Abstract

At the core of unsupervised learning, generative models provide a systematic framework to understanding real-world data from various domains in a probabilistic manner. Among many possible desiderata of generative models, density estimation, data generation, and representation learning are widely regarded as the three most wanted properties. In recent years, with the rapid development of deep neural networks and computational hardware, the field of deep generative models has witnessed dramatic advancement in all three aspects, significantly outperforming traditional generative models.

Despite the success, existing neural architectures and training objectives still face many fundamental drawbacks. With these challenges in mind, this thesis focuses on developing novel neural architectures and training objectives that are highly expressive, allow for efficient optimization, and can scale to a large amount of data for generative modeling.

Notably, to better exploit the optimization advantage of Transformer to capture long-term dependency, we propose Transformer-XL, which integrates segment-level recurrence into self-attention without disrupting the temporal coherence. Further, to combine the benefits of autoregressive and denoising auto-encoding based language pretraining, we propose XLNet, which relies on a permutation language modeling objective to maximize the expected log-likelihood of a sequence w.r.t. all possible permutations of the factorization order and hence capture bidirectional context. By further integrating ideas from Transformer-XL, XLNet consistently outperforms previous best language pretraining method under the same training condition, and achieves the state-of-the-art performance when scaled up. In addition, to further exploit the effectiveness of language pretraining, we propose a more efficient self-attention architecture Funnel-Transformer, which compresses the hidden state sequence to a shorter length and hence reduces the computation cost. With sequence compression, Funnel-Transformer allows one to trade the sequential resolution of the hidden state sequence for a deeper or wider model, leading to substantial gains under the same amount of computation as measured by the FLOPs.

http://zihangdai.github.io/docs/draft.pdf