Dr. Eric Nalisnick
Assistant Professor
Informatics Institute
University of Amsterdam

Address:
Science Park 900
1098 XH Amsterdam
email: e.t.nalisnick@uva.nl

Dissertation Review

Title: ***Data Representations in Generative Modeling***

Author of Dissertation: Kamil Deja, MSc.

1. Setting & Motivation

Generative models are an increasing important topic in modern AI and already have consequences for broader society. Models such as GPT and Stable Diffusion have demonstrated and forced us to consider the full range of benefits and dangers presented by generative models. For example, the benefits range from new creative technologies to automated assistants capable of a wide range of tasks to help with scientific discoveries. The harms are numerous as well: the generation of false news and mass propaganda, disrupting fair evaluation in education, and the careless violation of privacy and copyright. Autoencoder- and Diffusion-based architectures represent two of the most popular architectures in modern generative models. This dissertation does important work to understand (i) the information / structure encoded in the latent representations of these models, (ii) the wider utility of the representations, and (ii) the stability of the representation / its information when the model is fine-tuned on additional or different data. Topic (iii) will be especially important in the near future as people wish to apply foundation models to an increasingly wide range of tasks by way of finetuning. The dissertation respectively terms these three topics as “structure”, “applications”, and “consolidation”. Overall, I find the dissertation timely, well-motivated, and in conversation with the research frontier in terms of both applications and theory.

2. Novelty and Impact

2.1 End-to-End Sinkhorn Autoencoder

The contribution of the End-to-End Sinkhorn Autoencoder with Noise Generator (E2E-SAE) is to essentially define two encoders, one that maps random noise into the latent space and another that deterministically maps the sample into the same latent space. This presents a clever way to separate representation learning from the prior, as representation learning can happen on the deterministic path. Moreover, when combined with the stochastic encoder, the architecture still remains a proper generative model, as the deterministic encoder should generate outputs that follow

the aggregate prior implicitly defined by the other encoder. Moreover, this prior is also updated by the Sinkhorn loss for extra flexibility.

The model is clearly useful and competitive with the state-of-the-art, as demonstrated by the experiments that range from image to calorimeter generations. Thus, the model is clearly meets the bars for impact and novelty. My primary critique of the methodology is in regards to its ability to perform representation learning. I thought this was the primary motivation for the dual encoders, as described above, but I do not see any results that analyze the method's ability to perform good representation learning.

2.2 Analyzing Generative and Denoising Capabilities of Diffusion-Based Deep Generative Models

Diffusion models have achieved impressive success in their ability to generate high-quality, realistic, even seemingly creating images. While they are foundationally well-defined and understood as a continuous diffusion process, their computational pipeline consists of 1000+ steps. Unlike precursors models like VAEs and GANs, the subcomponents of diffusions models—if they even exist—-are hard to identify, let alone determine their role in image generation. This work aims to address this open problem, understanding the role of groups of steps of the diffusion.

In what I consider to be the highlight result of the dissertation, empirical analysis is conducted to show that diffusion models have two distinct phases: a generation phase and a denoising phase. The former is comprised of the first 80-90% of steps of the model, meaning that the model uses most of its capacity to translate noise into a structured image. The remaining 10-20% of steps are used to denoise the image in a fairly input-agnostic way. This perspective is further validated by successfully replacing the denoting steps with a denoising autoencoder, showing that this modular structure is indeed present. My only comment on this work is that, while the work on Generalized DAEs by Bengio et al. [2013] is mentioned, I thought their proposed 'walk-back' training procedure might be quite related to the proposed DAED model, as it also applies several steps of noise to the input before passing it into an autoencoder.

2.3. Learning Data Representations with Join Diffusion Models

This work continues the theme of probing and understanding diffusion models, further investigating the information contained in the intermediate states. Specifically, this work considers defining a joint distribution (so called 'hybrid model') that models both the features and a label, thereby doing both unsupervised density estimation as well as classification. The authors propose a clever, U-Net-based

model that sits in the middle of the diffusion process—not at the end, like previous approaches have done—and transports the intermediate states across the levels of the 'U'. The classifier can then use the features obtained at various points of the diffusion. Impressively, this model demonstrates all of the characteristics that a hybrid model promises: pure classification and generation, semi-supervised learning, and transfer learning. Moreover, the authors do this with a sensible, natural objective, not needing the 'hacks' that previous models have needed to balance generation and classification.

2.4. BinPlay: A Binary Latent Autoencoder for Generative Replay Continual Learning

Continual learning (CL)—the ability for a model to learn from a stream of new tasks, just as humans do, yet still retain past information—is a hallmark goal of artificial intelligence. Generative replay is a popular technique for CL, allowing the model to revisit previous important datapoints so as to not forget the information they provided. The work described in this section presents an interesting and practical take on the problem by formulating a generative model with a discrete latent space. Binary codes representing past data points can be stored and then fetched and passed to the decoder to obtain the full feature vector, which then can be passed to the classifier for replay. The experiments demonstrate strong classification performance in the CL setting and also report the memory footprint, verifying the motivation to use binary codes. While beyond the scope of the work, it would have been interesting to see the work better exploit the strong structure of binary codes to, e.g. organise the latent space into semantic concepts or some other hierarchy.

2.5. Multiband VAE

In this chapter, the dissertation describes a new procedure for unsupervised CL in which a local model is first trained on the task and then a global 'translator' model aggregates the local encodings. The motivation for such an approach is to cope with catastrophic forgetting—a key problem when performing CL. In short, the translator is trained with replay in order to prevent such forgetting. The experiments demonstrate that the model is competitive-to-superior to other SOTA generative CL models (in terms of FID score). This demonstrates the utility of this sensible approach. My only critique is that I wonder if such an approach will be less sample efficient, especially for later tasks, as the model cannot use information from previous tasks to help with learning later tasks.

3. Evaluating the Written Document

Overall, the document is in the appropriate format, providing the necessary context and background information before diving into specific works in the middle chapters. The document closes with a summary and research questions for future work, which are appropriate, if a bit short-sighted. I have only two remarks for improvements: Regarding the E2E-SAE model, I found the description rather algorithmic in that it describes what the model it doing in a procedural manner. However, I think a more conceptually 'cleaner' presentation is to separately define the generative model, the inference model, and the training procedure. This makes clear when one could make alternative choices for each of these building blocks. Secondly, the switch to continual learning is somewhat abrupt. Perhaps this could be better integrated into the background on generative models. Alternatively, it could be kept in its current place but with more, earlier discussion about the connections between generative modelling and continual learning (what's currently in 7.2). Yet, in summary, I found the dissertation a pleasure to read. See the appendix for detailed comments on minor points.

4. Conclusions

The reviewed dissertation of Kamil Deja, MSc., meets the requirements for doctoral dissertations by the Act on Scientific Degrees and Academic Title of 14 March 2003 (Journal of Laws No. 65) as it presents novel, impactful concepts that push forward the discipline of Computer Science. This fact is further supported by the prestigious and selective venues in which the dissertation's material has already been published (e.g. *Neural Information Processing Systems* 2022, International Joint Conference on Artificial Intelligence 2022). Any critical remarks presented above should not detract from my incontrovertibly positive assessment. **I request that the doctoral degree be awarded to Kamil Deja, MSc, and due to the prestige and impact the work has already gathered, I recommend the degree be awarded with honors.**

Please do not hesitate to contact me at *e.t.nalisnick@uva.nl* if you have any further questions.

Sincerely,

Eric Thomas Nalisnick, PhD

Appendix: Detailed Remarks

p 13: *"Normalising flows (Rezende and Mohamed, 2015) and Glows (Kingma and Dhariwal, 2018) are explicitly trained to map original data samples into a lower-dimensional manifold through invertible operations"*: Firstly, Glow is a type of normalizing flow, and secondly, I find the statement "map…samples into a lower-dimensional manifold" a bit misleading since flows preserve dimensionality. This does happen by introducing inductive biases such as multi-scale architectures, but the core mathematics does not allow this.

Eq 2.1, 2.2, 2.3, 2.4, and others: loss functions / optimization objectives should usually be written as functions of the parameters that one would optimize (theta and phi, in this case 2.1)

p 26 *"we cannot simply propagate a gradient through a random node."*: This is an overstatement since the reparameterization trick wasn't the first method to differentiate through samples. For instance, Kingma & Welling could have used the existing score function estimator (REINFORCE). The contribution is more of providing a practical, low-variance gradient estimator (with the plus that, when combined with the variational approximation, made the overall computation graph look like an autoencoder).

p 28, 48, 61, 85, and others: You give several references in NAME [YEAR] format but look like they should be in [NAME, YEAR] format.

p 34: The reference to my paper should be ICLR 2019, not 2018. Moreover, I'm not sure my paper is the best one to cite when describing the effect of the aggregated posterior. My contribution is more about the aggregated posterior's effect on OOD detection. I first learned about the aggregated posterior from the VAE papers that re-write the ELBO to include a KL term between the aggregated and per-data-point posteriors, for the purposes of representation learning. This was also discussed in the VampPrior paper, which pre-dates my ICLR 2019 paper.

Eq 4.3 and 4.4: Could be written in one line instead of two.

Eq 9.1, 9.2, 9.3: These expressions should be set equal to something