TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Conditional Diffusion Models and Guidance

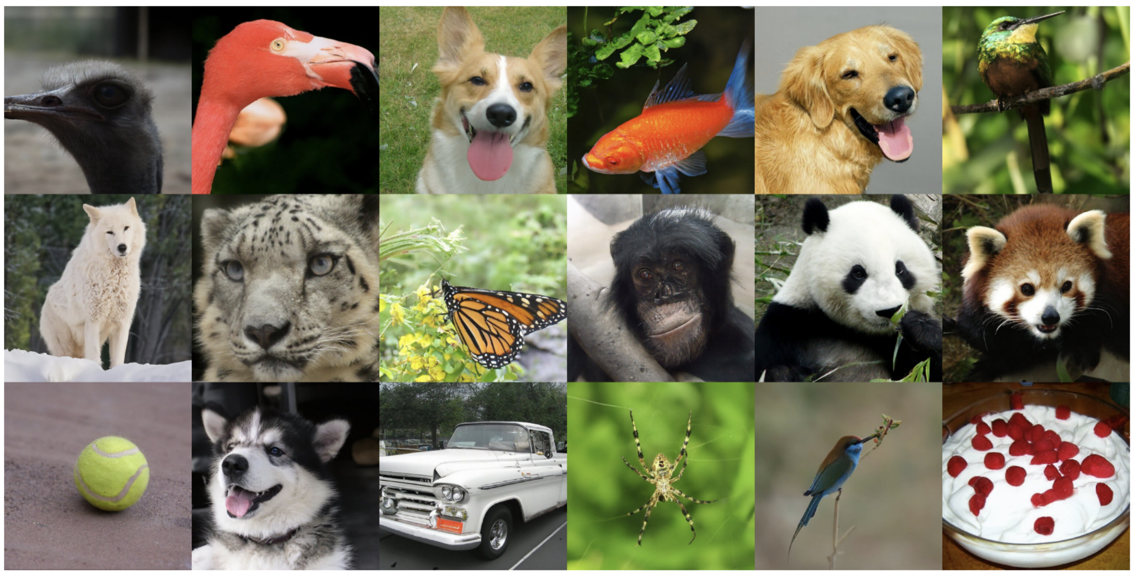
Conditional Diffusion Models and Guidance

Deep unsupervised learning using nonequilibrium thermodynamics
Sohl-Dickstein et al., 2015.

Denoising Diffusion Probabilistic Models (DDPM)
Ho, Jain and Abbeel, (Berkeley, May 2021)

Diffusion Models Beat GANs on Image Synthesis

Dharwali and Nichol (OpenAI, May 2021)



Conditional Diffusion Models

We assume training data consisting of (x, y) pairs and we want to generate from the distribution $P(y|x)$. For example class-conditional image generation.

Previous approaches, such as StyleGAN, have trained a model (a GAN) for each class.

Here we will train a single model which takes the class label as input.

Conditional Diffusion Models

An obvious approach is to pass the conditioning information x to the image generator.

Unfortunately this natural approach to conditioning generates poor images.

It remains true that generating high quality images requires "guidance".

There are two forms of guidance — classifier guidance and self-guidance.

Classifier Guidance

We assume a distribution on pairs (x, y).

We also assume **a classifier** $P(x|y)$. For example x might be the ImageNET label for image y.

We use $p(y|x) \propto P(y)P(x|y)$.

We will generate an image by using $P(x|y)$ to "guide" generation from the unconditional model $\epsilon(z_\ell, \ell)$.

$$z(t - \Delta t) \;=\; z(t) + \eta \left(\nabla_z \ln P_t(z) + s\nabla_z \;\ln P(x|z)\right)$$

Here s is called the scale of the guidance.

Classifier Guidance

$$z(t-\Delta t) \;=\; z(t) + \eta\left(\nabla_z \ln P_t(z) + s\nabla_z\; \ln P(x|z)\right)$$

$$\nabla_z \ln P_t(z) \;=\; \frac{E[y|t,z]-z}{t}$$

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

However, increasing s decreases diversity so we have a diversity/quality trade off.

Other Improvements

Various architectural choices in the U-Net were optimized.

These improvements are used in DALLE-2.

Classifier-Free Diffusion Guidance

Ho and Salimans, (Google Brain, December 2021)

Classification diffusion guidance uses a classification model $P(x|y)$.

This paper introduces "classifier-free" diffusion guidance.

Classifier-free diffusion guidance is used in DALLE-2.

Classifier-Free Diffusion Guidance

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

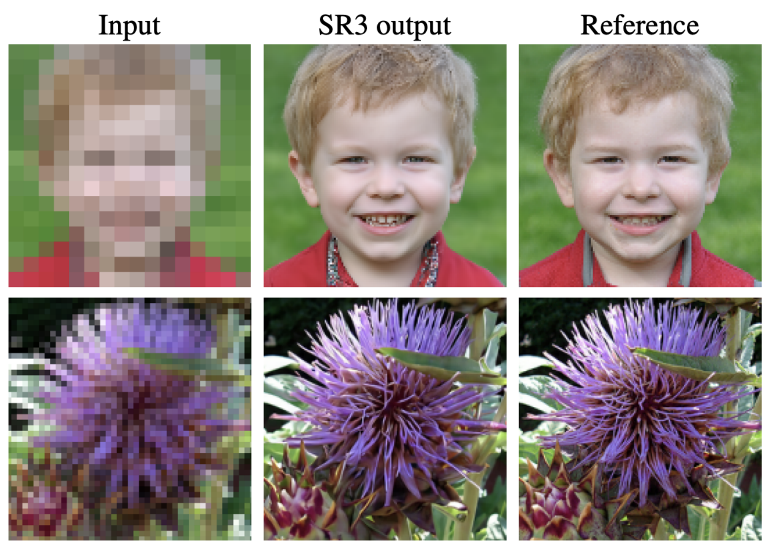
The prior then uses

$$z(t - \Delta t) \;=\; z(t) + \eta\left(s\nabla_z \ln P_t(z|x) - (s-1)\nabla_z \;\ln P(z|\emptyset)\right)$$

Image Super-Resolution via Iterative Refinement

Saharia, Ho et al., April 2021

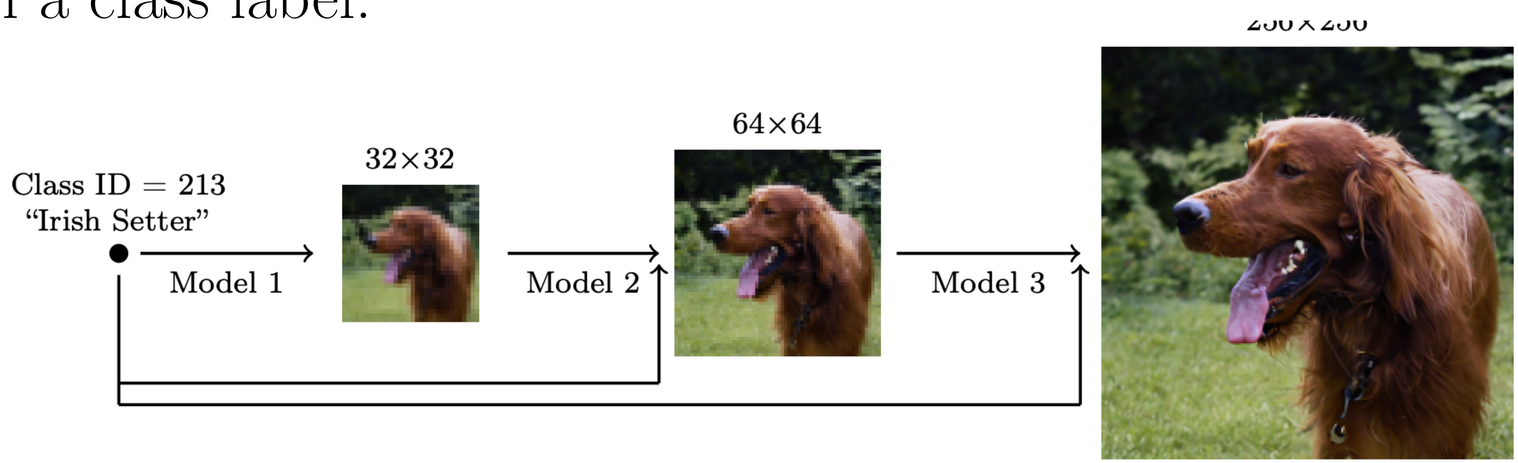
They construct a super-resolution diffusion model as conditional model for pairs for pairs (x, y) with x is a downsampling of y.



Cascaded Diffusion Models ...

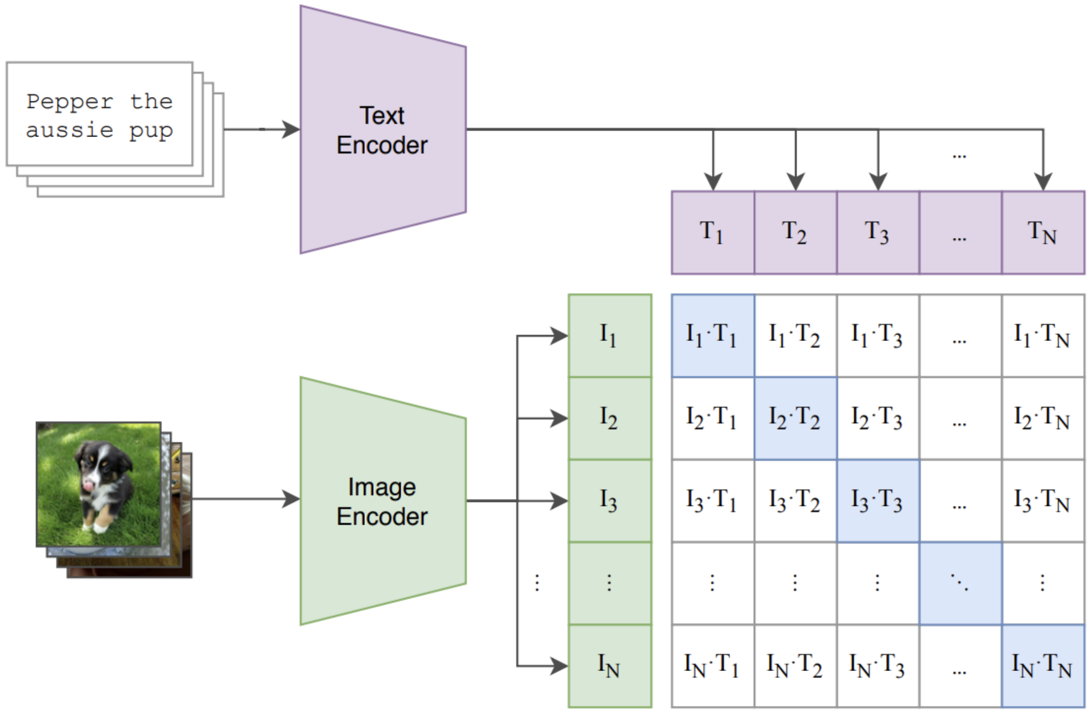
Ho, Saharia et al, May 2021

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



This architecture is used in DALLE-2.

CLIP Does Contrastive Coding



CLIP is used in DALLE-2 and in DALLE-2's predicessor GLIDE.

GLIDE: Towards Photorealistic Image Generation ...

Nichol, Dhariwal, Ramesh, et al., December 2021

GLIDE compares two forms of diffusion guidance.

(a) Classifier-free guidance based on comparing conditioned and unconditioned decoding directions.

(b) Classifer guidance based on CLIP.

Classifier-free (self-guided) GLIDE

Classifier-free GLIDE does not use CLIP.

The classifier-free guidance differs from the original version in that here we are conditioning on text rather than as Imagenet labels.

The text is transformed to a feature vector by a transformer before being fed to the prior.

CLIP-guided GLIDE

Let $C_I(y)$ be the CLIP vector for image y and let $C_T(x)$ be the CLIP vector for text x.

CLIP-based Glide approximates uses

$$\ln P(z|x) \approx C_T(x)^\top C_I(z)$$

CLIP is re-trained to handle noised images.

Upsamling

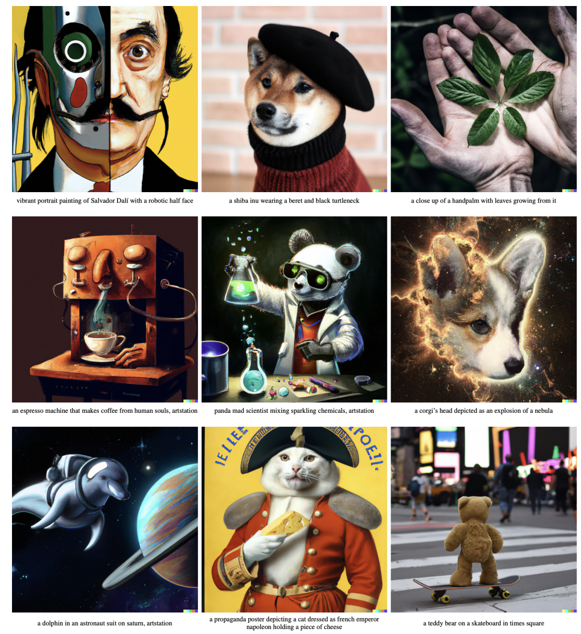
Both GLIDE versions use diffusion upsampling to go from 64×64 to 256×256.

The GLIDE paper concludes that the classifer-free model taking raw text as input is superior to the CLIP-guided model.

DALL·E-2

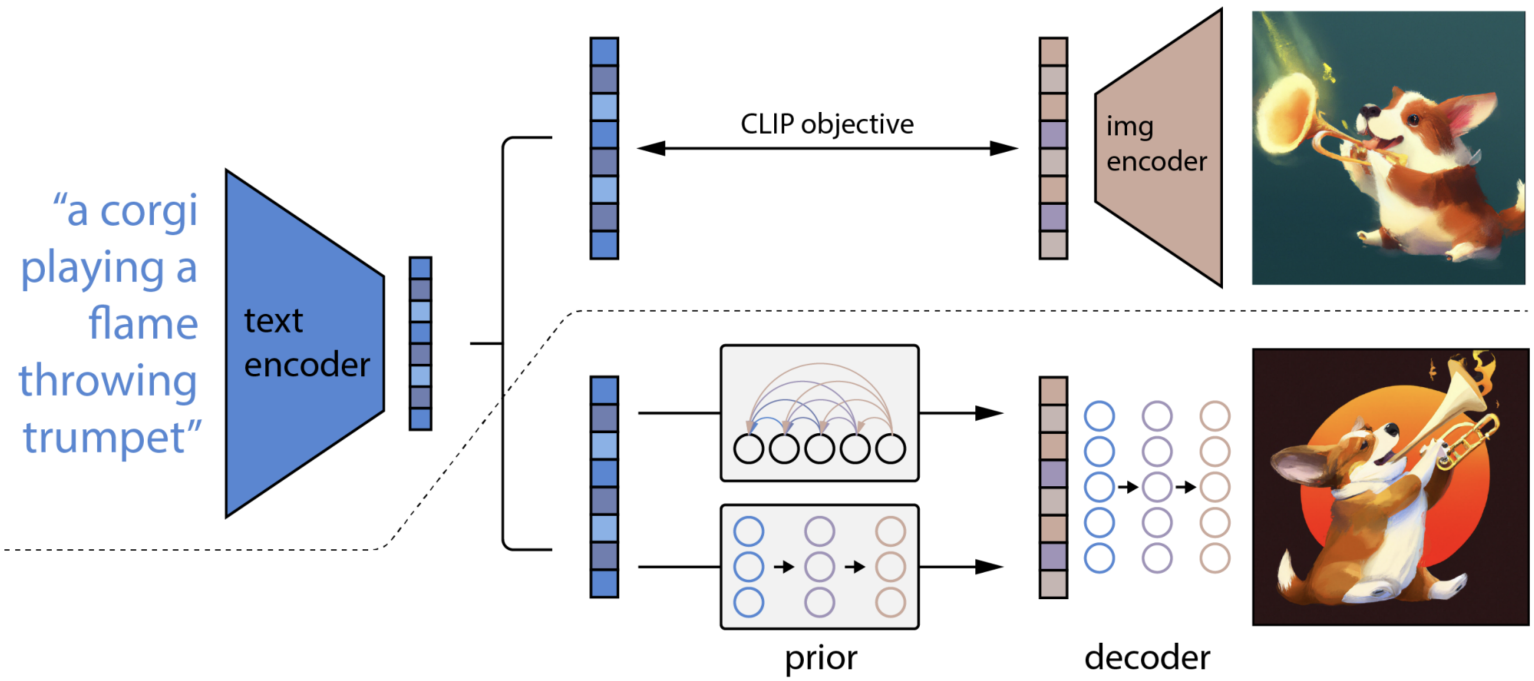
Ramesh, Nichol, Dhariwal, et al., March 2022

panda mad scientist mixing sparkling chemicals, artstation



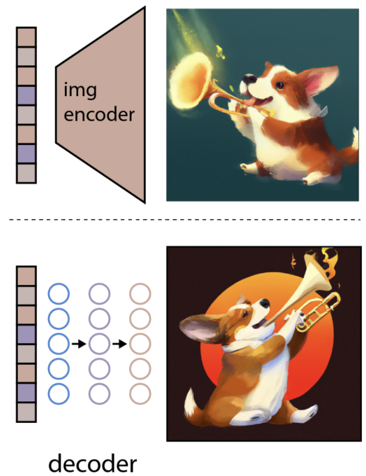
CLIP-guided DALLE-2 is similar in quality to self-guided GLIDE but is more diverse.

DALL·E-2



This figure is misleaning. The lines in the figure do not correspond to the actual data paths of DALLE-2.

A Conditional Image Auto-Encoder



Let $C_I(y)$ denote the CLIP embedding of image y.

$C_I(y)$ is taken to be the encoder of a VAE for y given x.

$P(C_I(y)|x)$ is the optimal prior for this auto-encoder.

$P(y|C_I(y), x)$ is the optimal decoder.

In DALLE-2 the prior and the generator both see the text x.

END