Supplementary materials for: Deliberative Explanations: visualizing network insecurities

Pei Wang and Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{pew062, nvasconcelos}@ucsd.edu

Appendix 1: Attribution Maps

An attribution map can be created by performing a first-order Taylor series expansion of the mapping g_p at each location (i, j). Performing the Taylor expansion in the neighborhood of a reference tensor of feature responses $\mathbf{A}^0 \in \mathbb{R}^{W \times H \times D}$ leads to

$$m_{i,j}^p = g_p(\mathbf{A}^0) + [\nabla g_p(\mathbf{A}^0)]_{i,j}^T [\mathbf{a}_{i,j} - \mathbf{a}_{i,j}^0] + \mathcal{O}^n \tag{1}$$

Several attribution approaches in the literature [8, 3, 5, 2, 4, 6] can be seen as implementing a map of this form, under the assumptions that $\mathbf{A}^0 = \mathbf{0}$ and g_p is a linear mapping that passes through the origin, i.e. $g_p(\mathbf{0}) = \mathbf{0}$ and $\nabla g_p(\mathbf{A}) = \nabla g_p(\mathbf{A}^0)$. In this case,

$$m_{i,j}^p = [\nabla g_p(\mathbf{A})]_{i,j}^T \mathbf{a}_{i,j}. \tag{2}$$

Different approaches use maps of this type, focusing on alternative ways to compute the gradient [1], e.g. using maps of the form

$$m_{i,j}^p = \langle \nabla g_p(\mathbf{A}) \rangle^T \mathbf{a}_{i,j}. \tag{3}$$

where $\langle \nabla g_p(\mathbf{A}) \rangle$ is some gradient average [3, 6]. Rather than this, we consider the second-order Taylor series expansion of g_p

$$m_{i,j}^p = g_p(\mathbf{A}^0) + [\nabla g_p(\mathbf{A}^0)]_{i,j}^T [\mathbf{a}_{i,j} - \mathbf{a}_{i,j}^0] + \frac{1}{2} [\mathbf{a}_{i,j} - \mathbf{a}_{i,j}^0]^T [\mathbf{H}(\mathbf{A}^0)]_{i,j} [\mathbf{a}_{i,j} - \mathbf{a}_{i,j}^0] + \mathcal{O}^n \tag{4}$$

where $\mathbf{H}(\mathbf{A}^0) = \nabla^2 g_p(\mathbf{A}^0)$ is the Hessian matrix of g_p at $\mathbf{A}^0$. We then make similar assumptions, namely that $\mathbf{A}^0 = \mathbf{0}$ and $g_p(\mathbf{0}) = \mathbf{0}$. However, rather than assuming that g_p is linear, we assume that it is a second order function and that $\mathbf{A} \approx \mathbf{0}$, from which it follows that $\nabla g_p(\mathbf{A}^0) \approx \nabla g_p(\mathbf{A})$ and $\mathbf{H}(\mathbf{A}^0) \approx \mathbf{H}(\mathbf{A})$. This leads to

$$m_{i,j}^p = [\nabla g_p(\mathbf{A})]_{i,j}^T \mathbf{a}_{i,j} + \frac{1}{2} \mathbf{a}_{i,j}^T [\mathbf{H}(\mathbf{A})]_{i,j} \mathbf{a}_{i,j}. \tag{5}$$

While the use of gradient and Hessian averages, as in (3), could also be used, we have not yet considered such variants.

Appendix 2: Attribute Assignment

The parts and attributes of the CUB200 dataset [7] are listed in Table 1.

Appendix 3: More Success Cases

More success case images are shown in Figure 1 for ADE20K and 2 for CUB200.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Parts	Attributes
back	back color, back pattern
beak	bill shape, bill length, bill color
belly	belly color, belly pattern
breast	breast color, breast pattern
crown	crown color, forehead color, head pattern
forehead	forehead color, head pattern
left/right eye	eye color, head pattern
left/right leg	leg color
left/right wing	wing color, wing shape, wing pattern
nape	nape color
tail	tail shape, upper tail color, under tail color, tail pattern
throat	throat color, head pattern

Table 1: Attributes assignments on CUB200 [7]

References

[1] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. A unified view of gradient-based attribution methods for deep neural networks. In *NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning*. ETH Zurich, 2017.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[4] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

[7] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Figure 1: Success case of deliberative visualizations for images from ADE20K [9].

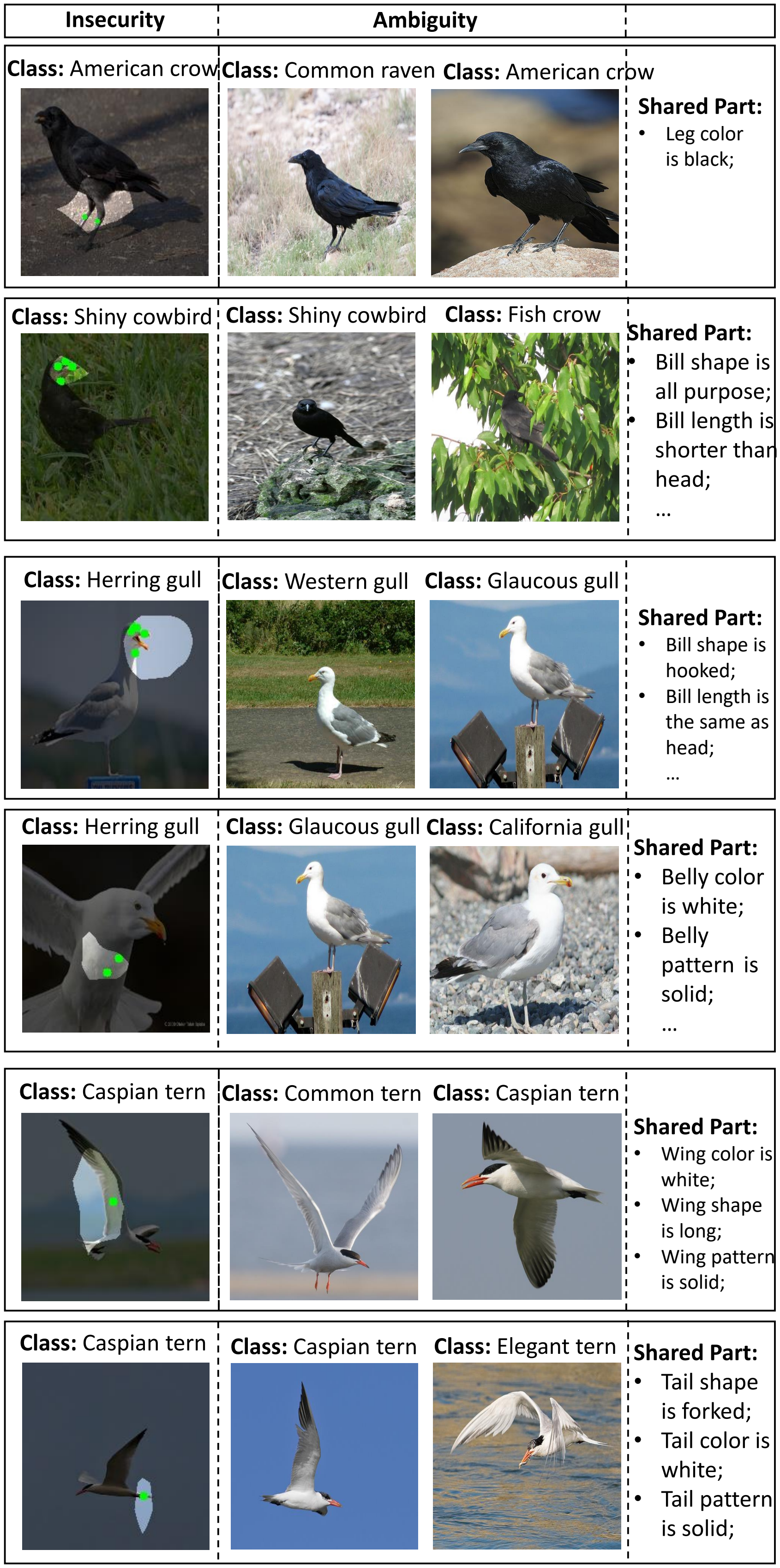

Figure 2: Success case of deliberative visualizations for images from CUB200 [7].