Chapter 7

Synthesis

A second category of applications distinct from prediction is synthesis. It consists of fitting a density model to training samples and providing means to sample from this model.

7.1 Text generation

The standard approach to text synthesis is to use an attention-based, autoregressive model. A very successful model proposed by Radford et al. [2018], is the GPT which we described in § 5.3.

This architecture has been used for very large models, such as OpenAI's 175-billion-parameter GPT-3 [Brown et al., 2020]. It is composed of 96 self-attention blocks, each with 96 heads, and processes tokens of dimension 12,288, with a hidden dimension of 49,512 in the MLPs of the attention blocks.

When such a model is trained on a very large dataset, it results in a Large Language Model (LLM), which exhibits extremely powerful properties. Besides the syntactic and grammatical structure of the language, it has to integrate very diverse knowledge, e.g. to predict the word following "The capital of Japan is", "if water is heated to 100 Celsius degrees it turns into", or "because her puppy was sick, Jane was".

This results in particular in the ability to solve few-shot prediction, where only a handful of training examples are available, as illustrated in Figure 7.1. More surprisingly, when given a carefully crafted prompt, it can exhibit abil-

I: I love apples, O: positive, I: music is my passion, O: positive, I: my job is boring, O: negative, I: frozen pizzas are awesome, O: **positive,**

I: I love apples, O: positive, I: music is my passion, O: positive, I: my job is boring, O: negative, I: frozen pizzas taste like cardboard, O: **negative,**

I: water boils at 100 degrees, O: physics, I: the square root of two is irrational, O: mathematics, I: the set of prime numbers is infinite, O: mathematics, I: gravity is proportional to the mass, O: **physics,**

I: water boils at 100 degrees, O: physics, I: the square root of two is irrational, O: mathematics, I: the set of prime numbers is infinite, O: mathematics, I: squares are rectangles, O: **mathematics,**

Figure 7.1: *Examples of few-shot prediction with a 120 million parameter GPT model from Hugging Face. In each example, the beginning of the sentence was given as a prompt, and the model generated the part in bold.*

ities for question answering, problem solving, and chain-of-thought that appear eerily close to high-level reasoning [Chowdhery et al., 2022; Bubeck et al., 2023].

Due to these remarkable capabilities, these models are sometimes called foundation models [Bommasani et al., 2021].

However, even though it integrates a very large body of knowledge, such a model may be inad-

equate for practical applications, in particular when interacting with human users. In many situations, one needs responses that follow the statistics of a helpful dialog with an assistant. This differs from the statistics of available large training sets, which combine novels, encyclopedias, forum messages, and blog posts.

This discrepancy is addressed by fine-tuning such a language model (see § 3.6). The current dominant strategy is Reinforcement Learning from Human Feedback (RLHF) [Ouyang et al., 2022], which consists of creating small labeled training sets by asking users to either write responses or provide ratings of generated responses. The former can be used as-is to fine-tune the language model, and the latter can be used to train a reward network that predicts the rating and use it as a target to fine-tune the language model with a standard Reinforcement Learning approach.

7.2 Image generation

Multiple deep methods have been developed to model and sample from a high-dimensional density. A powerful approach for image synthesis relies on inverting a diffusion process. Such a generative model is referred to, somehow incorrectly, as a diffusion model.

The principle consists of defining analytically a process that gradually degrades any sample, and consequently transforms the complex and unknown density of the data into a simple and well-known density such as a normal, and training a deep architecture to invert this degradation process [Ho et al., 2020].

Given a fixed T, the diffusion process defines a probability distribution over series of $T+1$ images as follows: sample x_0 uniformly from the dataset, and then sequentially sample $x_{t+1} \sim p(x_{t+1} \mid x_t), t = 0, \ldots, T-1$, where the conditional distribution p is defined analytically and such that it gradually erases the structure that was in x_0. The setup should degrade the signal so much that the distribution $p(x_T)$ has a known analytical form which can be sampled.

For instance, Ho et al. [2020] normalize the data to have a mean of 0 and a variance of 1, and their

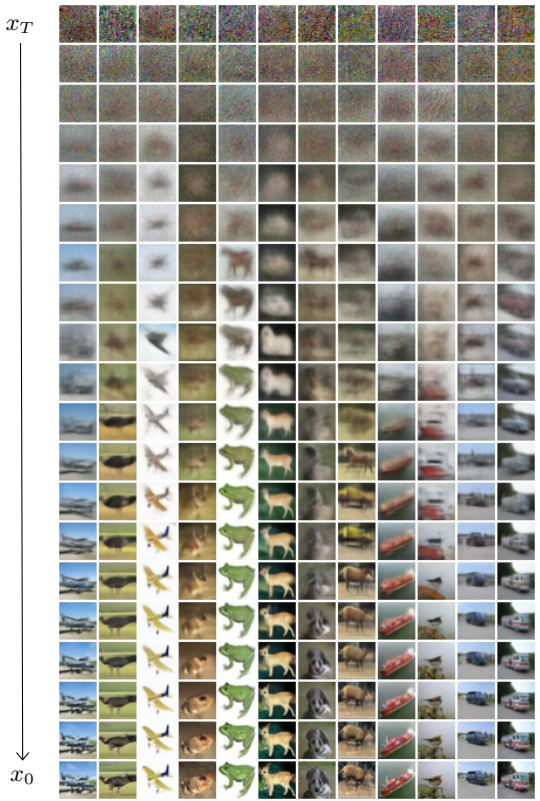

Figure 7.2: *Image synthesis with denoising diffusion [Ho et al., 2020]. Each sample starts as a white noise x_T (top), and is gradually de-noised by sampling iteratively $x_{t-1} \mid x_t \sim \mathcal{N}(x_t + f(x_t, t; w), \sigma_t)$.*

diffusion process consists of adding a bit of white noise and re-normalizing the variance to 1. This process exponentially reduces the importance of x_0, and x_t's density can rapidly be approximated with a normal.

The denoiser f is a deep architecture that should model and allow sampling from $f(x_{t-1}, x_t, t; w) \simeq p(x_{t-1} \mid x_t)$. It can be shown, thanks to a variational bound, that if this one-step reverse process is accurate enough, sampling $x_T \sim p(x_T)$ and denoising T steps with f results in x_0 that follows $p(x_0)$.

Training f can be achieved by generating a large number of sequences $x_0^{(n)}, \ldots, x_T^{(n)}$, picking a t_n in each, and maximizing

$$\sum_n \log f\left(x_{t_n-1}^{(n)}, x_{t_n}^{(n)}, t_n; w\right).$$

Given their diffusion process, Ho et al. [2020] have a denoising of the form:

$$x_{t-1} \mid x_t \sim \mathcal{N}(x_t + f(x_t, t; w); \sigma_t), \qquad (7.1)$$

where σ_t is defined analytically.

In practice, such a model initially hallucinates structures by pure luck in the random noise, and

then gradually builds more elements that emerge from the noise by reinforcing the most likely continuation of the image obtained thus far.

This approach can be extended to text-conditioned synthesis, to generate images that match a description. For instance, Nichol et al. [2021] add to the mean of the denoising distribution of Equation 7.1 a bias that goes in the direction of increasing the CLIP matching score (see § 6.6) between the produced image and the conditioning text description.