GAN Dissecting aGAN

Tripp Isbell
Auburn University
cai0004@auburn.edu

Lalithya Kuntamukkala
Auburn University
lzk0034@auburn.edu

James Lee
Auburn University
jyl0003@auburn.edu

Abstract

We explore the GAN Dissection method and stuff.

NOTE: this is a big work in progress (particularly introduction and stuff), skip to sections 2,3,4 and particularly 4.2 to see current progress.

1 Introduction

Much research in the field of Deep Learning has worked towards the goal of visualizing and understanding deep neural network architectures (cite much research). Introduce Network Dissection as a means of understanding semantic representation of Convolutional Neural Networks (cite net dissect).

The field of generative models is a-boomin', and from it many unique models have emerged, including (cite unique models). A dominant/popular model for image generation is the Generative Adversarial Network (cite GANs) and many GAN variants (cite GAN variants, or maybe don't, probably don't blow all our citations so early). All of this research/progress has spawned new research into understanding generative models. And maybe go into some research on generative models (with citations). Since the generative models make use of neural networks, a lot of the research into network visualization and understanding can be applied. Bau et al. [2019] adapt their method from Bau et al. [2017] and build on it to understand and reason about GANs.

cite spam to be copied: Bau et al. [2017] Bau et al. [2019] Karras et al. [2017] Karras et al. [2018] Goodfellow et al. [2014] Nguyen et al. [2016] Yosinski et al. [2015]

2 GAN Dissection

The paper presents a framework to visualize and understand GANs at the unit-, object-, and scene-level. It aims to understand how different objects and classes of objects are encoded by the GAN.

A group of interpretable units that are closely related to object concepts are identified using a segmentation-based network dissection method. The causal effect of interpretable units is quantified by measuring the ability of interventions to control objects in the output .The contextual relationship between these units and their surroundings is examined by inserting the discovered object concepts into new images.

The structure of the featuremap is studied in two different phases

1. Dissection: Identify the classes that have an explicit representation
2. Intervention: For the represented classes identified through dissection, we identify causal sets of units and measure causal effects between units and object classes by forcing sets of units on and off

2.1 Applications / Hypotheses

In Bau et al. [2019] during the continuous intervention phase used to identify sets of causal units, they add a regularization term to find a minimal set of causal units. We hypothesize that this minimization is similar to increasing the interpretability / disentanglement of the network measured and studied in Bau et al. [2017]. So we wish to study the effect of other GAN architectures on this.

Another question they raise in the paper is what is making the GAN veto certain interventions while accepting others. Why can we remove certain objects but not remove others (like windows in hotel bedrooms)? Why can we insert a door on a building but not in the sky? How does the GAN represent this stronger association between certain concepts when generating its output?

Something peculiar that they note in the paper is that the causal units seem most disentangled around layer 4 (in the proGAN that they dissect that is, this is highly likely to vary among different architectures). This seems unusual when compared to their Network Dissection results, which found (high level) human-interpretable concepts to most closely agree with the units at the small end of the network, which seems intuitive since that's where the higher level concepts would be represented. If we then think of the generator of a GAN as an inverse classifier, mapping latent vectors to images, it would seem the most interpretable units would lie somewhat close to the beginning. We think that investigation into this might offer insight into their question of how the network vetoes certain interventions and allow others, as perhaps the concept correlations are generated at the higher level of the network (and not the concepts themselves).

One way we could investigate this is by manipulation of latent vectors. (cite research on latent space manipulation, addition, etc which might be interesting to investigate with GAN dissection)

Another way of looking at it, that sort of follows that idea of a generator as an "inverse classifier" (wherein the classes are some continuous value in a latent space) is the idea of a the generator as a decoder. If we could train a network to encode the latent vectors given the GAN output, would we see the same phenomenon in its layer interpretability? (In sort of a reverse order of a VAE, though it may be easier and just as interesting to look at a VAE itself). If so, this would point to some sort of architecture-independent, inevitable emergence in the way concepts are represented through the network, dependent on the relationship between the chosen latent space (of which we have control) and the paired outputs.

3 StyleGAN

As an extended version of ProGan, StyleGAN allows for more control in the image synthetic process. This is possible in the ways that the StyleGAN generator was redesigned. Through the use of a mapping network of eight fully connected layers and getting a learned constant, this allows the model to generate a vector that can reduce the correlation between features. Then, learned affine transformations then specialize the vector(denoted at w in the paper) to styles that then control the adaptive instance normalization(AdaIn) where each feature map is normalized separately, and then scaled and biased using scalar components from the style. Lastly, "explicit noise inputs" are then introduced to generate stochastic detail. Compelling enough, this design led to improved generated image quality. Combined with that the generator architecture makes it possible to control the image synthesis through scale-specific modifications to the styles, we believe that this Gan model is a good candidate for our project because it is an interesting alternative extension to ProGan and that we are able to utilize tools that the Gan Dissection paper used as well. Lastly, we believe that studying and performing gan dissection on this model could lead to interesting results.

4 Methods

Our methods for this project involve researching other GAN models we think might yield interested results when studied under the GAN Dissection Bau et al. [2019] framework. We find a pre-trained GAN and a segmented dataset as a candidate to be examined. We then apply the GAN dissection tool released by Bau et al. [2019] at `http://github.com/CSAILVision/GANDissect` which carries out dissection on the specified model, and this basically carries out the heavy lifting. The tool also produces metrics and visualizations of sets of units with high agreement with specified concepts.

4.1 Progressive GANs

We first replicate one of the experiments of Bau et al. [2019], using a Progressive GAN (Karras et al. [2017]) trained on LSUN living room images, we use the GAN dissection tool to identify units which are highly activated by certain concepts. We analyze the results of this and make sure that they align with the results provided in Bau et al. [2019]. The metrics provided by the tool will act as a sort of baseline to compare the results on other networks. Bau et al. [2019] note that GAN dissection should not be used for qualitative comparisons across GAN architectures, but through comparison we seek to reason about why their differences might be so.

4.2 Challenges / current progress

As of right now (3/18), we've been working on getting the tool set up to run the dissection. We tried running with CUDA on a personal desktop GPU (running the above experiment) and after getting dependencies working, ran into memory issues running the tool. We plan to try again using a smaller batch size, but will likely need to use cloud computing or the Auburn GPU server. One of our group members has access to a GPU in their lab but the lab is closed down due to... circumstances.

4.3 StyleGAN

Same stuff

5 Results

More to come.

6 Analysis

More to come.

7 Discussion

More to come.

Acknowledgments

References

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Computer Vision and Pattern Recognition*, 2017.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL `http://arxiv.org/abs/1710.10196`.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. URL `http://arxiv.org/abs/1812.04948`.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML)*, 2015.