GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

Summary and review by - Purvanshi Mehta

September 24, 2019

1 Summary

The paper proposes a method to understand and interpret Generative Adversarial Networks (GAN). The major contribution of the paper is twofold: Firstly, a group of interpretable units (feature maps) are identified and interpreted for a GAN. Secondly, the causal effect of these interpretable units is quantified on various classes of objects.

Let $\mathbf{r}$ represents the feature maps of a given layer $\mathbf{L}$ and c be any class from a pool of classes C. To understand how the information is represented in $\mathbf{r}$ at a given position $\mathbf{P}$, $\mathbf{r}$ is decomposed into two parts: $r_{U,P}$ and $r_{Ubar,P}$. The generation of an object belonging to a particular class c at a position P is mainly due to $r_{U,P}$ while $r_{Ubar,P}$ has no effect on it. This capability is learnt well by the GANs which is demonstrated well through an experiment in the paper. It is shown in the paper that a GAN architecture learns well to generate a door on various scenes where there is a possibility of a door but fails to generate a door in between the clouds. To identify a unit u has learnt representation for a class c, the unit is upsampled first and then a threshold is applied over it to maximize the quality of information present. This idea is taken from previous work of Bau et. al (2017) where many units were able to approximately locate emergent object classes when the units were upsampled and thresholded. The IoU is calculated between the units and the class c using a semantic segmentation network to understand how well a unit u matches with a class c taken from a pool of classes C. However, a unit that correlates highly (high IoU score) with an object class c isn't necessarily the cause of the output which can be due to several units and the authors went on to propose a causal measure to quantify it in their work.

The causal effect of a unit is measured by turning the unit ON and OFF and computing the overall average effect on the generation of a class . Averaging the effects over all the locations and images gives the average causal effect (ACE) of a unit on the generation of a given class. As already seen that several units can be responsible for the generation of a particular class, the authors went on to identify a set of units U that maximize the ACE in their work. The authors introduced a term α that keeps a track of contribution for every unit to maximize causal effects. The units are then ranked based on the values of α, and causal effect can be controlled by the value of α.

Finally, the authors also demonstrate that knowing important units that contribute to the generation of a particular object/class can help in improving

the generation capabilities of a GANs. All the above experiments and visualization was released in an open-source visualization framework, easy to use and interpret.

2 Review

For dissecting the GAN, all the units for every layer are taken and individually IoU is calculated to match a concept class. Then units are thresholded and divided into U and U_{bar}. This process is highly inefficient both in terms of time and memory. The bottleneck of every layer input units identification, ranking and grouping can be minimized by taking the units of the last layer of a network, calculate the IoU for every unit and then rank them and categorize them for every class c. Then use the approach of guided backpropagation (or more efficiently alternating backpropagation as provided by Han et. al. (2017)) to get similar units from every previous layer corresponding to the generation of one particular class. The bottleneck of memory and time can be reduced efficiently.

Another interesting observation when playing around with the demo provided by the authors, the GAN was limiting the distortion of images in many unrealistic ways (like drawing a door in the clouds). However, when trying to draw some grass in the clouds which are equally unlikely, the results were a bit different which again raises the concern of maximizing the quality of information through a threshold. The GAN is enforcing the relationship between objects and hence it suppresses the generation of a door in the sky. But the authors failed to establish a relationship between the layers of a GAN which can be established with a guided backpropagation approach as mentioned above.

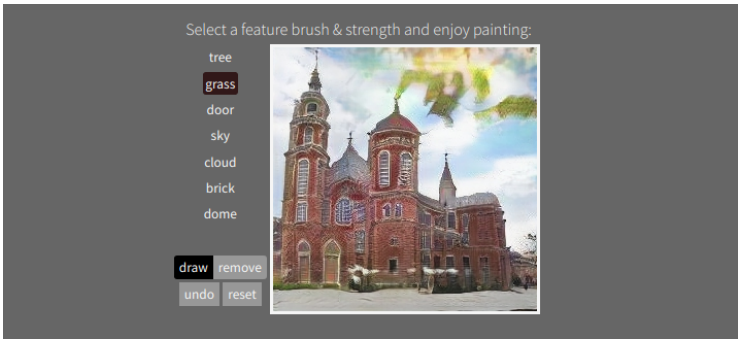

Figure 1: Generation of grass in clouds using the demo provided by the authors

3 References

Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6541-6549).

Han, T., Lu, Y., Zhu, S. C., Wu, Y. N. (2017, February). Alternating back-propagation for generator network. In Thirty-First AAAI Conference on Artificial Intelligence.