GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

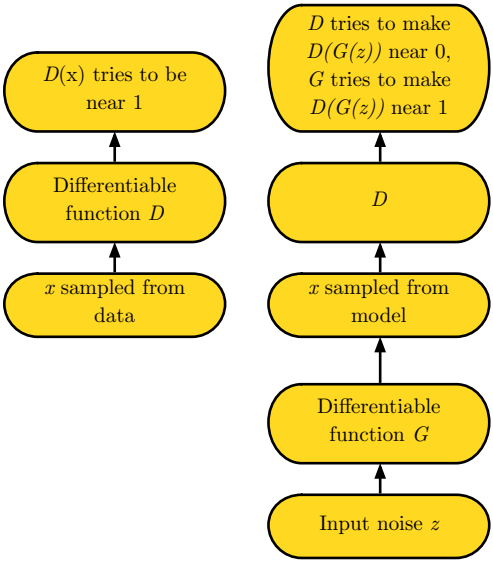
Martin Šafránek

Faculty of Information Technology
Czech Technical University in Prague

Parts

- GAN overview
- GAN dissection (paper)

Adversarial Nets Framework



Minimax Game

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

- 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
- assumes GAN generator is implemented as Convolutional neural network
- presents that learned GAN contains variables for doors, trees, ...
- shows how we 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

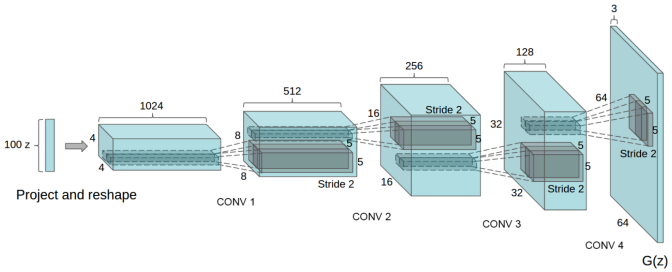
(e) Identify GAN units that cause artifacts

(f) Bedroom images with artifacts

(g) Ablating “artifact” units improves results

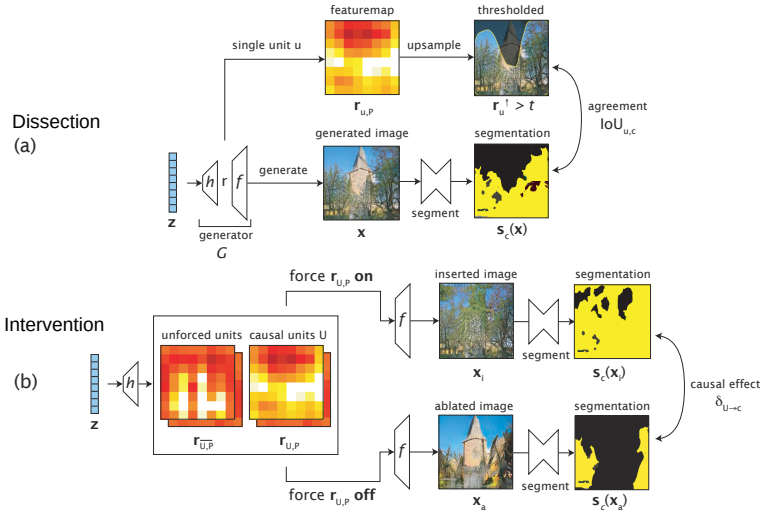
Terminology

- tensor r – output from the layer of a generator (featuremap)
  - r has all the data neccessary to produce output image x
- unit – channel of the featuremap
- concept $\mathbf{c} \in \mathbb{C}$ (chair, door, ...)
  - segmented using external tool (~300 concepts)
- $\mathbf{r}_{\mathbb{U},P} = (r_{U,P}, r_{\overline{U},P})$ – featuremap factorization



GAN generator layers [Source]

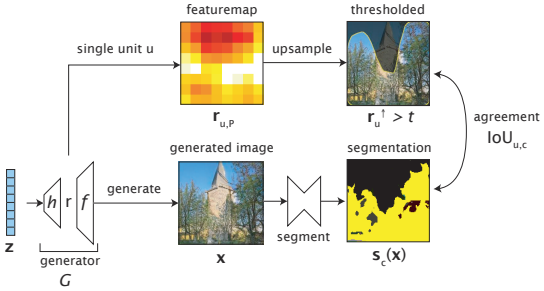
Architecture



Dissection 1/2

- which units correlate with which concept
- semantic segmentation $s_c(x)$ of generated image x

1. Upsample unit to output image resolution
   - using bilinear interpolation
2. Threshold unit
   - constant $t_{u,c}$ same for each concept c
   - learned by maximizing Information Quality Ratio (IQR) on validation set



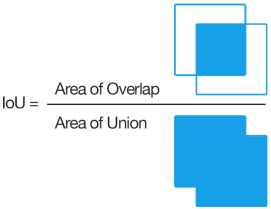
Dissection architecture

Dissection 2/2

3 Compute IoU score

- label each concept with unit that matches it best (largest IoU score)

$$\text{IoU}_{u,c} \equiv \frac{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \wedge \mathbf{s}_c(\mathbf{x}) \right|}{\mathbb{E}_{\mathbf{z}} \left| (\mathbf{r}^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \vee \mathbf{s}_c(\mathbf{x}) \right|}$$

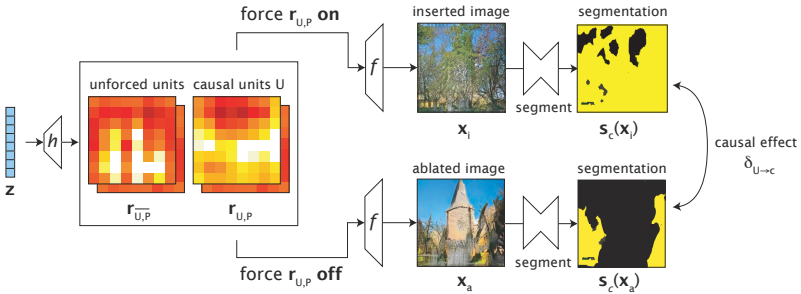


Thresholding unit #65 layer 3 of a dining room generator matches 'table' segmentations with IoU=0.34.

Thresholding unit #37 layer 4 of a living room generator matches 'sofa' segmentations with IoU=0.29.

Intervention 1/3

- unit that correlates highly with a generated ouput might not cause it's generation
- intervention identifies combination of units that cause object generation
  - select set of units U and concept c from image $\rightarrow$ force U on/off $\rightarrow$ is c still present ?



Intervention architecture

Intervention 2/3

1. force units U on position P on/off
   - an object is caused by U if the object appears in x_i and disappears from x_a

Original image : $\mathbf{x} = G(\mathbf{z}) \equiv f(\mathbf{r}) \equiv f(\mathbf{r}_{\mathrm{U,P}}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

Image with U ablated at pixels P : $\mathbf{x}_a = f(\mathbf{0}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

Image with U inserted at pixels P : $\mathbf{x}_i = f(\mathbf{k}, \mathbf{r}_{\overline{\mathrm{U,P}}})$

(a) Original image (b) Units that match trees (dissection) (c) Ablating tree units (d) Inserting tree units

Intervention 3/3 — Finding sets of units with high ACE

2. Measure average causal effect (ACE)

- ACE: $\delta_{\mathrm{U}\to c} \equiv \mathbb{E}_{\mathbf{z},\mathrm{P}}[\mathbf{s}_c(\mathbf{x}_i)] - \mathbb{E}_{\mathbf{z},\mathrm{P}}[\mathbf{s}_c(\mathbf{x}_a)]$
  - $s_c(x)$ segmentation indicating the presence of class c in the image x at P
- ineffective, as there are $\binom{d}{|U|}$ subsets for every $d \in \{1, \ldots, |U|\}$
- instead optimize $\alpha \in [0,1]^d$

$$\begin{aligned}
&\text{Image with partial ablation at pixels P :} && \mathbf{x}'_a = f((\mathbf{1}-\boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}}) \\
&\text{Image with partial insertion at pixels P :} && \mathbf{x}'_i = f(\boldsymbol{\alpha} \odot \mathbf{k} + (\mathbf{1}-\boldsymbol{\alpha}) \odot \mathbf{r}_{\mathbb{U},\mathrm{P}},\ \mathbf{r}_{\mathbb{U},\overline{\mathrm{P}}}) \\
&\text{Objective :} && \delta_{\boldsymbol{\alpha}\to c} = \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_i)\right] - \mathbb{E}_{\mathbf{z},\mathrm{P}}\left[\mathbf{s}_c(\mathbf{x}'_a)\right],
\end{aligned}$$

- optimize with L2 loss: $\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}}(-\delta_{\boldsymbol{\alpha}\to c} + \lambda||\boldsymbol{\alpha}||_2)$
- rank units by α^*_u and achieve stronger casual effect

Results

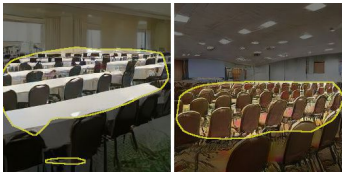
Interpretable units for different scene categories

conference rm

table #96 iou=0.30

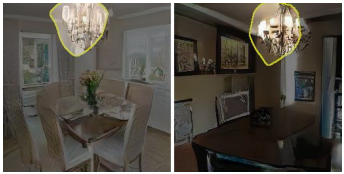
person-b #91 iou=0.21

seat #83 iou=0.13



dining room

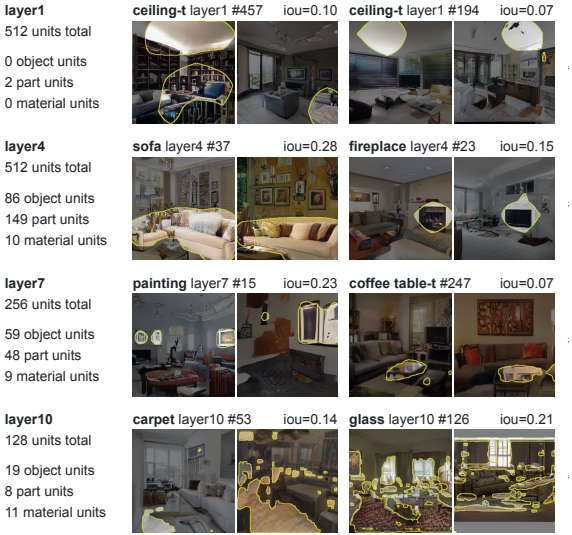
chandelier-l #184 iou=0.21



chair-l #456 iou=0.19

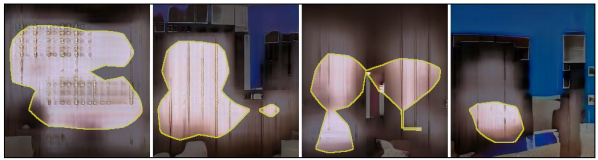
table #89 iou=0.31

Interpretable units for different network layers



Diagnosing and improving GANs

- identify artifacts with human annotation
- 10 minutes to locate 20 artifact-causing units
- ablate found artifacts



(a) Example artifact-causing units

(b) Bedroom images with artifacts

(c) Ablating “artifact” units improves results

Locating casual units with ablation

- force erasing/size reduction
- Does GAN learns patterns ? eg „*All bedrooms must have windows.*“

Figure 10: Comparing the effect of ablating 20 window-causal units in GANs trained on five scene categories. In each case, the 20 ablated units are specific to the class and the generator and independent of the image. In some scenes, windows are reduced in size or number rather than eliminated, or replaced by visually similar objects such as paintings.

Characterizing contextual relationships via insertion

- force insertion of features into specific locations in scenes
- GAN forces realtionships eg „*Doors can't be added to sky.*“, because the choice is vetoed later

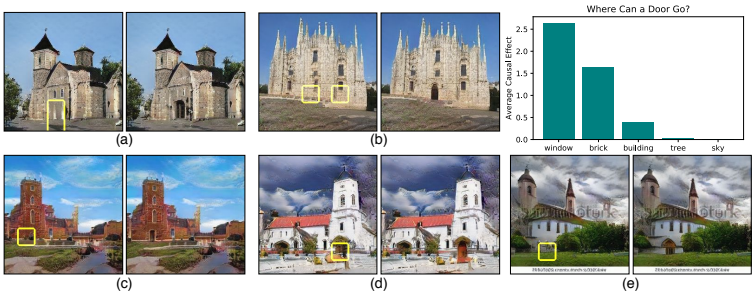
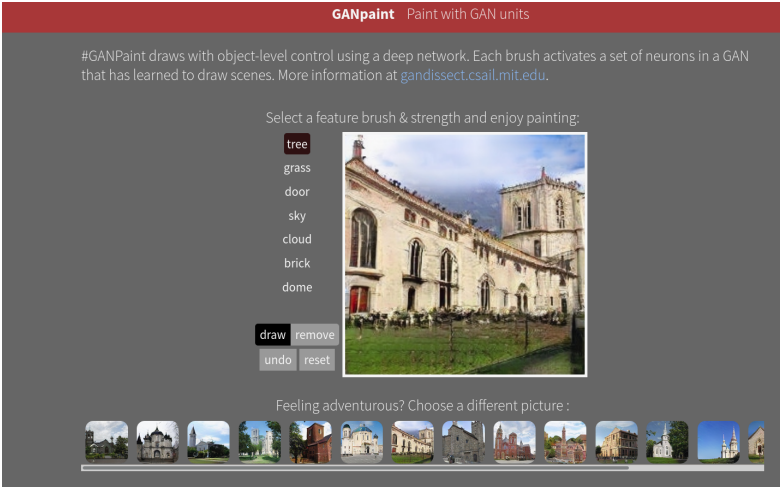


Figure 11: Inserting door units by setting 20 causal units to a fixed high value at one pixel in the representation. Whether the door units can cause the generation of doors is dependent on its local context: we highlight every location that is responsive to insertions of door units on top of the original image, including two separate locations in (b) (we intervene at left). The same units are inserted in every case, but the door that appears has a size, alignment, and color appropriate to the location. Emphasizing a door that is already present results in a larger door (d). The chart summarizes the causal effect of inserting door units at one pixel with different contexts.

Interactive toolkit



https://gandissect.csail.mit.edu/

Further reasearch suggested by authors

1. „*Why can a door not be inserted in the sky ?*“
2. „*How does GAN suppress the signal in the later layers ?*“
3. „*What are the relationships between layers of a GAN ?*“

Questions ?