Zero-Shot Translation using Diffusion Models

Eliya Nachmani*1,2 and Shaked Dovrat*2

Abstract—In this work, we show a novel method for neural machine translation (NMT), using a denoising diffusion probabilistic model (DDPM) [1], adjusted for textual data, following recent advances in the field. We show that it's possible to translate sentences non-autoregressively using a diffusion model conditioned on the source sentence. We also show that our model is able to translate between pairs of languages unseen during training (zero-shot learning).

I. INTRODUCTION

Deep generative models can be categorized in the following categories: (i) Flow-based models, such as Glow [2] (ii) Autoregressive models, e.g. Transformer, for language modeling [3], (iii) GAN [4] based models, such as, WaveGAN [5] for speech and StyleGAN [6] for vision application. (iv) VAE [7] based models, e.g. VQ-VAE [8] and NVAE [9], and (v) Diffusion Probabilistic Models [10] such as ADM [11].

Diffusion probabilistic models achieve comparable and superior results to other deep generation models such as WaveGrad for speech synthesis [12] and ADM for image generation [11].

The underline architecture of the diffusion probabilistic models is a chain of Markov latent variables. The data flows in two directions: (i) the diffusion process, and (ii) the denoising process. The denoising process is the inference process which generates the data starting from Gaussian noise. The diffusion process is the training process which learns to transform data samples into Gaussian noise.

In the seminal work of Hoogeboom et al. [13], a diffusion model for categorical variable was introduced. The paper shows that the original diffusion process, which is suitable for continuous data such as speech and and ordinal data such as images, can model discrete categorical data. They trained a diffusion network on the language modeling task.

In this work we propose a diffusion model for neural machine translation. Furthermore, we show that the proposed model has some capabilities of zero-shot translation. To our knowledge, we are the first to perform conditional text generation using a diffusion model.

II. RELATED WORK

In [10] Sohl-Dickstein et al. introduce the diffusion process. The diffusion process takes the variational distribution \(q(x_t|x_{t-1})\) and adds Gaussian noise at each time step where \(t \in \{1, \ldots, T\}\), \(x_0\) is the original data point and \(x_T\) is completely noise.

In this section we will recap the multinomial diffusion process as defined by Hoogeboom et al. [13] for categorical data. We denote \(x_t\) as a 1-hot vector with \(K\) categories. \(x_0\) is the data point, and \(q(x_t|x_{t-1})\) is the diffusion model that gradually adds a small amount of noise at each step. At \(t = T\), \(x_T\) is almost completely noise. The opposite direction \(p(x_{t-1}|x_t)\) is a learnable distribution that denoises the data. The diffusion model is optimized with the variational bound on negative log likelihood:

\[
\log P(x_0) \geq E_{x_1,...,x_T \sim q} \left[ \log p(x_T) + \sum_{t=1}^{T} \log \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right]. \tag{1}
\]

Sohl-Dickstein et al. [10] use \(x_0\) as condition and show that Eq.1 becomes:

\[
\log P(x_0) \geq E_{q} \left[ \log p(x_0|x_1) - KL(q(x_T|x_0)p(x_T)) - \sum_{t=2}^{T} KL(q(x_{t-1}|x_t,x_0)p(x_{t-1}|x_t)) \right] \tag{2}
\]

where \(KL(q(x_T|x_0)p(x_T)) \approx 0\) if the diffusion trajectory \(q\) is defined well. The variational distribution \(q(x_t|x_{t-1})\) is defined as follows:

\[
q(x_t|x_{t-1}) = C(x_t)(1 - \beta_t)x_{t-1} + \beta_t/K \tag{3}
\]

where \(\beta_t\) is the probability to sample from the uniform distribution. Using a Markov chain property, one can get the closed form to sample \(x_t\) from \(x_0\):

\[
q(x_t|x_0) = C(x_t)[\tilde{\alpha}_t x_0 + (1 - \tilde{\alpha}_t)/K] \tag{4}
\]

where \(\tilde{\alpha}_t\) and \(\alpha_t\) are defined in the same manner as in the original DDPM [1], i.e. \(\alpha_t = 1 - \beta_t\) and \(\tilde{\alpha}_t = \prod_{r=1}^{t} \alpha_r\).

One can further relax the closed form:

\[
q(x_{t-1}|x_t,x_0) = C(x_{t-1})[\hat{\theta}/K] \tag{5}
\]

where

\[
\hat{\theta} = [\alpha_t x_t + (1 - \alpha_t)/K] \circ [\tilde{\alpha}_{t-1} x_0 + (1 - \tilde{\alpha}_{t-1})/K] \tag{6}
\]

Hoogeboom et al. [13] predicts a probability vector for \(\tilde{x}_0\) from \(x_t\). They parametrize \(p(x_{t-1}|x_t)\) from \(q(x_{t-1}|x_t, \tilde{x}_0)\),

---

1Facebook AI Research
2Tel-Aviv University
*Equal contribution
where \(x_0\) is approximated with a neural network \(\hat{x}_0 = \mu(x_1,t)\). Denote

\[
\theta_{\text{post}}(x_t, x_0) = \tilde{\theta} \sum_{k=1}^{K} \tilde{\theta}_k
\]

(7)

Then, the variational lower bound Eq.2 becomes:

\[
\log P(x_0) \geq E_q \left[ \sum_k x_{0,k} \log \hat{x}_{0,k} - \sum_{t=2}^{T} \text{KL}(C(\theta_{\text{post}}(x_t, x_0)))|C(\theta_{\text{post}}(x_t, \hat{x}_0))) \right]
\]

(8)

It is worth to mention the work by Austin et al. [14] which improves Hoogeboom et al. [13] by introducing corruption processes with uniform transition probabilities. They use transition matrices that mimic Gaussian kernels in continuous space and show that using different transition matrix leads to improved results in text generation.

**Data Processing** The diffusion model requires inputs to be of fixed length. Thus, we use padding and truncation of all sentences to a fixed length \(L\). Sentences are padded with a special token \([PAD]\).

Two special language tokens are added to each input sentence, the first indicates the source language and the second indicates the target language. This is used to accelerate convergence and to allow for zero-shot learning, given pairs of languages unmet during training.

**IV. Experiment Setup**

**A. Datasets**

We used three datasets in total, all of which are from WMT. We trained our net on WMT14 [15] DE-EN and WMT14 FR-EN jointly, and from both directions, meaning German to English, English to German, French to English and English to French. We downsampled the larger French dataset in each epoch in order to have the same number of German-English and French-English samples.

Lastly, we used the WMT19 [16] DE-FR for evaluation only, in order to test the method’s zero-shot learning performance.

**B. Evaluation Metrics**

We used common machine translation evaluation metrics: Corpus level BLEU, SacreBLEU, TER and chrF. We used the official SacreBLEU implementation [17] with default parameters.

**C. Baselines**

Current state-of-the-art results for the WMT14 translation tasks perform \(~ 35\) BLEU and SacreBLEU for the German-English translation and \(~ 45\) for the French-English translation. All of these methods use some type of a transformer with autoregressive decoding, use very large models, and some use extra data.

**D. Hyperparameter Settings**

We used the ADAM optimizer [18], and tuned the learning rate, batch size and gamma parameters. Eventually, we used a learning rate of \(5e-4\), gamma value of 0.9 and batch size
The vocabulary size is an important hyper parameter that we tuned.

### Table I

| Tasks   | BLEU | SacreBLEU | TER | chrF |
|---------|------|-----------|-----|------|
| DE→EN   | 7.17 | 8.13      | 93.1| 34.7 |
| EN→DE   | 3.54 | 4.54      | 102.3| 33.5 |
| FR→EN   | 8.62 | 9.93      | 88.4| 37.8 |
| EN→FR   | 7.56 | 9.02      | 90.7| 37.5 |
| DE→FR   | 4.17 | 5.06      | 94.7| 31.4 |
| FR→DE   | 2.96 | 4.04      | 98.1| 31.4 |

### Table II

| V. Size | DE→EN | EN→DE | FR→EN | EN→FR |
|---------|-------|-------|-------|-------|
| 1024    | 5.60  | 3.18  | 7.23  | 6.73  |
| 2048    | 7.92  | 4.42  | 9.83  | 8.99  |
| 4096    | 8.13  | 4.54  | 9.93  | 9.02  |
| 8192    | 6.76  | 4.00  | 8.89  | 7.75  |

SacreBLEU results with different vocabulary sizes. We see that $K = 4096$ gives the best results, which is not on the edge of the values chosen. This indicates it’s a sweet spot for the vocabulary size tradeoff.

### V. Results

Results for the different translation tasks are depicted in Table I. The results are unsatisfactory, implying the method is currently not suitable for the translation task. Results for the zero-shot translation tasks (WMT19) show that some generalization to unmet pairs of languages was possible, but because the overall performance of the system is low, it is hard to estimate if the method transforms well to zero-shot learning.

Results for the vocabulary size tuning is depicted in Table II, suggesting a vocabulary of size $K = 4096$ is closest to the optimal value in this case.

Qualitatively speaking, results quality vary, and overall we see an expected correlation between the difficulty of inputs and the quality of the translation. Nonetheless, some observations are hard to explain, such as relatively good translations for seemingly hard sentences and relatively bad translations for seemingly easy sentences. Table III shows four randomly selected samples from the test set, one from each task (ordered pair of languages).

### VI. Discussion

#### A. Learning the Transition Matrix

One idea we had was to learn the transition matrices that determine the probabilities of noise changing one token to another. In the described "vanilla" implementation, all probabilities of change are uniform. Diffusion models for continuous or ordinal data use Gaussian noise, which gives higher probabilities to small changes, resulting in a much easier learning ground for the denoising optimization procedure. This advantage is lost when using the uniform distribution for categorical data. Austin et. al. [14] was able to improve on that by using non-uniform noise distributions.

Following this idea, we aimed to learn the noise distribution jointly with the diffusion model. Later we found out it was infeasible, since the learning procedure uses pre-computed powers of the transition matrix to enable fast learning. Specifically, for each training iteration at some $t$, this would require the computation of the $t$th power of a $K \times K$ matrix, where $K \sim 2^{10}$ and $t \sim 1000$. This makes the technique infeasible.

#### B. Conclusions

In this work, we tried to solve a thoroughly researched NLP task, MNT, using a recent and very promising method, DDPMs, for the first time (to our knowledge). This method has the potential to generate text with high performance in a non-autoregressive way. Although DDPMs achieve state-of-the-art results in generating both continuous and ordinal data, it is yet to show competing results for categorical data such as text. We hoped to show that it can give reasonable results for non-autoregressive translation.

### References

[1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” arXiv preprint arXiv:2006.11239, 2020.

[2] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” arXiv preprint arXiv:1807.03039, 2018.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998-6008.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” arXiv preprint arXiv:1406.2661, 2014.
### Table III

| Sample   | Sentence                                                                 | Lang. | SacReBLEU |
|----------|---------------------------------------------------------------------------|-------|-----------|
| 1st Input| je sais qu’il voudrait une garantie de quatre ans.                        | FR    | -         |
| 1st Reference | i know he would like a four - year guarantee.                          | EN    | -         |
| 1st Prediction | i know he need a guarantee for four years.                        | EN    | 17.47     |
| 2st Input | the ecb’s sole mandate has always revolved around inflation, therefore mario draghi and his team have all the more reason to take action at their meeting next week. | EN    | -         |
| 2st Reference | le mandat unique de la bce a toujours porte sur l’inflation, donc mario draghi et son equipe ont davantage de raisons d’agir lors de la reunion de la semaine prochaine. | FR    | -         |
| 2st Prediction | l’ile systeme unique de la bce est derriere le trend en phoque, afin ou mario draghi et son mont sont plus justifies de prendre en contact a la session prochaine. | FR    | 17.10     |
| 3rd Input | zwei kinder haben in uruguay den mord eines 11 - jahrigen eingestanden. | DE    | -         |
| 3rd Reference | two children have confessed to the murder of an 11 - year - old in uruguay. | EN    | -         |
| 3rd Prediction | they had spent a 118’old increased blood in uruguay about alleged murders abandone treating two children. | EN    | 6.84      |
| 4th Input | town council delighted with solid budget                              | EN    | -         |
| 4th Reference | gemeinderat freut sich uber soliden haushalt                            | DE    | -         |
| 4th Prediction | der stadtrat erfullt einen beliebten haushalt                          | DE    | 8.12      |

Randomly selected samples from our model. 2nd sample shows a relatively good translation for a long sentence, and the 3rd sample shows a failed translation for a seemingly easier sentence.

---

[5] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” *arXiv preprint arXiv:1802.04208*, 2018.

[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[7] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[8] A. Razavi, A. v. d. Oord, and O. Vinyals, “Generating diverse high-fidelity images with vq-vae-2,” *arXiv preprint arXiv:1906.00446*, 2019.

[9] A. Vahdat and J. Kautz, “Nvae: A deep hierarchical variational autoencoder,” *arXiv preprint arXiv:2007.03898*, 2020.

[10] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[11] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *arXiv preprint arXiv:2105.05233*, 2021.

[12] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “Wavegrad: Estimating gradients for waveform generation,” *arXiv preprint arXiv:2009.00713*, 2020.

[13] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forró, and M. Welling, “Argmax flows and multinomial diffusion: Towards non-autoregressive language models,” *arXiv preprint arXiv:2102.05379*, 2021.

[14] J. Austin, D. Johnson, J. Ho, D. Tarlow, and R. v. d. Berg, “Structured denoising diffusion models in discrete state-spaces,” *arXiv preprint arXiv:2107.08006*, 2021.

[15] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand et al., “Findings of the 2014 workshop on statistical machine translation,” in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 12–58.

[16] L. Barrault, O. Bojar, M. R. Costa-Jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmi, et al., “Findings of the 2019 conference on machine translation (wmt19),” in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 2019, pp. 1–61.

[17] M. Post, “A call for clarity in reporting BLEU scores,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: [https://www.aclweb.org/anthology/W18-6319](https://www.aclweb.org/anthology/W18-6319)

[18] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.