TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

The DALL·E-2 Code Base

Improved Denoising Diffusion Probabilistic Models

Nichol and Dhariwal, February 2021

This paper provides a method for training an "uncertinty level" for each color channel of each pixel.

Later papers in the code base use these uncertainty levels to weight guidance strength for each color channel of each pixel in "guided diffusion".

Guided diffusion with channel-level guiding strength is used in DALLE-2.

Getting Per-Pixel Decoder Uncertainty

Per-pixel decoder yncertainty will be estimated by optimizing the VAE bound on cross-entropy loss.

These papers call it the variational lower bound (VLB) rather than the ELBO.

The paper is written from the perspective of simply optimizing the VLB.

Why Optimize the VLB?

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**

Rewriting the VLB

For a progressive VAE with layers $z_0, \ldots, z_L$ where $z_0 = y$ the VLB is

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

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

Rewriting the VLB

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

Rewriting the VLB

$$
\begin{aligned}
-\ln p_{\text{gen}}(z_0) \;&\leq\; E_{\text{enc}} - \ln p_{\text{pri}}(z_L) - \sum_{\ell} \ln \frac{p_{\text{dec}}(z_{\ell-1}|z_\ell)}{p_{\text{enc}}(z_{\ell-1}|z_\ell, z_0)} - \ln \frac{p_{\text{dec}}(z_{\ell-1}|z_0)}{p_{\text{enc}}(z_\ell|z_0)} \\
&=\; E_{\text{enc}} - \ln \frac{p_{\text{pri}}(z_L)}{p_{\text{enc}}(z_L|z_0)} - \sum_{\ell \geq 2} \ln \frac{p_{\text{dec}}(z_{\ell-1}|z_\ell)}{p_{\text{enc}}(z_{\ell-1}|z_\ell, z_0)} - \ln p_{\text{dec}}(z_0|z_1)
\end{aligned}
$$

$$
=\; E_{\text{enc}} \left\{ \begin{array}{l} KL(p_{\text{enc}}(z_L|z_0), p_{\text{pri}}(z_L)) \\ + \sum_{\ell \geq 2} KL(p_{\text{enc}}(z_{\ell-1}|z_\ell, z_0), p_{\text{dec}}(z_{\ell-1}|z_\ell)) \\ - \ln p_{\text{dec}}(z_0|z_1) \end{array} \right.
$$

Rewriting the VLB

$$-\ln p_{\text{gen}}(z_0) \;\leq\; E_{\text{enc}} \begin{cases} KL(p_{\text{enc}}(z_L|z_0), p_{\text{pri}}(z_L)) \\ +\sum_{\ell \geq 2} KL(p_{\text{enc}}(z_{\ell-1}|z_\ell, z_0), p_{\text{dec}}(z_{\ell-1}|z_\ell)) \\ -\ln p_{\text{dec}}(z_0|z_1) \end{cases}$$

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

Nichol and Dhariwal compute $-\ln p_{\text{dec}}(z_0|z_1)$ 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 Per-Channel Decoder Variances

We now introduce a decoder network $\tilde{\sigma}_\Psi(z_\ell, \ell) \in R^d$ to give the decoder noise level.

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

The decoder noise network $\tilde{\sigma}_\Psi(z_\ell, , \ell) \in R^d$ is trained with the VLB objective.

This improves the value of the VLB.

Optimizing Per-Channel Decoder Variances

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

One can interpret $\tilde{\sigma}(z_\ell, \ell)[i]$ is a level of uncertainty in the decoder estimate of $\epsilon(z_\ell, \ell)[i]$.

The more uncertain the model $\epsilon(z_\ell, \ell)$ the more guidance should be used in adjusting it.

Diffusion Models Beat GANs on Image Synthesis
Dharwali and Nichol, May 2021

This paper introduces guided diffusion.

A form of guided diffusion is used in DALLE-2.

Diffusion Models Beat GANs on Image Synthesis

Dharwali and Nichol, May 2021

Guided diffusion is introduced as an approch to class-conditional image generation for ImageNet.

Class-Conditional Generation

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

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

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

We assume a model of $P(x|y)$ where y is an image and x is a class label.

We want $P(y|x)$.

$$P_\Phi(y|x) = \frac{P(y)P(x|y)}{P(x)} \propto P(y)P(x|y)$$

Score-matching interprets $\epsilon(z_\ell, \ell)$ as $-\nabla_z \ln p(z)$.

Using the Score Matching Interpretation

We now want

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

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

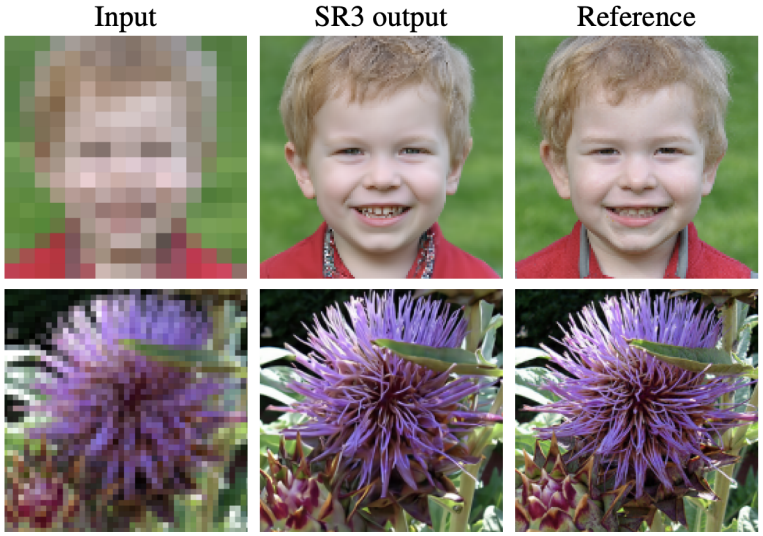
Other Improvements

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

Image Super-Resolution via Iterative Refinement

Saharia et al., April 2021

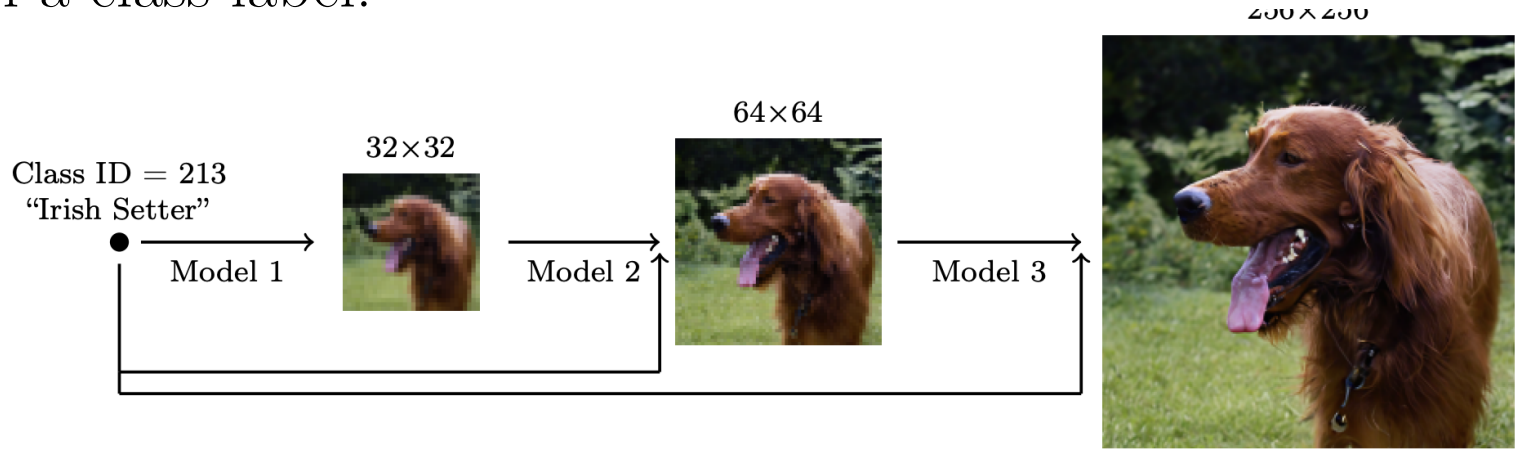
They construct a super-resolution diffusion model as conditional model for pairs for pairs (x, y) with x is a downsampling of y.



Cascaded Diffusion Models ...

Ho et al, May 2021

A series of super-resolution diffusion models each conditioned on a class label.



Classifier-Free Diffusion Guidance

Ho and Salimans, December 2021 (NeurIPS workshop)

We assume training data consisting of (x, y) pairs and we want to generate from the distribution $P(y|x)$. For example generating images from text.

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 $\epsilon(z_\ell, \ell, x)$.

This paper proposes a modification to this naive approach which seems to help.

Classifier-Free Diffusion Guidance

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

We take a score matching interpretation:
They then use

$$\hat{\epsilon}(z_\ell, \ell, x) = \epsilon(z_\ell, \ell, x) - \alpha\epsilon(z_\ell, \ell, \emptyset) \quad \alpha > 0$$

This drives the image away from being generic and strengthens the dependence on x.

END