GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

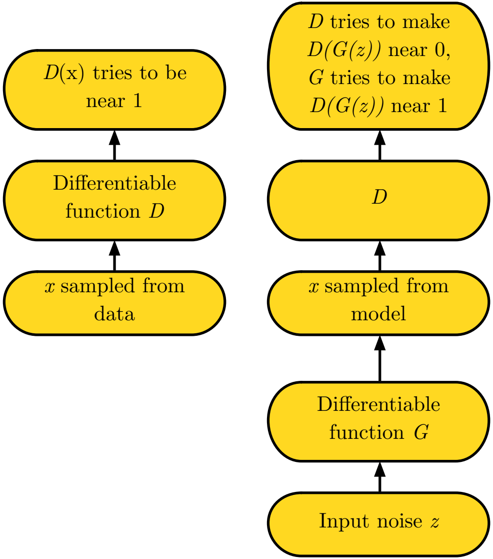
Martin Šafránek

Faculty of Information Technology
Czech Technical University in Prague

Parts

- GAN overview
- GAN dissection (paper)

Adversarial Nets Framework



D(x) tries to be near 1
Differentiable function D
x sampled from data
D tries to make $D(G(z))$ near 0, G tries to make $D(G(z))$ near 1
D
x sampled from model
Differentiable function G
Input noise z

Minimax Game

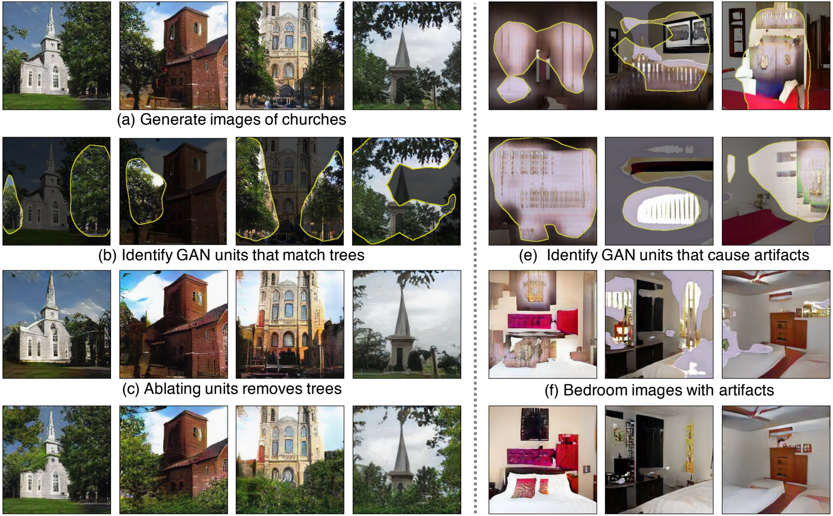
$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{x \sim p_{\mathrm{data}}} \log D(x) - \frac{1}{2}\mathbb{E}_{z} \log\left(1 - D(G(z))\right)$$

$$J^{G} = -J^{D}$$

- Generator minimizes the log-probability of the discriminator being correct
- Equilibrium if the discriminator is unable to differentiate between real and generated input

The paper

- presents method for visualizing and understanding GAN
- learned GAN contains variables for doors, trees, ...
- can interactively manipulate objects in a scene



(a) Generate images of churches

(b) Identify GAN units that match trees

(c) Ablating units removes trees

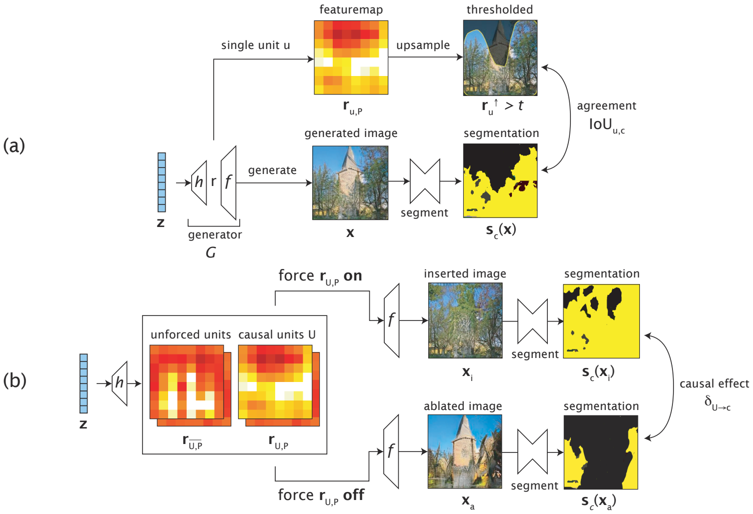
(d) Activating units adds trees

(e) Identify GAN units that cause artifacts

(f) Bedroom images with artifacts

(g) Ablating “artifact” units improves results

Dissection architecture



(a)
featuremap
thresholded
single unit u
upsample
$\mathbf{r}_{u,\mathbb{P}}$
$\mathbf{r}_u^{\uparrow} > t$
agreement
$\mathrm{IoU}_{u,c}$
generated image
segmentation
generate
h
r
f
segment
$\mathbf{z}$
generator
G
$\mathbf{x}$
$\mathbf{s}_c(\mathbf{x})$
(b)
force $\mathbf{r}_{U,\mathbb{P}}$ on
inserted image
segmentation
unforced units
causal units U
f
segment
$\mathbf{x}_i$
$\mathbf{s}_c(\mathbf{x}_i)$
causal effect
$\delta_{U\rightarrow c}$
$\mathbf{z}$
h
$\mathbf{r}_{\overline{U},\mathbb{P}}$
$\mathbf{r}_{U,\mathbb{P}}$
ablated image
segmentation
f
segment
force $\mathbf{r}_{U,\mathbb{P}}$ off
$\mathbf{x}_a$
$\mathbf{s}_c(\mathbf{x}_a)$