Denoising Diffusion Probabilistic Models

拡散確率モデルによる高品質画像生成

Jonathan Ho, Ajay Jain, Pieter Abbeel
University of California, Berkeley
{jonathanho, ajayj, pabbeel}@berkeley.edu

要約 (Abstract)

我々は、拡散確率モデル（diffusion probabilistic models）を用いた高品質な画像生成の結果を示します。このモデルは、非平衡熱力学の観点から着想を得た潜在変数モデルの一種です。最良の結果は、拡散確率モデルとラングヴィンダイナミクスに基づくノイズ除去スコアマッチングとの新しい関係に基づき設計された重み付き変分境界を用いて訓練することで得られました。また、このモデルは進行的な情報損失を伴う圧縮スキームを自然に受け入れ、自動回帰的復号化の一般化として解釈できます。

CIFAR-10 データセット上では、Inception スコア 9.46 および最先端の FID スコア 3.17 を達成し、256 × 256 サイズの LSUN データセットでは ProgressiveGAN に匹敵するサンプル品質を達成しました。我々の実装は GitHubで公開されています。

導入 (Introduction)

最近、ディープ生成モデルの進展により、さまざまなデータモダリティにおいて高品質なサンプルが生成されています。これには、生成的敵対ネットワーク（GANs）[?]、自動回帰モデル [?]、フロー [?]、および変分オートエンコーダー（VAEs）[?] などが含まれ、印象的な画像や音声サンプルを生成しています。また、エネルギーベースのモデルやスコアマッチングにおける顕著な進展もあり、これらのモデルは GAN に匹敵する画像を生成できることが示されています [?, ?]。

本研究では、拡散確率モデル（以下、拡散モデルと略称）[?] の進展について紹介します。このモデルは、有限時間内にデータと一致するサンプルを生成するように訓練された、変分推論に基づくパラメータ化されたマルコフ連鎖です。この連鎖は、データにノイズを徐々に加える「拡散プロセス」を逆転させることで機能します。ノイズが小さなガウス分布で構成されている場合、サンプリング連鎖の遷移を条件付きガウス分布に設定するだけで十分であり、これにより単純なニューラルネットワークによるパラメータ化が可能となります。

これまでの研究では、拡散モデルが高品質なサンプルを生成する能力を持つことは示されていませんでした。本研究では、拡散モデルが実際に高品質なサンプルを生成可能であり、時には他の生成モデルよりも優れていることを示します。また、特定のパラメータ化により、訓練中の複数のノイズレベルにわたるノイズ除去スコアマッチングと等価であることを明らかにしました。この新たな発見に基づき、変分境界が簡略化され、従来の拡散モデルよりも優れた結果を達成しました。

本研究では、CIFAR-10 および LSUN ベンチマークデータセットで拡散モデルの性能を検証しました。非条件付きモデルであるにもかかわらず、拡散モデルは CIFAR-10 において文献中の多くのクラス条件付き生成モデルと同等、またはそれを上回る性能を示しました。

背景 (Background)

拡散モデルは、次の形式で定義される潜在変数モデルです：

$$p_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_{1:T}$$

ここで、$x_1, \ldots, x_T$ はデータ x_0 と同次元の潜在変数であり、$x_0 \sim q(x_0)$ です。

逆過程（reverse process）と呼ばれる共分布 $p_\theta(x_{0:T})$ は、学習されたガウス遷移から成るマルコフ連鎖として定義されます。これは $p(x_T) = \mathcal{N}(x_T; 0, I)$ から始まります：

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \quad p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

以下のプロセスによる $q(x_{1:T}|x_0)$、すなわち順過程（拡散過程）は、一定の分散スケジュール $\beta_1, \ldots, \beta_T$ に従ってデータにガウスノイズを徐々に加えることで固定されます：

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

負の対数尤度に基づく通常の変分境界を最適化することで学習を行います：

$$\mathbb{E}[-\log p_\theta(x_0)] \leq \mathbb{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right]$$

モデルの詳細 (Model Details)

拡散モデルは、潜在変数モデルの一種であり、その目的は有限の時間内にデータと一致するサンプルを生成することです。以下に、モデルの構成要素を説明します。

順過程 (Forward Process)

順過程は、次の形式で固定されます：

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

ここで、β_t は分散スケジュールであり、データにノイズを徐々に加えるマルコフ連鎖を構成します。これにより、x_T は標準正規分布に近づきます。

逆過程 (Reverse Process)

逆過程は学習されたパラメータ θ を用いて定義されます：

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

ここで、μ_θ と Σ_θ はニューラルネットワークで表現される平均と分散です。

トレーニング目的 (Training Objective)

学習は次の変分境界を最小化することで行われます：

$$\mathbb{E}[-\log p_\theta(x_0)] \leq \mathbb{E}_q \left[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right].$$

この目的関数は KL ダイバージェンスとして書き換えられます：

$$L = D_{KL}(q(x_T|x_0)\|p(x_T)) + \sum_{t=2}^{T} D_{KL}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1).$$

パラメータ化 (Parameterization)

モデルの逆過程では、μ_θ を次の形式で定義します：

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0,$$

ここで、$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ です。この形式により、逆過程の精度が向上します。

簡略化された目的 (Simplified Objective)

トレーニングの効率化のため、以下の簡略化された目的が用いられます：

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t, x_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 \right],$$

ここで、ϵ は標準正規分布からサンプリングされるノイズです。

実験 (Experiments)

我々は、全ての実験において $T = 1000$ を設定し、ニューラルネットワーク評価の回数が以前の研究と一致するようにしました [**?**, **?**]。順過程の分散は、$\beta_1 = 10^{-4}$ から $\beta_T = 0.02$ まで線形に増加する定数として設定しました。この設定により、データが $[-1, 1]$ にスケールされた場合、逆過程と順過程がほぼ同じ形式を持ちながら、信号対雑音比が最小化されます（$L_T = D_{KL}(q(x_T|x_0)\|N(0, I)) \approx 10^{-5}$ ビット/次元）。

サンプル品質 (Sample Quality)

CIFAR-10 における Inception スコア、FID スコア、負の対数尤度（損失なし符号長）の結果を以下の表に示します。FID スコア 3.17 において、我々の非条件付きモデルは、文献中の多くのモデル（クラス条件付きモデルを含む）よりも優れたサンプル品質を達成しました。FID スコアは訓練データセットに基づいて計算されますが、テストセットに基づいて計算した場合でも 5.24 というスコアを達成し、文献中の多くの訓練セット FID スコアよりも優れています。

モデル	Inception スコア (IS)	FID スコア	負の対数尤度 (NLL)
我々のモデル (非条件付き)	9.46	3.17	-
SNGAN	8.22	21.7	-
StyleGAN2	9.74	3.26	-

Table 1: CIFAR-10 におけるサンプル品質比較

逆過程のパラメータ化と目的のアブレーション (Reverse Process Parameterization and Training Objective Ablation)

逆過程のパラメータ化において、μ_θ を ϵ の予測に再パラメータ化することで、ラングヴィンダイナミクスに類似した手法を利用することが可能になりました。この方法により、拡散モデルの変分境界が簡略化され、複数のノイズスケールにわたるスコアマッチングに類似した目的が得られます。

進行的符号化 (Progressive Coding)

進行的な非条件付き生成プロセスを実行し、ランダムビットからの進行的な復号化により画像を生成しました。これにより、逆過程中のサンプル品質を評価しました（図参照）。

[width=0.8]progressive$_s$*amples.png*

Figure 1: CIFAR-10 の進行的サンプリング品質

補間 (Interpolation)

CelebA-HQ 256×256 サンプルにおいて、異なる潜在表現間で補間を実行しました。これにより、モデルが学習した高次元表現の滑らかさを確認しました（図参照）。

[width=0.8]celeba$_i$*nterpolation.png*

Figure 2: CelebA-HQ サンプルの補間

結論 (Conclusion)

本研究では、拡散確率モデル（diffusion probabilistic models）を用いて高品質な画像生成を実現しました。また、拡散モデルと以下の手法との関連性を示しました：- マルコフ連鎖の変分推論を用いた訓練- ノイズ除去スコアマッチングおよびアニーリング・ランジュバン・ダイナミクス- エネルギーベースモデル（拡張による）- 自動回帰モデル- 進行的損失圧縮

拡散モデルは、特に画像データに対して優れた帰納的バイアスを持つことが示唆されており、他のデータモダリティや生成モデル、さらには機械学習システムの構成要素としての利用可能性が期待されます。