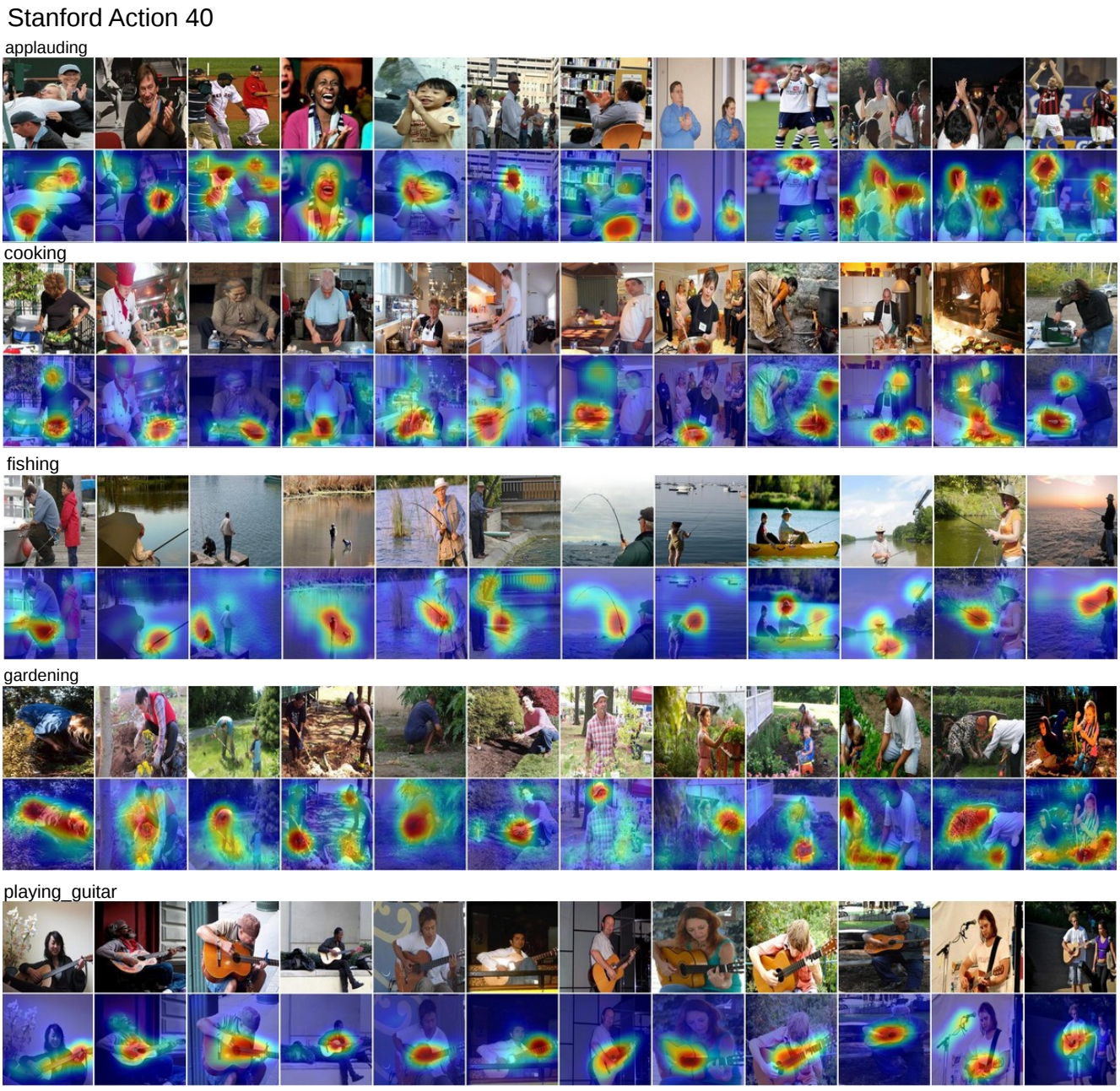

Figure 6. Examples of the class activation maps for 5 action classes from Stanford Action 40 dataset.

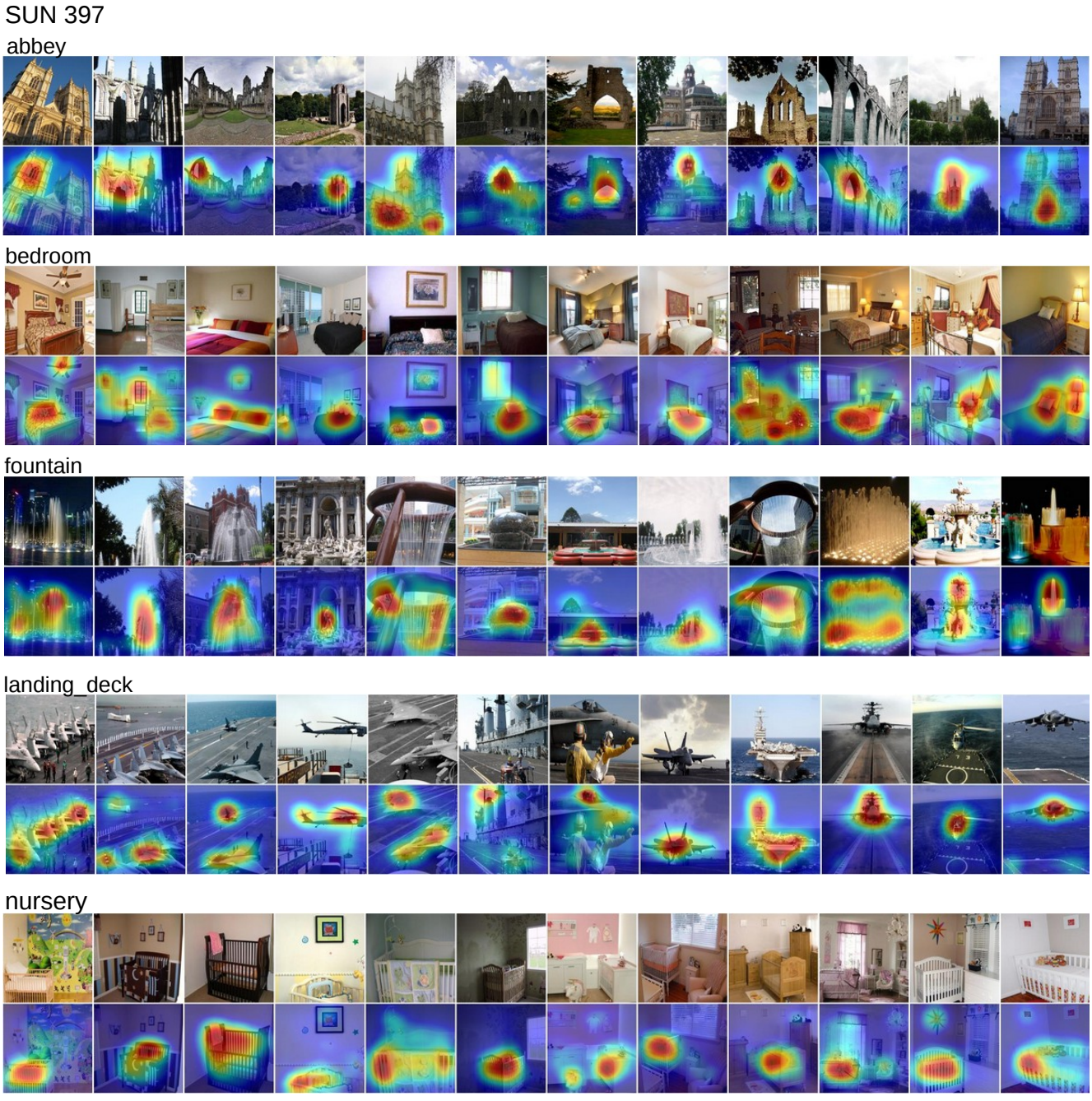

Figure 7. Examples of the class activation maps for 5 scene categoris from SUN397 dataset.

Figure 8. Examples of class-specific units from the CNN trained on ImageNet and the CNN trained on Places Database.