GAN DISSECTION: VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS

David Bau¹², Jun-Yan Zhu¹, Hendrik Strobelt²⁵, Bolei Zhou³, Joshua B. Tenenbaum¹, William T. Freeman¹⁴, Antonio Torralba¹²
¹Massachusetts Institute of Technology, ²MIT-IBM Watson AI Lab, ³The Chinese University of Hong Kong, ⁴Google Research, ⁵IBM Research

ICLR 2019

Presenter: Jack Lanchantin
Introduction

- $G: z \rightarrow x$, where $z \in \mathbb{R}$ and $x \in \mathbb{R}^{H \times W \times 3}$
- Tensor $r$ output from a particular layer of $G$: $r = h(z)$ and $x = f(r) = f(h(z)) = G(z)$
- $r$ certainly contains the information to deduce the presence of any visible class $c$ in the image
- Question is *how* the information about $c$ is encoded in $r$
Introduction

- In particular, we seek to understand whether \( r \) explicitly represents the concept \( c \) in some way where it is possible to factor \( r \) at locations \( P \) into components

\[
    r_{U,P} = (r_{U,P}, r_{\overline{U},P})
\]

where the generation of the object \( c \) at locations \( P \) depends mainly on the units \( r_{U,P} \) and is insensitive to the other units \( r_{\overline{U},P} \)

- Refer to each channel of the featuremap as a unit; \( U \) denotes the set of unit indices of interest and denotes its complement
- we will write \( U \) and \( P \) to refer to the entire set of units and feature map pixels in \( r \)
Characterizing Units by Dissection

- Quantify the spatial agreement between the unit U’s thresholded featuremap and a concept c’ segmentation with the following intersection-over-union (IoU) measure:

\[
\text{IoU}_{u,c} \equiv \frac{\mathbb{E}_z \left[ (r_{u,P} \uparrow > t_{u,c}) \land s_c(x) \right]}{\mathbb{E}_z \left[ (r_{u,P} \uparrow > t_{u,c}) \lor s_c(x) \right]}
\]

where \( \land \) and \( \lor \) denote intersection and union, \( t_{u,c} \) is a fixed threshold, and \( s_c(x) \) is a binary segmentation mask for concept c
Characterizing Units by Dissection
Characterizing Units by Dissection

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. 10 top 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.
Measuring Causal Relationships Using Intervention

- Which of those units are actually responsible for triggering the rendering of that object?
  - Correlation $\neq$ causation
  - Furthermore, any output will jointly depend on several parts of the representation (need to identify combinations)
Measuring Causal Relationships Using Intervention

- Recall that $r_{U,P}$ denotes the feature map $r$ at unit $U$ and location $P$.
- **Ablate** such unit by forcing $r_{U,P} = 0$.
- **Insert** such unit by forcing $r_{U,P} = c$, where $c$ is a big constant.
- Decompose $r$ into two parts ($r_{U,P}, r_{U,P}$), where $r_{U,P}$ are unforced components of $r$.

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

Image with $U$ ablated at pixels $P$: $x_a = f(0, r_{U,P})$

Image with $U$ inserted at pixels $P$: $x_i = f(c, r_{U,P})$
Measuring Causal Relationships Using Intervention

- Recall that $r_{U,P}$ denotes the feature map $r$ at unit $U$ and location $P$.
- **Ablate** such unit by forcing $r_{U,P} = 0$.
- **Insert** such unit by forcing $r_{U,P} = c$, where $c$ is a big constant.
- Decompose $r$ into two parts $(r_{U,P}, r_{U,P}^*)$, where $r_{U,P}^*$ are unforced components of $r$.

Original image: $x = G(z) \equiv f(r) \equiv f(r_{U,P}, r_{U,P}^*)$

Image with $U$ ablated at pixels $P$:

$x_a = f(0, r_{U,P}^*)$

Image with $U$ inserted at pixels $P$:

$x_i = f(c, r_{U,P}^*)$

➢ An object is caused by $U$ if the object appears in $x_i$ and disappears from $x_a$. 


Measuring Causal Relationships Using Intervention

- This causality can be quantified by comparing the presence of an object in $x_i$ and $x_a$ and averaging effects over all locations and images.
- Define the average causal effect (ACE) of unit U on the generation of on class c as:

$$\delta_{U \rightarrow c} \equiv \mathbb{E}_{z, p} \left[ s_c(x_i) \right] - \mathbb{E}_{z, p} \left[ s_c(x_a) \right]$$

where $s_c(x)$ denotes a segmentation indicating the presence of class c in image x at P.
Measuring Causal Relationships Using Intervention
Sets of Units with High Causal Effect

- Objects tend to depend on more than one unit.
- Thus we need to identify a set of units $U$ that maximize the average causal effect $\delta_{U\rightarrow c}$ for a class $c$. 
Finding sets of units with high average causal effect

- Given a representation $r$ with $d$ units, searching for a fixed-size set $U$ with high $\delta_{U \rightarrow c}$ requires $\binom{d}{|U|}$ operations.
- Instead, we optimize a continuous intervention $\alpha \in [0, 1]^d$, where each dimension $\alpha_u$ indicates the degree of intervention for unit $u$. 

Finding sets of units with high average causal effect

- We maximize the following average causal effect formulation \( \delta_{\alpha \rightarrow c} \):

\[
\begin{align*}
\text{Image with partial ablation at pixels } P : & \quad x'_a = f((1 - \alpha) \odot r_{\text{U,P}}, r_{\text{U,P}}) \\
\text{Image with partial insertion at pixels } P : & \quad x'_i = f(\alpha \odot c + (1 - \alpha) \odot r_{\text{U,P}}, r_{\text{U,P}}) \\
\text{Objective :} & \quad \delta_{\alpha \rightarrow c} = \mathbb{E}_{z,P} [s_c(x'_a)] - \mathbb{E}_{z,P} [s_c(x'_i)],
\end{align*}
\]

where \( r_{\text{U,P}} \) denotes the all-channel featuremap at locations \( P \), \( r_{\text{U,P}} \) denotes the all-channel featuremap at other locations \( \bar{P} \), and \( \alpha \) applies a per-channel scaling vector to the featuremap \( r_{\text{U,P}} \).

- \( \alpha^* = \arg\min_\alpha (-\delta_{\alpha \rightarrow c} + \lambda ||\alpha||_2) \)
Finding sets of units with high average causal effect
A unit is counted as a class predictor if it matches a supervised segmentation class with pixel accuracy > 0.75 and IoU > 0.05 when upsampled and thresholded.
Comparing Layer Differences

<table>
<thead>
<tr>
<th>Layer</th>
<th>Units Total</th>
<th>Object Units</th>
<th>Part Units</th>
<th>Material Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>layer1</td>
<td>512</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>layer4</td>
<td>512</td>
<td>89</td>
<td>159</td>
<td>7</td>
</tr>
<tr>
<td>layer7</td>
<td>256</td>
<td>52</td>
<td>69</td>
<td>17</td>
</tr>
<tr>
<td>layer10</td>
<td>128</td>
<td>18</td>
<td>9</td>
<td>13</td>
</tr>
</tbody>
</table>
Ablating Artifacts

(f) Bedroom images with artifacts

(g) Ablating “artifacts” units improves results

Ablate the 20 artifact-causing units out of 512 units in layer4.
## Ablating Artifacts

<table>
<thead>
<tr>
<th>Fréchet Inception Distance (FID)</th>
<th>Human preference score</th>
<th>original images</th>
</tr>
</thead>
<tbody>
<tr>
<td>original images</td>
<td>52.87</td>
<td></td>
</tr>
<tr>
<td>“artifacts” units ablated (ours)</td>
<td>32.11</td>
<td>79.0%</td>
</tr>
<tr>
<td>random units ablated</td>
<td>52.27</td>
<td>50.8%</td>
</tr>
</tbody>
</table>
Ablating Objects

- **ablate person units**
- **ablate curtain units**
- **ablate window units**
- **ablate table units**
Inserting Objects