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

"Observations of hidden units in large deep neural networks have revealed that human-interpretable concepts sometimes emerge as individual latent variables within those networks"
include introductory network dissection talk because its precursor to GAN dissection but not too much since we covered it in class

Three step process:

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

Perhaps recap the process of gathering activation maps of each unit, determining quantile $P(\alpha_k > T_k) = 0.005$, upscaling low-resolution activation map to input-resolution annotation mask for a concept, thresholding the upscaled activation map by T_k, then evaluating against every concept c in the data set, and then intersection over union score (which has more intuitive understanding visually than formulaic, but it is metrically similar to mutual information)