DS530.V Paper Presentation Summary: Network Dissection: Quantifying Interpretability of Deep Visual Representations

Utku Acar
Ozyegin University
utku.acar@ozu.edu.tr

I. Problem Studied

This research studies the alignment between individual hidden units and a set of semantic ideas, addressing the challenge of quantifying the interpretability of Convolutional Neural Networks (CNNs). Its goal is to assess how interpretable structure is emerging in deep networks and how it can affect disentangled representations.

II. Contribution of the Paper

Network Dissection, a wide paradigm for analyzing the interpretability of CNN hidden representations, is proposed in the study. The research investigates interpretability-related assumptions by analyzing the latent representations of networks trained to complete separate supervised and self-supervised training tasks. This study looks at how training iterations, network width and depth, batch normalization, and dropout affect interpretability. Using the Broadly and Densely Labeled Dataset (Broden), the proposed technique evaluates the semantics of hidden units at each intermediate convolutional layer, correlating them with notions that are understandable to humans. The method presented in the research for scoring the semantic meaning of hidden units is significant since it allows for the analysis of interpretability in deep visual representations. The strategy employed takes a deliberate approach to

III. Methodology Used

The following approaches were employed in the research:

A. Network Dissection Framework:

In order to measure the interpretability of Convolutional Neural Networks (CNNs), the research presents the Network Dissection framework, which operates by matching individual hidden units with a bunch of semantic segmentation filters. This framework is based on a three-step strategy that is similar to the methods that neuroscientists adopt to comprehend concerns regarding representation in biological neurons. These are ”Identifying a broad set of human-labeled visual concepts”, ”Gathering hidden variables’ response to known concepts”, and ”Quantifying alignment of hidden variable-concept pairs”.

B. Human Evaluation of Interpretations:

The study involves using Amazon Mechanical Turk (AMT) to conduct human evaluations of the unit interpretations discovered by the approach. Raters were shown images with highlighted patches indicating the most highly activating regions for each convolutional unit in AlexNet trained on Places205 and asked whether a particular label described most of the image patches. This human evaluation is used to validate the accuracy of the method’s unit interpretations.

C. Measurement of Axis-Aligned Interpretability:

The study tests the significance of assigning an interpretable idea to an individual unit by examining two theories about interpretability formation in hidden layer units. Hypothesis 1 suggests interpretable units emerge because interpretable concepts appear in most directions in representation space, challenging the notion that interpretations of single units on a natural basis may not be meaningful. Hypothesis 2 suggests interpretable alignments are unusual and emerge because learning converges to a special basis that aligns explanatory factors with individual units, suggesting the natural basis represents a meaningful decomposition learned by the network. The study aims to validate or refute these hypotheses through rigorous testing and analysis.

D. Network Architectures and Supervisions:

The research focuses on how network structure and training supervision impact the interpretability of learned representations. It investigates alternate Convolutional network models and supervisions via network dissection, focusing on the last convolutional layer of each CNN where semantic detectors are most prevalent. The goal is to comprehend how these variables influence the emergence and alignment of semantic concepts inside learned representations. The complete review shines a light on CNN models and training approaches that go beyond discriminative power, illustrating the disentangled nature of representations and the factors that influence their interpretability.

These methodologies are employed to systematically evaluate the interpretability of CNN representations and understand the factors that influence their interpretability.

IV. Results

The results show that interpretability is not an axis-independent phenomenon and that various training models and regularizations, such as dropout or batch normalization, can have a significant impact on the interpretability of hidden unit representations. Because of its feed-forward mechanism for transferring weights of shallow layers to deeper layers, ResNet152 was the best resulting model of all the models. The results reveal that representations at different levels can unravel different forms of meaning and that interpretability is an axis-aligned property of a representation that can be destroyed by rotation without affecting discriminative power. The study gives insights into the properties of CNN models and training methods that go beyond evaluating discriminative capabilities, providing information on the disentangled nature of the representations. The proposed paradigm has become a useful tool for interpretability.