TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

Progressive VAEs

Progressive VAEs

These slides were written “gedanken” (as a thought experient) while teaching this class in 2021. They were not based on any paper.

While written independently of diffusion models, these slides provide a good theoretical framework for diffusion models (currently very hot).

The original motivation for these slides was to provide a theoretically clean approach to a multi-layer VQ-VAE.

Progressive VAEs

We consider a VAE with layers of latent variables $z_1, \ldots, z_L$ and a population distribution on an observable variable y.

The encoder will define $P_{\mathrm{enc}}(z_1|y)$ and $P_{\mathrm{enc}}(z_{\ell+1}|z_\ell)$.

The decoder will define $P_{\mathrm{dec}}(z_{\ell-1}|z_\ell)$ and $P_{\mathrm{dec}}(y|z_1)$.

Following VQ-VAE, we will train the encoder and the decoder independent of any prior.

We then train a prior on the top layer latent variable. The top level prior and decoder allow us to sample y from the model.

Phase One Training

We train a encoders and decoders $\mathrm{enc}_1, \mathrm{dec}_1, \ldots, \mathrm{enc}_L, \mathrm{dec}_L$ where the distribution on $z_1, \ldots, Z_L$ is defined by y and the encoder.

$$\mathrm{enc}_1^*, \mathrm{dec}_1^* = \operatorname*{argmin}_{\mathrm{enc}_1, \mathrm{dec}_1} \; E_{y, z_1} \left[- \ln P_{\mathrm{dec}_1}(y | z_1) \right]$$

$$\mathrm{enc}_{\ell+1}^*, \mathrm{dec}_{\ell+1}^* = \operatorname*{argmin}_{\mathrm{enc}_{\ell+1}, \mathrm{dec}_{\ell+1}} \; E_{z_\ell, z_{\ell+1}} \left[- \ln P_{\mathrm{dec}_{\ell+1}}(z_{\ell-1} | z_\ell) \right]$$

If these encoders and decoders share parameters the shared parameters are influenced by all of the above training losses (this observation was added after seeing DALLE-2's diffision model).

Phase Two Training

$$\text{pri}^* = \underset{\text{pri}}{\text{argmin}}\;\; E_{z_L}\left[-\ln P_{\text{pri}}(z_L)\right]$$

Because of the autonomy of the encoder, the universality assumption implies that we get a perfect model of the population distribution on y.

Given the prior and the decoder we can sample images.

END