Tutorial: GAN Dissection

What is learned inside a GAN?

David Bau


We will go through a Jupyter notebook. Ideal machine has git, conda, and a GPU.

Generative Adversarial Networks

The generator learns to beat a discriminator...

Goal: make D think it’s real!

Input: random vector \( z \)

Generator G trained to fool D

Discriminator D trained to tell G’s fakes from real

In the end the generator can synthesize terrific images. But how does it work? What does it actually learn?

What do we learn?

Our Plan: GAN Dissection

1. Run the network
2. Collect activations
3. Analyze each unit

Each layer is a learned nonlinear convolution

Each representation is a set of channels

Each step increases data resolution, reduces depth

Input: random vector $z$

What happens inside?

Output: random realistic image $x$
You can play along. Tutorial setup...


git clone --branch tutorial https://github.com/CSAILVision/gandissect.git
cd gandissect
script/setup_env.sh  # Create a conda environment with dependencies
script/make_dirs.sh   # Create the dataset and dissect directories
source activate netd   # Enter the conda environment
pip install -v -e .     # Link the local netdissect package into the env
cd notebooks
jupyter notebook &     # Run jupyter

Then run the notebook called: dissect_progan.ipynb