Part IIB Project 2017/2018
Approaches to Understanding GANs

Aidas Liaudanskas (al747)
Girton College

Technical Abstract

The problem this work focuses on is that of unsupervised learning and generative models, namely Generative Adversarial Networks (GANs). We aim to investigate some of the outstanding questions in the field regarding the objective training metrics and interpretation of network dynamics during and after training. This work is primarily experimental and focuses on WGAN-GP version trained on visual data.

We review significant contributions in the GAN field and overview investigative studies in the Background section.

We compare the most popular GAN metrics - Inception Score (IS), Fréchet Inception Distance (FID), Kernel Inception Distance (KID) - to metrics based on the final layer activations of the critic network. Our key findings are that FID and KID are the most suitable metrics for tracking the convergence and comparing the quality of samples of different models, while metrics based on the activations of the critic are shown to be uninformative.

The section on Asymmetric Training investigates how the balance between the generator's and discriminator's capacities affects the training. We find that the most computationally efficient way to get quality samples is to have the critic at at least about half of generator's capacity, as the critic is the main contributor to the computational cost of the algorithm. There is no advantage in having the critic more powerful than the generator.

We trained 20 progressively more complex architectures in order to test how much more powerful the generator should be in order to reliably fool the critic, and found that the critic's decisions are only weakly correlated with the generator's capacity. As expected, a more potent generator tends to fool a simpler critic more often. The equilibrium for the critic's decisions was found to be at an offset from zero, especially for smaller architectures.

Lastly, we investigated which data features the GAN framework aims to replicate. For this test, we employed a well known deep network interpretation technique called LIME (Locally Interpretable Model-Agnostic Explanations) to explain which features the discriminator focuses on when judging real samples from fake samples. We found that the critic appears to mostly focus on the data features that make intuitive sense. Since the generator's goal is to fool the critic, this finding reassures that the framework

follows the intended objective and aims to replicate the data based on interpretable features.