GAN Dissection

Visualizing and Understanding Generative Adversarial Networks

David Bau[12] Jun-Yan Zhu[1] Hendrik Strobelt[23] Bolei Zhou[4] Joshua B. Tenenbaum[1] William T. Freeman[1] Antonio Torralba[12]

[1]Massachusetts Institute of Technology

[2]MIT-IBM Watson AI Lab

[3]IBM Research

[4]The Chinese University of Hong Kong

Outline

- Network Dissection
- GAN Dissection
- Results and Applications

A method of interpreting generative models
Focus is on the generator
Applied in two parts: dissection and intervention

Network Dissection

Three step process applied:

1. Identify a broad set of concepts (segmentation maps), could be specific objects, textures, colors, etc
2. Gather hidden variables' response to known concepts
3. Quantify alignment of hidden variable-concept pairs

”In a fully interpretable local coding such as a one-hot encoding, each variable will match with exactly one concept”

but partially nonlocal representations learned in interior layers, and emergent concepts often align with a combination of several hidden units

Ideally we’d like to have disentangled object representations in the network

So we need a set of human-interpretable concepts
Attained from semantic segmentation (picture)
This method seeks to measure agreement between *activation units* and (labeled concepts) attained from the segmentation Then quantify that agreement to identify highly-activated units for specific concepts as well as quantify the network interpretability as a whole

Method (of Network Dissect)

"Dissect" the network at a specific layer

Look through units of the feature map (depth slices of the output which share the same filter and look for the same feature)

For a specific image, obtain the activation at that unit by running it through the network

Upscale and threshold the activation map into a binary segmentation of its own

For each concept in the semantic segmentation of the image, measure alignment with between the concept and the binary segmentation

Alignment quantified using Intersection over Union (pictures)