# Импульсная диффузионная модель генерации речевых команд.

М.А. Прибыль, Н.В. Шмырёв

Диффузионные вероятностные модели шумоподавления или просто диффузионные вероятностные модели – это класс генеративных моделей, которые на каждом временном шаге добавляют гауссовый шум к данным (процесс прямой диффузии), а затем обучаются обращать процесс диффузии, чтобы получить исходные данные (процесс обратной диффузии).

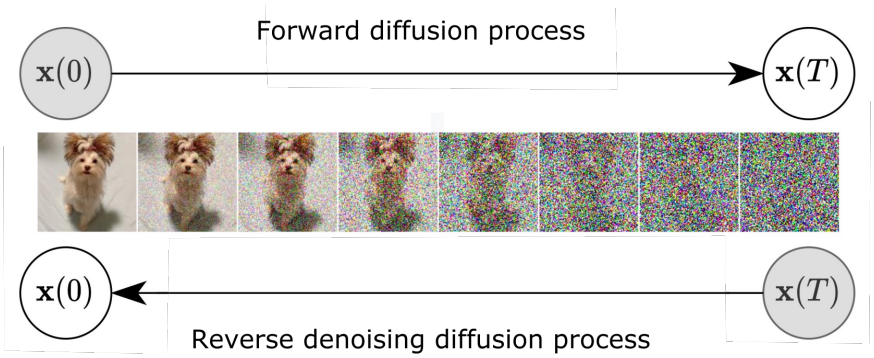



Рис. 1: Возмущение картинки гауссовым шумом.

Диффузионные вероятностные модели принципиально отличаются от всех предыдущих генеративных моделей. Интуитивно они направлены на разложение процесса генерации сэмпла на множество небольших шумоподавляющих шагов.

Впервые диффузионные модели были предложены в [11]. В [5] авторы предложили определенную параметризацию модели, которая не только упрощает обучение, но и способствует генерации высококачественных сэмплов, которые иногда даже превосходят сэмплы, сгенерированные другими видами генеративных моделей. После этой работы интерес к диффузионным вероятностным моделям вырос. Появились работы в таких областях, как аудиомоделирование [9], [2], видеомоделирование [21], прогнозирование временных рядов [18], а также преобразование текста в речь [16].

В работе [12] вместо возмущения данных конечным числом шумовых распределений авторы предложили использовать континуум распределений. Процесс преобразование данных в шум задается с помощью стохастического дифференциального уравнения, которое не зависит от данных и не имеет обучаемых параметров. Обратный процесс диффузии также описывается стохастическим дифференциальным уравнением, которое может быть получено из уравнения прямой диффузии, если известна оценка плотности маргинального распределения $\nabla_x \log p_t(x)$, как функции времени.

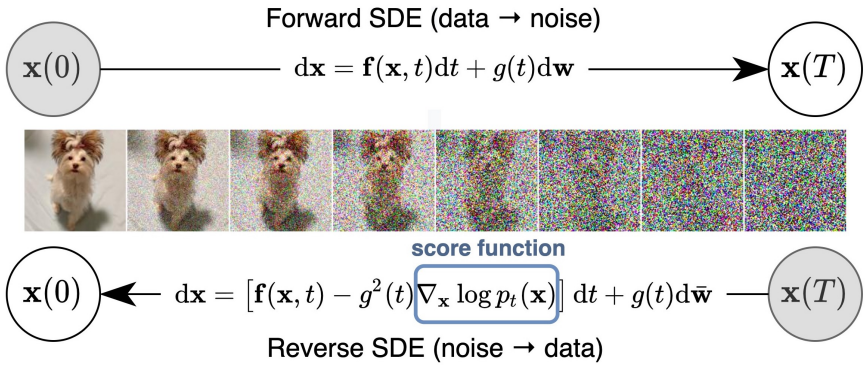



Рис. 2: Преобразование данных в простое распределение шума выполняется с помощью стохастического дифференциального уравнения.

Предлагаемый подход улучшает генеративные результаты и обеспечивает более эффективное сэмплирование по сравнению с диффузионными моделями с конечным числом шумовых распределений.

Основной недостаток диффузионных вероятностных моделей состоит в том, что они требуют много итераций для получения высококачественного сэмпла. Поэтому сэмплирование происходит на 2-3 порядка медленнее, чем у генеративно-состязательных нейронных сетей и вариационных автоэнкодеров. Для решения этой проблемы в [13] предложена неявная диффузионная вероятностная модель, в которой используется немарковский процесс прямой диффузии. В [10] предложен метод для преобразования многошагового процесса шумоподавления в одношаговый. Однако упомянутые методы уступают в качестве сгенерированных сэмплов.

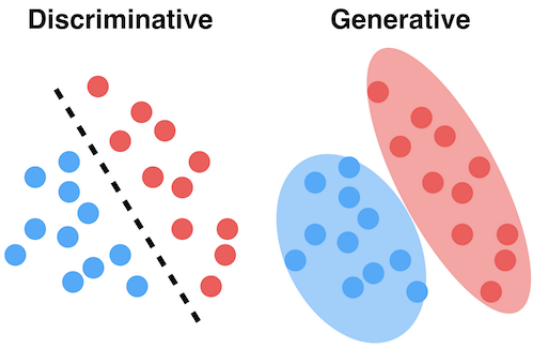
В [1] предложена импульсная диффузионная вероятностная модель для генерации изображений. Авторы предложили архитектуру, которая достигает генеративной производительности, сравнимой с производительностью не импульсной диффузионной вероятностной модели всего за 4 импульсных временных шага. Более того, предложенная модель превосходит другие генеративные модели на основе импульсных нейронных сетей в производительности, достигая 12-кратного и 6-кратного улучшения на базах CIFAR-10 и CelebA соответственно.

Успех генеративных моделей и, в частности, диффузионных моделей, объясняется тем, что они могут генерировать новые экземпляры данных, в отличии от дискриминативных моделей, который только различают виды данных.

С помощью диффузионных моделей можно:

1. Научиться работать с зашумлёнными и нечёткими данными (часто данные для тренировки неточно размечены) [3];

2. Научиться описывать сложные распределения для создании синтетических данных [4].




В качестве примеров диффузионных моделей отметим Stable Diffusion [15] для создания изображений, которая является наиболее популярной, а также Grad-TTS [16] для синтеза речи. Для моделирования текста диффузия пока не совсем применима, так как слишком сложные распределения приходится моделировать. Поэтому диффузионные модели сочетают с трансформерами [20].

Можно также провести параллели между генеративным моделированием, в частности, диффузионным, и работой мозга. Эксперименты показали, что человеческое поведение в высокой степени согласуется с вероятностным мышлением в сенсорной, моторной и когнитивной областях [17]. Существует несколько теорий, как мозг производит вероятностный вывод. Эту случайность в процессах мышления можно сравнить с вероятностной природой генеративных моделей. Поэтому нам было интересно попробовать применить идеи генеративного моделирования в импульсных нейронных сетях.

Для генерации звуковых команд мы используем импульсную диффузионную модель из работы [1]. В архитектуры данной модели используется дискретная версия IF модели импульсного нейрона, которая описывается следующими уравнениями:

$$U(n) = V(n-1) + I(n),$$

$$S(n) = \Theta(U(n) - \vartheta_{th}),$$

$$V(n) = U(n)(1 - S(n)) + V_{reset}S(n),$$

где $n$– временной шаг, U(n) – мембранный потенциал до "сброса" . Когда мембранный потенциал $U(n)$ достигает порога $\vartheta_{th}$, нейрон генерирует импульс $S(n)$, $\Theta$ – функция Хевисайда. $V(n)$ представляет мембранный потенциал после генерации импульса, который "сбрасывается" до $V_{reset}$.

Гауссовый шум преобразуется энкодером в импульсную последовательность. Энкодер состоит из двух свёрточных слоев с пакетной нормализацией и одного слоя импульсных нейронов. Time embedding – это векторы, которые отмечают каждую 32 миллисекунду или каждый отсчет, чтобы нейросеть могла их отличать. Далее импульсная последовательность вместе с "time embedding" подаются на основную компоненту архитектуры –

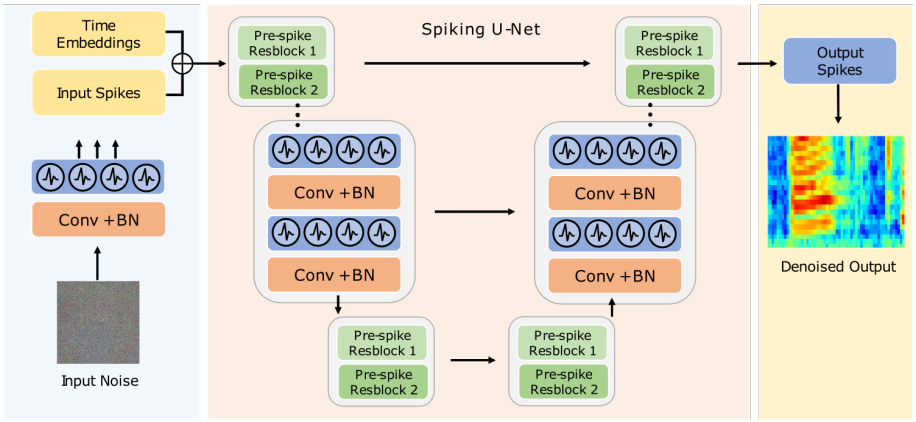



Рис. 3: Архитектура импульсной диффузионной вероятностной модели.

импульсную U-net. Она состоит из сужающегося пути (слева) и расширяющегося пути (справа). На каждом этапе сужающегося пути разрешение данных уменьшается, а количество каналов свойств увеличивается. Каждый шаг расширяющегося пути восстанавливает разрешение и уменьшает количество каналов. Эти два пути связаны пропускными соединениями (Skip Connection), которые необходимы для восстановления пространственной информации, потерянной при уменьшении разрешения. Сужающийся и расширяющийся пути состоят из нескольких остаточных блоков (Pre-Spike Residual Block), каждый из которых содержит импульсные слои, свёрточные слои и пакетную нормализацию. Пропускные соединения в остаточном блоке стабилизируют обновления градиента при обучении.

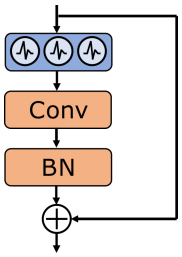



Рис. 4: Pre-Spike Residual block.

Импульсная U-net передает только импульсы, представленные векторами из нулей и единиц. Далее, импульсы из U-net передаются в декодирующий слой, который состоит из двух свёрточных слоев и специального слоя мембранного потенциала [6], где воспроизводится звук. Процесс сэмплирования состоит из множества шагов шумоподавления, где каждый шаг проходит через импульсную диффузионную вероятностную модель.

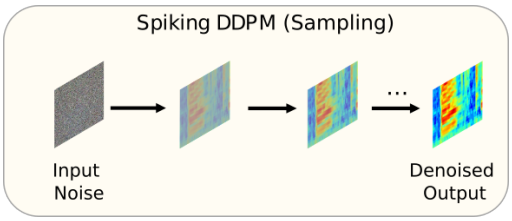



Рис. 5: Сэмплирование.

**Данные**

Для тренировки используется записи английских команд Google Speech Commands V2 [19]. Датасет состоит из коротких записей команд, всего 105000 записей, 35 различных слов, 2618 голосов. Перед тренировкой речевой сигнал преобразуется в мел-спектральные коэффициенты размерностью 32 и нарезается на несколько сегментов $32 \times 32$. То есть вычисляем энергию сигнала для каждого окна в мел-шкале и нарезаем на фрагменты длительностью 0.32 секунды.

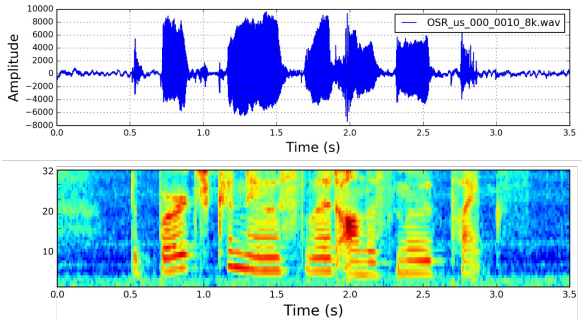



Рис. 6: Исходный сигнал и мел-спектральные коэффициенты.

**Вычисления**

Для тренировки используется SpikingJelly – пакет Python для обучения импульсных нейронных сетей, в котором алгоритм обучения основан на методе суррогатного градиента [14].

Тренировка производится на компьютере с видеокартой NVIDIA GTX 1080 и занимает 3 дня.

**Метрика**

Для оценки модели используется метрика Frechet Audio Distance (FAD) [7], которая является аналогом метрики FID для изображений и сравнивает наборы аудиофайлов по статистическим параметрам.

**Результаты**

В качестве сравнения используется: **VITS**(GAN) [8], **Grad-TTS**(диффузия) [16], натренированные на данных LibriTTS.

| Model | FAD |
|---|---|
| VITS | 10.5 |
| GRAD-TTS | 8.5 |
| **SDDPM** | 11.5 |

Таким образом, предложенный метод даёт сравнимое качество генерации при преимуществах импульсных нейронных сетей. В дальнейшем мы планируем увеличить скорость сэмплирования, создать большую модель на полноценных записях для синтеза произвольной речи и использовать полученную модель для практических задач.

Данная работа представлена на конференции `https://neuro.kaspersky.ru/conference/`.

## Литература

[1] J. Cao, Z. Wang, H. Guo, H. Cheng, Q. Zhang, R. Xu, Spiking Denoising Diffusion Probabilistic Models, *arXiv preprint arXiv: 2306.17046*, 2023.

[2] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi and W. Chan, Wavegrad: Estimating gradients for waveform generation, International Conference on Learning Representations, 2021.

[3] B. Frénay, A. Kabán, A comprehensive introduction to label noise, The European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Belgium 2014.

[4] Xu Guo, Yiqiang Chen, Generative AI for Synthetic Data Generation: Methods, Challenges and the Future, *arXiv preprint arXiv:2403.04190*, 2024.

[5] J. Ho, A. Jain and P. Abbeel, Denoising diffusion probabilistic models, 34th Conference on Neural Information Processing Systems, Vancouver, Canada, 2020.

[6] H. Kamata, Y. Mukuta, and T. Harada, Fully spiking variational autoencoder. In AAAI, volume 36, pages 7059–7067, 2022.

[7] K. Kilgour, M. Zuluaga, D. Roblek, M. Sharif, Frechet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms, *arXiv preprint arXiv:1812.08466v4*, 2019.

[8] J. Kim, J. Kong, J. Son, Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, *arXiv preprint arXiv:2106.06103v1*, 2021.

[9] Zh. Kong, W. Ping, J. Huang, K. Zhao, B. Catanzaro, DiffWave: A Versatile Diffusion Model for Audio Synthesis, International Conference on Learning Representations, 2021.

[10] E. Luhman, T.Luhman, Knowledge Distillation in Iterative Generative Models for Improved Sampling Speed, *arXiv preprint arXiv:2101.02388v1*, 2021.

[11] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep Unsupervised Learning using Nonequilibrium Thermodynamics, Proceedings of the 32nd International Conference on Machine Learning, V. 37, France, 2015.

[12] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A.k Kumar, S. Ermon, B. Poole, Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021.

[13] J. Song, Ch. Meng, S. Ermon, Denoising Diffusion Implicit Models, *arXiv preprint arXiv:2010.02502v4*, 2022.

[14] SpikingJelly, `https://spikingjelly.readthedocs.io/zh-cn/latest/index.html#index-en` .

[15] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Muller, J. Penna, R. Rombach, SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, *arXiv preprint arXiv:2307.01952* , 2023.

[16] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech, *arXiv preprint arXiv:2105.06337v2*, 2021.

[17] A. Pouget, J.M. Beck, W.J. Ma, P.E. Latham, Probabilistic brains: knowns and unknowns, Nature Neuroscience, 16, 2013.

[18] K. Rasul, C. Seward, I. Schuster and R. Vollgraf, Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting, International Conference on Machine Learning, 2021.

[19] P. Warden, Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,*arXiv preprint arXiv:1804.03209v1*, 2018.

[20] T. Wu, Zh. Fan, X. Liu, Y. Gong, Y. Shen, J. Jiao, Hai-Tao Zheng, J. Li, Zh. Wei, J. Guo, N. Duan, W. Chen, AR-Diffusion: Auto-Regressive Diffusion Model for Text Generation, *arXiv preprint arXiv:2305.09515* , 2023.

[21] R. Yang, P. Srivastava, S. Mandt, Diffusion Probabilistic Modeling for Video Generation, *arXiv preprint arXiv:2203.09481v5*, 2022.