Visual interpretability for deep learning: A survey

Fangfang Li

June 14, 2018

1. Disentangling convolutional neural network representations into decision trees

Zhang further proposed a decision tree to encode decision modes in fully connected layers. The decision tree is not designed for classification. As shown in Fig. 1, the method mines potential decision modes memorized in fully connected layers. The decision tree organizes these potential decision modes in a coarse-to-fine manner. Furthermore, this study uses the method proposed by Zhang *et al* [4]. Instead, it is used to quantitatively explain the logic for each CNN prediction; given an input image, we use CNN to make a prediction. Visualization of CNN representations in intermediate network layers. These methods either synthesize mainly the image that maximizes the score of a given unit in a pre-trained CNN, or invert feature maps of a conv-layer back to the input image. Please see Section 2 for detailed discussions. The decision tree tells people which ters in a convlayer are used for the prediction and how much they contribute to the prediction [1]. A heat map visualizes the spatial distribution of the top patterns in the layer of the explanatory graph with the highest inference scores. References to color refer to the online version of this figure.

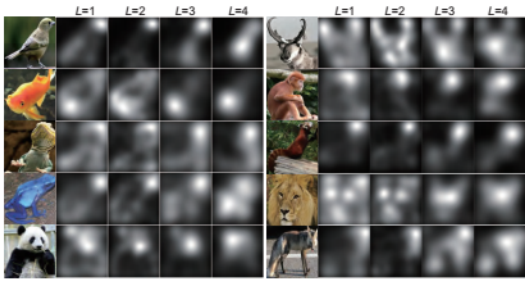

Figure 1. Heat maps of patterns.

As shown in Fig. 2, the method mines potential decision modes memorized in fully connected layers. The decision tree organizes these potential decision modes in a coarse-to-fine manner. Furthermore, this study uses the method proposed by Zhang.

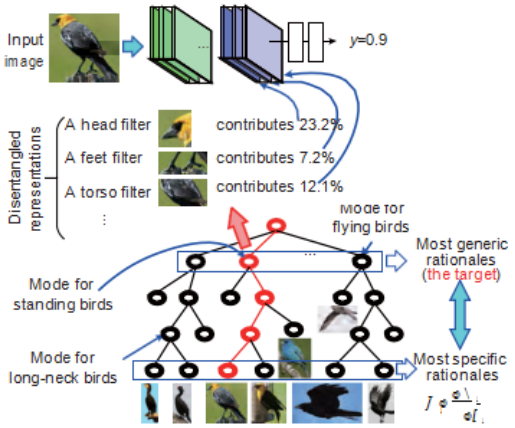

Figure 2. Decision tree that explains a convolutional neural network (CNN) prediction at the semantic level.

2. Learning neural networks with interpretable/disentangled representations

Almost all methods mentioned in Sections focus on the understanding of a pre-trained network. In this section, we review studies of learning disentangled representations of neural networks, where representations in middle layers are no longer a black box but have clear semantic meanings. Compared with the understanding of pre-trained networks, learning networks with disentangled representations present more challenges. Up to now, only a few studies have been published in this direction.

3. Prospective trends and conclusions

In this paper, we have reviewed several research directions within the scope of network interpretability [2]. Visualization of a neural unital patterns was the starting point of understanding network representations in the early years. Then, people have grad-

ually developed methods to analyze feature spaces of neural networks and diagnose potential representation hidden inside neural networks [3]. At present, disentangling chaotic representations of conv-layers into graphical models or symbolic logic has become an emerging research direction to open the black-box of neural networks. The approach for transforming a pre-trained CNN into an explanatory graph was proposed. It exhibited a signial ant efficiency in knowledge transfer and weakly-supervised learning.

End-to-end learning of interpretable neural networks, whose intermediate layers encode comprehensible patterns, is also a prospective trend. Interpretable CNNs have been developed, where each in high conv-layers represents a special object part.

Furthermore, based on interpretable representations of CNN patterns, semantic-level middle-to-end learning was proposed to speed up the learning process. Compared with traditional end-to-end learning, middle-to-end learning allows human interactions to guide the learning process and can be applied with a few annotations for supervision. In the future, we believe that the middle-to-end learning will continuously be a fundamental research direction. In addition, based on the semantic hierarchy of an interpretable network, debugging CNN representations at the semantic level will create new visual applications.

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015. 1

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[3] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *IEEE International Conference on Computer Vision*, pages 2056–2063, 2014. 2