GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba

Ing. Jakub Žitný
Faculty of Information Technology, Czech Technical University in Prague

Supervisor: doc. Ing. Pavel Kordík PhD.

April 6, 2020

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Introduction

- Dissertation topic: interpretability, explainability
  - Visualizing the Impact of Feature Attribution Baselines 10.23915/distill.00022
- Focus on: generative models, medical imaging (applications)
  - BraTS, KiTS, RA2, MURA
  - DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning

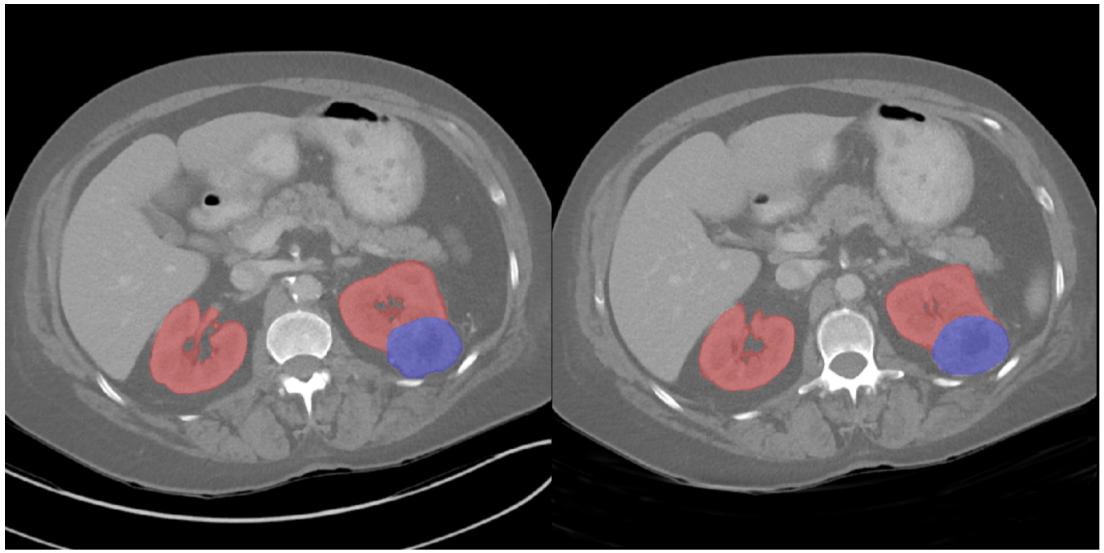


Figure: DC-GAN sample of tumour segmentation (KiTS dataset)

Motivation

How does GAN represent a our visual world internally?

What causes artifacts in GAN results?

How do architectural choices affect GAN learning?

Motivation

Does GAN contain internal variables that correspond to the objects that humans perceive?

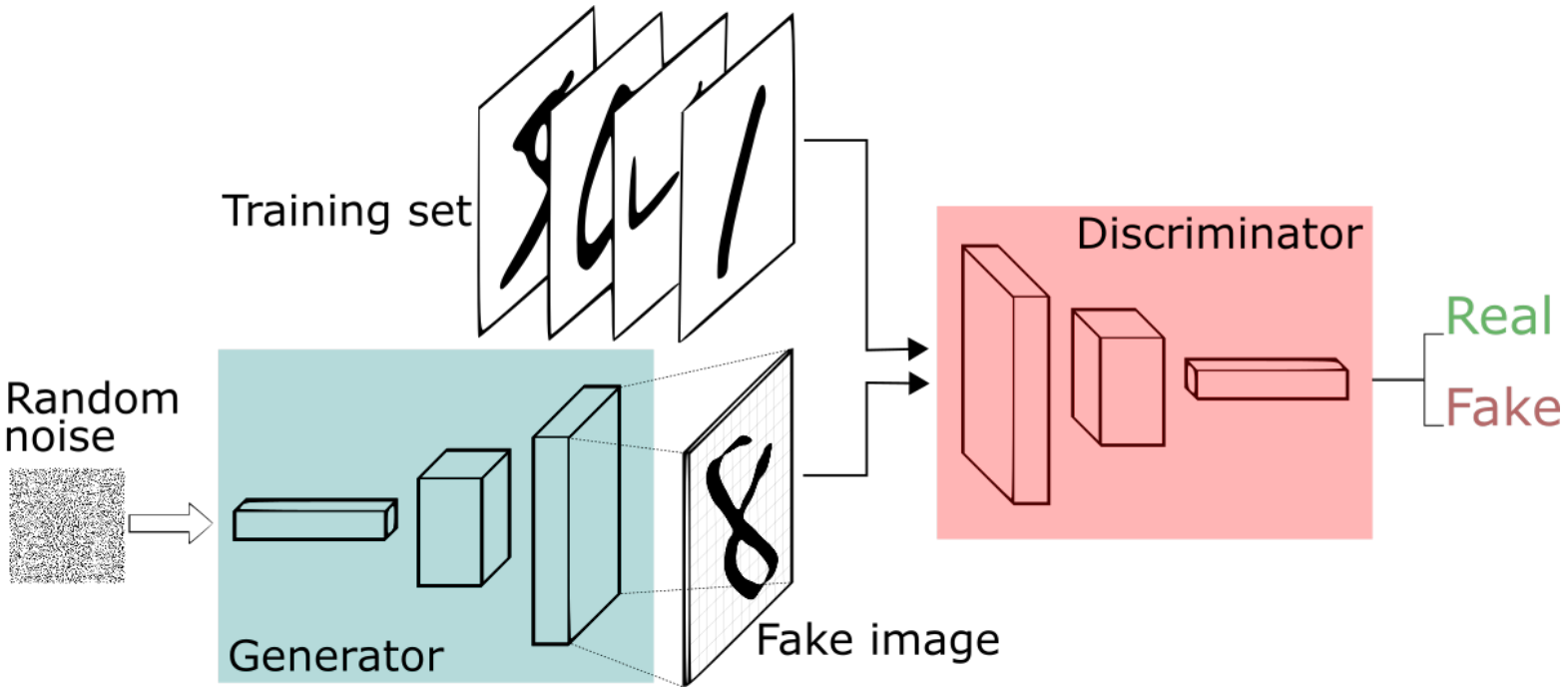
If so, do they cause the actual generation or they just correlate?

Previous work

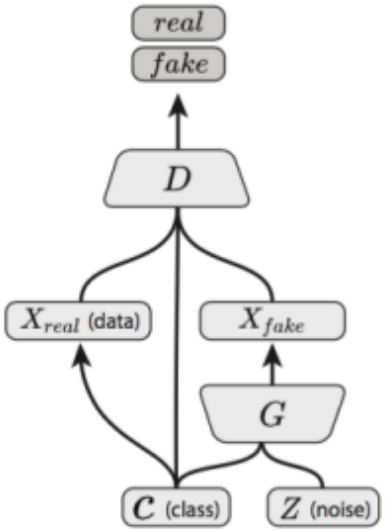
- **Network dissection: Quantifying interpretability of deep visual representations** (Bau, Zhou, et al., CVPR 17)
- **Unified perceptual parsing for scene understanding** (Zhou et al., ECCV 18)
- **Generative adversarial nets** (Goodfellow et al., NIPS 14)
- **Progressive growing of gans for improved quality, stability, and variation** (Karras et al., ICLR 18)

Generative Adversarial Networks

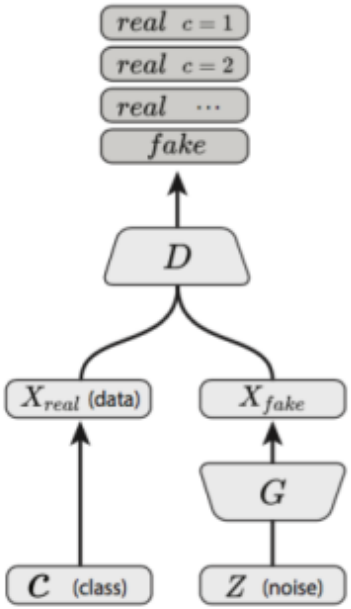
FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE



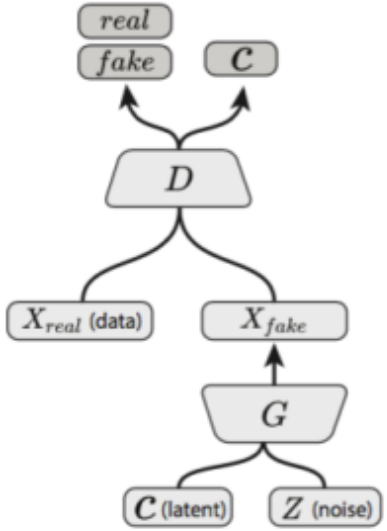
$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$



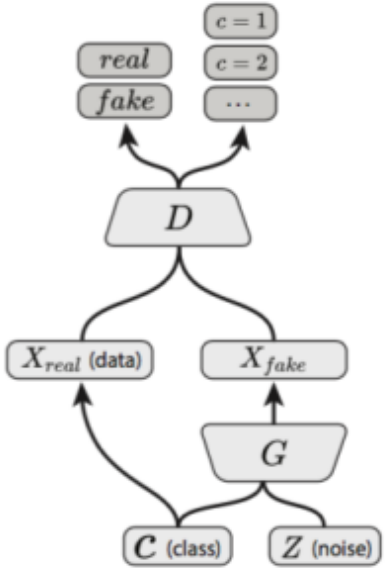
Conditional GAN
(Mirza & Osindero, 2014)



Semi-Supervised GAN
(Odena, 2016; Salimans, et al., 2016)



InfoGAN
(Chen, et al., 2016)



AC-GAN
(Present Work)

Method

1. The information is present, but how?
2. Characterizing units by dissection
3. Measuring causal relationships using intervention

- tensor **r** from layer from G

$$r = h(z)$$

- image **x** from random **z** through a composition of layers

$$x = f(r) = f(h(z)) = G(z)$$

- so **x** is a function of **r**

- feature map $\mathbf{r}_{U,\mathbb{P}}$
- universe of concepts $c \in \mathbb{C}$
- can we factor $\mathbf{r}$ at locations P?

$$\mathbf{r}_{\mathbb{U},P} = (\mathbf{r}_{U,P}, \mathbf{r}_{\overline{U},P})$$

- where P depends on $\mathbf{r}_{U,P}$ and not on $\mathbf{r}_{\overline{U},P}$

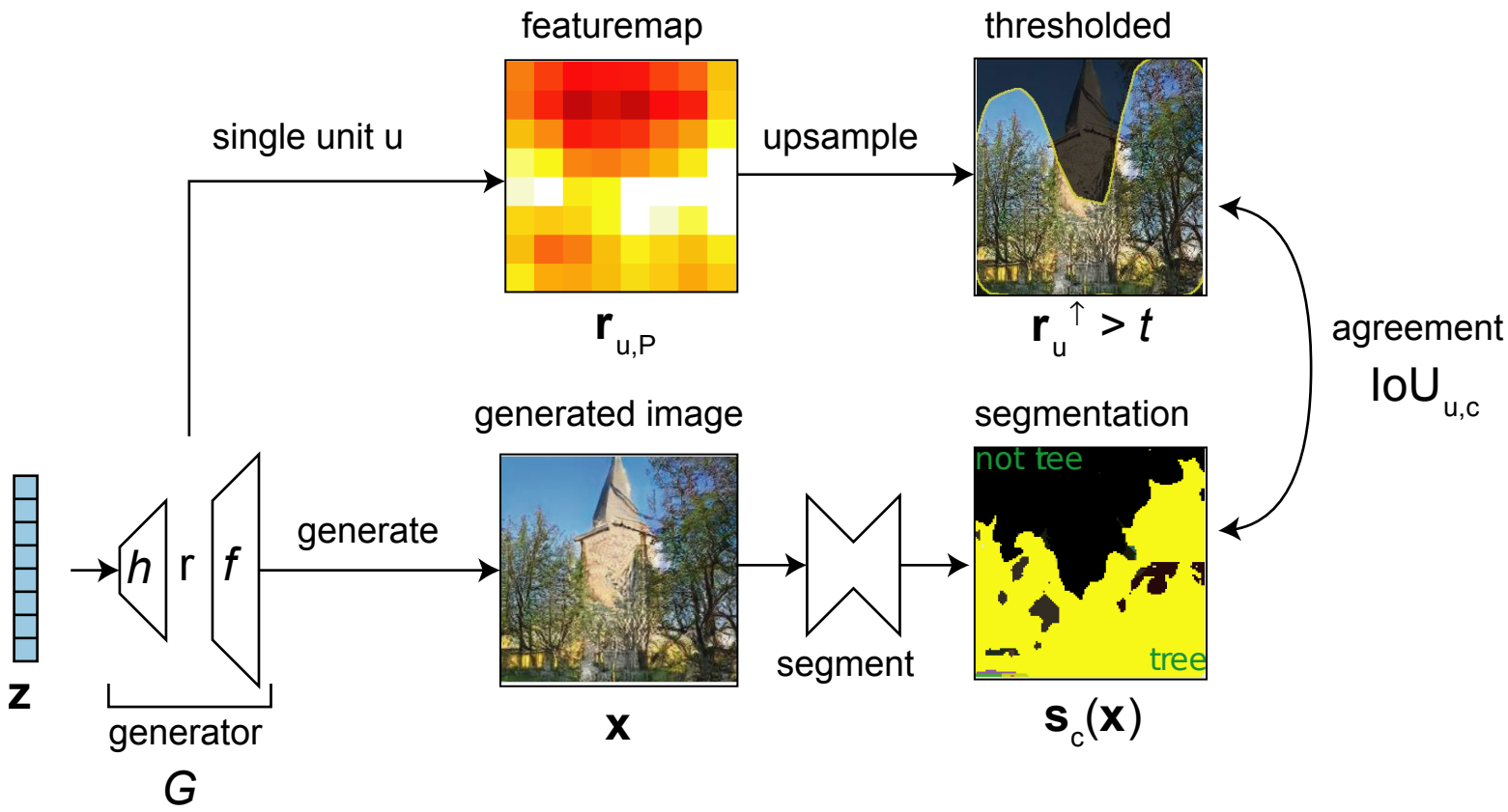


Figure: Which units correlate to a object class?

Characterizing units by dissection

Intersection-over-union measure for spatial agreement between unit **u**'s thresholded featuremap and **c**'s segmentation

$$IoU_{u,c} \equiv \frac{\mathbb{E}_z|(r^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \wedge s_c(x)|}{\mathbb{E}_z|(r^{\uparrow}_{u,\mathbb{P}} > t_{u,c}) \vee s_c(x)|}$$

$$t_{u,c} = \arg\max_t \frac{I(r^{\uparrow}_{u,\mathbb{P}} > s_c(x))}{H(r^{\uparrow}_{u,\mathbb{P}} > s_c(x)}$$

After we identified units that match closely with object class, we want to know which ones are responsible for triggering the rendering of the object.

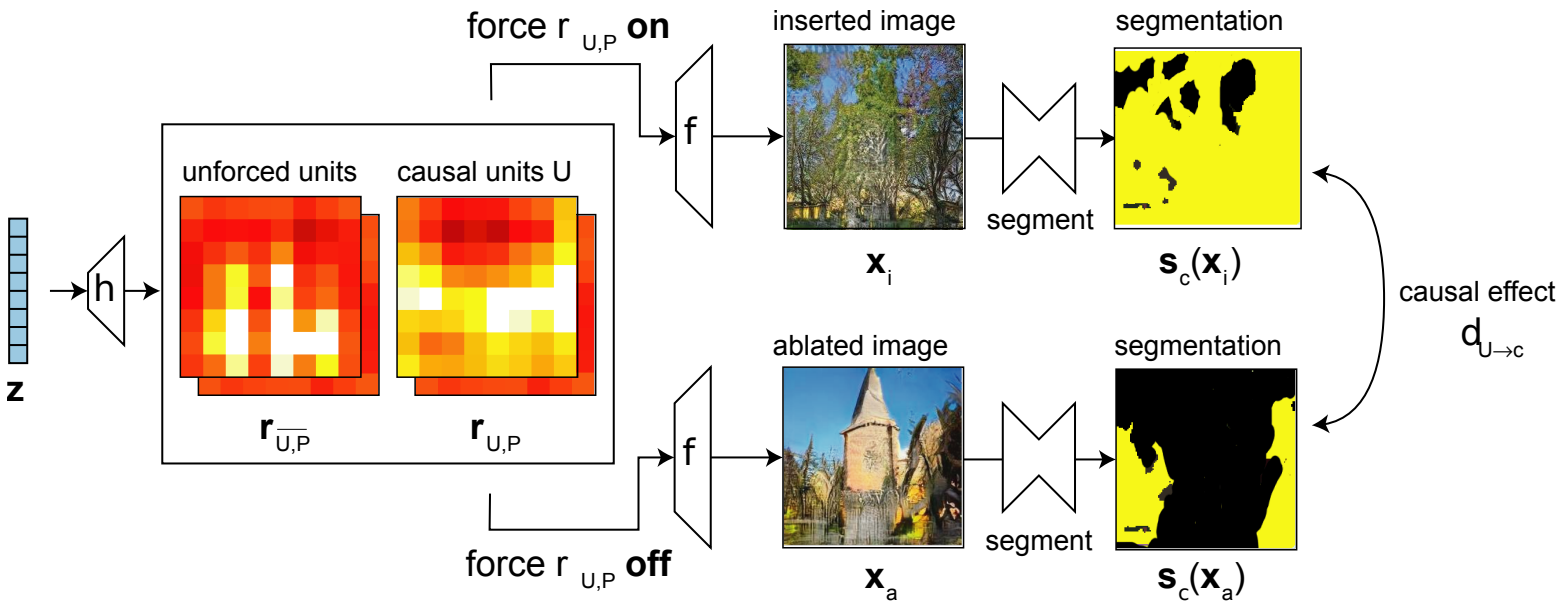


Figure: Insert and remove units and observe causality.

Causal relationships intervention

- Original image

$$x = G(z) \equiv f(r) \equiv f(r_{U,P} r_{\overline{U,P}})$$

- U ablated at P

$$x_a = f(0, r_{\overline{U,P}})$$

- U inserted at P

$$x_i = f(k, r_{\overline{U,P}})$$

- Average causal effect of units **u** on **c**

$$\delta_{U \to c} \equiv \mathbb{E}_{z,P}[s_c(x_i)] - \mathbb{E}_{z,P}[s_c(x_a)]$$

- Relaxed to partial ablations/insertions

$$x_a = f((1-\alpha) \odot r_{\mathbb{U},P}, r_{\mathbb{U},\overline{P}})$$

$$x_i = f(\alpha \odot k + (1-\alpha) \odot r_{\mathbb{U},P}, r_{\mathbb{U},\overline{P}})$$

- Optimize α, SGD, L2

$$\alpha^* = \arg\min_{\alpha}(-\delta_{\alpha \to c} + \lambda||\alpha||_2)$$

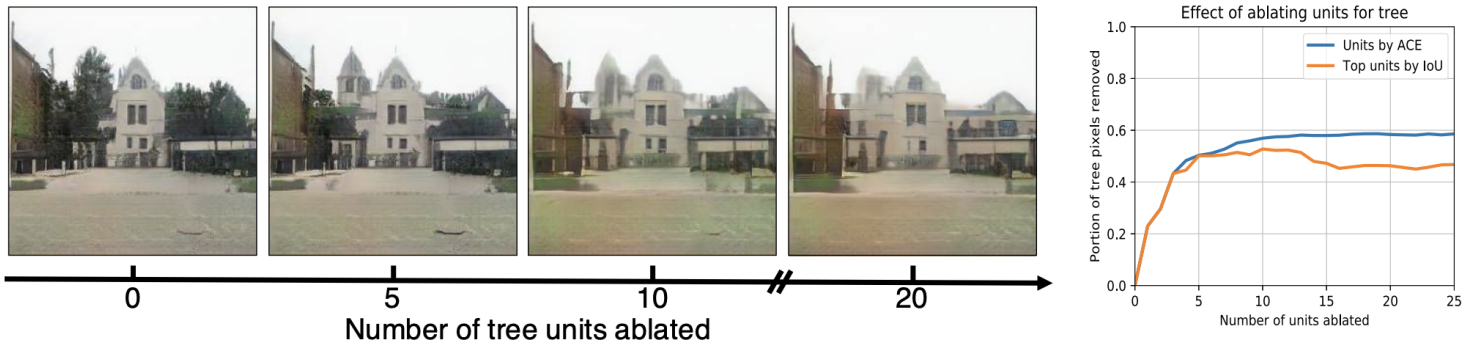
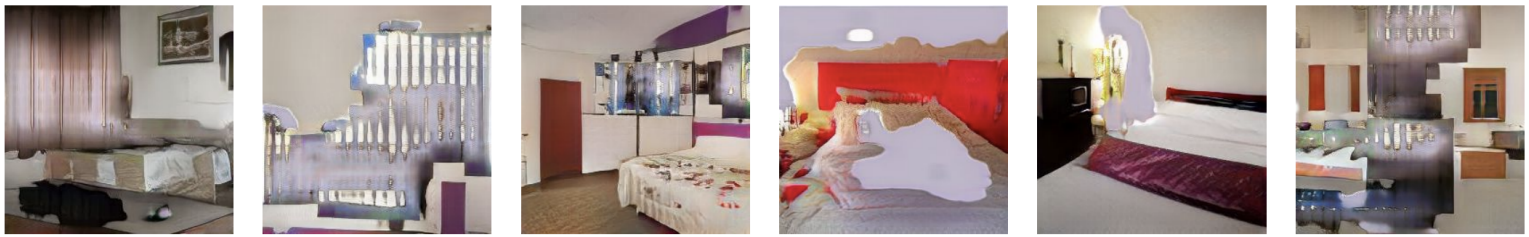


Figure 4: Ablating successively larger sets of tree-causal units from a GAN trained on LSUN outdoor church images, showing that the more units are removed, the more trees are reduced, while buildings remain. The choice of units to ablate is specific to the tree class and does not depend on the image. At right, the causal effect of removing successively more tree units is plotted, comparing units chosen to optimize the average causal effect (ACE) and units chosen with the highest IoU for trees.

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.

Figure 3: Visualizing the activations of individual units in two GANs. Top ten activating images are shown, and IoU is measured over a sample of 1000 images. In each image, the unit feature is upsampled and thresholded as described in Eqn. 2.



(a) original generated images without ablation

(b) ablating the 20 highest-FID units.

(b) ablating the 20 manually-identified untis.

Results

FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

- Practical implications
  - Debugging, monitoring, tracing
  - Controlling — tuning / composing GAN outputs
- Observations
  - Usually multiple units are responsible for generating an object
  - First has no units that match semantic objects
  - Later layers are dominated by low-level materials, edges and colors
  - Network learns the context of object location (e.g. windows can be on building, but not in the sky)

DEMO

Questions?

Thank you