拡散モデル（Diffusion Model）を理解するために

20B01392 松本侑真

2023 年 9 月 18 日

概要

近年、機械学習を応用した生成系 AI がブームである。生成系 AI を使うことで、プロンプトに入力したテキストデータから画像データなどを出力することができる。拡散モデルを用いた有名な生成系 AI の Stable Diffusion では、無料でプロンプトに入力した文字（例：かわいい犬）から、かわいい犬の画像を出力することができる。このような生成系 AI で用いられる拡散モデルについて理解するための基礎を説明する。

目次

1　生成モデルでは何を行っているのか

生成モデル（拡散モデル）は、事前に与えられた教師データを元にモデルを学習することで、未知の入力に対して最適化された出力をアウトプットすることができる。例えば、いろいろな人間の顔データを学習させた生成モデルを用いると、「40 歳のおじさん」や「20 代のアイドル」といった入力を元にして、学習されたモデルで生成した顔画像を出力する。

生成モデルの学習というのは、「モデルの入力から欲しいデータを生み出す確率分布」をいかにして見つけるかということである。すなわち、任意のデータ $\vec{x}$ は、とある 1 つの確率分布 $p(\vec{x})$ に従って生み出されるという仮定のもと、自分のモデルをその $p(\vec{x})$ に限りなく近づけることが生成モデルにおける学習である。すなわち、あらゆるデータが従う確率分布 $p(\vec{x})$ そのものを手に入れることができれば、$p(\vec{x})$ の値を出力するような $\vec{x}$ を見つけることが可能であるだろう。これによって、$p(\vec{x})$ =「20 代のアイドル」を満たす画像 $\vec{x}$ を出力することができる。

学習を行う前に得られているもの

モデルの学習を行うためには、そのための学習データが必要である。n 個の学習データ $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n$ が手元にあるとする。このとき、

$$p_0(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\vec{x} - \vec{x}_i) \tag{1.1}$$

を経験分布と呼ぶ。この経験分布は、n が十分に大きいとき、あらゆるデータを生み出す確率分布 $p(x)$ に非常に近いであろう：

$$p_0(\vec{x}) \approx p(\vec{x})\ 。 \tag{1.2}$$

なお、データを連続変数として扱う場合は $\delta(\vec{x} - \vec{x}_i)$ をデルタ関数、離散変数として扱う場合はクロネッカーのデルタとして扱う。

学習のために必要なもの

生成モデルにおいて学習させるための確率分布は自ら用意する必要がある。すなわち、パラメータ θ を用いて、

$$p(\vec{x}) \approx q_\theta(\vec{x}) \tag{1.3}$$

となるようなモデル $q_\theta(\vec{x})$ を見つけたい。モデル $q_\theta(\vec{x})$ の関数形は計算しやすい形を用いて、パラメータ θ を最適化することで $p(\vec{x})$ に近づけるというアプローチを取る。

2　ボルツマンマシン

ボルツマンマシンとは、モデル $q_\theta(\vec{x})$ が

$$q_\theta(\vec{x}) = \frac{1}{Z_\theta}\exp(-E_\theta(\vec{x}))\ 。 \tag{2.1}$$

と表されるものである。このような関数形を Gibbs-Boltzmann 分布（指数関数分布族）と呼ぶ。ここで、規格化定数 Z_θ は

$$Z_\theta = \sum_{\vec{x}} \exp(-E_\theta(\vec{x})) \tag{2.2}$$

である。このように定義された $q_\theta(\vec{x})$ を $p(\vec{x})$ に近づけることが目標である。しかし、$p(\vec{x})$ の形は誰も知らない。今手元にあるのは経験分布 $p_0(\vec{x})$ だけである。そのため、学習データの数 n が十分に多いとして、$q_\theta(\vec{x})$ を経験分布 $p_0(\vec{x})$ に近づけることを行う。

2.1 Kullback-Leibler 情報量

学習を行うためには、2 つの確率分布 $p(\vec{x})$, $q(\vec{x})$ の近さを定義する必要がある。この「近さ」は、データ $\vec{x}$ を入力した際の 2 つの確率分布間の「距離 $D(p \parallel q)$」を測ることで定義することができるだろう。このような関数 $D(p \parallel q)$ はどのように設定しても良いのだが、今回は、Kullback-Leibler divergence（KL 情報量）$D_{\mathrm{KL}}(p \parallel q)$ を導入する：

$$D_{\mathrm{KL}}(p(\vec{x}) \parallel q(\vec{x})) = \sum_{\vec{x}} p(\vec{x}) \ln \frac{p(\vec{x})}{q(\vec{x})} \text{。} \tag{2.3}$$

KL 情報量には

(i). $D(p \parallel q) \geq 0$（等号成立は $p = q$ のとき）
(ii). $D(p \parallel q) \neq D(q \parallel p)$（非対称性）

といった性質がある。特に、KL 情報量は 0 以上であるといった性質を良く用いる。この性質を 2 つの方法で証明してみる。

Gibbs の不等式を用いる方法

Gibbs の不等式とは、$x \geq 0$ において

$$x - 1 \geq \ln x \tag{2.4}$$

が成立する不等式である。

Jensen の不等式を用いる方法

3 隠れ変数を導入してリッチなモデルへ

4 制限ボルツマンマシン

5 マルコフ連鎖モンテカルロ法（MCMC）

6 交換モンテカルロ法

7 MCMC を用いないボルツマンマシン

8 変分オートエンコーダー

9 階層変分オートエンコーダ