Review-43

No MCMC for me: Amortized sampling for fast and stable training of energy-based models

Despite recent successes in Energy-based Models (EBMs), training of models remains an open problem. EBMs present limited scalability to high-dimensional data, unstable training and significant tuning in conjunction with domain expertise. To address these challenges, the work presents an entropy-regularized method for training EBMs which amortize MCMC sampling. The method adopts a fast variational approximation which demonstrates effectivness in training tractable likelihood models. Additionally, the proposed method, when combined with the recently proposed Joint Energy Model (JEM), presents faster and stable training while preserving its original performance.

MCMC methods presents a computational bottleneck due to their slow and unstable nature during training. Noise contrastive approaches, on the other hand, do not scale well to high-dimensional data. This allows to rethink EBM training from a practical viewpoint. The work reinterprets likelihood as a bi-level variational optimization problem which amortizes away MCMC sampling into a GAN-style generator. The generator is encouraged to have high entropy based on fast variational approximation. The variational maximum likelihood objective consists of an entropy regularization term combined with expected estimates from an auxilary sampler. Optimization of entropy regularization utilizes Gaussian reparameterization followed by gradient updates of the score function $\log q_{\phi}(x)$. This requires one to compute posterior $q_{\phi}(z|x)$ which is carried out using variational approximation with importance sampling. The approximation is adopted as a smple diagonal Gaussian generating samples x and is optimized using the Evidence Lower-BOund (ELBO) at each training iteration. Combination of steps highlighted above gives rise to Variational Entropy Regularized Approximate maximum likelihood (VERA) which is used to train the EBM.

VERA demonstrates sutability of the generative scheme by fitting tractable models of maximum likelihood estimates. Samples approximated by VERA on MNIST closely match exact samples from NICE model. Moreover, entropy regularization term of VERA presents lower bias per dimension when compared to Hamiltonian Monte-Carlo (HMC) estimator. Lastly, combination of VERA with JEM presents faster and training on CIFAR10, CIFAR100 AND SVHN benchmarks. Although the entropy regularization framework highlights a suitable alternative for training EBMs, its details need a more thourough understanding in light of its limitations. For instance, VERA when combined with JEM on CelebA benchmark for OOD detectin degrades the performance of EBM in comparison to standalone JEM. This indicates that certain aspects of variational approximation do not compare well to the effectiveness of efficient MCMC sampling. Additionally, it would be interesting to observe a more thorough analysis of sampling speed in comparison to alternative fast samplers such as HMC, rejection sampling, etc.

Variational approximation in conjunction with entropy regularization provides fast and stable training of EBMs on a wide variety of tasks. The setup can be further extended towards modern samplers and alternative approximation schemes which depict promise for future work. For instance, applications of VERA could be extended to collapsed samplers or autoregressive flows whererin MCMC and Gibbs sampling present computational bottlenecks for likelihood estimation.