TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Variational Auto-Encoders (VAEs)

Generative AI: Autoregression and GANs

For an autoregressive language model we can compute $P_{\mathrm{gen}}(y)$ and train a generative model by cross-entropy loss.

$$\mathrm{gen}^* = \operatorname*{argmin}_{\mathrm{gen}} \; E_{y \sim \mathrm{Pop}} \; -\ln P_{\mathrm{gen}}(y)$$

But it is not obvious how to this for continuous signals like sounds and images.

GANs replace the cross-entropy loss with an adversarial discrimation loss.

Generative AI for Continuous Data: Flow Models

$$\text{gen}^* = \underset{\text{gen}}{\text{argmin}}\; E_{y \sim \text{pop}(y)}\; -\ln p_{\text{gen}}(y)$$

Flow-based generative models work with Jacobians over continuous transformations (no ReLUs) and can be directly trained with cross-entropy loss.

But flow models have not caught on and we will not cover them.

Generative AI for Continuous Data: VAEs

A variational autoencoder (VAE) is defined by three parts:

- An encoder distribution $P_{\mathrm{enc}}(z|y)$.
- A "prior" distribution $P_{\mathrm{pri}}(z)$
- A generator distribution $P_{\mathrm{gen}}(y|z)$

VAE generation uses $P_{\mathrm{pri}}(z)$ and $P_{\mathrm{gen}}(y|z)$ (like a GAN).

VAE training uses a "GAN inverter" $P_{\mathrm{enc}}(z|y)$.

We will rely on expectatin notation and will not distinguish disctrete distributions from densities.

Cross-Entropy for Continuous Data: L_2 Loss

Define $p_{\text{gen}}(y|z)$ by

$$y = \hat{y}_{\text{gen}}(z) + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

We then get that

$$-\ln p_{\text{gen}}(y|z) = \frac{||\hat{y}_{\text{gen}}(z) - y||^2}{2\sigma^2} + \ln Z(\sigma)$$

For a fixed σ we can ignore $\ln Z(\sigma)$ and we get L_2 distortion loss.

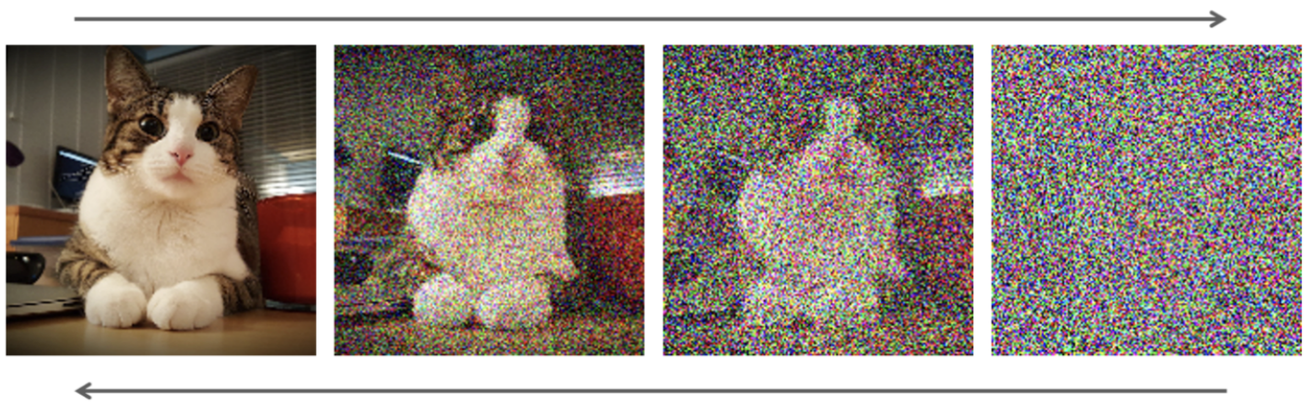
Cross-Entropy for Continuous Data: L_2 Loss

$$-\ln p_{\mathrm{gen}}(y|z) = \frac{||\hat{y}_{\mathrm{gen}}(z) - y||^2}{2\sigma^2} + \ln Z(\sigma)$$

When using L_2 distortion loss z should nearly specify y.

This is true in each step of a diffusion model.

Diffusion Model Preview



A diffusion model multi-step (Markovian) VAE where each encoder step adds a small amount of noise.

Diffusion Model Preview

Each step of a diffusion model is a VAE:

- $P_{\mathrm{enc}}(z|y)$ is defined by adding a small amount of noise to y.
- $P_{\mathrm{pri}}(z)$ is trained to model the marginal onto z of $\mathrm{Pop}(y)P_{\mathrm{enc}}(z|y)$.
- A "denoising" $\hat{y}_{\mathrm{gen}}(z)$ is computed by a U-Net.

Here z contains almost all the information in y.

Fixed Encoder Training

In a diffusion model the encoder is fixed.

$$\mathrm{pri}^*, \mathrm{gen}^* = \operatorname*{argmin}_{\mathrm{pri},\mathrm{gen}} \; E_{y\sim\mathrm{Pop}(y),z\sim\mathrm{enc}(z|y)} \left[-\ln P_{\mathrm{pri}}(z)P_{\mathrm{gen}}(y|z)\right]$$

This is a cross-entropy loss from a joint "population distribution" $P_{\mathrm{Pop,enc}}(y,z)$ to a model distribution $P_{\mathrm{pri,gen}}(y,z)$.

Assuming universality we get $P_{\mathrm{pri}^*,\mathrm{gen}^*}(z,y) = P_{\mathrm{Pop,enc}}(z,y)$ which implies $P_{\mathrm{pri}^*,\mathrm{gen}^*}(y) = \mathrm{Pop}(y)$.

Training the Encoder (The Bayesian Interpretation)

VAEs were originally motivated by a Bayesian interpretation:

- $P_{\mathrm{pri}}(z)$ is the Bayesian prior on hypothesis z.
- $P_{\mathrm{gen}}(y|z)$ is the propability of the "evidence" y given hypothesis z.
- $P_{\mathrm{enc}}(z|y)$ is a model approximating the Bayesian posterior on hypothesis z given evidence y.

The Bayesian motivation is to train $P_{\mathrm{enc}}(z|y)$ to approximate Bayesian inference.

Training the Encoder

$$\mathrm{enc}^*, \mathrm{pri}^*, \mathrm{gen}^* = \operatorname*{argmin}_{\mathrm{enc},\mathrm{pri},\mathrm{gen}} \; E_{y \sim \mathrm{Pop}, z \sim P_{\mathrm{enc}}(z|y)} \; \mathcal{L}(y, z)$$

$$\mathcal{L}(y, z) \; = \; -\ln \frac{P_{\mathrm{pri}}(z) P_{\mathrm{gen}}(y|z)}{P_{\mathrm{enc}}(z|y)}$$

Here we can hope to train the encoder to capture a causal origin for y.

Training the Encoder

Consider training P_{enc} while holding P_{pri} and P_{gen} fixed.

$$\begin{aligned}
\text{enc}^* &= \operatorname*{argmin}_{\text{enc}} \; E_{y\sim\text{Pop}(y),z\sim\text{enc}(z|y)} \; -\ln \frac{P_{\text{pri}}(z)P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)} \\
&= \operatorname*{argmin}_{\text{enc}} \; E_{y\sim\text{Pop}(y),z\sim\text{enc}(z|y)} \; -\ln \frac{P_{\text{pri,gen}}(y)P_{\text{pri,gen}}(z|y)}{P_{\text{enc}}(z|y)} \\
&= \operatorname*{argmin}_{\text{enc}} E_{y\sim\text{Pop}(y)} \; KL(P_{\text{enc}}(z|y), P_{\text{pri,gen}}(z|y)) + E_{y\sim\text{Pop}(y)} \left[-\ln P_{\text{pri,gen}}(y)\right]
\end{aligned}$$

Training $P_{\text{enc}}(z|y)$ to equal $P_{\text{pri,gen}}(z|y)$ can drive the KL term to zero.

Training $P_{\text{pri}}(z)$ and $P_{\text{gen}}(y|z)$ can drive the cross-entropy term to $H(\text{Pop})$.

The Evidence Lower Bound (ELBO)

The previous derivation can be applied to an arbitrary fixed value of y yielding.

$$\begin{aligned}\ln P_{\mathrm{pri,gen}}(y) &\geq E_{z\sim P_{\mathrm{enc}}(z|y)}\;\ln\frac{P_{\mathrm{pri}}(z)P_{\mathrm{gen}}(y|z)}{P_{\mathrm{enc}}(z|y)}\\ &= E_{z\sim P_{\mathrm{enc}}(z|y)}\left[-\mathcal{L}(y,z)\right]\end{aligned}$$

A Bayesian thinks of y as "evidence" for hypothesis z in the Bayesian model. This method of training $P_{\mathrm{enc}}(z|y)$ is called variational Bayesian inference.

Under the Bayesian interpretation the negative of the VAE loss is called the evidence lower bound (ELBO) .

Degrees of Freedom

$$\text{enc}^*, \text{pri}^*, \text{gen}^* = \underset{\text{enc,pri,gen}}{\text{argmin}} \; E_{y \sim \text{Pop}, z \sim P_{\text{enc}}(z|y)} \; \mathcal{L}(y, z)$$

$$\mathcal{L}(y, z) \;=\; -\ln \frac{P_{\text{pri}}(z) P_{\text{gen}}(y|z)}{P_{\text{enc}}(z|y)}$$

The objective is fully optimized whenever

$$P_{\text{pri}}(z) P_{\text{gen}}(y|z) = \text{Pop}(y) P_{\text{enc}}(z_y)$$

Any joint distribution on (y, z) optimizes the bound provided that the marginal on y is Pop.

Posterior Collapse

Under the Bayesian interpretation we would like z to provide useful information about (a causal origin of) y.

However the objective function only produces

$$P_{\mathrm{pri}}(z)P_{\mathrm{gen}}(y|z) = \mathrm{Pop}(y)P_{\mathrm{enc}}(z|y)$$

For language models the generator can assign a meaningful probability to a block of text y independent of z.

When we train a sentence encoder (a thought vector) as the latent valriable of a language model VAE we can get a constant (zero) thought vector.

This is called "posterior collapse".

The Reparameterization Trick

$$\mathrm{enc}^* = \operatorname*{argmin}_{\mathrm{enc}} \; E_{y\sim \mathrm{Pop}(y),\, z\sim P_{\mathrm{enc}}(z|y)} \left[-\ln \frac{P_{\mathrm{pri}}(z)P_{\mathrm{gen}}(y|z)}{P_{\mathrm{enc}}(z|y)}\right]$$

Gradient descent on the encoder parameters must take into account the fact that we are sampling from the encoder.

To handle this we sample noise ϵ from a fixed noise distribution and replace z with a determinstc function $z_{\mathrm{enc}}(y,\epsilon)$

$$\mathrm{enc}^*, \mathrm{pri}^*, \mathrm{gen}^* = \operatorname*{argmin}_{\mathrm{enc,pri,gen}} \; E_{y,\epsilon,z=\hat{z}_{\mathrm{enc}}(y)+\sigma\epsilon} \left[-\ln \frac{P_{\mathrm{pri}}(z)P_{\mathrm{gen}}(y|z)}{P_{\mathrm{enc}}(z|y)}\right]$$

The Reparameterization Trick

$$\mathrm{enc}^*, \mathrm{pri}^*, \mathrm{gen}^* = \operatorname*{argmin}_{\mathrm{enc},\mathrm{pri},\mathrm{gen}} \; E_{y,\epsilon,z=\hat{z}_{\mathrm{enc}}(y)+\sigma\epsilon} \left[-\ln \frac{P_{\mathrm{pri}}(z)P_{\mathrm{gen}}(y|z)}{P_{\mathrm{enc}}(z|y)}\right]$$

To get gradients we must have that $\hat{z}_{\mathrm{enc}}(y)$ is a differentiable function of the encoder parameters.

Optimizing the encoder is tricky for discrete z. Discrete z is handled effectively in EM algorithms and general vector quantization (VQ) methods.

The KL-divergence Optimization

$$
\begin{aligned}
\mathcal{L}(y) &= E_{z \sim P_{\mathrm{enc}}(z|y)} \left[-\ln \frac{P_{\mathrm{pri}}(z) P_{\mathrm{gen}}(y|z)}{P_{\mathrm{enc}}(z|y)} \right] \\
&= KL(P_{\mathrm{enc}}(z|y), P_{\mathrm{pri}}(z)) + E_{z \sim P_{\mathrm{enc}}(z|y)} \left[-\ln P_{\mathrm{gen}}(y|z) \right] \\
&= \frac{||\hat{z}_{\mathrm{enc}}(y) - \hat{z}_{\mathrm{pri}}||^2}{2\sigma^2} + E_{\epsilon} \, \frac{||y - \hat{y}_{\mathrm{gen}}(\hat{z}_{\mathrm{enc}}(y) + \epsilon)||^2}{2\sigma^2}
\end{aligned}
$$

A closed-form expression for the KL term avoids sampling noise.

EM is Alternating Optimization of the ELBO Loss

Expectation Maximimization (EM) applies in the (highly special) case where the exact posterior $P_{\mathrm{pri,gen}}(z|y)$ is samplable and computable. EM alternates exact optimization of enc and the pair (pri, gen) in:

$$\text{VAE:} \qquad \mathrm{pri}^*, \mathrm{gen}^* = \operatorname*{argmin}_{\mathrm{pri,gen}} \min_{\mathrm{enc}} E_{y,\; z \sim P_{\mathrm{enc}}(z|y)} \; -\ln \frac{P_{\mathrm{pri,gen}}(z,y)}{P_{\mathrm{enc}}(z|y)}$$

$$\text{EM:} \qquad \mathrm{pri}^{t+1}, \mathrm{gen}^{t+1} = \operatorname*{argmin}_{\mathrm{pri,gen}} \quad E_{y,\; z \sim P_{\mathrm{pri}^t,\mathrm{gen}^t}(z|y)} \; -\ln P_{\mathrm{pri,gen}}(z,y)$$

Inference (E Step)	Update (M Step)
$P_{\mathrm{enc}}(z|y) = P_{\mathrm{pri}^t,\mathrm{gen}^t}(z|y)$	Hold $P_{\mathrm{enc}}(z|y)$ fixed

END