Basic Information

Do GANs learn the distribution? Some Theory and Empirics.
Author: Sanjeev Arora, Andrej Risteski, Yi Zhang
ICLR 2018, Cited by 62.

Summary

In this paper, the authors proposed an approach to estimate the diversity of generated samples. The main contribution of this paper is that through this test, researchers can measure the quality of samples generated by GAN. Using this test method, the authors argue that GANs generate distributions with low support. The main idea is to apply the birthday paradox: duplicates will appear after $\sqrt{N} i.i.d$ samples if we have a uniform distribution through N support. The authors use the paradox inversely: if we sample a certain number of objects, the duplicate appearance probability will indicate the original support size. The suggested plan is to manually check for duplicates in a sample of size s and if duplicate exists, then estimate the size of the support. This test is expected to experimentally support the previous theoretical analysis by Arora et al. (2017). The further theoretical construction also shows that for encoder-decoder GAN architectures the distributions with low support can be very close to the optimum of the specific (BiGAN) objective. Experiments are conducted in this paper to study this behavior by varying the discriminator capacity and then estimating the support size using the idea described above. The results are interpreted to mean that mode collapse is strong in a number of state-of-the-art generative models. Bidirectional models (ALI, BiGANs) however demonstrate significantly higher diversity that DCGANs and MIX+DCGANs. Finally, the authors verify empirically the hypothesis that diversity grows linearly with the size of the discriminator.

Strong points

1. **Originality and contribution**

This is a very interesting area and exciting work. This paper made a significant contribution to the discussion of whether GANs learn the target distribution. The main idea behind the proposed test is very insightful. The main theoretical contribution stimulates and motivates much needed further research in the area. Given how little we know about the behavior of modern generative models, it is a good step in the right direction.

The other main contribution of the paper also showed that bidirectional GANs can also suffer from serious mode collapse. Through a very particular construction on the generator and encoder, Theorem 3 capture the underlying behavior of bidirectional GANs very well.

2. **Writing**

The paper is written well and the issues raised are well motivated and proper background is given. I really enjoy reading the sections from 1.1 to 2.2, these parts lead me to understand the motivation and help me understand the idea in a clear and insightful way.

Weak points

1. **Technical details**

1. The biggest issue with the proposed test is that it conflates mode collapse with non-uniformity. The authors do mention this issue, but do not put much effort into evaluating its implications in practice, or parsing Theorems 1 and 2. My current understanding is that, in practice, when the birthday paradox test gives a collision I have no way of knowing whether it happened because my data distribution is modal, or because my generative model has bad diversity. Anecdotally, real-life distributions are far from uniform, so this should be a common issue. I would still use the test as a part of a suite of measurements, but I would not solely rely on it. I feel that the authors should give a more prominent disclaimer to potential users of the test.

2. I think this method works largely due to the choose of the automatic measure of image similarity. As the authors stated, the Euclidean distance no longer works well on CIFAR-10. We have to choose a measure for each dataset. This make the method artificial. It also requires human visual inspection of the duplication. A concern will be, what if the measure is not convincible and top similar pictures are not actually the most similar pictures for human.

3. It seems that this paper fail to discuss about the coverage of the distribution. The proposed test is a measure of diversity, not coverage, so it does not discriminate between a generator that produces all of its samples near some mode and another that draws samples from all modes of the true data distribution. As long as they yield collisions at the same rate, these two generative models are equally diverse.