An Overview of Diffusion Models for Text Generation

Helena Čeović*, Marin Šilić*, Goran Delač*, Klemo Vladimir*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10 000 Zagreb, Croatia
helena.ceovic@fer.hr, marin.silic@fer.hr, goran.delac@fer.hr, klemo.vladimir@fer.hr

***Abstract*—Given the great success that diffusion models have achieved in generating various types of continuous data, including image, video and audio, there has been a growing interest in the application of these models to text generation. However, the discrete nature of text presents a challenge for diffusion models initially designed for application in a continuous feature space. The two main lines of work that aim to bring together diffusion models and natural language processing are focused on either defining the diffusion process in continuous space by converting discrete tokens to embeddings or defining the diffusion process in discrete space. These recent works attempt to combine diffusion models with leading sequence-to-sequence generation Transformer architecture as well as with existing pre-trained language models. In this work, we give a detailed overview of the approaches developed to date. We present and analyze the benefits and limitations that each model introduces, along with how they compare to the autoregressive models that dominate this field.**

Keywords*—*diffusion models, denoising diffusion models, text generation, deep generative modeling, natural language processing

I. INTRODUCTION

Latest text-conditional image generators such as DALL-E 2 [1] and Imagen [2] have introduced unprecedented improvements in the field of image generation. They enable production of high-resolution outputs for any possible textual prompt. These modern frameworks are mostly based on diffusion models, a family of generative iterative models that achieve state-of-the-art sample quality for continuous data.

An immediate response to such successes was inevitable. Once the immense potential of diffusion models in generating images was displayed, the interest shifted further towards exploiting their benefits and extending their application to other challenging generative tasks. Naturally, this includes the common challenge of natural language processing (NLP) - text generation. Text generation spans over a large portion of popular NLP tasks including machine translation, building conversational systems, and abstractive text summarization. Text generation is a type of language modeling problem that aims to produce plausible and coherent natural language texts by automatically learning from data.

Autoregressive models currently dominate the field of language modeling. The biggest advance in their development was the Transformer architecture, presented in [3]. Imposing results are later achieved with GPT-3 [4] on a wide range of text generation tasks. However, despite their extensive capabilities, the fixed generation order limits the autoregressive (AR) models' flexibility in many controllable settings, specifically those that require both left and right context [5]. Instead of causal attention in AR models, diffusion models can leverage bidirectionality by predicting all tokens in a sequence at once and potentially lead to more coherent samples [6].

The main challenge that emerges in applying diffusion models to text generation is the discrete nature of textual data. The existing solutions are built and specialized in the continuous data domain which involves generating images, audio, and video. Multiple design modifications are required to make these models suitable for natural language modeling. Previous work in this field was mainly developed in one of the following two directions. The first involves defining text diffusion models in the discrete state space whereas the second focuses on continuous diffusion models. Discrete diffusion models corrupt sentences and refine them on the token level or switch from one discrete value to another. Some of these models attain strong results on several language modeling tasks, but they still fall behind AR models in terms of coherence due to the inability to model semantic correlation [6]. Diffusion models for continuous domain were first introduced in [5] where they explain the approach of embedding the discrete text into a continuous latent space. By preserving the continuity of the input, we keep important properties such as classifier-free guidance and the ability to represent uncertainty at individual token level. However, due to the different nature of data, modifications are necessary to apply existing diffusion models to textual data.

The paper is organized as follows. Sect. II introduces related surveys in the diffusion domain. Sect. III includes the formal and mathematical definition of diffusion models. The aforementioned two leading lines of work are discussed in detail in Sect. IV and Sect. V by presenting the development of diffusion models in discrete and continuous space, respectively. Conclusion and discussion of possible future research directions are given in Sect. VI.

II. RELATED WORK

Given the rapid increase in interest and development in the research of diffusion models, surveys have begun to emerge that aim to give an overview of the latest occurrences in the research field and state-of-the-art results

as well as point out current dominant applications along with the remaining challenges and limitations. The most comprehensive and extensive surveys to date are given by Cao et al. [7] and Yang et al. [8]. Both of these surveys mention natural language processing as one of the key areas for application and further work, but only briefly discuss the current works in the field.

Following the successes achieved with diffusion models in computer vision and the numerous papers focused on the task, surveys such as [9] and [10] attempt to give an overview and categorize the different approaches and design choices as well as to identify their contributions and limitations. More specialized surveys have begun to appear such as [11] where the focus lies on the medical domain by providing an extensive overview of diffusion models used for medical image analysis.

We believe this to be the first survey that specifically concentrates on diffusion models for text generation.

III. Diffusion Models

The idea of diffusion models as generative models was first introduced in [12]. They present a novel perspective derived from non-equilibrium statistical physics instead of standard variational Bayesian methods. The proposed algorithm consists of a forward inference diffusion process and a reverse generative process. The forward trajectory converts any complex data distribution $q(\boldsymbol{x}_0)$ into a simple, tractable distribution by gradually adding Gaussian noise to the data according to a variance schedule $\beta_1, ..., \beta_T$ [13]:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{1}$$

with each transition $x_{t-1} \rightarrow x_t$ parametrized by:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}). \tag{2}$$

The reverse process predicts the noise of current time step t and denoises to previous state $\boldsymbol{x}_{t-1}$. It is defined as a Markov chain with learned Gaussian transitions starting at $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x_T}; \mathbf{0}, \boldsymbol{I})$ [13]:

$$p(\boldsymbol{x}_{0:T}) := p(\boldsymbol{x_T}) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t). \tag{3}$$

Each denoising transition $x_{t-1} \rightarrow x_t$ is parametrized by the model:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)). \tag{4}$$

The variational lower bound (VLB) is tractable and differentiable with respect to θ on $\log p_\theta(\boldsymbol{x}_0)$ [5]:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}_0) = \mathop{\mathbb{E}}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \Bigg[\log \frac{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_T)} + \\ \sum_{t=2}^{T} \log \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0, \boldsymbol{x}_t)}{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)} - \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) \Bigg]. \end{aligned} \tag{5}$$

Sohl-Dickstein et al. [12] also show that we can rewrite the training objective in terms of KL divergences and entropies [5]:

$$\begin{aligned} L = \mathop{\mathbb{E}}_{q(\boldsymbol{x}_0)} \Bigg[D_{KL}[q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T)] + \\ \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \Big[D_{KL}[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)] \Big] - \\ \mathop{\mathbb{E}}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)] \Bigg]. \end{aligned} \tag{6}$$

However, Ho et al. [13] show that better results are obtained by simplifying the training objective. To avoid the potential instability, they suggest a simplified training objective that expands and reweights each KL-divergence term in Eq. 6 to obtain the sum of mean-squared errors between the ground truth and its estimates [5]:

$$\mathcal{L}(\boldsymbol{x}_0) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} ||\mu_\theta(\boldsymbol{x}_t, t) - \widehat{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0)||^2, \tag{7}$$

where $\widehat{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0)$ is the mean of the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0, \boldsymbol{x}_t)$ which is a closed form Gaussian, and $\mu_\theta(\boldsymbol{x}_t, t)$ is the predicted mean of $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ computed by a neural network.

IV. Diffusion models in discrete space

Discrete diffusion models build on categorical distributions where each token has some probability to be corrupted [14]. Diffusion models with discrete state spaces were first considered in [12] and applied to binary 'heartbeat' data. They train a probabilistic model on simple binary sequences of length 20, where a 1 occurs in every fifth bin, and the remainder of the bins are 0. The proposed binary sequence learning via binomial diffusion process was nearly perfect.

Hoogeboom et al. [15] go on to introduce multinomial diffusion modeled directly on categorical random variables with transition matrices characterized by uniform probabilities [16]. The same form was described in [17] although with no empirical evaluations. Furthermore, Austin et al. [16] generalize the approach from [15] with a more structured categorical corruption process. They shape data generation and embed structure or domain knowledge into transition matrices used by the forward process without relaxing or embedding discrete data into continuous spaces. They explore several structured transition matrices in their experiments. One of them is a transition matrix with an absorbing state where each token either stays the same or transitions to a special mask token with some probability allowing for the corrupted tokens to be distinguished from the original ones. They explore using similarity in an embedding space to guide the forward process and construct a transition matrix that transitions more frequently between tokens that have similar embeddings while maintaining a uniform stationary distribution. This work shows the importance of transition matrix choice by outperforming various non-autoregressive baselines for text generation on

character-level text generation as well as scaling discrete diffusion models to large vocabularies and long sequence lengths. However, this D3PM class of models remains inferior in comparison to leading autoregressive models for text generation such as Transformer XL.

None of the aforementioned works have incorporated pre-trained language models in their proposed work on diffusion models for text generation. He et al. [18] bring to attention that pre-trained language models such as BERT [19] and BART [20] are pre-trained with denoising objectives and explores discrete diffusion models to integrate the pre-trained language model. They use the transition matrix with an absorbing state for BERT introduced in [16]. It is more challenging to control the degree of noise added in the discrete domain than it is in the continuous. It is important to take into consideration the linguistic differences among tokens in addition to avoiding generating the most frequently appearing tokens only to achieve a higher likelihood. They introduce a spindle schedule that generates noise for $\boldsymbol{x}_t$ conditioned on both $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_0$. By experimenting with several methods of incorporating the time step into pre-trained language models, the time-agnostic decoding method examines the best results. Experiments are conducted on unconditional text generation where DiffusionBERT outperforms [5], a nominal work described in Sect. V, and additionally proves its incompatibility with BERT. Results prove the efficiency of the spindle noise schedule that is independent for each token and depends on the token's frequency in the dataset. DiffusionBERT achieves significant improvement in perplexity results, but still falls behind competitive autoregressive models.

Qian et al. [21] propose a non-Markovian diffusion process where modalities are gradually added as the diffusion step t decreases. Additionally, to achieve smoother modality learning, the generative process uses a non-autoregressive process with residual glancing sampling [22]. Experiments are conducted on tasks of machine translation, paraphrase generation and image caption where DiffGLAT, combined with Directed-Acyclic-Transformer [23], manages to outperform the autoregressive Transformer in both accuracy and efficiency on top of several other high-performing works outlined in Sect. V, including [14], [24], [25]. Their method significantly reduces the decoding iterations when compared to a strong prior discrete diffusion model known as SUNDAE [26]. In [26], Savinov et al. present an unrolled denoising technique where they unroll the chain by sampling from the transition distribution and feed samples back into the input during training.

A different angle is taken in [27] by introducing an edit-based generative model for text generation which is based on the diffusion denoising method. Both corruption and denoising processes are based on Levenshtein operations: insert, delete, replace, and keep. Diffuser demonstrates results comparable to AR models on machine translation, summarization, and style transfer tasks. They also show that Diffuser is complementary with token-level autoregressive methods outperforming autoregressive baselines.

V. Diffusion models in continuous space

The development of diffusion models for text generation has gradually shifted from models in discrete space presented in Sect. IV to continuous diffusion models which induce continuous latent representations. The idea is to preserve the advantages of high-quality continuous diffusion models while adjusting to a new data domain and thus get closer to AR models' results.

The first to pursue this type of architecture were Li et al. [5] by introducing multiple modifications to the standard diffusion model. They start by defining an embedding function that maps discrete text into a continuous space to enable direct application of a continuous diffusion model to discrete text. Embedding function $EMB(w_i)$ maps each word in sequence $\boldsymbol{w}$ of length n to: $EMB(\boldsymbol{x}) = [EMB(w_1), ..., EMB(w_n)] \in \mathbb{R}^{nd}$, where d is the number of dimensions. Furthermore, based on conducted experiments, they propose a new design of the training objective Eq. 5 that jointly learns the diffusion model's parameters and word embeddings. The novelties to the original diffusion models include an additional Markov transition between discrete words $\boldsymbol{w}$ and $\boldsymbol{x}_0$. In the forward process, this transition marks the embedding process, parametrized by $q_\phi(\boldsymbol{x}_0|\boldsymbol{w}) = \mathcal{N}(EMB(\boldsymbol{w}), \sigma_o I)$. A proper method is required for the inverse process of rounding a predicted continuous $\boldsymbol{x}_0$ back to discrete text. They accomplish rounding by choosing the most probable word for each position following argmax $p_\theta(\boldsymbol{w}|\boldsymbol{x}_0) = \prod_{i=1}^{n} p_\theta(w_i|x_i)$, where $p_\theta(w_i|x_i)$ is a softmax distribution. They reparametrize the training objective to ensure that the model predicts $\boldsymbol{x}_0$ in every term with $\boldsymbol{x}_0$ lying directly on a word embedding. Diffusion-LM performs control by running iterative gradient updates with fluency regularization and multiple gradient steps on the continuous latent variables which improve performance and fluency of the text along with speeding up decoding. Diffusion-LM demonstrates successful control for six control tasks, including text infilling where it achieves results competitive with a fine-tuned autoregressive model for this task [28]. However, the authors point out several weaknesses of the model, namely high perplexity, slower decoding and slower training convergence.

Following the same line of work that conducts diffusion directly in a continuous token embedding space, but focusing on a wider scope of application and more diverse textual data, Strudel et al. [6] introduce their Self-conditioned Embedding Diffusion (SED) model. Unlike Li et al. [5], they do not learn the embedding matrix $\boldsymbol{E}$ due to the discovered empirical instability, potential unigram entropy drops and limited applicability. Thus, the training objective depends only on trainable readout weights. Their model is based on the self-conditioning technique introduced by Chen et al. [29] which adapts $\tilde{\boldsymbol{x}}_0$ estimates by passing the estimate obtained at the previous sampling

step as the input to the denoising network instead of only $\boldsymbol{x}_t$. Dieleman et al. [30] point out the resemblance of the self-conditioning technique to the unrolled denoising strategy [26]. SED also allows conditional text generation by applying the span masking strategy for infilling tasks and classifier-free guidance [31] on text data to alleviate the need for a separately-trained guide model. Although AR baselines of similar capacities outperform SED, the model still performs strongly and comparably. Authors point out limitations which include limited model tuning, sampling efficiency, relying on another model for diffusion in a pre-trained embedding space as well as the lack of appropriate metrics and baselines for certain tasks.

Dieleman et al. [30] explore continuous diffusion for categorical data by following the formalism from [32] to describe the corruption and the reverse process using differential equations. The proposed framework, Continuous diffusion for categorical data (CDCD), is based on the diffusion framework in [33]. Similarly to prior works in this section, they embed the discrete input into a continuous space and directly apply continuous diffusion to the embeddings. The proposed CDCD framework replaces the standard score matching function with score interpolation for diffusion model training allowing the use of familiar categorical cross-entropy loss function for training. They go on to learn the embeddings and the diffusion model jointly to embed the discrete input into a continuous space, as seen in [5]. The final component that makes up the CDCD framework is an active learning strategy called time warping which automatically adapts the distribution of noise levels sampled during training to maximize efficiency. Experiments have shown that the CDCD model can produce compelling samples for prompt completion and infilling, but overall lower performance on machine translation with autoregressive models of the same size.

The focus in [14] lies in conducting a study of challenges connected to embedding discrete textual data and utilizing continuous diffusion models to generate it. Unlike images and audio that have a fixed data space during training, the embedding space is learnable for discrete textual data which may cause a collapse of the denoising loss function and bring instability to the training of the model. They find that the rounding loss used by Li et al. [5] and Gong et al. [24] is not sufficient for alleviating this issue. Secondly, because of the imbalanced frequency of tokens in textual datasets, the learning of token embeddings diverges. Considering the diverse embedding scale of different words, it is suboptimal to add the same amount of noise to different embeddings. Lastly, when denoising an embedding from noise sampled from the normal Gaussian prior, the generation process may be distracted by other embeddings that are near the noise. Their proposed model, named Differmer, solves the described challenges by implementing the following methods. First, they propose an anchor loss training objective that uses the model prediction $\hat{\boldsymbol{x}_0}$ as input instead of the noisy embedding $\boldsymbol{x}_0$ and successfully regularizes embeddings and prevents loss collapse. The solution to the imbalanced embedding scales is implemented in the form of a layer normalization module on the top of the embedding layer which guarantees the uniform scale of tokens. As for the potential distraction in the diverse process, it is eliminated by adding an increasing scale of noise at each step. Experiments conducted on the tasks of machine translation and text summarization show the Differmer outperforming concurrent diffusion-based models including those presented in [5] and [30] as well as achieving comparable results with the autoregressive Transformer model.

Gong et al. [24] go on to extend the framework from [5] to a more generalized sequence-to-sequence setting, a key NLP aspect that covers various downstream tasks. The proposed DiffuSeq model introduces partial noising in the forward process by imposing noise only on the target sequence, but not on the source sequence. The reverse process can therefore impose the input as the condition when denoising. Since no additional classifiers are required to control the denoising process, it is considered classifier-free. Experiments are conducted on four sequence-to-sequence tasks: open domain dialogue, question generation, text simplification and paraphrase. DiffuSeq achieves comparable or higher quality results than competitive AR, iterative non-autoregressive, and large-scale pre-trained models, specifically demonstrating its ability to generate diverse, and hence high-quality sequences.

Yuan et al. [25] continue the research on diffusion models for sequence-to-sequence text generation. Unlike the encoder-only Transformers with partial denoising used in [24], their SeqDiffuSeq model is based on an encoder-decoder Transformers architecture with self-conditioning [29] and a token-level adaptive noise schedule. The encoder-decoder architecture has computational advantages during inference because the input sequences require only one forward computation through the encoder during the whole reverse process. They improve sequence-to-sequence text generation by introducing a heuristic that the difficulty of predicting the sample should increase linearly with respect to time steps. Their suggested noise schedule is different from the previous proposals including those in DiffusionBERT [18] and CDCD [30]. SeqDiffuSeq achieves significant acceleration in inference speed compared to DiffuSeq. Seq2Seq proves its generation quality by surpassing both DiffuSeq and several autoregressive baseline models on text simplification and paraphrase. On question generation, it is comparable with DiffuSeq whereas for open domain dialogue and machine translation it demonstrates poorer results than the AR baseline.

GENIE [34] can be considered a semi-non-autoregressive model and the first large-scale pre-trained language model based on diffusion. A new approach is proposed that combines the diffusion model and Transformers. They present a pre-training task named continuous paragraph denoise which predicts the noise added to continuous paragraphs in the current time step based on the paragraph context information and

the noisy paragraph information. Experiments show the effectiveness of the large scale pre-training by achieving comparable results to a pre-trained AR model on task of text summarization.

Another line of work includes employing the continuous process on surrogate representations of discrete data such as analog bits [29] and simplex [35]. Chen et al. [29] design a simple and generic approach to enable continuous state diffusion models to generate discrete data. The idea lies in representing the discrete data as binary bits which are then modeled into analog bits by continuous diffusion models. Analog bits are real numbers that share the same bimodal values as binary bits that represent discrete data. This process requires no discrete space or re-formulation of the continuous diffusion process. For sampling, the process is the same as in continuous diffusion models with the additional step of applying a thresholding operation which decodes the generated analog bits that are then converted into original discrete variables. They introduce the self-conditioning technique previously described in this section and often used in the following work. Their model achieves comparable results to an autoregressive Transformer baseline on the task of image captioning.

The SSD-LM presented in [35] introduces a different approach that aims to bring together the advantages of autoregressive and diffusion models by proposing a semi-autoregressive solution. It maintains the ability to train and generate variable-length sequences in addition to allowing refinement within the token block, thus maintaining the advantages of both autoregressive and diffusion models. Unlike the representation in [29] which can lead to extremely long sequences, they keep a subword based vocabulary with each token represented as a sequence of manually defined logits. SSD-LM is comparable to or outperforms competitive autoregressive baselines such as GPT-2 both in quality and diversity on unconstrained text generation. Still, several limitations of the model in comparison to autoregressive models are pointed out including lower sample efficiency, slower decoding speed and inflexibility of the decoding schedule.

VI. Conclusion

This paper gives a comprehensive overview of research conducted in the field of text generation using diffusion models. Following the impressive achievements of diffusion models in image generation, the desire to extend their application to other domains has grown rapidly. However, language modeling itself represents a challenging task due to the complexity and diversity of the natural language. Furthermore, the high-performing image generation models were designed for continuous data space, requiring significant modification to generate discrete textual data. We presented the most notable advancements in their development and how they compare to the current state-of-the-art autoregressive models.

We considered works developed in both discrete and continuous data domains, ranging from those focused on controllable text generation to unconditional and conditional text generation as well as sequence-to-sequence tasks. Although many of them examine respectable results, they are still mostly limited to either a specific task domain or language structure in order to perform comparably to autoregressive models.

Diffusion models offer a more flexible approach by abandoning the sequential generation of autoregressive models and generating simultaneous outputs through iterative refinement. However, they struggle to catch up with AR models' inference speed and efficiency. There are still multiple open questions addressed in many of the considered works, such as defining the noise schedule and the embedding space.

Finally, this is an active area of research aspiring to reach the breakthrough seen in the field of image generation. It is yet to be seen whether such performance is feasible, especially considering the existing state-of-the-art autoregressive models that are already capable of generating high-quality diverse texts. Another issue in this field is the lack of appropriate metrics and baselines for certain tasks, such as infilling. The discussed works show promising directions for diffusion models in text generation that are yet to be extended to their full potential.

References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022. [Online]. Available: https://arxiv.org/abs/2204.06125

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022. [Online]. Available: https://arxiv.org/abs/2205.11487

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[5] X. L. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," 2022. [Online]. Available: https://arxiv.org/abs/2205.14217

[6] R. Strudel, C. Tallec, F. Altché, Y. Du, Y. Ganin, A. Mensch, W. Grathwohl, N. Savinov, S. Dieleman, L. Sifre, and R. Leblond, "Self-conditioned embedding diffusion for text generation," 2022. [Online]. Available: https://arxiv.org/abs/2211.04236

[7] H. Cao, C. Tan, Z. Gao, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion model," 2022. [Online]. Available: https://arxiv.org/abs/2209.02646

[8] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," 2022. [Online]. Available: https://arxiv.org/abs/2209.00796

[9] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2209.04747

[10] A. Ulhaq, N. Akhtar, and G. Pogrebna, "Efficient diffusion models for vision: A survey," 2022. [Online]. Available: https://arxiv.org/abs/2210.09292

[11] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models for medical image analysis: A comprehensive survey," 2022. [Online]. Available: https://arxiv.org/abs/2211.07804

[12] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *CoRR*, vol. abs/1503.03585, 2015. [Online]. Available: http://arxiv.org/abs/1503.03585

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[14] Z. Gao, J. Guo, X. Tan, Y. Zhu, F. Zhang, J. Bian, and L. Xu, "Difformer: Empowering diffusion model on embedding space for text generation," 2022. [Online]. Available: https://arxiv.org/abs/2212.09412

[15] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling, "Argmax flows and multinomial diffusion: Learning categorical distributions," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 454–12 465. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/67d96d458abdef21792e6d8e590244e7-Paper.pdf

[16] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, "Structured denoising diffusion models in discrete state-spaces," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=h7-XixPCAL

[17] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020. [Online]. Available: https://arxiv.org/abs/2010.02502

[18] Z. He, T. Sun, K. Wang, X. Huang, and X. Qiu, "Diffusionbert: Improving generative masked language models with diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2211.15029

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019. [Online]. Available: https://arxiv.org/abs/1910.13461

[21] L. Qian, M. Wang, Y. Liu, and H. Zhou, "Diff-glat: Diffusion glancing transformer for parallel sequence to sequence learning," 2022. [Online]. Available: https://arxiv.org/abs/2212.10240

[22] L. Qian, H. Zhou, Y. Bao, M. Wang, L. Qiu, W. Zhang, Y. Yu, and L. Li, "Glancing transformer for non-autoregressive neural machine translation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1993–2003. [Online]. Available: https://aclanthology.org/2021.acl-long.155

[23] F. Huang, H. Zhou, Y. Liu, H. Li, and M. Huang, "Directed acyclic transformer for non-autoregressive machine translation," 2022. [Online]. Available: https://arxiv.org/abs/2205.07459

[24] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2210.08933

[25] H. Yuan, Z. Yuan, C. Tan, F. Huang, and S. Huang, "Seqdiffuseq: Text diffusion with encoder-decoder transformers," 2022. [Online]. Available: https://arxiv.org/abs/2212.10325

[26] N. Savinov, J. Chung, M. Binkowski, E. Elsen, and A. v. d. Oord, "Step-unrolled denoising autoencoders for text generation," 2021. [Online]. Available: https://arxiv.org/abs/2112.06749

[27] M. Reid, V. J. Hellendoorn, and G. Neubig, "Diffuser: Discrete diffusion via edit-based reconstruction," 2022. [Online]. Available: https://arxiv.org/abs/2210.16886

[28] C. Donahue, M. Lee, and P. Liang, "Enabling language models to fill in the blanks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2492–2501. [Online]. Available: https://aclanthology.org/2020.acl-main.225

[29] T. Chen, R. Zhang, and G. Hinton, "Analog bits: Generating discrete data using diffusion models with self-conditioning," 2022. [Online]. Available: https://arxiv.org/abs/2208.04202

[30] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, C. Durkan, C. Hawthorne, R. Leblond, W. Grathwohl, and J. Adler, "Continuous diffusion for categorical data," 2022. [Online]. Available: https://arxiv.org/abs/2211.15089

[31] J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022. [Online]. Available: https://arxiv.org/abs/2207.12598

[32] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020. [Online]. Available: https://arxiv.org/abs/2011.13456

[33] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," 2022. [Online]. Available: https://arxiv.org/abs/2206.00364

[34] Z. Lin, Y. Gong, Y. Shen, T. Wu, Z. Fan, C. Lin, W. Chen, and N. Duan, "Genie: Large scale pre-training for text generation with diffusion model," 2022. [Online]. Available: https://arxiv.org/abs/2212.11685

[35] X. Han, S. Kumar, and Y. Tsvetkov, "Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control," 2022. [Online]. Available: https://arxiv.org/abs/2210.17432