Visual interpretability for deep learning: A survey

Fangfang Li

June 12, 2018

Abstract

This paper reviews recent studies in understanding neural-network representations and learning neural networks with interpretable/disentangled middle-layer representations. Although deep neural networks have exhibited superior performance in various tasks, interpretability is always Achilles' heel of deep neural networks. At present, deep neural networks obtain high discrimination power at the cost of a low interpretability of their black-box representations. We believe that high model interpretability may help people break several bottlenecks of deep learning, learning from a few annotations, learning via human–computer communications at the semantic level, and semantically debugging network representations. We focus on convolutional neural networks (CNNs), and revisit the visualization of CNN representations, methods of diagnosing representations of pre-trained CNNs, approaches for disentangling pre-trained CNN representations, learning of CNNs with disentangled representations, and middle-to-end learning based on model interpretability. Finally, we discuss prospective trends in explainable artificial intelligence.

1. Introduction

Convolutional neural networks (CNNs) have achieved superior performance in many visual tasks, such as object classification and detection. However, the end-to-end learning strategy makes CNN representations a black box. Except for the final network output, it is difficult to understand the logic of CNN predictions hidden inside the network [1]. In recent years, a growing number of researchers have realized that high model interpretability is of significant value in both theory and practice, and have developed models with interpretable knowledge representations [2].

Finally, Zhang presented a method to discover potential, biased representations of a CNN. Figure 1 biased representations of a CNN trained to estimate face attributes. When an attribute usually co-appears with specific visual features in training images, CNN may use such co-appearing features to represent the attribute. When the co-appearing features used are not semantically related to the target attribute, these features can be considered as biased representations.

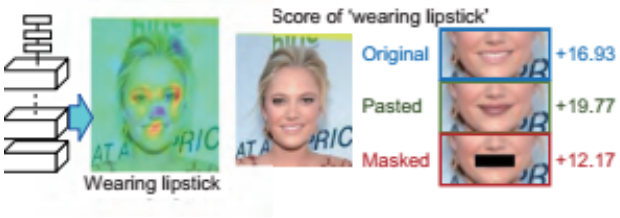

Figure 1. Biased representations in a convolutional neural network.

2. Disentangling convolutional neural network representations into explanatory graphs and decision trees

As shown in Figure 2, each filter in a high conv-layer of a CNN usually represents a mixture of patterns.

Figure 2. Feature maps of a ftlter obtained using different input images.

To visualize the feature map, the method propagates receptive fields of activated units in the feature map

back to the image plane [3]. In each sub-feature, the filter is activated by various part patterns in an image. This makes it difficult to understand the semantic meaning of a filter. References to color refer to the online version of this figure.

References

[1] 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

[2] S. M. Plis and V. D. Hjelm. Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience*, 11(8):229–235, 2014. 1

[3] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 81(3):85–117, 2015. 2