Interactive Classification for Deep Learning Interpretation

Angel Cabrera, Fred Hohman, Jason Lin, Duen Horng Chau
Georgia Institute of Technology
{alex.cabrera, fredhohman, jlin401, polo}@gatech.edu

Abstract

We present an interactive system enabling users to manipulate images to explore the robustness and sensitivity of deep learning image classifiers. Using modern web technologies to run in-browser inference, users can remove image features using inpainting algorithms to obtain new classifications in real time. This system allows users to compare and constrast what image regions humans and machine learning models use for classification.

1. Interactive Classification

Public trust in artificial intelligence and machine learning is essential to its prevalence and widespread acceptance. To create trust, both researchers and the general public have to understand why models behave the way they do. Existing research has used interactive data visualization as a mechanism for humans to interface with black-box machine learning models, revealing how models learn and behave [3, 7].

We have designed and developed an interactive system that allows users to experiment with deep learning image classifiers and explore their robustness and sensitivity. Users are able to remove selected areas of an image in real time with classical computer vision inpainting algorithms, including Telea [6] and PatchMatch [1], which allows them to ask a variety of "what if" questions by experimentally modifying images and seeing how the deep learning model reacts [2]. These interactions reveal a wide range of surprising results ranging from spectacular failures (e.g., a *water bottle* image becomes a *concert* when removing a person) to impressive resilience (e.g., a *baseball player* image remains correctly classified even without a glove, see. Fig. 1). The system also computes class activation maps on demand, which highlight the important semantic regions of an image a model uses for classification [8]. Combining these tools, users can develop qualitative insight into what a model sees and which features impact an image's classification.

Our system is web-based and uses the latest web technologies, including TensorFlow.js, React, and modern deep learning image classifiers: SqueezeNet [5] and MobileNet [4]. This investigation will help people explore the extent to which humans and machines think alike, and shed light on the advantages and potential pitfalls of machine learning. We demonstrate our system at CVPR 2018 for the audience to try it live. Our system is open-sourced at `https://github.com/poloclub/interactive-classification`. Watch a video demo at `https://youtu.be/llub5GcOF6w`.

Acknowledgments

This work was supported by NSF grants IIS-1563816, CNS-1704701, and TWC-1526254; NASA Space Technology Research Fellowship; and gifts from Google, Symantec, Yahoo, Intel, Microsoft, eBay, Amazon.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, 2009.

[2] F. Hohman, N. Hodas, and D. H. Chau. ShapeShop: Towards Understanding Deep Learning Representations via Interactive Experimentation. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2017.

[3] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2018.

[4] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.

[5] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size. *arXiv:1602.07360*, 2016.

[6] A. Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 2004.

[7] D. S. Weld and G. Bansal. Intelligible Artificial Intelligence. *arXiv:1803.04263*, 2018.

[8] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Modified Image

Class	↓ Confidence %	Absolute % Change
ballplayer, baseball player	48.34	+36.38
baseball	25.47	-41.13
racket, racquet	10.96	+6.20
tennis ball	1.77	-0.58
crutch	1.56	-0.66

Original Image

Class	Confidence %
baseball	66.60
ballplayer, baseball player	11.96
racket, racquet	4.76
tennis ball	2.35
crutch	2.22

Figure 1. The modified image (left) remains correctly classified as *baseball* when the ball, glove, and base are removed from the original image (right). The top five classification scores are tabulated underneath each image.

Modified Image

Class	↓ Confidence %	Absolute % Change
liner, ocean liner	51.48	+35.56
dock, dockage, docking facility	38.05	-13.95
drilling platform, offshore rig	6.53	-21.52
container ship, containership, container vessel	1.46	+1.27
fireboat	0.97	+0.85

Original Image

Class	Confidence %
dock, dockage, docking facility	52.00
drilling platform, offshore rig	28.05
liner, ocean liner	15.92
schooner	1.59
pirate, pirate ship	1.40

Figure 2. The modified image (left), originally classified as *dock* is misclassified as *ocean liner* when the masts of a couple boats are removed from the original image (right). The top five classification scores are tabulated underneath each image.