GAN DISSECTION: VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS (Summary)

David Bau , Jun-Yan Zhu,Joshua B. Tenenbaum [1]
davidbau@csail.mit.edu,junyanz@csail.mit.edu,jbt@csail.mit.edu
Antonio Torralba, William T. Freeman[1]
torralba@csail.mit.edu,billf@csail.mit.edu
Hendrik Strobelt[2]
hendrik.strobelt@ibm.com
Bolei Zhou[3]
bzhou@ie.cuhk.edu.hk

[1] Massachusetts Institute of Technology, Boston
[2] IBM Research, Cambridge MA
[3] The Chinese University of Hong Kong

1 Abstract

In this work, the authors have presented an analytic framework to visualize and understand Generative Adversarial Networks(GANs) at the unit, object, and scene-level. They identified a group of interpretable units that are closely related to object concepts using a segmentation-based network dissection method. Then, they quantified the causal effect of interpretable units by measuring the ability of interventions to control objects in the output. This work has examined the contextual relationship between these units and their surroundings by inserting the discovered object concepts into new images. They have shown several practical applications that are enabled by their framework, from comparing internal representations across different layers, models, and datasets, to improving GANs by locating and removing artifact-causing units, to interactively manipulating objects in a scene.

2 Summary

To a human observer, a well-trained GAN appears to have learned facts about the objects in the image, for example, a door can appear on a building but not on a tree. We wish to understand how a GAN represents such a structure. Do the objects emerge as pure pixel patterns without any explicit representation of objects such as doors and trees, or does the GAN contain internal variables that correspond to the objects that humans perceive? If the GAN does contain variables for doors and trees, do those variables cause the generation of those objects, or do they merely correlate? How are relationships between objects represented?. One can imagine two general ways that a computer algorithm could generate a beautiful realistic looking scene, either It could learn about objects and compose the scene out of objects it knows or it could memorize a scene that it saw in the training data and render one just like it. State-of-the-art GANs are still not perfect. They sometimes produce images that are quite unrealistic. Here, in this paper, the progressive GAN trained on 3,033,042 bedroom images, the largest LSUN data set, and it still outputs images with visible artifacts. They appear about 5% of the time as quoted by the authors in this work. We want to know, what causes these mistakes?. This question is the motivation behind this paper “GAN Dissection”.

The basis for the study is three variants of progressive GANs trained on LSUN scene datasets. To understand what’s going on inside these GANs, the authors develop a technique involving a combination of dissection and intervention. Refer Figure [2] [3].

Given a trained segmentation model (i.e., a model that can map pixels in an image to one of a set of predefined object classes), we can dissect the intermediate layers of the GAN to identify the level of agreement between individual units and each object class. The segmentation model used in the paper was trained on the ADE20K scene dataset and can segment an input image into 336 object classes, 29 parts of large objects, and 25 materials.Dissection reveal units that correlate with the appearance of objects of certain classes. Two different types of intervention help us to understand this relationship. First, we can ablate(switch them off) those units and see if the correlated objects disappear from an image in which they were previously present. Second, we can force the units on and see if the correlated objects appear in an image in which they were previously absent.

Figure [1] in the paper provides an excellent overview. To characterize units by dissection, they took an upsampled and thresholded feature map of a unit and compare it to the segmentation map of a given object class.

Figure 1: Here we can see (a) a set of generated images of churches and (b) the results of dissection identifying GAN units matching trees. When we ablate those units (c) the trees largely disappear, and when we deliberately activate them (d) trees reappear.(f)images with artifacts . (g)identify the GAN units that cause those artifacts (e) ablate them, (g)remove unwanted artifacts from generated images

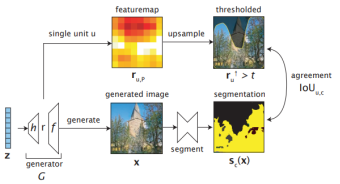

Figure 2: Dissection

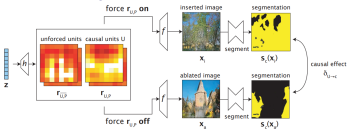

Figure 3: Intervention

The extent of agreement is captured using an IoU (intersection-over-union) measure. They took the intersection of the thresholded image and the pixels defined as belonging to the segment class and divide it by their union. The result shows what fraction of the combined pixels are correlated with the class.The following examples show units with high IoU scores for the classes table and sofa. Figure [4]

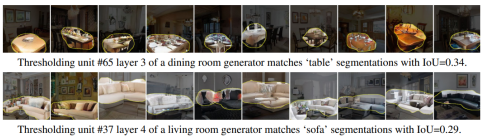

Figure 4: Visualizing the activations of individual units in two GANs

3 Conclusion

Units emerge that correlate with instances of an object class, with diverse visual appearances. The units are learning abstractions. The set of all object classes matched by units of a GAN provides a map of what a GAN has learned about the data.