Paper review for "Grathwohl et al.: Your classifier is secretly an energy based model and you should treat it like one"

EECS 598 Paper Review - Week 2 - Changyuan Qiu

This paper tries to bridge the gap between discriminative models and generative models by combining an energy-based model with a classifier network. The core approach of the paper, referred to as JEM (**J**oint **E**nergy-based **M**odel), is a re-interpretation of the logits in the classification problems that it define an energy based model for the joint density distribution among input and labels ($p(x, y)$) from the logits, and then uses that to compute $p(x)$ and $p(x \mid y)$.

While much of the prior work (much of which dedicated to invertible neural network architectures) attempting to improve the discriminative performance of generative models still underperform their purely discriminative counterparts, as reported by the authors in results of experiments, JEM not only achieves performance rivaling SOTA on both discriminative and generative tasks (JEM 92.9 % vs. Wide-Resnet 95.8% for discriminative tasks on CIFAR10), but also performs well in many downstream tasks including calibration, out-of-distribution detection and robustness to adversarial examples.

One anomaly phenomenon that the paper does not discuss and address is that while on CIFAR10 the performance gap between JEM and SOTA is approximately 2.9%, on CIFAR100 that gap is approximately 7.8%. This is really a large difference and some sort of analysis should be included.