Application of a score-based generative modeling with stochastic differential equations to street figures pattern recognition

Emilien Jemelen†
ENSAE - IP Paris
emilien.jemelen@ensae.fr

Christophe Morau†
ENSAE - IP Paris
christophe.morau@ensae.fr

Abstract

Generative modeling has become increasingly important in various fields, such as computer vision, natural language processing, and drug discovery for instance. This paper presents an application of the work of (Song et al., 2021), a novel approach to generative modeling based on stochastic differential equations (SDEs) that transform a complex data distribution to a known prior distribution by slowly injecting noise, and a corresponding reverse-time SDE that transforms the prior distribution back into the data distribution by slowly removing the noise.

This approach has the advantage of using an explicit time-continuous mapping between the two distributions, making it more interpretable than Generative Adversarial Networks (GANs) and easy to sample from by using a general-purpose SDE solver.

We applied the approach of (Song et al., 2021) to generate street figures images using the Street View House Numbers (SVHN) dataset. Generating such images can be useful in Machine Learning fields which require massive data mining of numbers photos that are blurred or taken against daylight. Our results show that SDE-based generative modeling can produce decent street figures images with a wide variety of colors and contrasts and different levels of pattern clarity, and that further studies on the topic could bring promising results.

1 Introduction

Generative modeling is becoming increasingly important in many fields : for example, among other purposes, creating data from noise proves to be useful for the imputation of missing data, images synthesis in Computer Vision, texts generation in Natural Language Processing and molecule patterns generation for drugs discovery.

In this paper, we relied heavily on the work of (Song et al., 2021), who developed a novel and record-breaking approach to smoothly transform a complex data distribution to a known prior distribution and then reverse a Stochastic Differential Equation (SDE) to transform the prior distribution back into the data distribution.

By modeling the transition from the data to the noise with a SDE, (Song et al., 2021) propose an approach with an explicit time-continuous mapping between the two distributions. As a consequence, SDE-based models have the advantage of being more interpretable than Generative Adversarial Networks (GANs, another well-known class of generative modeling), which do not propose explicit mappings and can be more difficult to interpret. Additionally, sampling with a SDE-based model is not really challenging, for any general-purpose SDE solver can be used to integrate the reverse-time SDE.

In this paper, we developed a methodology to apply the approach of (Song et al., 2021) to street figures images generation using the Street View House Numbers (SVHN) dataset. These images are colorized with a wide variety of colours and contrasts, and with different levels of pattern clarity (some figures are blurred). Therefore, this application is more complex than the one based on the MNIST dataset proposed by *Song* in an illustrative tutorial of the paper: we aim at evaluating the robustness of such a simple implementation to images that are no longer consistant in scale and colours.

†With equal contribution

Fig. 1 : instances of SVHN street figures

2 Method

2.1 From score-based models to SDEs

Score-based models are a special instance of generative models, among implicit generative models (GANs...) and likelihood-based models. Like the latter, score-based model are interested in learning the distribution of the original image in order to generate new samples based on this law. A key difference however is that, while likelihood-based models use the maximum likelihood to learn the distribution from the samples, score-based models are looking for its derivative: the score. Unlike likelihood-based models, score-based models do not need a normalization constant to be tractable, which makes them computationally more efficient. Score-matching techniques ((Hyvarinen, 2005), (Vincent, 2011)) can be used to retrieve the score from the samples and Langevin dynamics ((Parisi, 1981), (U. Grenander, 1994)) are then used to draw samples from it.

However, score-based models perform badly on regions of the distribution where few samples are available, as the estimation of the score is less reliable. This is an important issue since, in high-dimension, the starting point of our Langevin dynamics is likely in a low density region and the procedure will then derail from its start. To solve this issue, (Y. Song, 2019) and (Y. Song, 2020) proposed to add noise at different scales to the data: larger noise allows to better cover low-density region while lower noise causes less corruption. The score of each noise-perturbed distribution is estimated and annealed Langevin dynamics ((Y. Song, 2019), (Y. Song, 2020), (A. Jolicoeur-Martineau, 2021)), a sequential version of Langevin dynamics that progressively decreases noise is used to generate samples.

The idea of (Song et al., 2021) was to perturb the data continuously instead of using several discrete scales. The noise scale would continuously increase over $[0, T]$, which could be seen as an interval of time, and the random variable $x(t)$, which follows the random perturbation at time t, follows a stochastic process.

2.2 Perturbing data with SDE

To add noise to the original data, (Song et al., 2021) indeed construct a stochastic diffusion process $(X_t)_{t\in[0,T]}$ indexed by a continuous time variable $t \in [0, T]$, such that $X_0 \sim p_0$, the distribution of the data without noise, and $X_T \sim p_T$, the prior distribution, which corresponds to the distribution of the data with maximum noise. From time 0 to time T, the noise intensity is continuously increased by a Brownian motion of variance equal to t at time t.

$(X_t)_{t\in[0,T]}$ can then be seen as a diffusion process modeled as the solution of an Itô SDE :

$$\boxed{dX_t = f(X_t, t).dt + g(X_t, t).dW_t \quad \textbf{(A)}}$$

Remarks on (A) :

For all $t \in [0, T]$, $f(., t) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a vector-valued function called the *drift* coefficient of X_t, since it represents the time-deterministic evolution of process X.

For all $x \in \mathbb{R}^d$, $g(x, .)$ is a real-valued function called the *diffusion* coefficient : it represents the stochastic evolution of process X. (Song et al., 2021) demonstrate that the real-valued case can be extended to the $\mathbb{R}^d$ scenario with extra lemmas.

For ease of representation in the modeling of the problem, the authors of the paper make the assumption that the *diffusion* coefficient $g(x, .)$ is independent of the value of $x, \forall x \in \mathbb{R}^d$. As a consequence, (A) can be alternatively expressed as : $dX_t = f(X_t, t).dt + g(t).dW_t$

W is Levy process (independent and stationary increments, right continuous with left limit sample paths, stochastic continuity property) without any jump and a continuous part characterized by the distribution $W_t \sim \mathcal{N}(0, t), \forall t \in [0, T]$.

(Øksendal, 2003) showed that (A) has a unique solution X as long as the coefficient functions are Lipschitz in time for $g(.)$ and both in state and in time for $f(.,.)$.

Hereafter, $p_t(X_t)$ is defined as the probability density of X_t, and $p_{s,t}(X_t|X_s)$ the transition kernel from X_s to X_t, where $0 \leq s < t \leq T$.

*

* *

This forward SDE introduces stochastic modifications to the original distribution of the data p_0 through a continuous timeline modeling until reaching a multivariate Gaussian distribution p_T. The generative part of the model comes from the ability to go backwards and produce accurately samples from distribution p_0. The ability to reverse the time-continuous SDE is the key to generative modeling with SDE.

2.3 Denoising perturbed data with reverse-SDE

By perturbing the original data X_0 with a SDE of the form $dX_t = f(X_t, t).dt + g(t).dW_t$ (with some additional assumptions regarding f and g), it is possible to obtain samples of $X_T \sim p_T$.

The main result brought by (Anderson, 1982) is that by starting from samples of $X_T \sim p_T$, it is possible to *reverse* the stochastic process X and obtain samples $X_0 \sim p_0$. Formally, (Anderson, 1982) states that - by using a backward Itô integration method - the reverse of a diffusion process is also a diffusion process $\bar{X}$ given by the *reverse-time* SDE :

$$\boxed{\begin{array}{l} d\bar{X}_t = \\ {[f(\bar{X}_t, t) - g(t)^2.\nabla_{\bar{X}_t} log(p_t(\bar{X}_t))].dt} \\ +g(t).d\bar{W}_t \quad \textbf{(B)} \end{array}}$$

Where $\bar{W}$ is a standard Wiener process when time flows backwards from time T to time 0, and dt is an infinitesimal negative timestep.

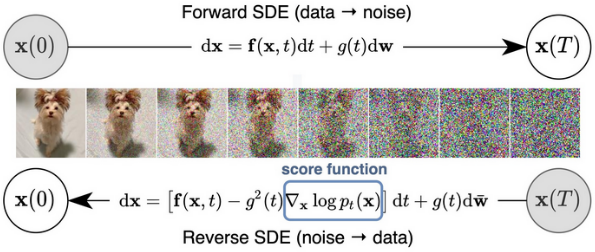
Forward SDE (data → noise)
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
$\mathbf{x}(T)$
score function
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
$\mathbf{x}(T)$
Reverse SDE (noise → data)

Fig. 2 (taken from (Song et al., 2021)) : Overview of score-based generative modeling through SDEs

To sample from p_0 by simulating from p_T (Gaussian) and deriving the reverse diffusion process from (B), it is necessary to compute - at each time $t \in [0, T]$ - the score of each marginal distribution :

$$\nabla_{\bar{X}_t} log(p_t(\bar{X}_t))$$

2.4 Estimating the score of the marginal distribution

Relying on the works of (Hyvarinen, 2005) and (Song et al., 2019), (Song et al., 2021) propose to estimate the score of each marginal distribution $\nabla_{\bar{X}_t} log(p_t(\bar{X}_t))$ by training a score-based model on samples with *score matching*.

Concretely, a score-based model $s_\theta(X_t, t)$ function of a parameter θ is defined, and the value of parameter θ is a solution of the following convex problem (if the map $\theta \mapsto s_\theta(X_t, t)$ is strictly convex, the solution is unique) :

$$\boxed{\begin{array}{l} \theta_t^* \in \underset{\theta \in \Theta \text{ convex set}}{argmin} \; (\mathbb{E}_t[\lambda(t).\mathbb{E}[X_0].\mathbb{E}[X_t|X_0]. \\ ||s_\theta(X_t, t) - \nabla_{X_t} log(p_{0,t}(X_t|X_0))||^2]) \\ \qquad\qquad\qquad \textbf{(C)} \end{array}}$$

Remarks on (C) :

$\lambda : [0, T] \rightarrow \mathbb{R}_+^*$ is a positive weighting function. In practice, (Song et al., 2021) recommend choosing λ such that : $\forall t \in [0, T], \lambda(t) \propto \frac{1}{\mathbb{E}[||\nabla_{X_t} log(p_{0,t}(X_t|X_0))||_2^2]}$

At each time $t \in [0, T]$, the score of the marginal distribution $\nabla_{\bar{X}_t} log(p_t(\bar{X}_t))$ is estimated by $s_{\theta_t^*}(X_t, t)$.

(Song et al., 2021) affirm that with sufficient data and model capacity, score matching ensures that $s_{\theta_t^*}(X_t, t)$ equals $\nabla_{\bar{X}_t} log(p_t(\bar{X}_t))$ for almost all $t \in [0, T]$.

*

* *

Solving (C) requires knowing the transition kernel $p_{0,t}(X_t|X_0)$: (Sarkk and Solin, 2019) showed that when $f(., t)$ is affine, the transition kernel is a Gaussian distribution with tractable mean and variance in most closed-forms cases. In the case of general SDEs, (Øksendal, 2003) showed that the transition kernel can be obtained by solving Kolmogorov's forward equation.

Alternatively, (Song et al., 2021) propose a *sliced score matching* method for model training which bypasses the computation of the transition kernel.

3 Experiments Protocol

All the code implementation is available at our Google Colab workspace[1].

3.1 Interest of the SVHN dataset

In the implementation of their findings[2], (Song et al., 2021) generate samples with the MNIST dataset, a dataset of black and white figures images.

To extend this implementation, we chose to apply the score-based generative modeling with SDE to the Street View House Numbers (SVHN) data which contain colorized street numbers with a variety of contrasts and definitions.

The images in the SVHN dataset are under a 32x32 pixels format, with RGB colours (x3). Compared to the MNIST data, generating SVHN-like data seems more challenging due to the higher dimensionality of the data (colorized) and the fact that the number signal is often weaker in the SVHN data (some numbers are blurred, see *Fig. 1*). As a consequence, it seems interesting to test whether the score-based generative modeling with SDE is able to cope with these two additional characteristics.

3.2 Estimation of the score

In practice, (Song et al., 2021) estimate $t \in [0, T] \mapsto s_{\theta_t^*}(X_t, t)$ with a time-dependent score-based model built upon a U-net architecture.

To train the model and find weights such that it be a good approximation of $t \in [0, T] \mapsto s_{\theta_t^*}(X_t, t)$, we use Pytorch Adam in-built optimizer. The loss function used is the objective function of (C), which is tractable according to (Sarkk and Solin, 2019) since - as in (Song et al., 2021) - we use an affine function $f(., t)$.

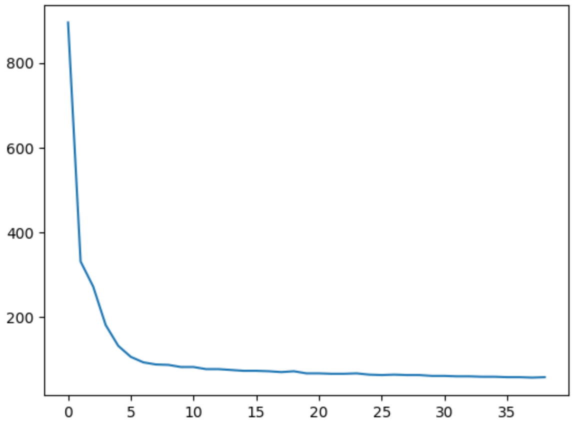

Fig. 3 : evolution of the per-epoch training loss (40 epochs)

After 160 epochs of training, the loss seemed to be stuck around a value of 33. The problem being convex (at least theoretically), being stuck at a local mode means that the global minimum should not be far.

3.3 Intents to improve the training

Assuming that the training loss being stuck at a value of 33 means that the optimum of the problem is close, we implemented the Newton algorithm, which is a type of gradient descent under some strong assumptions (two times differentiable objective, initialization close enough to the optimum, strong assumptions regarding the behaviour of the Hessian function) and with step $i \in \mathbb{N}^*$:

$$x_{i+1} = x_i - \nabla^2 f(x_i)^{-1}.\nabla f(x_i)$$

The advantage of Newton algorithm is that the two-times differentiability allows for greater smoothness when approaching the optimum, avoiding the potentially infinite oscillations around the optimum due to too big step size.

[1] https://colab.research.google.com/drive/1esuK-P3sOANjoaOlYPTagpg9Gtl80KjJ?usp=sharing

[2] https://yang-song.net/blog/2021/score/

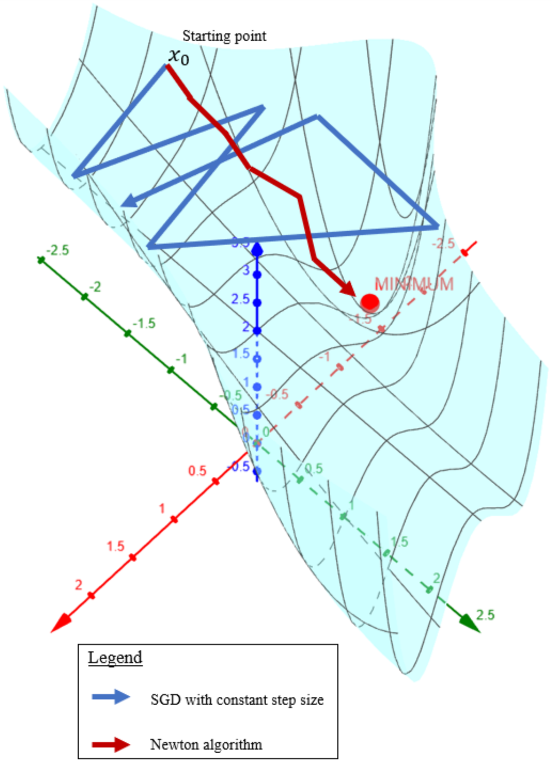

Fig. 4 : illustration of the interest of Newton algorithm in the case of narrow slopes region (Geogebra)

Yet, running steps of the Newton algorithm did not improve our training results any further. A possible interpretation is that we were already very close to the global minimum, at least in terms of objective function.

Another idea to improve the training was to split the SVHN dataset in clusters of colours. Indeed, the number signal is seemingly not linked to the colours of the image. As such, the colours in the SVHN data can be seen as noise for the purpose of generating numbers signals, which is our goal. Therefore, doing the training of the model separately on images with homogeneous colours structure could make it easier for the model to capture during training and then generate the numbers signals.

We launched Kmeans centroid-based clustering algorithm on the data transformed to normed histograms of colours intensities. We tested the algorithm with K = 2, 3 and 4. After an empirical study of the results, it seemed that 3 clusters allowed Kmeans to capture a significant segmentation of the data, whereas 2 clusters was a too simplistic separation and 4 clusters a too unclear one.

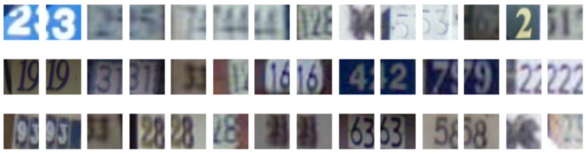

Fig. 5 : random images of each cluster (row n°k = cluster k)

Comment on the colours segmentation :

Cluster 1 seems to contain light colours and white, whereas cluster 2 contains darker and bluer images, cluster 3 being an intermediate cluster with not too bright neither too dark/blue images

*

*   *

Despite the fact that the model seems to have learnt the colours code of each cluster (see *Fig. 6* below : images generated from cluster 1 are brighter, *etc*...), the result of separate trainings performed one the 3 clusters of images is still very similar to the result of the training performed on the whole SVHN dataset.

Indeed, the best value of 33 for the training loss on the whole dataset was neither beaten by the Newton algorithm nor by the per-cluster training.

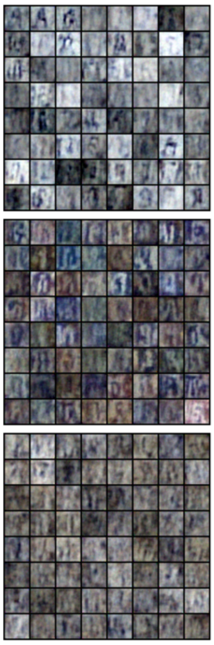

Fig. 6 : images generated from models trained separately on each cluster (grid n°k = cluster k)

3.4 Using the score to generate samples

After considering some options for alternative training, our best result in terms of loss was reached by the training on the whole dataset.

With the model trained on the whole training dataset, we used an Euler Maruyama sampler with Langevin predictor-corrector (PC) - which is a general purpose SDE solver method detailed by (Song et al., 2021) - to reverse the SDE and thus generate images from our prior Gaussian distribution.

The Euler-Maruyama method is a numerical solver which approximates trajectories from SDEs with a fixed time step. In addition, following the advice of (Song et al., 2021), we applied a MCMC method called Langevin corrector to find better time steps when reversing the SDE.

4 Results

By reversing the SDE from our prior Gaussian distribution, the score-based generative modeling accomplished the task of generating SVHN-like data at a pace of around 10 images per second.

It appears that the generated data contain a variety of colours which respects the original colour code of the SVHN data (dominance of blue, white and brownish colours) and the original definitions distribution (some numbers are blurred, and some can be read easily).

However, the generative modeling with SDE seems to struggle to generate some high definition street numbers.

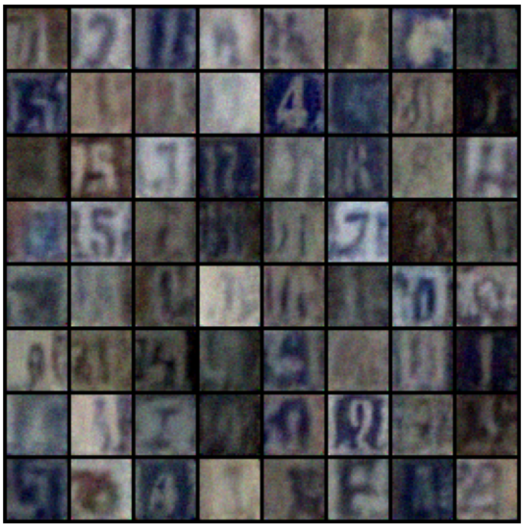

Fig. 7 : images generated from the model trained on the whole dataset

5 Conclusion/discussion

This paper presented a new application of generative modeling based on a stochastic differential equation (SDE). Compared to other approaches such as Generative Adversarial Networks (GANs), the SDE-based model has the advantage of having an explicit mapping and being more interpretable. Moreover, sampling with a SDE-based model is not very challenging in terms of computation cost.

This paper applied this approach to generate street figures images using the Street View House Numbers (SVHN) dataset. By introducing stochastic modifications to the original data distribution through a continuous timeline modeling, the SDE-based model was able to generate samples from a multivariate Gaussian distribution p_T. The ability to reverse the time-continuous SDE allowed for accurate generation of figures from noised data, which could be useful in data mining from blurred or low-quality photos. Overall, this approach provides an alternative for generative modeling with potentially broad applications in various fields.

To improve the training of the U-net and go further in the application of (Song et al., 2021) to SVHN data, performing a per cluster training with clusters of street numbers with homogeneous definition could be interesting, notably to generate more accurately street numbers of high definition and thus generate a wider range of street numbers.

References

G. Parisi. 1981. Correlation functions and computer simulations. *Nuclear Physics B, Vol 180(3), pp. 378–384. Elsevier.*

Brian D.O. Anderson. 1982. Reverse-time diffusion equation models. *Stochastic Processes and their Applications Volume 12, Issue 3, May 1982, Pages 313-326.*

M.I. Miller U. Grenander. 1994. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological), Vol 56(4), pp. 549–581. Wiley Online Library.*

Bernt Øksendal. 2003. Stochastic differential equations. *Stochastic differential equations, pp. 65–84. Springer.*

Aapo Hyvarinen. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research, April 2005, Pages 695–709.*

P. Vincent. 2011. A connection between score matching and denoising autoencoders. *Neural computation, Vol 23(7), pp. 1661–1674. MIT Press.*

Simo Sarkk and Arno Solin. 2019. Applied stochastic differential equations. *volume 10. Cambridge University Press.*

S. Ermon. Y. Song. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems, pp. 11895–11907.*

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. 2019. Sliced score matching: A scalable approach to density and score estimation. *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019, pp. 204, 2019a.*

S. Ermon. Y. Song. 2020. Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

I. Mitliagkas R.T.d. Combes. A. Jolicoeur-Martineau, R. Piche-Taillefer. 2021. Adversarial score matching and improved sampling for image generation. *International Conference on Learning Representations.*

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456.*