Generative Adversarial Networks

Matt Amodio
Advanced Topics in Machine Learning and Data Mining

January 31, 2018

Deep neural networks have achieved dramatic successes in modeling probability distributions over complex data spaces. Early work focused on discriminant tasks, where a prediction over an output distribution is made when presented with a sample input. Sampling from the distribution of seen inputs or generating plausible but unseen, novel points can not be done with these models. Generative Adversarial Networks (GANs)[1] address these tasks by training two distinct networks alternatively. In the image domain, GANs can create images that are not in the training set it learned from but are indistinguishable from real images to the human eye.

One disadvantage of vanilla GANs is that the choice of the prior distribution p(z) can limit the expressivity of the model. Adversarial Autoencoders (AAEs)[2] built on this approach by augmenting the generator network, turning it into a full autoencoder. Instead of mapping samples from the prior directly to the input space, AAEs map the input to a latent space that the discriminator enforces to be indistinguishable from the prior distribution. Then, the generator's decoder must be able to reconstruct the original input from its representation in the latent space, ensuring that information is preserved. Since the generator's encoder/decoder can learn arbitrary mappings to/from the prior distribution, the choice of prior plays a much less important role. AAEs can be sampled from by simply decoding samples from the known prior. [2] also extends the AAE framework to supervised and semi-supervised settings, where additional information can be used as well.

GANs and AAEs provide powerful ways of learning and sampling from a complex probability distribution. DiscoGANs[3] use the same general concept to not only learn one probability distribution, but learn two and a mapping between them. They use two sets of GANs, one to map from the first domain to the second domain, and another to do the reverse. Inputs from the first domain are mapped to the second, where a discriminator tries to distinguish them from true samples from the second domain. Then, the inputs originally mapped from the first domain are mapped back to the second, where a reconstruction penalty enforces that they look the same as they originally did. This autoencoder aspect makes sure that information is preserved in the mapping, while the adversarial aspect makes sure that the mapping aligns the two domains. Instead of autoencoders aligning a distribution with itself, DiscoGANs use a similar idea to align two different distributions.

Spotlight Question: We have seen adversarial training used to teach several type of generative models (from one domain to another, within one domain, both to and from two domains). Can adversarial training be used to teach any other generative models, or even in completely different ways?

Reading materials:

1. Goodfellow, Ian, et al. "Generative adversarial nets." *A*dvances in neural information processing systems. 2014.
2. Makhzani, Alireza, et al. "Adversarial autoencoders." arXiv:1511.05644 (2015).
3. Kim, Taeksoo, et al. "Learning to discover cross-domain relations with generative adversarial networks." arXiv:1703.05192 (2017).