Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning [1]

Maxime BONNIN, Paul CHAUVIN, Ewen MICHEL

May 24, 2023

MVA: Generative modeling

Introduction

- State-of-the-art generative models for discrete data have limitations in terms of computation and slow generation.
- Diffusion models offer a solution to these limitations by enabling modeling of higher-dimensional data with parallel sampling steps.
- However, current diffusion models struggle to generate discrete/categorical data effectively.
- The proposed approach introduces analog bits and additional techniques to enable continuous-state diffusion models to generate high-quality discrete data.

Method - context

Introduction to Diffusion Models:

- Diffusion models learn state transitions from noise to data distribution.
- Forward transition equation: $x_t = \sqrt{\gamma(t)} \cdot x_0 + \sqrt{1 - \gamma(t)} \cdot \epsilon$.
- Diffusion models cannot directly handle discrete/categorical data.

Learning Process of Diffusion Models:

- Instead of modeling x_t to $x_{t-\Delta}$, learn $f(x_t, t)$ to predict x_0.
- Estimate $x_{t-\Delta}$ using x_t and estimated $\tilde{x}_0$.
- Training based on denoising with regression loss.

Sample Generation in Diffusion Models:

- Perform reverse state transitions from x_T to x_0.
- Apply denoising function f iteratively to estimate x_0 at each state x_t.
- Transition rules specified in DDPM [2] or DDIM [3].

Method - Analog Bits

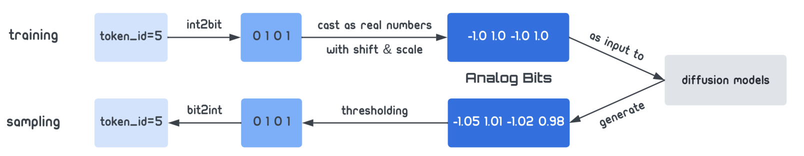


Figure 1: Bit Diffusion

- Novel approach: Representing discrete data variables using a continuous representation.
- Analog bits: Introducing real numbers trained to capture bimodal characteristics.
- Bridging the gap between continuous and discrete data representation.

- Real number values are mapped to their digit-wise binary representations in $[0,1]^d$.
- The approach of the authors : example in $\mathbb{R}^3$ of analog bits representation :

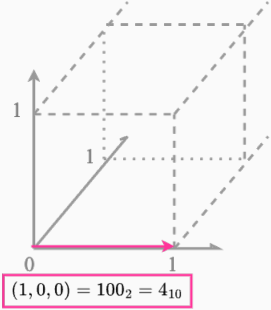


Figure 2: Representation of 4 in analog bits, values of $[0,1]^d$

Method - Analog bits

- Real number values are mapped to their digit-wise binary representations in $[0, 1]^d$.
- The approach of the authors : example in $\mathbb{R}^3$ of analog bits representation :

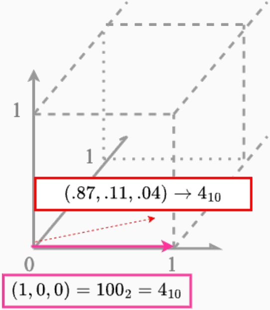


Figure 3: Representation of 4 in analog bits, values of $[0, 1]^d$

Quantization and continuous models

- Goal is for model to learn to output bimodal representation.

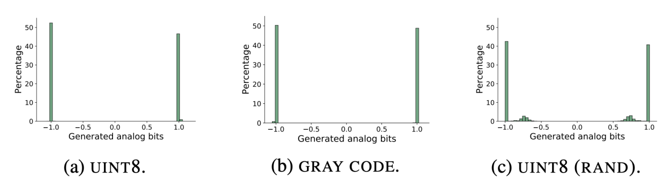


Figure 4: Histogram of the model's output before quantization

- Their continuous model behaves in an almost discrete manner.
- Otherwise, their continuous model would hardly learn due to stochastic loss feedback.

Quantization and continuous models

- Otherwise, their continuous model would hardly learn due to stochastic loss feedback.

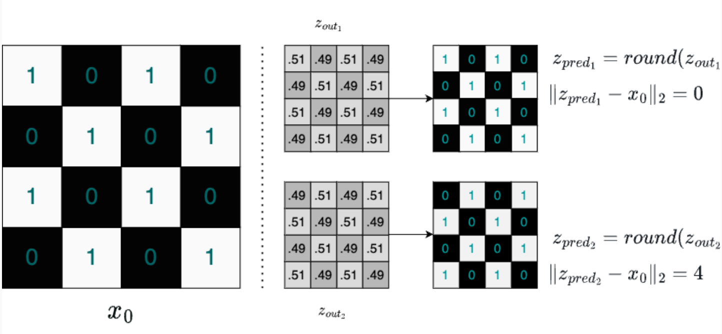


Figure 5: Opposite L_2 loss for near-indistinguishable outputs

Quantization and continuous models, other possible limits

- Loss of the semantic property of pixel space :
  - close pixels have close appearances
  - images are continuous by nature

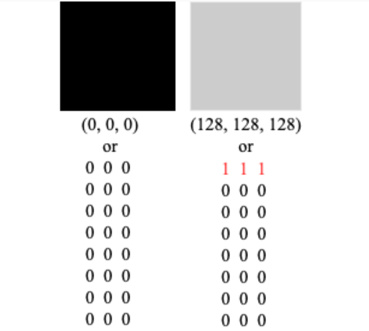


Figure 6: Example of two pixels distant in RGB space, close in analog bits space

Quantization and continuous models, other possible limits

- Loss of the semantic property of pixel space :
  - close pixels have close appearances
  - images are continuous by nature
- All dimensions of analog bits space are not equivalent in importance :

$$00000001_2 - 00000000_2 << 10000000_2 - 00000000_2$$

$$1 << 128$$

Quantization and continuous models, other possible limits

- Adressed in the paper: study of Random assignation of analog bits to the powers of 2 of the number.

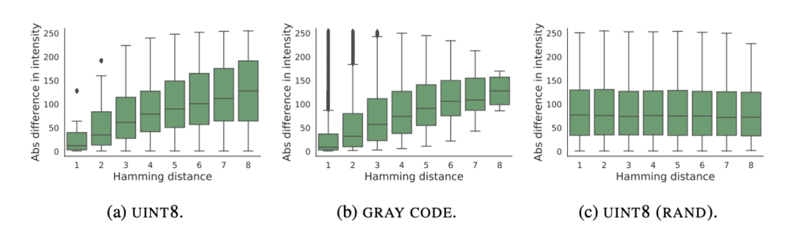
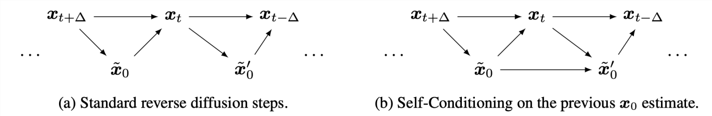


Figure 7: Study of correlation of bit values to pixel intensity

Method - Self-conditioning



- Improves diffusion models by directly conditioning the model on previously generated samples during the iterative sampling process, leading to enhanced sample quality.
- In the typical diffusion sampling process, the previous estimate of x_0 is discarded when estimating x_0 for a new time step. The implementation involves concatenating x_t with the previously estimated $\tilde{x}_0$. This incurs a negligible additional cost during sampling.
- Training the denoising function $f(x_t, \tilde{x}_0, t)$ involves making some changes to the training process. $\tilde{x}_0$ is set to 0 with a certain probability, while at other times, it is estimated as $f(x_t, 0, t)$ and used for Self-Conditioning. The estimated $\tilde{x}_0$ is not backpropagated, resulting in a minimal increase in training time (less than 25%).

Asymmetric Time Intervals

Impact of Time Step Parameter:

- Traditional symmetric time intervals used in Bit Diffusion models.
- Authors propose asymmetric time intervals to enhance sampling quality.

Asymmetric Time Intervals for Improved Sampling:

- Sampling process with $f(x_t, t_0)$, where $t_0 = t + \xi$ and ξ is a small non-negative time difference parameter.
- Demonstrated improvement in sampling quality without changing the training process.

Experiments conducted in the paper

- Datasets: CIFAR-10 [4] and IMAGENET 64x64 [5]
- Evaluation metric: Fréchet Inception Distance (FID) [**?**]
- Discrete image generation: we consider three discrete encoding for sub-pixels: UINT8, GRAY CODE, and UINT8 (RAND).
  Architecture: U-Net [6]
  - CIFAR-10 [4]: Single channel dimension of 256, 3 stages (with 3 residual blocks). 51M parameters. Dropout of 0.3
  - IMAGENET [5]: base channel dimension of 192, multiplied by 1,2,3,4 in 4 stages and 3 residual blocks per stage. 240M parameters
- Image Captioning: Tokenisation using a vocabulary of size 32K. And then encode each token using 15 analog bits. Architecture:
  - Pre-trained image encoder using the object detection task.
  - Randomly initialized 6-layer Transformer decoder with 512 dimension per layer.
- Adam Optimizer [7]

Fréchet Inception Distance

- **Fréchet Distance** measure the **discrepancy** or **dissimilarity**
- Use of **InceptionV3** hidden features as **content** and **style** extractor
- **FID**: Fréchet Distance with the use of hidden features extracted
- Lower FID values indicate better image quality and similarity

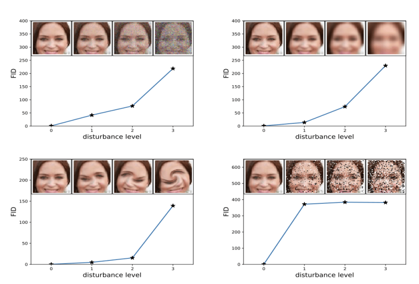


Figure 8: Example of How Increased Distortion of an Image Correlates with High FID Score.

Related work and importance of the paper

- **Autoregressive Models** work well on discrete image generation for small image resolution but very challenging to scale these approaches to data with large dimensions.
- Advantages of Continuous **Diffusion Models** (for discrete data) over 'standard' diffusion models using discrete state space: they are more flexible. There is also the benefit of Binary Encoding with Analog Bits (decoding easier and more robust).
- Limitations of **Normalizing Flows** [8] for Categorical Data: strict invertible restrictions on network architecture, thus limiting their capacity.
- Performance Comparison with **VAE** [9] and **GAN** [10]: not explored yet in the paper
- self-conditioning shares similarities with Self-Modulation in **GANs** [10] and **SUNDAE** [11] Techniques
- **Basically, unique work that incorporates Self-Conditioning and Asymmetric Time Intervals**

Our reproduction of the experiments - settings

- CIFAR-10 dataset: for its small size
- U-Net: 4 stages, 3 residual blocks, and 32 channels dimension for 9.1M params
- Training set up:
  - Adam optimizer with Exponential Moving Average
  - Learning rate: 10^{-4}
  - 100K steps (24 hours)
  - 64 batch size
  - 16-mixed precision
- FID score on CIFAR training set and 5K generated samples

Our reproduction of the experiments - results

Figure 9: Two samples generated with a U-Net BitDiffusion trained using self-conditioning on Cifar-10 after 70,000+ epochs

- Good visual results while training only 100K steps
- Classes like frogs, dogs and horses can be guessed sometimes.
- Not too much diversity in generated samples

Our reproduction of the experiments - results

- Training of BitDiffusion with self-conditionning and without

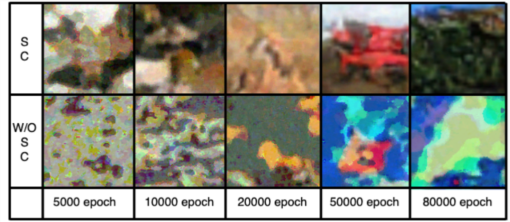


Figure 10: Comparison between images generated by our model with and without Self-Conditioning (SC), after respectively 5,000, 10,000, 20,000, 50,000 and 70,000 epochs.

Our reproduction of the experiments - results

Model type	SC	No SC
FID score	186.3	394.9

Table 1: FID score on Cifar-10 test set using BitDiffusion with and without self-conditionning

Training setup comparison with the paper:

- **Hyperparameters and training setup:** 100K vs 1.5M steps
- **Sample size to evaluate FID:** 5K vs 50K
- **Model complexity:** 9.1M vs 51M parameters

Conclusion

- Multiple concurrent contributions in this paper:
  - **Analog bits**: a technique to cast discrete data to higher-dimensional continuous space.
  - **Self-conditioning**: a technique to better guide generation at training and inference time.
  - **Asymmetric time intervals**: a technique impacting the t time parameters at sampling time, reducing the amount of artefacts.

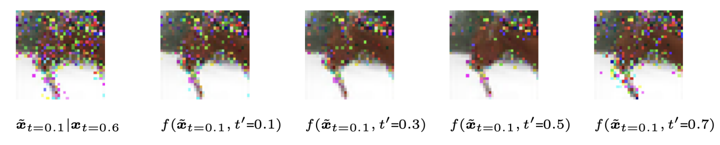


Figure 11: Improvement of their intermediate generation results using asymmetric time intervals at sampling time

References i

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton.
Analog bits: Generating discrete data using diffusion models with self-conditioning, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel.
Denoising diffusion probabilistic models.
In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon.
Denoising diffusion implicit models, 2022.

Alex Krizhevsky.
Learning multiple layers of features from tiny images.
University of Toronto, 05 2012.

References ii

- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei.
  Imagenet: A large-scale hierarchical image database.
  In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox.
  U-net: Convolutional networks for biomedical image segmentation, 2015.
- Diederik P. Kingma and Jimmy Ba.
  Adam: A method for stochastic optimization, 2017.
- Diederik P Kingma and Max Welling.
  Auto-encoding variational bayes, 2022.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling.
Improved variational inference with inverse autoregressive flow.
In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial networks, 2014.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord.
Step-unrolled denoising autoencoders for text generation, 2022.