Visual Analytics for Interpretability on Deep Neural Networks

Deep neural networks are widely adopted to solve high-stakes problems. However, yet still in practice, they often remain “black boxes”. Unfortunately, this lack of interpretability raises the vital concern about deep learning adoption, since people will not use neural networks if they cannot trust them to behave in reasonable ways [1]. Even if an uninterpretable model is deployed, it could have a detrimental impact on serious situations including security and medical diagnosis; without understanding the model, people would not know how to fix it when it fails.

Existing works often explain how models make a decision for a single data instance, which tends to fall into a narrow aspect of the model operation and lose track of the big picture. For example, if an image of a dog with a ball in its mouse is given, many neuron activation based methods can misjudge that the ball is one of the important features in the classification as dog simply because several ball related neurons are highly activated [2]. General interpretability on deep neural networks, understanding how a model has learned the entire classes instead of a single image, is fundamental to get the bigger picture about the model. However, the features from all input data are commonly represented as high-dimensional tensors, which are challenging to utilize.

In this work, we propose two visual analytic techniques for general interpretability on convolutional neural networks, both of which are based on neuron activations. First, to help users observe how the neurons react to inputs, we visualize the distribution of neuron activations as a high-level summary. Second, we visualize and explain how the models recognize concepts (e.g. face) throughout the neurons and their interaction.

Observation on neuron activations. Observation is the first step in understanding phenomena. For easier exploration and observation of the activation patterns, we compress activation values of all images for all neurons in all layers into a distribution representation as a horizontal density bar graph. A horizontal bar for each neuron uses opacity to encode the density of the distribution. Users also can explore the neuron activations of multiple sets of inputs at a time (e.g. all cats and a wrongly predicted cat), which works well for large datasets where instance-by-instance analysis would be too time consuming.

Internal mechanism. To gain deeper interpretation of models with an observable pattern, it is necessary to delve into the inner mechanism that causes it. With the first approach, we can find out that specific neurons are strongly activated in general by a given class. However, it is difficult to know what the observation means or and why it happens. We provide a visual summarization of a model’s recognition process for a class as a graph, where nodes are highly activated neurons for the class and each neuron is visualized by an interpretable concept (e.g. a neuron is corresponded to face) [3]. The graph shows what concept-neurons are important and how they interact with each other to build up complicated concepts for an entire class. For example, in a graph for the cat class, it shows that face neurons are highly interacting with eyes neurons and ears neurons. The important neurons and the interactions between them are calculated from the activation maps with all input images.

These two visual analytic techniques help users understand how deep neural network models react to inputs and why it happens inside of the models. Both of them aggregate neuron activations of all input images, which are for general interpretation on the models.

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. "Why should I trust you?: Explaining the predictions of any classifier." *KDD*, 2016.

[2] Liu, M., Liu, S., Su, H., Cao, K., & Zhu, J. Analyzing the noise robustness of deep neural networks. *IEEE VAST*, 2018.

[3] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. Distill, 2(11):e7, 2017.