Bayesian generative models can flag performance loss and bias

Anonymous Author(s)
Affiliation
Address
email

Abstract

Generative models are popular for medical imaging tasks such as anomaly detection, feature extraction, data visualization, or image generation. Since they are parameterized by deep learning models, they are often sensitive to distribution shifts and unreliable when applied to out-of-distribution data, creating a risk of, e.g. underrepresentation bias. This behavior can be flagged using uncertainty quantification methods for generative models, but their availability remains limited. We propose SLUG: A new UQ method for VAEs that combines recent advances in Laplace approximations with stochastic trace estimators to scale gracefully with image dimensionality. We show that our UQ score – unlike the VAE's encoder variances – correlates strongly with reconstruction error and racial underrepresentation bias for dermatological images.

1 Introduction

Variational Autoencoders (VAEs) stand out as one of the most widely used generative models in medical imaging because of their capacity to learn semantic, low-dimensional latent spaces. VAEs are often used to analyze and manipulate key characteristics of high-dimensional data, e.g. for data visualization [3], data generation [8], and anomaly detection [14].

Despite these advantages, generative models are parameterized with modern Deep Neural Networks (DNNs), which struggle with out-of-distribution (OOD) [13]. To tackle OOD performance in predictive models, uncertainty quantification (UQ) has emerged as an important tool [9], where the prediction is endowed with an associated uncertainty. This has proven useful for detecting silent failures and ensuring that unexpected outcomes do not occur. While UQ techniques have been proposed and studied with promising success for discriminative models [1], their applicability to generative models such as VAEs is underexplored. The adoption of a Bayesian approach can help address this issue; however, current Bayesian generative models tend to be computationally expensive, difficult to tune, or rely on uncorrelated posterior approximations [12, 2].

We propose a novel epistemic UQ method for VAEs building on the *Sketched Lanczos Uncertainty (SLU)* algorithm recently proposed for discriminative models [11]. SLU computes a rank-k approximation of the *generalized Gauss-Newton (GNN)* matrix, which captures the epistemic uncertainty according to a Laplace approximation [7]. However, SLU scales quadratically with the output dimension, which is intractable in image-generative models. Our proposed *Sketched Lanczos Uncertainty Global (SLUG)* measure overcomes this challenge using scalable stochastic trace estimators [6] to produce a per-image score.

We demonstrate SLUG's ability to detect underrepresentation bias in dermatological images. It is well known that dark skin tones are severely underrepresented in public datasets and that DNNs-based

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Medical Imaging meets EurIPS: MedEurIPS 2025. Do not distribute.

systems tend to reproduce and amplify this bias [5, 10]. Our experiments show that SLUG strongly correlates with the performance of the VAE and can serve to flag both bias and performance loss.

In short, **we contribute** a novel UQ method for VAEs that capture out-of-distribution data and demonstrate its utility in flagging errors and underrepresentation bias using two publicly available real-world dermatology datasets: Fitzpatrick17k [5], and PASSION [4].

2 Method

Given the VAE $f_{\phi,\theta}$ with encoder and decoder parameters $(\phi, \theta) \in \mathbb{R}^p$, we define its Jacobian $\mathbf{J}_{\{\phi,\theta\}}$ with respect to parameters and construct the Generalized Gauss-Newton (GGN) matrix as $\mathbf{G}_{\{\phi,\theta\}} = \sum_{i=1}^n \mathbf{J}_{\{\phi,\theta\}}(x_i)^T \mathbf{H}(x_i) \mathbf{J}_{\{\phi,\theta\}}(x_i)$, where $\mathbf{H}(x_i)$ is the Hessian of the loss w.r.t. to the neural network output and $\{x_i \in \mathbb{R}^{W \times H \times C}\}_{i=1}^n$ the training set.

The GGN commonly appears as the inverse covariance of the *linearized Laplace approximation (LLA)* to the true posterior [7]. Currently, LLA is the most promising Bayesian posterior approximation [7], but it is, unfortunately, intractable for generative models as its computational cost scales quadratically with the generated data dimension, specifically $\mathcal{O}((WHC)^2 p)$ for a network with p parameters.

Recently, Miani et al. [11] developed a sketching-based algorithm to evaluate the associated predictive uncertainty, which scales logarithmically with p. The resulting *Sketched Lanczos Uncertainty (SLU)* algorithm, however, still scales quadratically with the image dimension, making it impractical for VAEs. Our approach extends SLU to scale gracefully to large images.

2.1 Scaling to VAEs: Sketched Lanczos Uncertainty Global score (SLUG)

We based our proposed score on the SLU algorithm [11]. Let $\mathbf{U}$ denote the matrix containing the leading eigenvectors of the GGN, then the SLU approximates the predictive variance of the linearized Laplace approximation with $\mathbf{I} - \mathbf{U}\mathbf{U}^\top$ is covariance, i.e., for a specific pixel (w, h, c),

$$\text{SLU}_{w,h,c}(x) = e_{whc} \mathbf{J}_{\theta^*}(x)(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{J}_{\theta^*}(x)^\top e_{whc}^\top, \tag{1}$$

where e_{whc} is a one-hot encoding vector that selects the pixel position to compute the uncertainty.

SLU approximates this predictive uncertainty using several tricks from randomized numerical linear algebra. Unfortunately, even SLU does not scale to neural networks with high-dimensional outputs like those in generative models. Producing one predictive variance per generated pixel requires $\mathcal{O}(WHC)$ SLU invocations, which is practically prohibitive. Our main interest is in measuring a scalar uncertainty score for a generated image, and we choose the sum of per-pixel predictive variances, which we denote the *Sketched Lanczos Uncertainty Global (SLUG)* score. We use a stochastic trace estimator [6] to estimate the SLUG score,

$$\begin{aligned} \text{SLUG}(x) &= \sum_{w,h,c} \text{SLU}_{w,h,c}(x) \qquad (2) \\ &= \text{Tr}(\mathbf{J}_{\theta^*}(x)\ (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\ \mathbf{J}_{\theta^*}(x)^\top) \qquad (3) \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \epsilon_s \mathbf{J}_{\theta^*}(x)\ (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\ \mathbf{J}_{\theta^*}(x)^\top \epsilon_s^\top, \qquad (4) \end{aligned}$$

where $\epsilon_s \sim \mathcal{N}(0, \mathbf{I})$. This can be implemented using only S invocations of SLU.

3 Experiments and results

Datasets. In our experiments, we use two dermatology datasets. The VAE is trained on the **Fitzpatrick17k** dataset [5], which includes 16,577 images labeled with Fitzpatrick skin types (FST). To evaluate the impact of skin tone representation in the training data, we create three subsets of 1,668 images each: **Dataset A – Light** (100% FST 1–2), **Dataset B – Mixed** (50% FST 1–2 and 50% FST 5–6), and **Dataset C – Black** (100% FST 5–6). Two separate test sets of 512 images each are sampled for lighter and darker skin tones. For external validation and to assess model bias, we use the **PASSION** dataset [4], which contains 4,901 dermatology images from Sub-Saharan countries with darker skin tones (FST 3–6).

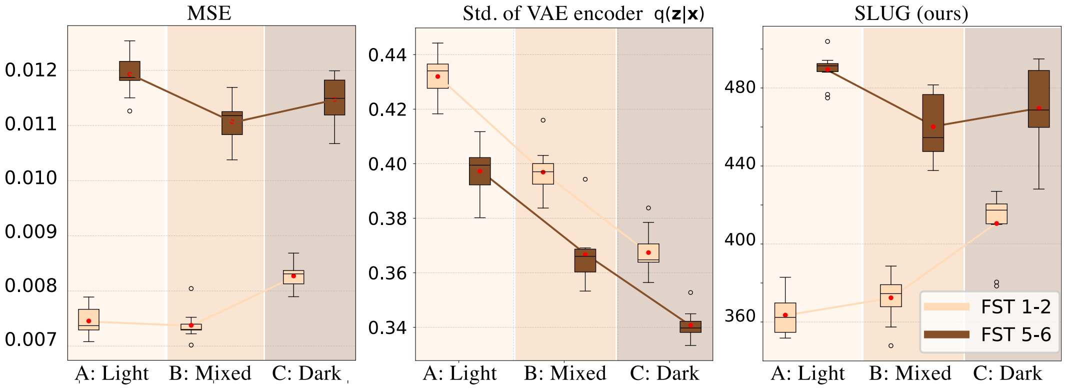

Figure 1: On Fitzpatrick17k, the performance on light and dark skin tones changes with their representation. The VAE encoder uncertainty is a poor indicator, while SLUG follows performance across groups and training scenarios.

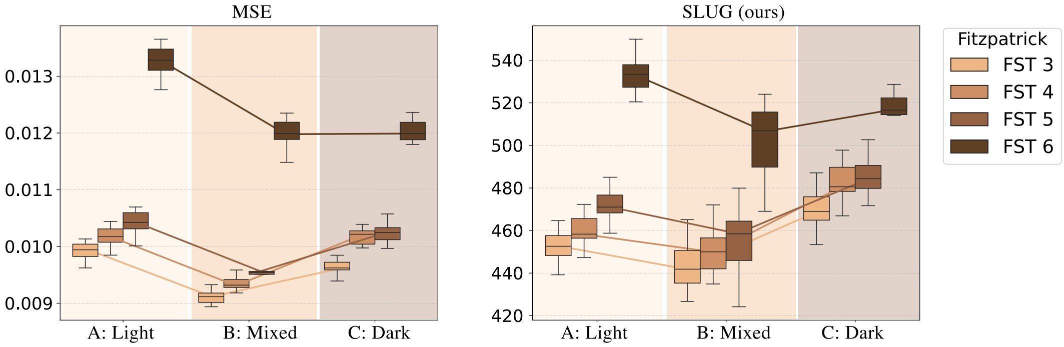

Figure 2: On the external PASSION dataset, we see again how reduced MSE is flagged by increased SLUG uncertainty across dark skin tone groups.

Racial bias. Fig. 1 shows that SLUG effectively captures racial bias in the Fitzpatrick17k dataset—unlike the VAE's latent uncertainty. We also evaluate the trained models in Fitzpatrick17k on the external PASSION dataset. The results reveal a consistent pattern: performance degrades as skin tone darkens (see Fig. 2). Notably, our SLUG score also captures this racial bias in the external dataset, highlighting its utility as a metric to flag bias.

4 Conclusion

This work highlights the urgent need for precise and scalable UQ for generative models. Despite the widespread use of generative AI, we still lack a reliable mechanism to ensure its trustworthiness. **We demonstrate that epistemic UQ can warn of performance loss, and detect bias**.

We expect that these results will motivate the recognition of epistemic uncertainty as an essential tool for generative models. While our SLUG method for VAEs can capture bias and performance loss in dermatological images, it will be valuable to see how epistemic uncertainty and bias interact in other large generative models, such as diffusion models.

5 Potential Negative Impact Statement

This paper studies the relationship between epistemic uncertainty and racial bias in dermatology using generative models, evaluated on the Fitzpatrick17k and a Sub-Saharan dataset. The analysis is limited to racial categories present in these datasets and does not consider other sources of bias, such as gender, age, or data origin. These unexamined factors may also influence uncertainty and model performance.

The models are evaluated on curated datasets and not tested under out-of-distribution or real-world clinical conditions. As a result, their robustness and reliability in practice remain unclear. Further evaluation in real-world and OOD settings is necessary to assess clinical safety.

While the method proposed may help identify racial bias, they do not prevent misuse or biased deployment. Ensuring fairness in clinical applications requires broader efforts, including careful validation, and responsible use by practitioners and developers.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. In: *Information fusion* 76 (2021), pp. 243–297.

[2] Erik Daxberger and José Miguel Hernández-Lobato. “Bayesian variational autoencoders for unsupervised out-of-distribution detection”. In: *arXiv preprint arXiv:1912.05651* (2019).

[3] Jan Ehrhardt and Matthias Wilms. “Autoencoders and variational autoencoders in medical image analysis”. In: *Biomedical Image Synthesis and Simulation*. Elsevier, 2022, pp. 129–162.

[4] Philippe Gottfrois, Fabian Gröger, Faly Herizo Andriambololoniaina, Ludovic Amruthalingam, Alvaro Gonzalez-Jimenez, Christophe Hsu, Agnes Kessy, Simone Lionetti, Daudi Mavura, Wingston Ng’ambi, et al. “PASSION for Dermatology: Bridging the Diversity Gap with Pigmented Skin Images from Sub-Saharan Africa”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 703–712.

[5] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. “Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1820–1828.

[6] Michael F Hutchinson. “A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines”. In: *Communications in Statistics-Simulation and Computation* 18.3 (1989), pp. 1059–1076.

[7] Alexander Immer, Maciej Korzepa, and Matthias Bauer. “Improving predictions of Bayesian neural nets via local linearization”. In: *International conference on artificial intelligence and statistics*. PMLR. 2021, pp. 703–711.

[8] Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. “Generating synthetic data for medical imaging”. In: *Radiology* 312.3 (2024), e232471.

[9] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. “Second opinion needed: communicating uncertainty in medical machine learning”. In: *NPJ Digital Medicine* 4.1 (2021), p. 4.

[10] Miguel López-Pérez, Søren Hauberg, and Aasa Feragen. “Are generative models fair? a study of racial bias in dermatological image generation”. In: *Scandinavian Conference on Image Analysis*. Springer. 2025, pp. 389–402.

[11] Marco Miani, Lorenzo Beretta, and Søren Hauberg. “Sketched Lanczos uncertainty score: a low-memory summary of the Fisher information”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. 2024, pp. 23123–23154.

[12] Marco Miani, Frederik Warburg, Pablo Moreno-Muñoz, Nicki Skafte, and Søren Hauberg. “Laplacian autoencoders for learning stochastic representations”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 21059–21072.

[13] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 427–436.

[14] Hansen Wijanarko, Evelyne Calista, Li-Fen Chen, and Yong-Sheng Chen. “Tri-VAE: Triplet Variational Autoencoder for Unsupervised Anomaly Detection in Brain Tumor MRI”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 3930–3939.