Guided Language Diffusion

Classifier-free conditional text generation with DDPM

Victor Hobe-Gelting

Abstract

This paper explores classifier-free guidance in diffusion language models. As a recent development, diffusion models are being actively researched and developed. Classifier-free guidance, showing promising results in image generation has not yet been applied in the same way on diffusion language models. Emotion targets are injected into the diffusion model to guide the denoising process toward the target control constraints. Although results show the principal concept has an effect on the generated outputs, the results are still inconclusive.

1 Introduction

Rapid advances in the abilities of generative models in recent years have led to the topic of „Artificial Intelligence“ now being broadly and controversially discussed in mainstream society. Deep generative models now have the ability to generate high quality output in a variety of domains like image, natural language, speech and music generation. To the human eye these outputs often are indistinguishable from ‚real‘ output. Despite the impressive capabilities of generative models, there are ongoing challenges, e.g. model size and connected to that, environmental impact are in many cases linked to the challenge of controlling output and mitigating the trade off between fidelity and mode coverage.

As other domains, natural language has seen rapid development in recent years and models can generate high quality text (Radford et al 2019, Zhang et al., 2022). Despite the notable increases in output quality, controlling the text generation process has been notoriously difficult – with many complex or combined control tasks not being practically achievable(Li et al., 2022). Model size has been increased to hundreds of billions of parameters in order to raise output quality by increasing implicit representation and thus output quality, yet even the largest models are hard to control and for many applications output has to be filtered in order to ensure that it adheres to both task presented and quality. Toxic language for example has often proven to be challenging to eliminate from a model's output.

Some continuous space domains, most notably image generation have seen the advent of a new architecture, where a model is trained to reverse a noising process that was previously applied to the data. As this denoising loop is an iterative process, it has been proven to lend itself to guiding the output. The improvement of quality or the reduction in parameter count, or both at the same time, can in some cases be linked to diffusion models, maybe the most popular example being Dalle-2.

With the advent of classifier guidance, diffusion models have recently exhibited controllable generation of high quality samples. Further research explored classifier-free guidance where the diffusion model generates controlled outputs independent of an external control mechanism (Marcus et al., 2022), (Nichol et al., 2021). In this domain image diffusion models have shown to be highly expressive and controllable and outperforming classifier guided diffusion (Dhariwal et al., 2021).

2 Background

2.1 Diffusion models

Diffusion models employ a forward process, which generates training data and a learnt backward process. Inspired by non-equilibrium thermodynamics, diffusion models define a process as a Markov chain of diffusion steps to iteratively add gaussian noise to data according to:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathrm{I})$$

This means that the noising chain approaches random data $T \rightarrow \infty, x_T \approx \mathcal{N}(0, \mathrm{I})$

Image 1: Visualising the forward diffusion process, stepwise adding noise

The model is then trained on singular steps of the reverse process

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \textstyle\sum_\theta(x_t, t))$$

predicting x_{t-1} from x_t.

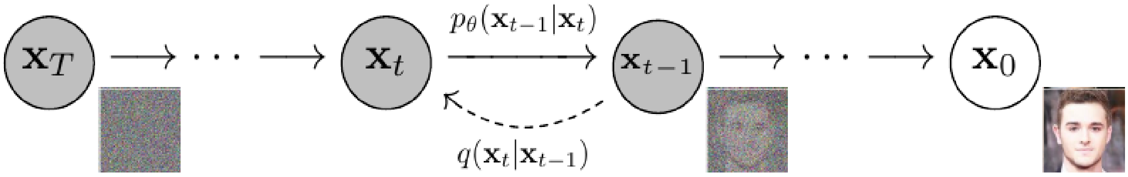
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

Figure 1: graphical representation of an unconditional diffusion model (Source: Ho, Jain et Abbel, 2020)

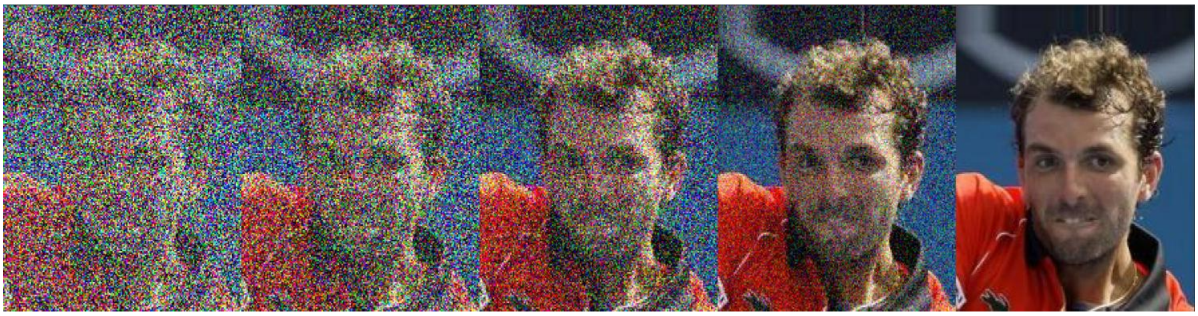

Image 2: Visualising the backward diffusion denoising process

These models learn to denoise previously noised training data in this iterative process, which is at the core of the generative process, as the model starts from random noise:

$$p_\theta(x_T) \approx \mathcal{N}(0, \mathrm{I})$$

via the learned iterative denoising process.

This means, when fed with random input, they can „denoise“ this into an output which resembles the training data – in other words they generate random but meaningful output. This randomly generated output has little value as it must be controllable in order to generate more specific output. As this generative denoising process is iterative, it lends itself to guidance. This can either be achieved by means of an external classifier or internally by classifier-free guidance.

Diffusion models seem to be robust in training (Li et al., 2022) and do not suffer from mode collapse like GANs do.

The diffusion process is in some ways similar to how we humans create text, paintings, code or any other sufficiently complex creational task. Writing a book for example is done in a kind of denoising process too. An author starts with nothing, not much different from randomness. Then the creative process starts, with a rudimentary idea, or more concrete with the first character typed. This transitory state of a single character does not resemble a novel, but it is closer to being a novel than a blank page. The first character is extended to the first word, the first sentence, paragraph, chapter and so on. With each step the work reduces the distance to what we call a novel. As the work progresses, the content is not static, it gets manipulated, shuffled, corrected and rewritten – denoised. It would be hard to imagine writing a book cover to cover, with the cursor only ever moving forward, not manipulating anything that already is created. This task would be infinitely harder for us humans. It seems that this in a certain way might hold true for machines too.

2.2 Classifier Guided Diffusion

Classifier guidance is a recent method employed to achieve similar control of outputs as e.g. low temperature sampling applied in other generative architectures. Classifier guidance employs a conditional reverse process, where the unconditional denoising process

$$p_\theta(x_{t-1}|x_t)$$

generates a transitory state. In order to condition the generation process the transitory state is then updated with the gradient of the log likelihood

$$\nabla_{x_{t-1}}\, log\, p_\phi(y|x_{t-1})$$

supplied by a classifier

$$p_\phi(y|x_{t-1})$$

that has been trained on noisy data. If one imagined the task of creating a cat image, the process could be described as the diffusion model denoising a step and asking the classifier what it thinks. As the classifier knows that the final output should be a cat image, it computes the loss with regard to the label ‚cat‘, then it supplies the gradient to the diffusion model, effectively telling it „It’s not a cat yet, but if you move a tiny step into this direction, it will be a tiny bit more cat-like“.This effectively trades off mode coverage and sample fidelity(Ho, Salimans, 2021).

Contrary to early results on class conditioning for image generation, where the condition was added at more and more points throughout the model, in order to trying to ‚convince‘ the model to take the condition into account to a satisfying level, classifier guidance has shown to lead to excellent samples (Dhariwal et al., 2021). As the classifier model is independent from the diffusion model, this lends itself to plug and play approaches. Furthermore it also facilitates conditioning the diffusion process on more than one control task as the gradients supplied by multiple classifiers can be combined to update the transitory state.

However, classifier guidance also comes at a cost. The diffusion model relies on externally supplied gradients for the transitory state updates, thus involving at least two models, which both have to calculate output for each reverse diffusion step at inference. Here size of classifier and with it the ability to successfully classify noisy data has to be traded off against the compute and time cost at training as well as inference time.
As these classifiers have to classify noisy data, it is not possible to employ pretrained models, so classifiers generally have to be trained with a similar process to the diffusion model training, where data is corrupted with varying amounts of noise.
With multiple control tasks the complexity and compute at inference and training rises with $O_{(n)}$ with respect to the number of control tasks. As such a system of a diffusion model with multiple classifiers does not get trained on shared objectives, the different objective gradients might oppose each other more than in a scenario where a system is trained on joint objectives. The joint objective should lead to a more efficient representation of the sub-objectives with regard to each other, making multiple objective transitory state updates more targeted and less erratic.

2.3 Classifier-free Guidance

Classifier-free guidance does not rely on any outside input to the transitory states. In order to achieve this, the diffusion model must not only learn the reverse diffusion process, but also a representation of the condition. The diffusion model is trained on two tasks in order to achieve this. As described by Ho and Salimans (2021), at training a conditional diffusion model

$p_\theta(x_{t-1}|x_t, y)$

the condition y is replaced with a null label $\emptyset$ with a fixed probability. Then the transitory states are extrapolated from $p_\theta(x_{t-1}|x_t, \emptyset)$ through $p_\theta(x_{t-1}|x_t, y)$ by s according to:

$$\hat{p}_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t, \emptyset) + s \cdot (p_\theta(x_{t-1}|x_t, y) - p_\theta(x_{t-1}|x_t, \emptyset))$$

This extrapolation pushes the transition even further into the direction where the conditional model would assign a higher probability of the output satisfying this condition. To revert back to the example of the diffusion model having to create an image of a cat, this would mean that now our model knows what a cat is. It will generate one transitory state with the

knowledge in mind that it is to create an image of a cat. Then it also performs one denoising step without paying any attention to its knowledge of what a cat is. The following can be understood as the model's cat amplifier, it compares both transitory states, seeing, if it does some certain denoising, the image is more cat-like than with the other denoising option it created. Instead of just picking the better one, the conditional transitory state, it now steps from the unconditional state to the better, conditional state – and then continues to walk another couple steps into that direction.
Eliminating the need for an external classifier as the model's embedded representation of the condition is leveraged also has advantages in areas where it is either hard to classify the conditions or where the transitory noisy states introduce classification challenges.

As with classifier guidance, this also has its limitations. The model is not a plug and play architecture, if additional criteria were to be introduced, the model would have to be retrained.

3 Related Work

Diffusion models are among the more recent additions of the family of model architectures, introduced by Sohl-Dickstein et al. in 2015. Over the last couple of years diffusion models have seen as steep a rise in interest as in development.

Most work has been carried out on image generation tasks. With the implementation of output control with classifier guidance (Dhariwal, Nichol, 2021) these models showed their potential to rival GANs on the creation of photorealistic images.

Classifier-free guidance as introduced by Ho and Salimans (2021) has seen much interest, Nichol et al., (2022) have implemented classifier-free image generation with text guidance.

So far the application of diffusion models on natural language have been limited (Li et al., 2022). Contrary to image data, text is discrete in nature, (Austin et al., 2021) and (Hoogenboom et al., 2021, 2022) have studied text diffusion in discrete space.
Recently (Li et al., 2022) have applied diffusion to text generation by mapping the discrete text space to a continuous space with learned text embeddings and proven that this approach can generate controlled output of high quality and thus is a feasible approach for complex control tasks - outperforming state of the art models on some of these tasks.

4 Experimental Setup

As the usefulness of generated text is directly related to the amount of control that can be exerted at inference. Control tasks for text have shown to be non-trivial as output of even the largest models could not be controlled to complex targets. The advent of diffusion models in discrete space domains has seen notable improvements in controlling output.
This paper tries to explore classifier-free guidance as described by Ho and Salimans (2021) on text generation, which to the knowledge of the author has not yet been applied to text generation.

With recent research now extending into the discrete space of text, continuous diffusion has shown early potential despite being applied to discrete data.
This paper explores the implementation of classifier-free continuous diffusion to generate controlled output according to emotional targets that were injected in form of a probability density across the four labels „Anger“, „Joy“, „Optimism“ and „Sadness“. These were then encoded and added to the timestep encoding of the diffusion process.

The model is the same 90M parameter transformer based model as Li et al. (2022) employed for their controllable text generation. In order to implement the classifier-free diffusion, the model was adapted to update the transitory state according to

$$\hat{\epsilon}_\theta(x_{t-1}|x_t, y) = \epsilon_\theta(x_{t-1}|x_t, \emptyset) + s \cdot (\epsilon_\theta(x_{t-1}|x_t, y) - \epsilon_\theta(x_{t-1}|x_t, \emptyset))$$

at inference.

4.1 Data

The model was trained on the ROCStories dataset which contains 100k 5-sentence common-sense short stories. The individual topics logically follow everyday topics.
To obtain labelled data the ROCStories dataset was classified with Twitter-roBERTa-base for Emotion Recognition, a 125M parameter transformer based model that was trained on ~58M tweets and fine tuned for emotion recognition (Barbieri et al., 2020).
Furthermore the dataset was split into individual sentences and again classified to obtain more narrowly defined label scopes for training the emotion recognition of the model. Both generated datasets seem less than ideal for emotion recognition, more work should be carried out on better suited training data like tweets or dialogues.

4.2 Training and Hyperparameters

The joint objectives of the conditional denoising process and the unconditional denoising process were implemented with a 0.15 dropout probability. For masking the conditions a Bernoulli distribution was applied to the emotion probability density input tensor. The model was trained with the end to end training objective in order to learn the word embeddings, which according to (Li et al., 2022) has been superior to random embeddings on the ROCStories. The noise schedule has significant influence on diffusion models as they share the parameters for the diffusion steps, the sqrt noise schedule $\bar{\alpha}_t = 1 - \sqrt{t/\mathrm{T} + c}$ (where c is a constant representing initial noise) proposed by Li et al. (2022) for text generation. The model was trained for 400.000 steps on the ROCStories Dataset. With a single NVIDIA GeForce RTX 3090 training time was about 22 hours. Further training on the split ROCStories dataset is ongoing.

4.3 Classifier-free guidance

The generation of text with classifier-free guidance according to the emotion probability density targets was done over 200 denoising steps with the DDPM sampling algorithm. For the last 20 steps guidance was masked out as this has shown to reduce errors when mapping the continuous representation to discrete space again, leading to less unknown tokens in the results. This setting also produced more coherent text in general. In order to reduce inference pases, instead of making one forward pass to derive the conditional epsilon

$\epsilon_\theta(x_{t-1}|x_t, y)$

and a second pass to derive the unconditional epsilon

$\epsilon_\theta(x_{t-1}|x_t, \emptyset)$,

both passes are done in parallel by halving the effective batch size. The first half of the batch size is given the target condition whereas the second half receives a copy of the first halfs input but with masked target conditions. The transitory state update is then performed by splitting the transitory prediction and after the calculation of $\hat{\epsilon}_\theta(x_{t-1}|x_t, y)$ duplicating the result again. As there are no conditional updates of the transitory state for the last 20 steps, the second half of the final output is simply a duplicate of the first half and is thus discarded.

5 Results

The outputs of the model were classified with Twitter-roBERTa-base for Emotion Recognition in order to obtain emotion recognition labels. The mean squared error and cross entropy loss between the target labels and the output labels could then be calculated. A batch of 1024 target labels was randomly picked from the set of ≈465k emotion labels derived from the split ROCStory data. The distribution of target and output labels is shown in (Table 1)

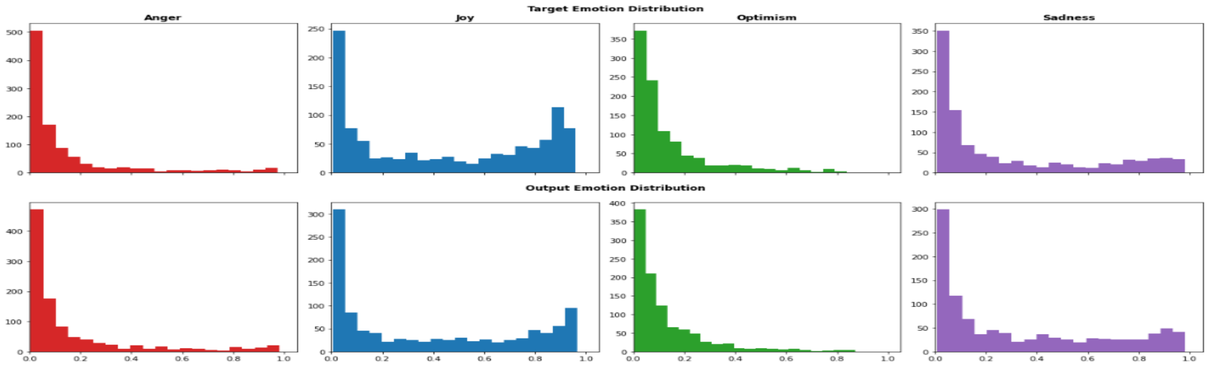

Table 1: Distribution of 1024 random target labels and the corresponding output emotion label values with $s = 1.2$

This batch was run with multiple scaling factors $s[1, 1.2, 1.5, 2, 3]$

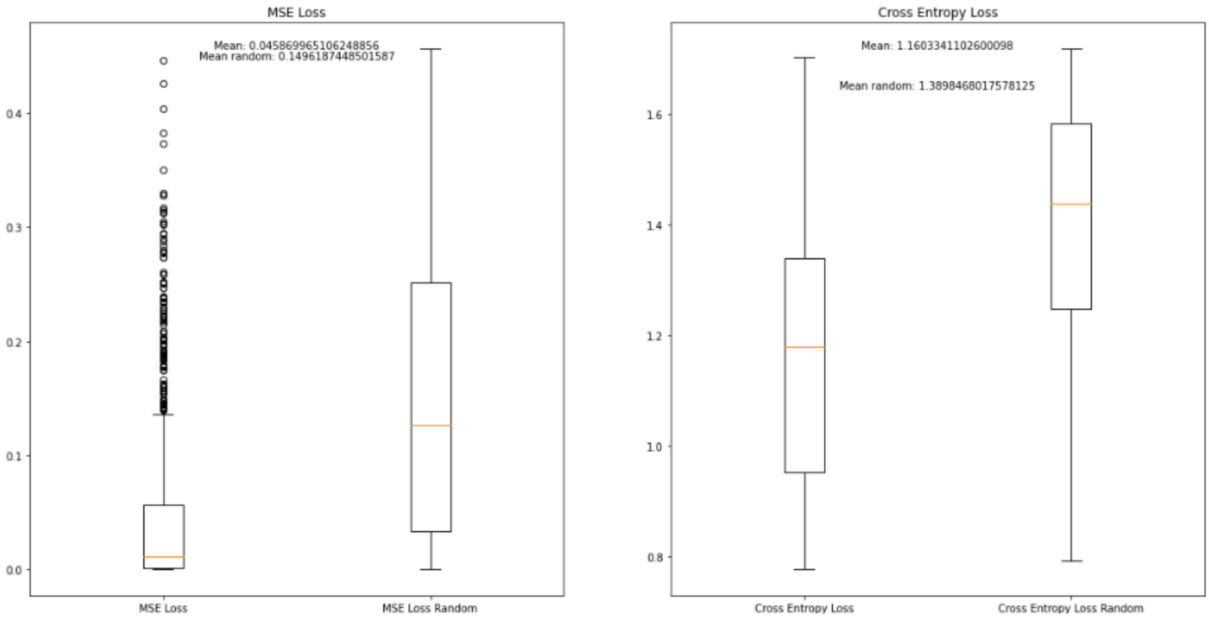

Table 2: Loss at $s = 1$, effectively being simple class conditioning

Sample output text with $s = 1$ and emotion['anger' 0.0511, 'joy' 0.7454, 'optimism' 0.1236, 'sadness' 0.0797]:

"John had left his shirt out of the day He had never wanted frozen lunch Gradually it was running back on Suddenly , he looked into the window to see he could hold a different one So he bought a white UNK "

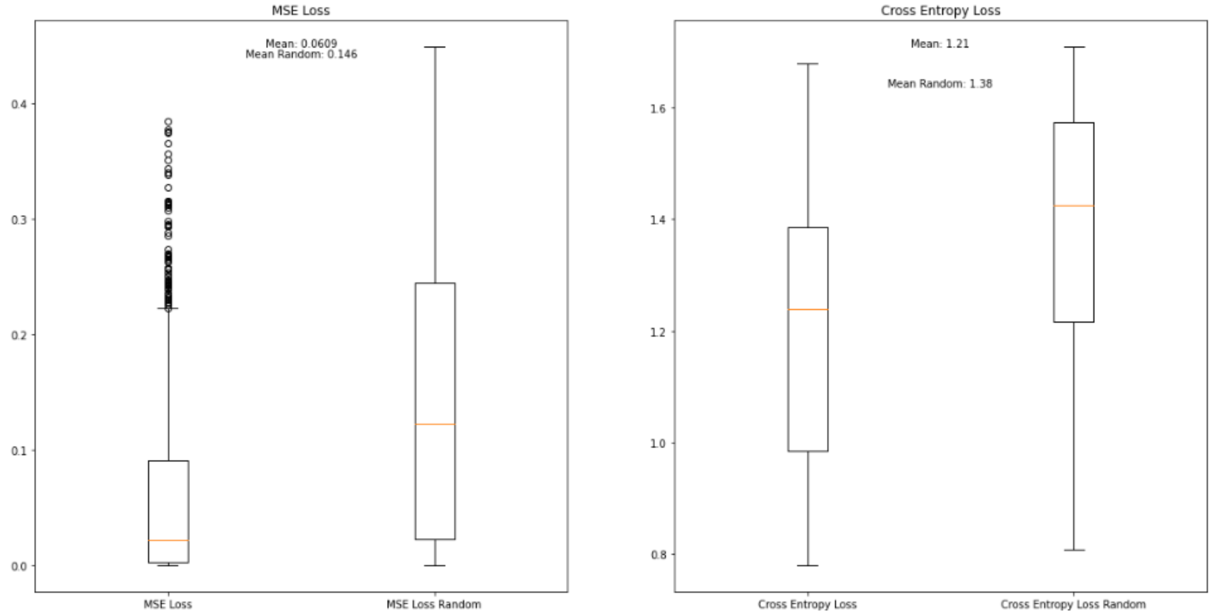

Table 3: Loss at $s = 1.2$

Sample text with $s = 1.2$ and the same emotion target:

"John had left his UNK out of the store He had never had one before The store was running over UNK Suddenly , he stopped into the kitchen When he bought twenty food he opened the store to buy a new lamp "

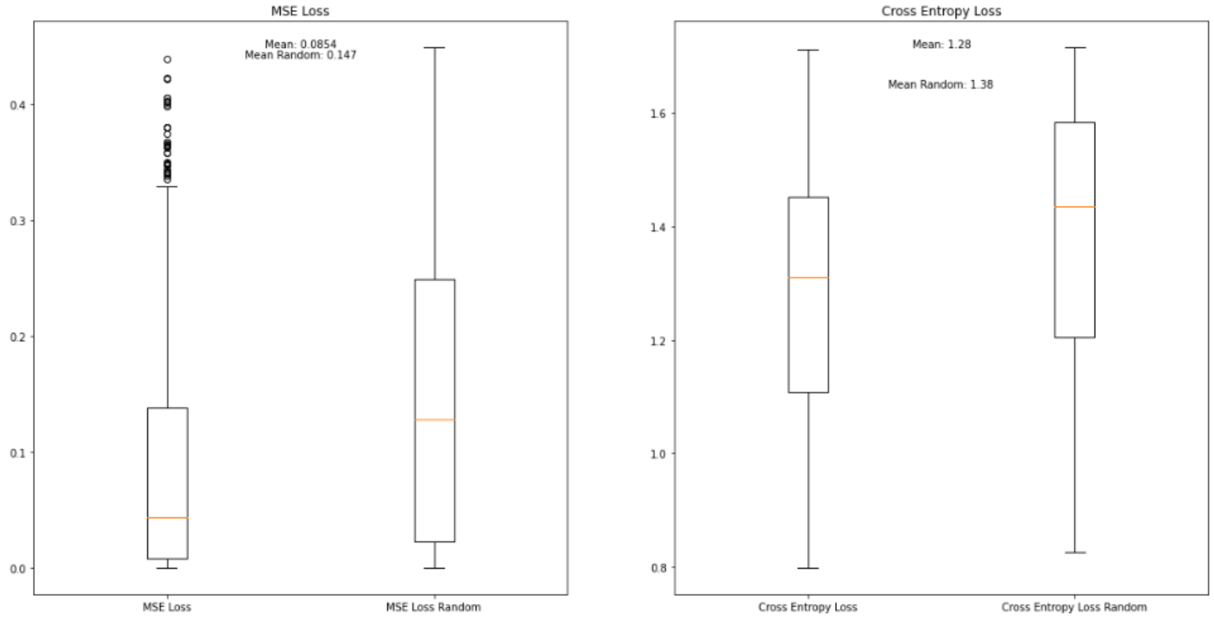

Table 4: Loss at $s = 1.5$

Sample text with $s = 1.5$ and the same emotion target:

"John had left flavor tonight. This Earl that Jim coffee could solve , frozen lunch in 2013 , bone offered him on. So , Jim searched for the dog , last , and John a different visit. John UNK too for him he hard , which became a him. Andy said. '"

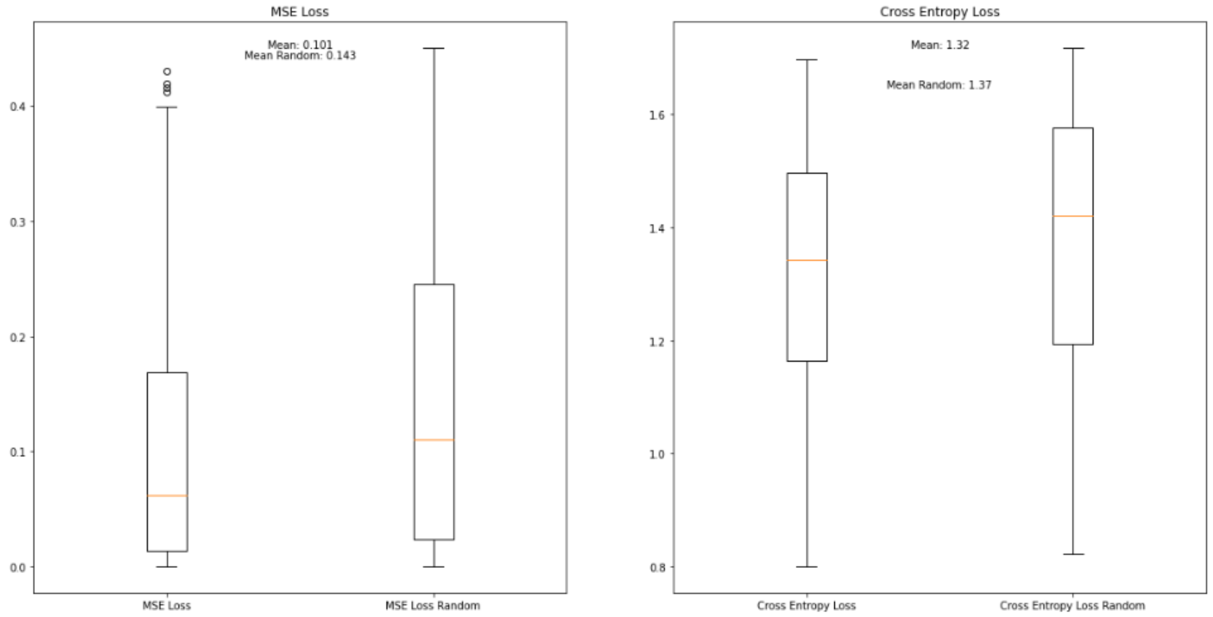

Table 5: Loss at $s = 2$

Sample text with $s = 2$ and the same emotion target:

"John taught school , tonight. Christmas career that Jim to speak mostly , large parties around sleep rolls , built him Ireland. While play Jim searched for the string , last , and Bo He reprimanded arrow in UNK UNK too studies public Bob hard , Eve poetry studies.. Andy said the publisher gave his string not Dawn '"

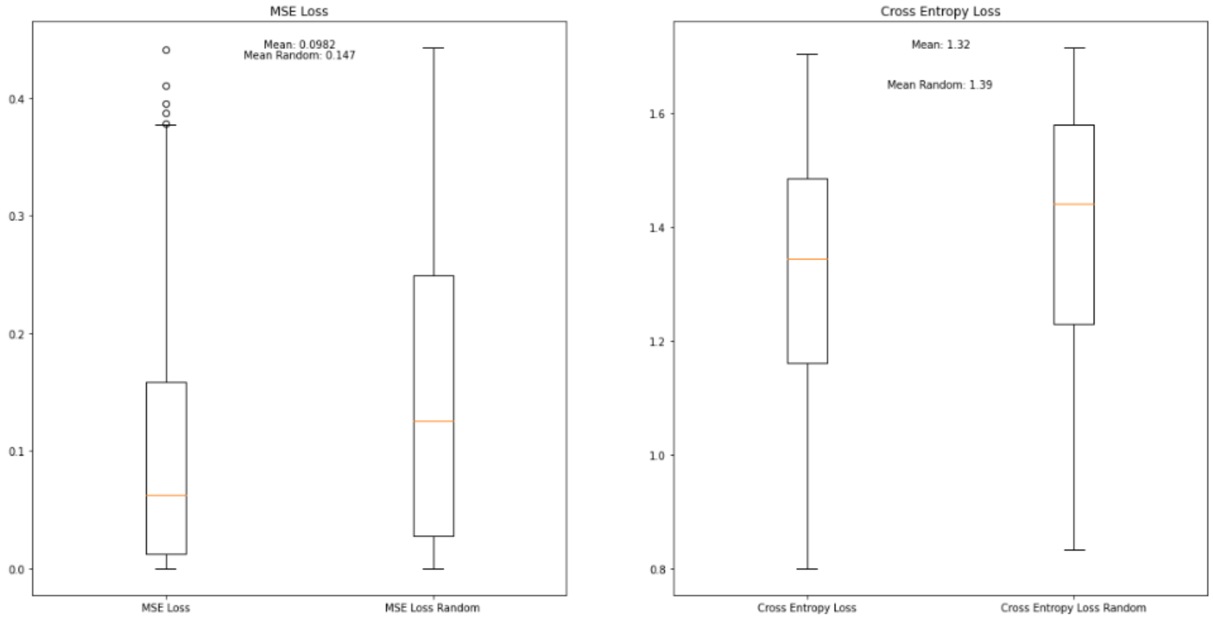

Table 6: Loss at $s = 3$

Sample text with $s = 3$ and the same emotion target:

"Bob married else , , attractive Earl asked most two - snowing Ronnie Marshall the paper board for bone built away on While play Sadie Being the Carrie base , if , and Thomas suffered confronted arrow around UNK UNK too found off Bob could one poetry followed And Charlie assured the excuse UNK bought picked by Sean was"

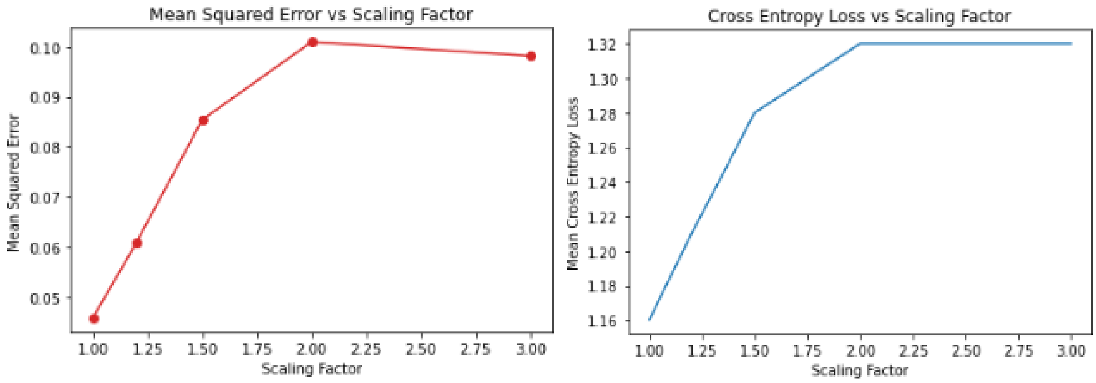

Table 6: Mean MSE and CE with respect to s

The scaling factor s has a substantial influence on the output quality, showing quickly rising losses at higher values and correspondingly deteriorating output coherence. With too high a scaling factor, the model is not able to satisfy the two objectives simultaneously, and returns to outputting gibberish which according to the conditional denoising has a high probability to fit the output label criteria. Interestingly, it fails at this too, which might be due to the fact that Twitter-roBERTa-base for Emotion Recognition is the larger model and was trained on several magnitudes more text.

6 Conclusion, Limitations and Outlook

Implementing classifier-free guidance for text generation does in principle seem to be an option to control output of a diffusion language model without having to rely on an external classifier. The results are however limited in their significance as not enough testing has been carried out.

The strong susceptibility of the model to changes of the scaling factor s can also be seen in image diffusion models like glid-3 (glid-3). If the scaling factor is picked too high, the update on the transitory state can be interpreted as moving from the unconditional transitory state in direction of the conditional transitory state but overshooting this so far that the realm of sensible representations is quickly left behind.

Here the model seems to react even stronger to changes than image diffusion models do. This might be connected to applying a diffusion model to discrete space data. It suggests that more research is needed on the method of transitory state updates. Maybe an approach of combining the transitory update direction with a direction matching some nearest neighbours might improve sample quality.

It could also indicate that the continuous space of the language data embeddings still exhibits very different properties compared to continuous space data like images. It might prove insightful to look for other examples of transitions from discrete space to continuous space and vice versa.

The models susceptibility to changes of the scaling factor might however also be related to the type and location of the target embedding, simply combining this with the timestep embedding might be a questionable choice. Testing different types and locations of these target embeddings is needed to shed light to the specifiies of language diffusion models in regard to classifier-free guidance.

One approach that was at least rudimentarily tested, is to employ a schedule for the scaling factor. Preliminary tests suggest that reducing or eliminating s at lower timesteps improves results. Either a simple stepped schedule or an asymptotic function should be further investigated. This finding sparks the question if the same effect might be seen in other domains like image or audio generation.

Diffusion models are an exciting addition to the pantheon of model architectures and might prove to be one of the influential steps, like CNN, Autoencoder and Transformer architectures before them.

Having already raised the bar on several tasks, the potential of this architecture has been shown. With diffusion models still being a new development and research still picking up, further improvements are likely to be seen. Diffusion models have some striking advantage compared to other generative model architectures. They seem robust in training, although –

as with all architectures – attention must be paid to hyperparameters, especially with diffusion language models these seem to be less robust to adjustments than in other domains. Further timestep resampling and stochastic processes on noise might still yield significant returns. Contrary to architectures that have matured through the amounts of research done on them, gains could still be significant. A further advantage is that diffusion models exhibit excellent mode coverage and are not susceptible to mode collapse. The strength of mode coverage might be explained as a consequence of the type of transient state update. Regardless of the source for the update, external classifier or internal guidance, this can to a certain extent be interpreted as a mode extender. More pronounced in the case of classifier-free guidance. High scaling factors leading to a similar effect as overexposed photographs might be interpreted as extending the mode space beyond the actual maximum, just as more light hits the film or sensor than can be mapped to the medium, leading to results outside the actual spectrum.

One area of research might be the comparison of regulatory or divergence effects of classifier guidance versus classifier-free guidance. Intuitively the classifier should be superior in pulling the transitory states back into the actual spectrum, acting as an attractor toward full mode, effectively bounding the model prediction. The central question here is if similar effects can be achieved without relying on an external classifier.

Diffusion models however have the disadvantage of the employed denoising process being iterative, leading to significantly more compute at inference, which is critical as the importance of efficiency at inference generally far outweighs the training efficiency aspect. Although work has been carried out and is ongoing to reduce the denoising steps needed, there will be a limit to this reduction given the markovian nature of the diffusion model approach.

Another drawback is that each new control task leads to either a completely new model, with potentially adjusted architecture, or at least to retraining. The author could not find any research on the question if the previous model could be applied like a pretrained model after the addition of a control task, congruent to fine tuning pretrained models in other applications.

An approach that combines classifier-free guidance with classifier guidance could be possible and feasible in areas with many control tasks. In such an ensemble several classifier-free diffusion models could act on each other as classifiers, updating the transitory state according to the combined predictions and then synchronising this transitory state update across the diffusion models. Compared to classifier guidance this could potentially reduce the ensemble complexity and compute at inference significantly and simultaneously retain at least some degree of a plug and play architecture as offered by the classifier guidance approach.

7 References

(Radford et al., 2019)

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. https://openai.com/blog/better-language-models/, 2019, accessed 17 Aug 2022.

(Zhang et al., 2022)

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

(Li et al., 2022)

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-LM improves controllable text generation. ArXiv preprint arXiv:2205.14217.

(Hoogenboom et al., 2021)

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. arXiv preprint arXiv:2102.05379, 2021.

(Hoogenboom et al., 2022)

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg and Tim Salimans. Autoregressive diffusion models. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=Lm8T39vLDTE, accessed 16 Aug 2022.

(Austin et al., 2021)

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, 2021. URL https://openreview.net/forum?id=h7-XixPCAL, accessed 16 Aug 2022.

(Ho, Jain et Abbel, 2020)

Ho, J., Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*,*33*, pp.6840-6851.

(Ho, Salimans, 2021)

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI, accessed 16 Aug 2022.

(Sohl-Dickstein et al,. 2015)

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265, 2015.

(glid-3)

https://github.com/Jack000/glid-3, accessed 17 Aug 2022

(Barbieri et al., 2020)

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1644–1650, Online, November. Association for Computational Linguistics.

(Nichol et al., 2021)

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. and Chen, M., 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

(Marcus et al., 2022)

Marcus, G., Davis, E. and Aaronson, S., 2022. A very preliminary analysis of DALL-E 2. *arXiv preprint arXiv:2204.13807*.

(Dhariwal et al., 2021)

Dhariwal, P. and Nichol, A., 2021. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34, pp.8780-8794.