Quan-shi Zhang, Song-chun Zhu, 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1): 27-39.
https://doi.org/10.1631/FITEE.1700808

Visual interpretability for deep learning: a survey

Key words: Artificial intelligence; Deep learning; Interpretable model

Corresponding author: Quan-shi ZHANG
E-mail: zhangqs@ucla.edu

ORCID: http://orcid.org/0000-0002-6108-2738

Motivation

- Although deep neural networks have exhibited superior performance in various tasks, interpretability is always the Achilles' heel of deep neural networks. We believe that high model interpretability may help people break several bottlenecks of deep learning, e.g., 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.

The Survey

- In this study, we conduct a survey of current studies in understanding neural-network representations and learning neural networks with interpretable/disentangled feature representations. We can roughly define the scope of the review into the following five research directions:

  - Visualization of CNN representations in intermediate network layers.
  - Diagnosis of CNN representations.
  - Disentanglement of the ‘mixture of patterns’ in each filter of CNNs.
  - Building explainable models.
  - Semantic-level middle-to-end learning via human–computer interaction.

Visualization of convolutional neural network representations

- Visualization of filters in a CNN is the most direct way to explore visual patterns hidden inside a neural unit.
  - Gradient-based methods are the mainstream of network visualization.
  - The up-convolutional network is another typical technique to visualize CNN representations.
  - In addition, Bau et al. (2007) proposed a method to accurately compute the image-resolution receptive field of neural activations in a feature map.

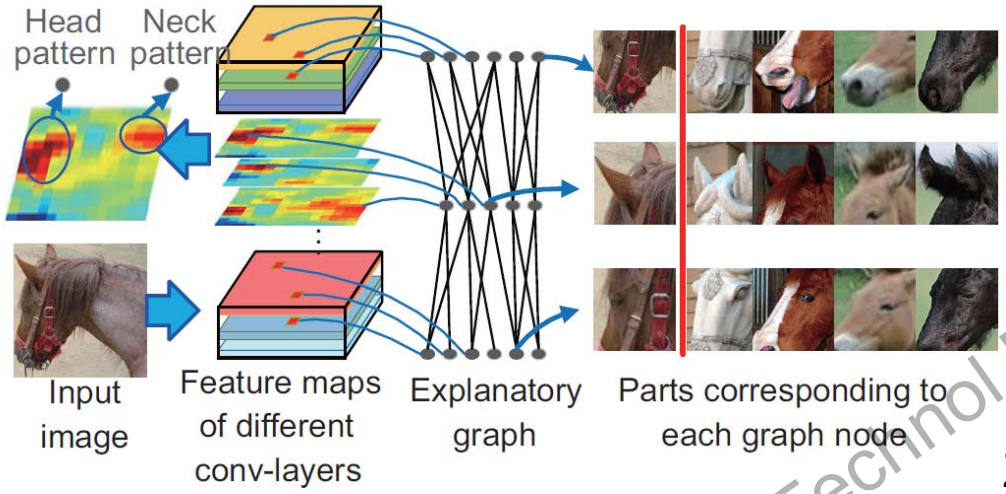
Bau D, Zhou B, Khosla A, et al., 2017. Network dissection: quantifying interpretability of deep visual representations. IEEE Conf on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2017.354.

Diagnosis of convolutional neural network representations

- Some studies analyze CNN features from a global view, e.g., exploring semantic meanings of each filter and analyzing the transferability of filter representations in intermediate conv-layers.
- The second research direction extracts image regions that directly contribute the network output for a label/attribute to explain CNN representations of the label/attribute.
- The estimation of vulnerable points in the feature space of a CNN is also a popular direction for diagnosing network representations.
- The fourth research direction is to refine network representations based on the analysis of network feature spaces.
- Finally, Zhang et al. (2018) presented a method to discover potential, biased representations of a CNN.

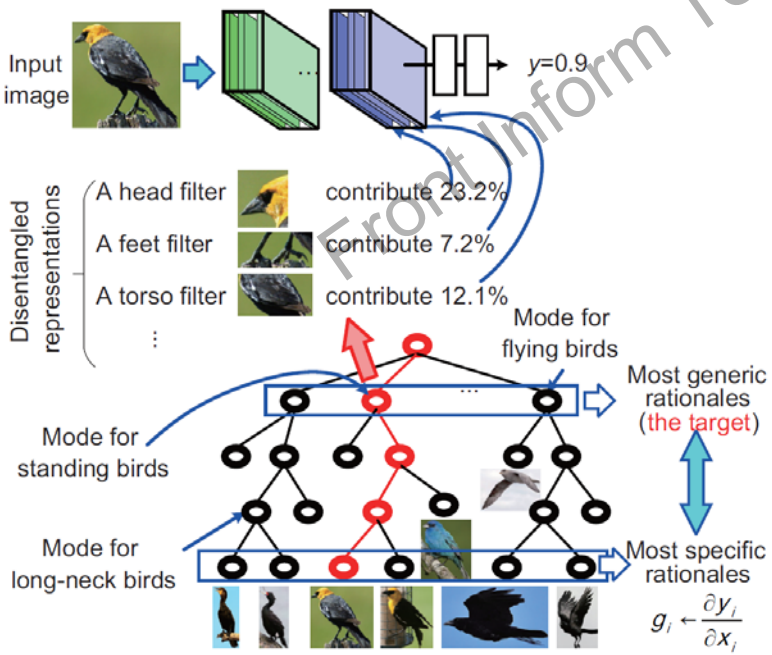
Zhang Q, Wang W, Zhu SC, 2018. Examining CNN representations with respect to dataset bias, AAAI. https:// arXiv:1710.1057

Disentangling CNN representations into explanatory graphs and decision trees



Disentangling convolutional neural network representations into explanatory graphs

Zhang Q, Cao R, Shi F, et al., 2018. Interpreting CNN knowledge via an explanatory graph, AAAI.



Learning neural networks with interpretable/disentangled representations

Zhang Q, Yang Y, Wu YN, et al., 2018. Interpreting cnns via decision trees. https:// arXiv:1802.00121

Learning neural networks with interpretable/disentangled representations

- Interpretable convolutional neural networks
  - Zhang Q, Wu YN, Zhu SC, 2018. Interpretable convolutional neural network, in CVPR.
- Interpretable region-based convolutional neural network
  - Wu TF, Li X, Song X, et al., 2017. Interpretable R-CNN, https:// arXiv:1711.05226.
- Capsule networks
  - Sabour S, Frosst N, Hinton GE, Dynamic routing between capsules, in Proc. NIPS, 2017.
- Information maximizing generative adversarial nets
  - Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P, Infogan: interpretable representation learning by information maximizing generative adversarial nets, in NIPS 2016.

Network interpretability for middle-to-end learning

- Active question-answering for learning And-Or graphs
  - Based on the semantic And-Or representation, Zhang et al. developed a method to use active question-answering to semanticize neural patterns in conv-layers of a pre-trained CNN and built a model for hierarchical object understanding.
  - Zhang Q, Cao R, Wu YN, et al., 2017. Mining object parts from CNNs via active question-answering, in CVPR.
- Interactive manipulations of convolutional neural network patterns
  - Let a CNN be pre-trained using annotations of object bounding boxes for object classification. Zhang et al. have explored an interactive method to diagnose knowledge representations of a CNN, to transfer CNN patterns to model object parts.
  - Zhang Q, Cao R, Zhang S, et al., Interactively transferring CNN patterns for part localization, https:// arXiv:1708.01783.

Conclusions

- In this paper, we have reviewed several research directions within the scope of network interpretability. Visualization of a neural unit's patterns was the starting point of understanding network representations in the early years. Then, people have gradually developed methods to analyze feature spaces of neural networks and diagnose potential representation flaws hidden inside neural networks. At present, disentangling chaotic representations of conv-layers into graphical models or symbolic logic has become an emerging research direction. End-to-end learning interpretable neural networks, whose intermediate layers encode comprehensible patterns, are also a prospective trend. Furthermore, based on network interpretability, semantic-level middle-to-end learning was proposed to speed up the learning process.