Improved Depth Estimation of Bayesian Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper proposes improvements over earlier work by Nazareth and Blei [1] for estimating the depth of Bayesian neural networks. Here, we propose a discrete truncated normal distribution over the network depth to independently learn its mean and variance. Posterior distributions are inferred by minimizing the variational free energy, which balances the model complexity and accuracy. Our method improves test accuracy in the spiral data set and reduces the variance in posterior depth estimates.

1 Introduction

Determining the optimal neural network architecture for a given problem is a challenging task, typically involving manual design iterations or automated grid searches. Both approaches are time-consuming and resource-intensive. A critical aspect of this process is balancing the model's complexity to prevent overfitting while ensuring high accuracy.

The seminal work of Nazareth and Blei [1] introduced a variational inference scheme to network depth estimation. By treating the layer depth of the model as a latent variable, they can infer its posterior distribution. Importantly, their variational free energy provided an excellent objective for balancing the model complexity against the model accuracy.

Although the approach presented in [1] offers a refreshing perspective, some areas could be improved. For instance, using a truncated Poisson distribution for layer depth results in the mean and variance being approximately equal, which can lead to significant uncertainty in determining the appropriate number of layers, especially for networks of increasing depth and complexity. Moreover, although the methodology in [1] is based on variational principles, certain simplifying assumptions undermine the probabilistic nature of their model. Specifically, the first-order linearization approximation over expectations neglects uncertainties over the parameters.

This paper focuses exclusively on Bayesian neural networks and builds on the work by [1], addressing the aforementioned areas of improvement. Specifically, we make the following contributions:

- We propose a discrete truncated normal distribution over the number of hidden layers of a Bayesian neural network, enabling variance reduction in the posterior estimates of the appropriate number of layers;
- Parameter estimation and structure learning are jointly performed by minimization of the variational free energy, explicitly taking the uncertainties over variables into account.

In Section 2 the probabilistic model is specified, after which the inference procedure is elaborated in Section 3. Section 4 discusses the results obtained, and Section 5 concludes the paper.

Submitted to Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (BDU at NeurIPS 2024). Do not distribute.

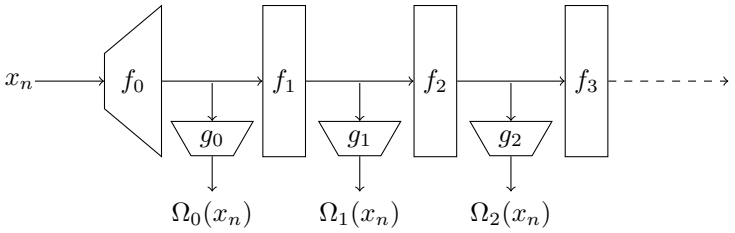

Figure 1: Visualization of the non-linearity Ω_L in (2). Deeper models reuse parts of shallower models.

2 Model specification

Let $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ be a dataset of N labeled observations. We define the likelihood function of a Bayesian neural network as

$$p(y_n \mid x_n, \theta, L) = \mathcal{N}(y_n \mid \Omega_L(x_n), \Sigma), \tag{1a}$$
$$p(y_n \mid x_n, \theta, L) = \mathrm{Cat}(y_n \mid \sigma(\Omega_L(x_n))), \tag{1b}$$

for regression and classification, respectively. $\mathcal{N}(\cdot \mid \mu, \Sigma)$ represents a normal distribution with mean μ and covariance Σ and $\mathrm{Cat}(\cdot \mid p)$ is a categorical distribution with event probabilities p, with $\sigma(\cdot)$ denoting the softmax function. The underlying non-linearity Ω_L is parameterized by parameters θ, is visualized in Figure 1 and is defined as the composition

$$\Omega_L = g_L \circ f_L \circ f_{L-1} \circ \cdots \circ f_1 \circ f_0, \tag{2}$$

with input transformation f_0, latent transformations $\{f_l\}_{l=1}^L$ and output transformations $\{g_l\}_{l=0}^L$.

We treat the model depth $L \in \mathbb{N}_0$ as an unknown variable. Therefore, a suitable discrete prior must be selected, with limited support and enabling efficient inference. The truncated Poisson distribution proposed in [1] has a variance and support that grows in network depth, preventing it from converging to a single value for the depth.

Alternative discrete distributions suffer from similar problems, such as the negative binomial distribution whose variance is always larger or equal to its mean. Others do not have continuous parameters, such as the hypergeometric distribution with integer parameters. For the categorical distributions used in [2], the support needs to be bounded. The generalized Poisson distribution [3] enables situations where its mean exceeds its variance, however, in those situations the distribution quickly becomes ill-defined [4]. Furthermore, the Conway-Maxwell-Poisson distribution does not require closed-form expressions for its normalization constant [5, 6].

Here, we propose to use a discrete truncated normal distribution, whose mean and variance are decoupled, which enables us to model both over- and under-dispersed distributions. Let $\mathcal{N}_{\geq 0}(x \mid \mu, \sigma^2) \triangleq \mathcal{N}(x \mid \mu, \sigma^2)\mathbb{1}[x \mid x \geq 0]$ denote a normal distribution truncated to the positive real line. Based on this truncated normal distribution, we define the prior over L as its discrete counterpart

$$p(L) = \int_L^{L+1} \mathcal{N}_{\geq 0}(l \mid \mu_L, \sigma_L^2)\, \mathrm{d}l \qquad \text{for } L \in \mathbb{N}_0. \tag{3}$$

We intentionally do not choose a discrete Gamma distribution [7] here, despite its positive domain, because computing derivatives to the shape parameter after truncation is difficult due to the presence of the lower incomplete gamma function in its cumulative density function.

To complete the model specification, the prior over the parameters is chosen to fully factorize as

$$p(\theta \mid L) = \prod_{\vartheta_{g_L} \in \theta_{g_L}} \mathcal{N}(\vartheta_{g_L} \mid \mu_\vartheta, \sigma_\vartheta^2) \prod_{l=0}^{L} \prod_{\vartheta_{f_l} \in \theta_{f_l}} \mathcal{N}(\vartheta_{f_l} \mid \mu_\vartheta, \sigma_\vartheta^2), \tag{4}$$

where an explicit distinction in made between the parameters in the input and hidden layers $\{\theta_{f_l}\}_{l=0}^L$, which are shared amongst different model depths, and in the depth-specific output layers $\{\theta_{g_l}\}_{l=0}^L$.

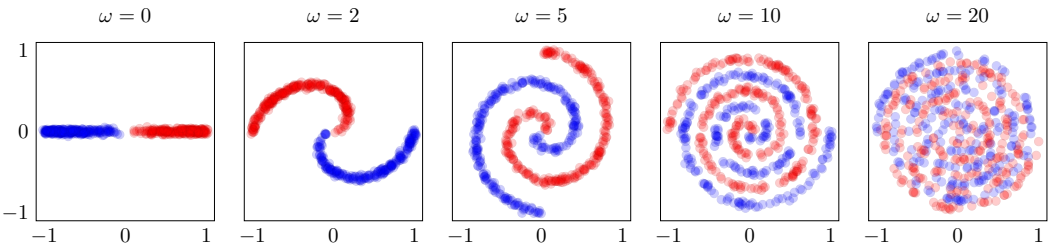

Figure 2: Spiral datasets for different rotation speeds ω, generated according to Appendix A.1.

With the model specified, the next step involves specifying the variational posterior distribution. We factorize the variational posterior distribution as

$$q(\theta, L) = q(L) \prod_{\vartheta_{g_L} \in \theta_{g_L}} \mathcal{N}(\vartheta_{g_L} \,|\, \hat{\mu}_\vartheta, \hat{\sigma}^2_\vartheta) \prod_{l=0}^{L} \prod_{\vartheta_{f_l} \in \theta_{f_l}} \mathcal{N}(\vartheta_{f_l} \,|\, \hat{\mu}_\vartheta, \hat{\sigma}^2_\vartheta). \tag{5}$$

To retain tractability, we further truncate the variational posterior distribution over L to its lower and upper quantiles defined by p_l, p_u to ensure a limited support by defining

$$\mathcal{N}^{[p_l,p_u]}_{\geq 0}(x \,|\, \mu, \sigma^2) \triangleq \propto \mathcal{N}_{\geq 0}(x \,|\, \mu, \sigma^2) \mathbb{1}\left[x \,\middle|\, p_l \leq \int_0^x \mathcal{N}_{\geq 0}(z \,|\, \mu, \sigma^2)\,\mathrm{d}z \leq p_u\right]. \tag{6}$$

Using this expression the variational posterior distribution over the network depth is formulated as

$$q(L) = \int_L^{L+1} \mathcal{N}^{[p_l,p_u]}_{\geq 0}(l \,|\, \hat{\mu}_L, \hat{\sigma}^2_L)\,\mathrm{d}l \qquad \text{for } L \in \mathbb{N}_0. \tag{7}$$

where the $\hat{\cdot}$ accent identifies the variational parameters in (6) and (7).

3 Probabilistic inference

Estimation of the variational posterior distributions, which encompasses both parameter estimation and structure learning, is achieved by minimization of the variational free energy

$$\begin{aligned} \mathrm{F}[p, q] &= \mathbb{E}_{q(L,\theta)}\left[\ln \frac{p(y, \theta, L \,|\, x)}{q(\theta, L)}\right], \\ &= \mathbb{E}_{q(L)}\left[\ln \frac{q(L)}{p(L)} + \mathbb{E}_{q(\theta \,|\, L)}\left[\ln \frac{q(\theta \,|\, L)}{p(\theta \,|\, L)} + \sum_{n=1}^{N} \ln p(y_n \,|\, x_n, \theta, L)\right]\right], \end{aligned} \tag{8}$$

where the expectation over parameters can be further decomposed as

$$\mathbb{E}_{q(\theta \,|\, L)}\left[\ln \frac{q(\theta \,|\, L)}{p(\theta \,|\, L)}\right] = \sum_{\vartheta_{g_L} \in \theta_{g_L}} \mathrm{KL}\left[q(\vartheta_{g_L}) \| p(\vartheta_{g_L})\right] + \sum_{l=0}^{L} \sum_{\vartheta_{f_l} \in \theta_{f_l}} \mathrm{KL}\left[q(\vartheta f_l) \| p(\vartheta_{f_l})\right]. \tag{9}$$

Although the expectation over the network depth seems computationally involved, the limited support as a result of the truncation in (7) reduces this operation to a finite summation as $\mathbb{E}_{q(L)}[f(\cdot \,|\, L)] = \sum_{l \in \mathrm{supp}\{q(L)\}} q(l) f(\cdot \,|\, l)$. Furthermore, since hidden layers are reused in networks of varying depth as illustrated in Figure 1, most computations can be reused in computing the expected log-evidence.

4 Experiments

All experiments[1] have been implemented in Julia [8] to explore its excellent metaprogramming capabilities as required by the dynamic nature of the unbounded models. We closely follow the

[1] All experiments are anonymously available at `https://anonymous.4open.science/r/DepthEstimationBNN-NeurIPS`.

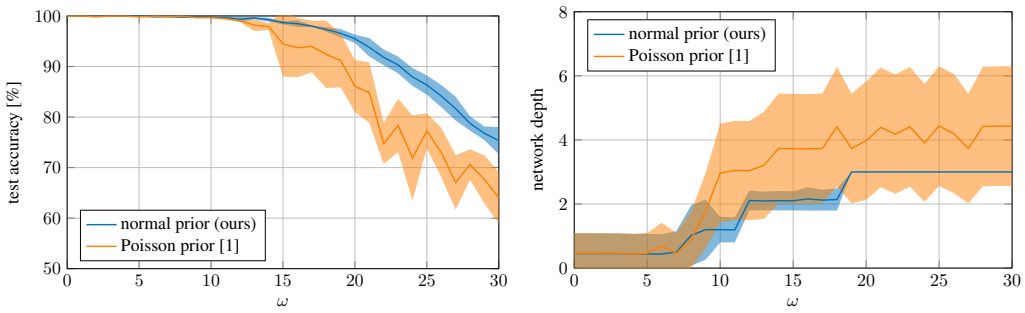

Figure 3: (Left) Test accuracy on the spiral classification task for varying rotation speeds ω. Solid lines represent the average accuracy over five independent runs, with shaded areas indicating one standard deviation ($\pm\sigma$). The discrete truncated normal distribution shows accuracy improvements across all rotational speeds compared to the Poisson-based model in [1]. (Right) Means and standard deviations of the posterior distributions over network depth, shown for the first run, with similar trends across other runs. As expected, the variance of the Poisson-based model increases at larger depths, while the normal distribution converges to a single depth.

experimental design of [1] and generate a train, validation and test set of 1024 samples each of the spiral dataset [1, 9] for binary classification as described in Appendix A.1. This dataset is parameterized by a rotation speed ω, which captures the difficulty of the dataset as shown in Figure 2.

The input layer $f_0 : \mathbb{R}^2 \to \mathbb{R}^{32}$ and latent layers $f_l : \mathbb{R}^{32} \to \mathbb{R}^{32}\, \forall\, l \geq 1$ each consist of a linear transformation followed by a LeakyReLU [10]. The output layers only involve a linear transformation $g_l : \mathbb{R}^{32} \to \mathbb{R}^2\, \forall\, l \geq 0$, where the non-linearity appears in (1b). We compare our approach to [1] which uses a $\mathrm{Poisson}(0.5)$ prior, where the variational posterior distribution is initialized by the $\mathrm{Poisson}(1.0)$ distribution, truncated to the 0.95-quantile. We select a similarly shaped normal distribution ($\mu_L = 0, \hat{\mu}_L = 0, \sigma_L = 1.15$ and $\hat{\sigma}_L = 1.8$), whose truncation is defined by $p_l = 0.025$ and $p_u = 0.975$. Appendix A.2 shows the resemblance between these priors.

We jointly learn the parameters of the probabilistic model and its variational posterior through stochastic variational inference [11] by minimizing the variational free energy in (8) using the Adam optimizer [12] until convergence. Appendix A.3 specifies the hyperparameter settings. Inference in the model is performed using Bayes-by-backprop [13] with local reparameterization [14]. The model that achieves the lowest variational free energy on the validation set is saved and evaluated on the test set by forming predictions according to

$$p(y^\star \mid x^\star) \approx \mathbb{E}_{q(\theta, L)} \left[p(y^\star \mid x^\star, \theta, L) \right]. \tag{10}$$

Figure 3 shows the achieved predictive accuracy on the test set and the inferred posterior distributions over the model depth. From this we conclude that the discrete truncated normal distribution outperforms the Poisson distribution on the spiral classification task. The normal-based model achieves a higher accuracy, which becomes increasingly significant when the complexity of the data increases. Furthermore, as expected, the posterior distribution over the model depth in the normal-based model has a reduced variance in comparison to the Poisson-based model, as its mean and variance are naturally decoupled during training. In practice this leads to computational savings when making predictions using (10) as the narrow support of $q(L)$ requires less output layers g_l to be active.

5 Discussion and conclusion

This paper introduces a discrete truncated normal distribution for modeling the depth of a Bayesian neural network and demonstrates how to infer its posterior distribution through the minimization of variational free energy. Compared to methods using a Poisson prior [1], our approach results in reduced variance in posterior estimates and improved test accuracy on the spiral classification task.

The results presented in this paper show promising improvements in estimating the depth of Bayesian neural networks. However, additional experiments are required involving more complex models and tasks. Network width estimation and parameter pruning [13, 15–17] offer valuable opportunities for further expanding the methodology presented here.

References

[1] Achille Nazaret and David Blei. "Variational Inference for Infinitely Deep Neural Networks". In: *Proceedings of the 39th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, 2022, pp. 16447–16461.

[2] Javier Antoran, James Allingham, and José Miguel Hernández-Lobato. "Depth Uncertainty in Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 10620–10634.

[3] P. C. Consul and G. C. Jain. "A Generalization of the Poisson Distribution". In: *Technometrics* 15.4 (1973). Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality], pp. 791–799. ISSN: 0040-1706.

[4] P. C. Consul and M. M. Shoukri. "The generalized poisson distribution when the sample mean is larger than the sample variance". In: *Communications in Statistics - Simulation and Computation* 14.3 (1985), pp. 667–681. ISSN: 0361-0918, 1532-4141.

[5] Peter Boatwright, Sharad Borle, and Joseph B. Kadane. "A Model of the Joint Distribution of Purchase Quantity and Timing". In: *Journal of the American Statistical Association* 98.463 (2003). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 564–572. ISSN: 0162-1459.

[6] Peter Boatwright et al. "Conjugate analysis of the Conway-Maxwell-Poisson distribution". In: *Bayesian Analysis* 1.2 (2006). Publisher: International Society for Bayesian Analysis, pp. 363–374. ISSN: 1936-0975, 1931-6690.

[7] Subrata Chakraborty and Dhrubajyoti Chakravarty. "Discrete Gamma Distributions: Properties and Parameter Estimations". In: *Communications in Statistics - Theory and Methods* 41.18 (2012), pp. 3301–3324. ISSN: 0361-0926, 1532-415X.

[8] Jeff Bezanson et al. "Julia: A Fresh Approach to Numerical Computing". In: *SIAM Review* 59.1 (2017). Publisher: Society for Industrial and Applied Mathematics, pp. 65–98. ISSN: 0036-1445.

[9] Kevin Lang and Michael Witbrock. "Learning to Tell Two Spirals Apart". In: *Proceedings of the 1988 Connectionist Models Summer School*. Connectionist Models Summer School. Pittsburgh, Pensylvania, USA: Morgan Kauffman Publishers, 1988.

[10] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: *Proceedings of the 30th International Conference on Machine Learning*. International Conference on Machine Learning. Atlanta, Georgia, USA, 2013.

[11] Matthew D. Hoffman et al. "Stochastic Variational Inference". In: *Journal of Machine Learning Research* 14.4 (2013), pp. 1303–1347. ISSN: 1533-7928.

[12] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR*. 3rd International Conference on Learning Representations, ICLR. San Diego, CA, USA, 2015.

[13] Charles Blundell et al. "Weight uncertainty in neural networks". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. Vol. 37. ICML'15. Lille, France: JMLR.org, 2015, pp. 1613–1622.

[14] Diederik P. Kingma, Tim Salimans, and Max Welling. "Variational Dropout and the Local Reparameterization Trick". In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015.

[15] Alex Graves. "Practical Variational Inference for Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc., 2011.

[16] Eric Thomas Nalisnick. "On Priors for Bayesian Neural Networks". PhD thesis. UC Irvine, 2018.

[17] Jim Beckers et al. "Principled Pruning of Bayesian Neural Networks Through Variational Free Energy Minimization". In: *IEEE Open Journal of Signal Processing* 5 (2024), pp. 195–203. ISSN: 2644-1322.

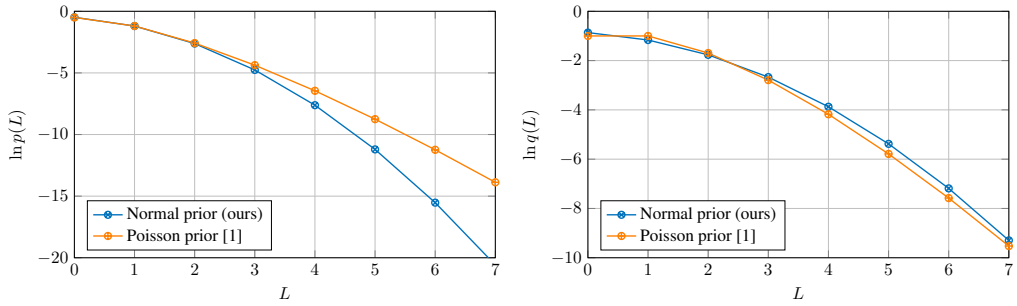

Figure 4: Log probability mass function of the (left) prior distribution over the model depth used in [1] and of the discrete truncated normal distribution used in this paper; and of the (right) initial variational posterior distributions over the model depth.

A Experimental details

This appendix outlines the implementation details corresponding to experiments in Section 4.

A.1 Data generation

The spiral dataset used in the experiments of Section 4 and visualized in Figure 2 are generated according to the following sampling procedure[2]:

$$t_n \sim \text{Uniform}([0, 1]) \tag{11a}$$

$$u_n = \sqrt{t_n} \tag{11b}$$

$$y_n \sim \text{Uniform}(\{-1, 1\}) \tag{11c}$$

$$x_n \sim \mathcal{N}\left(\begin{bmatrix} y_n u_n \cos\left(\omega u_n \frac{\pi}{2}\right) \\ y_n u_n \sin\left(\omega u_n \frac{\pi}{2}\right) \end{bmatrix}, 4 \cdot 10^{-4} \mathrm{I}_2\right) \tag{11d}$$

A.2 Prior selection

For selecting the prior and initial variational posterior distributions in the experiments of Section 4, we manually align the discrete truncated distribution with the Poisson distributions. Figure 4 shows a comparison of the log probability mass function of both functions as comparisons. Most important are the segments with a high log-probability, where the priors align relatively well from visual inspection. It should be noted that some discrepancies are inevitable, but at the same time negligible as these distributions only serve as a starting point and can be optimized over.

A.3 Training procedure

Below we describe the training procedure. Here we tried to stay as close to the experimental design of [1] as possible.

For each run we set a random seed equal to the run index. We then independently sample a train, validation and test set consisting of 1024 samples each for $\omega = 0, 1, \ldots, 30$ according to Appendix A.1. We use the following hyperparameters

- Prior on the model depth: $p(L) = \text{Poisson}(L \,|\, 0.5)$ or $p(L) = \int_L^{L+1} \mathcal{N}_{\geq 0}(l \,|\, 0, 1.15^2)\, \mathrm{d}l$.
- Initialization of variational posteriors: $q(L) = \text{Poisson}^{[0,0.95]}(L \,|\, 1)$ or $q(L) = \int_L^{L+1} \mathcal{N}_{\geq 0}^{[0.025,0.975]}(l \,|\, 0, 1.8^2)\, \mathrm{d}l$.
- Optimizer: Adam [12] with default hyperparameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$).

[2]The variance in the last step seems to differ from the original description in [1, Appendix B.1], however, the value reported there (0.02) refers to the standard deviation of the normal distribution, as verified with their publicly available experiments.

- Learning rate: 0.005 (0.0005 for $\hat{\mu}_L$ and $\hat{\sigma}_L^2$, and for the rate parameter of the posterior Poisson distribution $\hat{\lambda}_L$).
- Number of epochs: 20.000.
- Batch size: 256 (randomly shuffled per epoch).
- Leaky ReLU: $\max(\alpha x, x)$, where $\alpha = 0.1$.
- Reparameterization: strictly positive parameters are transformed using the $\mathrm{softplus}$ function for unconstrained optimization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction clearly specifies the problem this paper addresses and explicitly states the contributions of the paper. The abstract highlights these contributions and the results obtained from the experiments in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion states that more rigorous experiments are required. Additional experimentation is currently ongoing. Yet for the scope of this extended abstract, the authors deemed the results of the current experiments interesting enough to share. Furthermore, assumptions are clearly mentioned throughout the paper and experimental details are provided in Appendix A and in the complementary code repository. Although details regarding the computational efficiency are more clearly addressed in the motivating work of [1], we only highlight their key findings at the end of Sections 3 and 4 due to space limitations. We note here that there are little concerns regarding the computational scalability.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include any theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Great care has been taken in describing all details of the experimental design. This can be seen in 1) the length and accurateness of Section 2, 2) the most important experimental details provided in Section 4, 3) the clear overview of additional hyperparameters and data generation in Appendix A and 4) the complementary code repository. The authors highly value reproducibility and are therefore open to any form of feedback to improve upon this.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The complementary code repository allows for reproducing the results in the paper, including baseline evaluations and data generation. We have taken care to thoroughly test all code in the repository and have fixed random seeds to aid reproducibility.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and testing details are available in Section 4, Appendix A and in the public code repository. The Appendix contains all of the implementation details, such as hyperparameters, to retain readability in the main text.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experiments are performed over 5 independent runs to report the corresponding confidence intervals. The caption of Figure 3 describes that the plotted intervals concern one standard deviation error bands ($\pm\sigma$).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Computer resources are not explicitly mentioned as the experiments are executed on a personal laptop and hence do not require vast resources. The experiments also do not report execution times as these are not so relevant in the scope of the paper's contributions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The contributions presented in this paper provide tools to improve the model design cycle in general, which is task-agnostic and hence does not directly have a societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Synthetic datasets have been manually generated and experimental models do not provide a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We closely follow the experimental design of [1] which we clearly and properly credit for this. All other resources are developed by the authors themselves and do not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The code accompanying the paper is clearly documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.