TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

Improving Diffusion Models

Improved Denoising Diffusion Probabilistic Models

Nichol and Dhariwal, February 2021

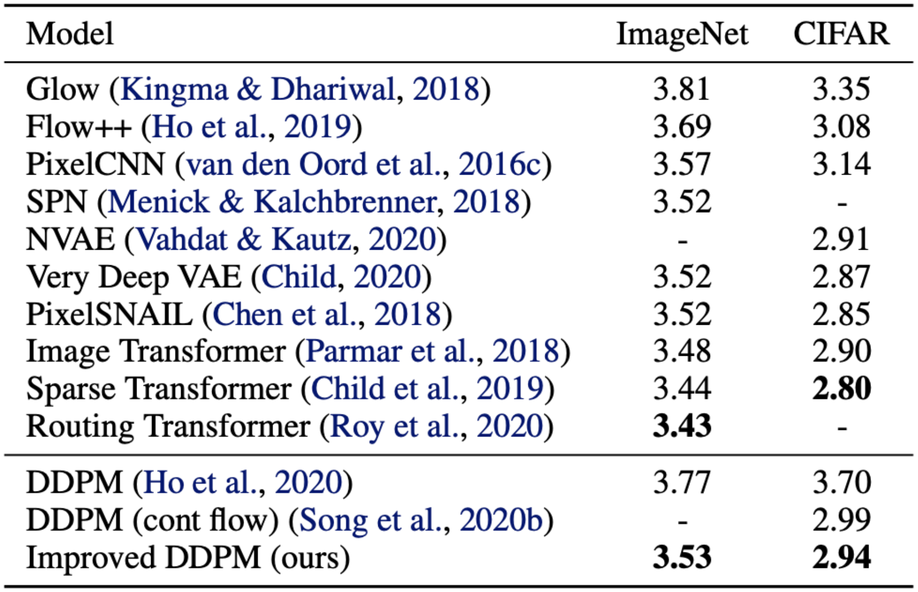
We can compare any two models of a distribution by computing upper bounds on cross-entropy loss for each model.

Since gradient descent on corss entropy (GPT-3) is so successful, maybe we shuld also be doing **graduate student descent** on cross entropy.

In other words, cross entropy may be an undervalued metric for comparing different systems trained with different architectures.

Improved Cross-Entropy Loss

For image models the cross entropy is generally refered to as negative log likelihood (or NLL) and is measured in bits per image channel.



Model	ImageNet	CIFAR
Glow (Kingma & Dhariwal, 2018)	3.81	3.35
Flow++ (Ho et al., 2019)	3.69	3.08
PixelCNN (van den Oord et al., 2016c)	3.57	3.14
SPN (Menick & Kalchbrenner, 2018)	3.52	-
NVAE (Vahdat & Kautz, 2020)	-	2.91
Very Deep VAE (Child, 2020)	3.52	2.87
PixelSNAIL (Chen et al., 2018)	3.52	2.85
Image Transformer (Parmar et al., 2018)	3.48	2.90
Sparse Transformer (Child et al., 2019)	3.44	**2.80**
Routing Transformer (Roy et al., 2020)	**3.43**	-
DDPM (Ho et al., 2020)	3.77	3.70
DDPM (cont flow) (Song et al., 2020b)	-	2.99
Improved DDPM (ours)	**3.53**	**2.94**

Reducing the variance of the VLB

The community seems to be now calling the ELBO the VLB (variational lower bound). Perhaps a slight to Bayesians.

For a progressive VAE with layers $z_0, \ldots, z_L$ where $z_0 = y$ the VLB (ELBO) gives

$$-\ln p_{\mathrm{gen}}(z_0) \;\leq\; E_{\mathrm{enc}} \;\; - \ln \frac{p_{\mathrm{gen}}(z_L, \ldots, z_0)}{p_{\mathrm{enc}}(z_1, \ldots, z_L | z_0)}$$

$$= \; E_{\mathrm{enc}} - \ln p_{\mathrm{pri}}(z_L) - \sum_{\ell} \frac{\ln p_{\mathrm{dec}}(z_{\ell-1} | z_\ell)}{\ln p_{\mathrm{enc}}(z_\ell | z_{\ell-1})}$$

Reducing the variance of the VLB

$$
\begin{aligned}
-\ln p_{\mathrm{gen}}(z_0) \;&\leq\; E_{\mathrm{enc}}\;\; -\ln \frac{p_{\mathrm{gen}}(z_L,\ldots,z_0)}{p_{\mathrm{enc}}(z_1,\ldots,z_L|z_0)} \\
&=\; E_{\mathrm{enc}} - \ln p_{\mathrm{pri}}(z_L) - \sum_{\ell} \frac{\ln p_{\mathrm{dec}}(z_{\ell-1}|z_\ell)}{\ln p_{\mathrm{enc}}(z_\ell|z_{\ell-1})}
\end{aligned}
$$

For a diffusion model this expression can be converted into a form involving KL divergences between Gaussians which can be calculated analytically.

Measuing Cross Entropy

$$
\begin{aligned}
-\ln p_{\mathrm{gen}}(z_0) \;&\leq\; E_{\mathrm{enc}} - \ln p_{\mathrm{pri}}(z_L) - \sum_\ell \ln \frac{p_{\mathrm{dec}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_\ell|z_{\ell-1})} \\
&=\; E_{\mathrm{enc}} - \ln p_{\mathrm{pri}}(z_L) - \sum_\ell \ln \frac{p_{\mathrm{dec}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_\ell|z_{\ell-1}, z_0)} \\
&=\; E_{\mathrm{enc}} - \ln p_{\mathrm{pri}}(z_L) - \sum_\ell \ln \frac{p_{\mathrm{dec}}(z_{\ell-1}|z_\ell) p(z_{\ell-1}|z_0)}{p_{\mathrm{enc}}(z_\ell, z_{\ell-1}|z_0)} \\
&=\; E_{\mathrm{enc}} - \ln p_{\mathrm{pri}}(z_L) - \sum_\ell \ln \frac{p_{\mathrm{dec}}(z_{\ell-1}|z_\ell) p_{\mathrm{enc}}(z_{\ell-1}|z_0)}{p_{\mathrm{enc}}(z_{\ell-1}|z_\ell, z_0) p_{\mathrm{enc}}(z_\ell|z_0)}
\end{aligned}
$$

Measuring The Cross Entropy

$$
\begin{aligned}
-\ln p_{\mathrm{gen}}(z_0) \;&\leq\; E_{\mathrm{enc}} - \ln p_{\mathrm{pri}}(z_L) - \sum_{\ell} \ln \frac{p_{\mathrm{dec}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_{\ell-1}|z_\ell, z_0)} - \ln \frac{p_{\mathrm{dec}}(z_{\ell-1}|z_0)}{p_{\mathrm{enc}}(z_\ell|z_0)} \\
&=\; E_{\mathrm{enc}} - \ln \frac{p_{\mathrm{pri}}(z_L)}{p_{\mathrm{enc}}(z_L|z_0)} - \sum_{\color{red}{\ell\geq 2}} \ln \frac{p_{\mathrm{dec}}(z_{\ell-1}|z_\ell)}{p_{\mathrm{enc}}(z_{\ell-1}|z_\ell, z_0)} - \ln p_{\mathrm{dec}}(z_0|z_1) \\
&=\; E_{\mathrm{enc}} \left\{ \begin{array}{l} KL(p_{\mathrm{pri}}(z_L), p_{\mathrm{enc}}(z_L|z_0)) \\ \\ +\sum_{\ell\geq 2} KL(p_{\mathrm{dec}}(z_{\ell-1}|z_\ell), p_{\mathrm{enc}}(z_{\ell-1}|z_\ell, z_0)) \\ \\ -\ln p_{\mathrm{dec}}(z_0|z_1) \end{array} \right.
\end{aligned}
$$

Measuring The Cross Entropy

$$-\ln p_{\mathrm{gen}}(z_0) \;\leq\; E_{\mathrm{enc}} \left\{ \begin{array}{l} KL(p_{\mathrm{pri}}(z_L), p_{\mathrm{enc}}(z_L|z_0)) \\ \\ + \sum_{\ell > 1} KL(p_{\mathrm{dec}}(z_{\ell-1}|z_\ell), p_{\mathrm{enc}}(z_{\ell-1}|z_\ell, z_0)) \\ \\ - \ln p_{\mathrm{dec}}(z_0|z_1) \end{array} \right.$$

All of the KL-divergences can be computed analytically from Gaussians. This reduces the variance in estimating the bound.

Nichol and Dhariwal compute $-\ln p_{\mathrm{dec}}(z_0|z_1)$ by treating each image channel as a discrete set of 256 values and computing the probability that a draw from the computed Gaussian rounds to the actual discrete value.

Optimizing the Decoder Variances

$$\mathrm{dec}(z_\ell, \ell) = \frac{1}{\sqrt{1-\sigma_\ell^2}}\left(z_\ell - \sigma_\ell\, \epsilon(z_\ell, \ell)\right) + \tilde{\sigma}_\ell \delta \qquad \delta \sim \mathcal{N}(0, I)$$

Nichol and Dhariwal train the decoder noise $\tilde{\sigma}_\ell$ with the VLB objective.

This significantly improves the value of the VLB (not surprising).

Optimizing the Decoder Variances

$$\mathrm{dec}(z_\ell, \ell) = \frac{1}{\sqrt{1-\sigma_\ell^2}}\left(z_\ell - \sigma_\ell\,\epsilon(z_\ell,\ell)\right) + \tilde{\sigma}_\ell \delta \qquad \delta \sim \mathcal{N}(0, I)$$

Training $\tilde{\sigma}_\ell$ by optimizing the VLB gives $\tilde{\sigma}_\ell < \sigma_\ell$ which makes the decoder more deterministic.

We can think of $\tilde{\sigma}_\ell$ as secifying the numerical precision of the decoder output.

Making the decoder more deterministic (higher precision) reduces the required number of levels from 1000 to 50 while still achieving good cross entropy loss.

Optimizing the Decoder Variances

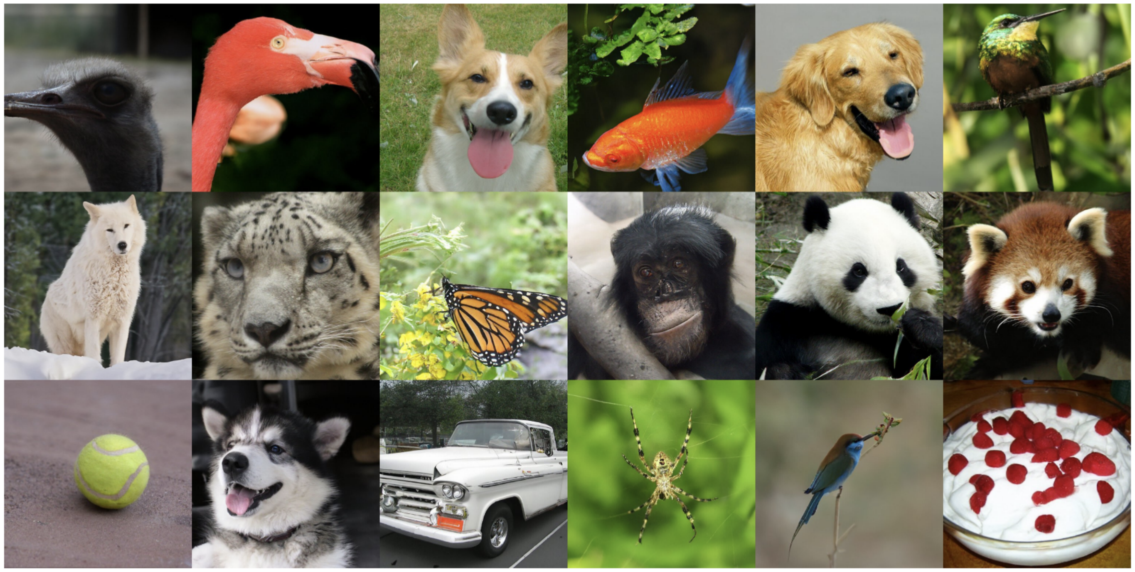
Song, Meng and Ermon, October 2021, found independently that deterministic decoding allows the number of levels to be reduced while achieving similar FID scores.

The deterministic decoding version is called DDIM (denoising diffusion implicit models) as opposed to the original version DDPM (denoising diffusion probabilistic models).

But decoding over infinite precision reals destroys the VLB.

Diffusion Models Beat GANs on Image Synthesis

Dharwali and Nichol, May 2021



Class-Conditional Generation

Consider training a model Φ on pairs (x, y) using

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}} \; -\ln \; P_\Phi(y|x)$$

We are interested in the case where y is an image or sound wave and we consider sampling y from $P_\Phi(y|x)$.

Here we consider the case where x is a class label (as in a class conditional GAN).

We assume that we can train a model of $P(x|y)$ — for example an ImageNet classifier.

Class-Conditional Generation

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; -\ln \; P_\Phi(y|x)$$

We assume a model of $P(x|y)$.

Sampling y from $P_\Phi(y|x)$ can be done by sampling the pair (y,x) from a the joint distribution defined by $p_\Phi(y)P(x|y)$.

Using the Score Matching (thermodynamic) Interpretation

Compare the decoding update

$$z_{\ell-1} = \frac{1}{\sqrt{1-\sigma_\ell^2}}\left(z_\ell - \sigma_\ell\,\epsilon(z_\ell,\ell)\right) \;+\; \tilde{\sigma}_\ell \delta$$

with the thermodynamic interpretation

$$z_{t+1} = z_t + \nabla_z\,\mathrm{score}(z) + \sigma\delta$$

Using the Score Matching (thermodynamic) Interpretation

$$z_{\ell-1} = \frac{1}{\sqrt{1-\sigma_\ell^2}}\left(z_\ell - \sigma_\ell\ \epsilon(z_\ell,\ell)\right)\ +\ \tilde{\sigma}_\ell\delta$$

$$z_{t+1} = z_t + \nabla_z \operatorname{score}(z) + \sigma\delta$$

Ignoring the expansion term we get that $-\sigma_\ell\epsilon(z_\ell,\ell)$ is identified with $\nabla_z \operatorname{score}(z)$.

Using the Score Matching (thermodynamic) Interpretation

$$z_{\ell-1} = \frac{1}{\sqrt{1-\sigma_\ell^2}}\left(z_\ell - \sigma_\ell\,\epsilon(z_\ell,\ell)\right) \;+\; \tilde{\sigma}_\ell \delta$$

$$z_{t+1} = z_t + \nabla_z\,\mathrm{score}(z) + \sigma\delta$$

Taking $\mathrm{score}(z,x) = \mathrm{score}(z) + \ln P(x|z)$, and introducing a hyperparameter s, we can do an update on the joint score as

$$\mathrm{dec}(z_\ell,\ell) = \frac{1}{\sqrt{1-\sigma_\ell^2}}\left(z_\ell - \sigma_\ell\,\epsilon(z_\ell,\ell) + s\nabla_z \ln P(x|z)\right) + \tilde{\sigma}_\ell \delta$$

Using the Score Matching (thermodynamic) Interpretation

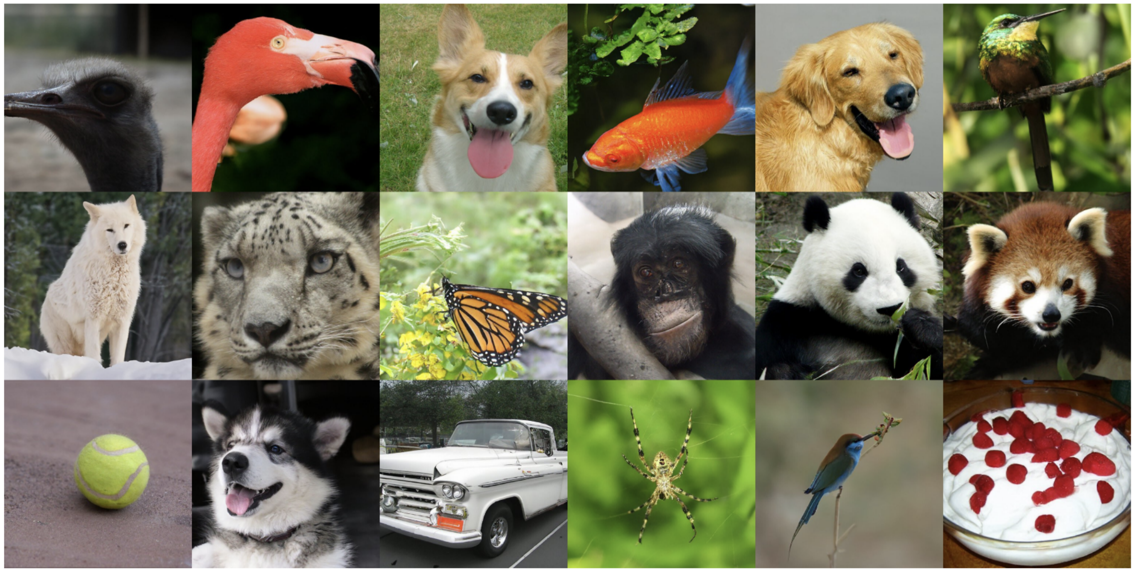
$$\mathrm{dec}(z_\ell, \ell) = \frac{1}{\sqrt{1-\sigma_\ell^2}} \left(z_\ell - \sigma_\ell\, \epsilon(z_\ell, \ell) + s \nabla_z \ln P(x|z)\right) + \tilde{\sigma}_\ell \delta$$

Empirically it was found that $s > 1$ is needed to get good class specificity of the generated image.

Other Improvements

Various architectural choices in the U-Net were optimized based on FID score (not NLL).

Voila: Class-Conditional Image Generation



Classifier-Free Diffusion Guidance

Ho and Salimans, December 2021 (NeurIPS workshop)

We assume training data consisting of (x, y) pairs.

An obvious approach to conditional diffusion models $P(y|x)$ is to draw a pair (x, y) and pass the conditioning information x to the decoder $\epsilon(z_\ell, \ell, x)$ when decoding for $z_0 = y$.

This simple modification to the decoder seems to be insufficient (no one is using the naive approach).

Instead people have modified the naive approach by modeling $P(x|z)$ in terms of the conditional decoder.

Classifier-Free Diffusion Guidance

As in the naive approach we pass the conditional information to the decoder and train $\epsilon(z_\ell, \ell, x)$ to generate the y associated with x.

But 5% of the time we set $x = \emptyset$ where $\emptyset$ is a fixed value unrelated to the image.

We then interpret generation conditioned on $\emptyset$ to model the marginal distribution on y.

Classifier-Free Diffusion Guidance

We now observe

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \propto \frac{p(y|x)}{p(y)}$$

This allows to take score$(x|y) =$ score$(y|x) -$ score(y).

Following the thermodynamic interpretation we interpret $-\epsilon(z_\ell, \ell, x)$ as ∇_z score$(z_\ell|x)$

This gives

$$\nabla_z \ln P(x|z) = \epsilon(z_\ell, \ell, \emptyset) - \epsilon(z_\ell, \ell, x)$$

Classifier-Free Diffusion Guidance

The classifier-guided decoding

$$\mathrm{dec}(z_\ell,\ell) = \frac{1}{\sqrt{1-\sigma_\ell^2}}\left(\begin{array}{l} z_\ell - \sigma_\ell\epsilon(z_\ell,\ell) \\ + s\nabla_z \ln P(x|z)\end{array}\right) \;+\; \tilde{\sigma}_\ell\delta$$

can be written as

$$\mathrm{dec}(z_\ell,\ell) = \frac{1}{\sqrt{1-\sigma_\ell^2}}(z_\ell - \sigma_\ell\hat{\epsilon}(z_\ell,\ell)) \;+\; \tilde{\sigma}_\ell\delta$$

with

$$\hat{\epsilon}(z_\ell,\ell) = \epsilon(z_\ell,\ell) - s'\nabla_z \ln P(x|z)$$

Classifier-Free Diffusion Guidance

$$\hat{\epsilon}(z_\ell, \ell) = \epsilon(z_\ell, \ell) - s'\nabla_z \ln P(x|z)$$

$$\nabla_z \ln P(x|z) = \epsilon(z_\ell, \ell, \emptyset) - \epsilon(z_\ell, \ell, x)$$

gives

$$\hat{\epsilon}(z_\ell, \ell) = \epsilon(z_\ell, \ell, \emptyset) - s'(\epsilon(z_\ell, \ell, \emptyset) - \epsilon(z_\ell, \ell, x))$$

where we take $s' > 1$ so that we are repelled from $\epsilon(z_\ell, \ell, \emptyset)$

Classifier-Free Diffusion Guidance

$$\hat{\epsilon}(z_\ell, \ell) = \epsilon(z_\ell, \ell, \emptyset) - s'(\epsilon(z_\ell, \ell, \emptyset) - \epsilon(z_\ell, \ell, x))$$

I am not convinced by this derivation but the result is intuitive — drive away from images in general and toward the conditioned distribution to increase the effect of the conditioning.

It seems likely to me that this hurts NLL while strengthening use of the condition information.

GLIDE: Towards Photorealistic Image Generation ...

Nichol, Dhariwal, Ramesh, et al., March 2022

Slides under development.

END