ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Distillation of a tractable model from the VQ-VAE

Armin Hadžić, Milan Papež and Tomáš Pevný *

Artificial Intelligence Center, Czech Technical University in Prague.
Karlovo náměstí 13, 121 35 Prague, Czech Republic.

Abstract. Deep generative models with a discrete latent space, such as the Vector-Quantized Variational Autoencoder (VQ-VAE), offer excellent data generation capabilities, but—due to the large size of their latent space—their probabilistic inference is deemed intractable. We demonstrate that the VQ-VAE can be *distilled* into a tractable model by selecting a subset of latent variables with high probability under the prior. We frame the distilled model as a probabilistic circuit, and show that it preserves the expressiveness of the VQ-VAE while providing tractable probabilistic inference. Experiments illustrate competitive performance in both density estimation and conditional generation tasks, challenging the view of the VQ-VAE as an inherently intractable model.

1 Introduction

Deep generative models provide a versatile framework for learning statistical patterns in diverse data modalities [1]. They excel at generating new samples; however, they cannot answer even the most basic inference tasks without approximations. This issue arises because deep neural networks make analytical integration—a crucial part of many inference tasks—infeasible [2]. For this reason, we refer to deep generative models as intractable probabilistic models.

Discrete latent variable models. A complete probabilistic description of a discrete latent variable model is given by the following marginal probability distribution:

$$p(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}), \tag{1}$$

where $\mathbf{x} \coloneqq \{x_1, \ldots, x_S\} \in \mathcal{X}$ denotes observations, and $\mathbf{z} \coloneqq \{z_1, \ldots, z_M\} \in \mathcal{Z}$ are discrete latent variables. The prior probability mass function, $p(\mathbf{z})$, describes our beliefs in each value of $\mathbf{z}$. The conditional distribution $p(\mathbf{x}|\mathbf{z})$ models $\mathbf{x}$ conditionally on a different set of parameters for each value of $\mathbf{z}$, i.e., $p(\mathbf{x}|\mathbf{z}) \coloneqq p(\mathbf{x}|\theta_{\mathbf{z}})$. The latent space, $\mathcal{Z}$, i.e., the set of all latent variable configurations, is a finite structured set with exponential cardinality [3, 4].

Probabilistic circuits. Probabilistic circuits (PCs)[5] are *tractable probabilistic models* that resolve this issue and thus provide closed-form solutions to a wide range of inference tasks, including marginalization, conditioning, and expectation. PCs are deep computational graphs composed of computational *units*:

*The authors acknowledge the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO and the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics".

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

input, *product*, and *sum* [5]. We denote the set of inputs of a unit u as $\text{in}(u)$. The input unit $p_u(\mathbf{x}_u)$ computes a predefined, parametrized probability distribution. Sum and product units receive the outputs of other units as inputs. The sum unit computes the weighted sum of its inputs $p_u(\mathbf{x}_u) \coloneqq \sum_{i \in \text{in}(u)} w_i p_i(\mathbf{x}_i)$, where $w_i \in \mathbb{R}_{\geq 0}$ and $\sum_{i \in \text{in}(u)} w_i = 1$ are the weight parameters, and the product unit computes the product of its inputs, $p_u(\mathbf{x}_u) \coloneqq \prod_{i \in \text{in}(u)} p_i(\mathbf{x}_i)$. If the computational graph follows certain structural constraints, then PCs are tractable [5]. PCs are an instance of (1). Recent work has begun exploring methods that combine the efficient inference of PCs with the expressive power of deep neural networks [6]. This paper extends this line of work by investigating the tractability of the vector-quantized variational autoencoder (VQ-VAE) [7].

VQ-VAEs. The VQ-VAE is a deep generative model that compresses input data into a discrete latent representation. The model is similar to the variational autoencoder [8], but differs primarily in its vector quantization block. This block uses a collection of latent embedding vectors, $\{\mathbf{e}_i\}_{i=1}^K$, and is referred to as the *codebook*, where $\mathbf{e}_i \in \mathbb{R}^D$, with D the codeword length, and K is the codebook size. The encoder network $\text{E} : \mathbb{R}^{d \times h \times w} \to \mathbb{R}^{D \times H \times W}$, compresses $\mathbf{x}$ into a continuous latent variable, where HW are dimensions of the latent space. After the compression, the continuous latent variable is mapped to a discrete latent variable, $\mathbf{z} \in \mathcal{Z}$, by a nearestneighbor search in the codebook using Euclidean distance. It is a common practice that the VQ-VAE's prior, $p(\mathbf{z})$, is learned via an autoregressive model, such as PixelCNN [9]. Viewing the discrete latent variable as a sequence of indices, the prior is modeled as $p(\mathbf{z}|c) \coloneqq \prod_{i=1}^{HW} p(z_i|\mathbf{z}_{<i}, c)$, where $\mathbf{z}_{<i}$ are all indices before i in the row-major order, and $c \in \mathcal{C}$ is a high-level data description represented as a latent vector (e.g., a class label in supervised learning). VQ-VAEs are an example of the discrete latent variable model in (1), where the PixelCNN prior parametrizes $p(\mathbf{z}) \coloneqq \sum_c p(c)p(\mathbf{z}|c)$, and the decoder defines $p(\mathbf{x}|\mathbf{z})$ using a deep neural network.

Definition 1. *The discrete latent space of a VQ-VAE is defined as* $\mathcal{Z} \coloneqq \{1, 2, \ldots, K\}^{H \times W}$*, yielding a latent space of an exponential size,* $|\mathcal{Z}| \coloneqq K^{HW}$.

However, the size of the latent space of the VQ-VAE (Definition 1) makes probabilistic inference tasks intractable [3, 4]. In practice, only a fraction of the discrete latent space tends to contribute to the model outputs (e.g., its likelihood) [10]. Therefore, we propose a novel approach to transform the VQ-VAE into a tractable probabilistic model (i.e., a PC) by distilling a subset of the most relevant latent variables. The subset is identified using two complementary approaches: (i) random sampling, which requires enumeration of all latent variables, making it exhaustive and computationally expensive; and (ii) a beam search, which trades off optimality for computational efficiency. We demonstrate that the distilled model based on the beam search delivers a competitive performance to other tractable probabilistic models in the context of learning a probability distribution of images.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2 Distilling mixture models

VQ-VAE intractability. Computing $p(\mathbf{x}|\mathbf{z})$ for all $\mathbf{z} \in \mathcal{Z}$ is very expensive since $p(\mathbf{x}|\mathbf{z})$ is typically parametrized by a large neural network and the size of the latent space is exponentially high (Definition 1). Consequently, training the VQ-VAE via the exact likelihood (1) or the traditional ELBO [8] objectives is infeasible. To deal with this problem, the training is done by a heuristic loss function [7]. These facts also imply that computing any probabilistic inference queries with (1) is intractable.

Model distillation. We propose to address this intractability issue by distilling the VQ-VAE model into a mixture of tractable distributions. We refer to this model as a distilled model (DM). The key idea is to identify a subset of the most relevant latent variables, $\bar{\mathcal{Z}} \subset \mathcal{Z}$, and construct the DM as follows:

$$\hat{p}(\mathbf{x}) \coloneqq \sum_{\mathbf{z} \in \bar{\mathcal{Z}}} \frac{1}{|\bar{\mathcal{Z}}|} p(\mathbf{x}|\mathbf{z}). \tag{2}$$

We discuss two ways to construct $\bar{\mathcal{Z}}$, but, first, we state assumptions that ensure tractability of the DM.

Assumption 1. $\mathbf{x}$ *is conditionally independent given* $\mathbf{z}$*, i.e.,* $p(\mathbf{x}|\mathbf{z}) \coloneqq \prod_{i=1}^{S} p(x_i|\mathbf{z})$*, and each* $p(x_i|\mathbf{z})$ *is a tractable distribution (e.g., Gaussian, categorical).*

Under Assumption 1, the discrete latent variable model (1) reveals the connection between VQ-VAEs and PCs. Indeed, there can be many equivalent representations between these two models, depending on a specific architecture of a PC. However, the simplest one is that the decoder $p(\mathbf{x}|\mathbf{z})$ is seen as a product unit whose children are input units, and that the PixelCNN prior, $p(\mathbf{z})$, represents the weights of a sum unit, which is directly the sum in (2). To make the resulting DM tractable, we also have to satisfy the following assumption.

Assumption 2. *The number of components in* (2) *is kept computationally feasible, i.e.,* $N \ll K^{HW}$*.* [1]

Random sampling. One way to construct the latent set $\bar{\mathcal{Z}}$ is to enumerate $p(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{Z}$, and then sample a distinct set of latent variables from the class-conditioned VQ-VAE prior, $\bar{\mathcal{Z}} = \{\mathbf{z}_i | c_i \sim \mathcal{U}(\mathcal{C}), \mathbf{z}_i \sim p(\mathbf{z}|c_i)\}_{i=1}^{N}$. The DM built from randomly sampled latent variables is called a DM via Random Sampling (DMRS). However, the exhaustive enumeration is impractical due to the large latent space (Definition 1).

Beam search. To overcome the drawbacks of DMRS, $\bar{\mathcal{Z}}$ can be constructed by a guided search through the latent space. This traversal of $\mathcal{Z}$ identifies the most probable, informative regions without the exhaustive enumeration. Viewing the $H \times W$ latent grid as a row-major sequence of HW tokens over a vocabulary of size K allows the application of sequence search techniques, such as beam search (BS). BS is commonly employed to maintain tractability in large search spaces by trading off completeness and optimality [11]. The latent space of VQ-VAE (Definition 1) is one such example where BS can be applied due to

[1] The original model in (1) is recovered for $N = K^{HW}$.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

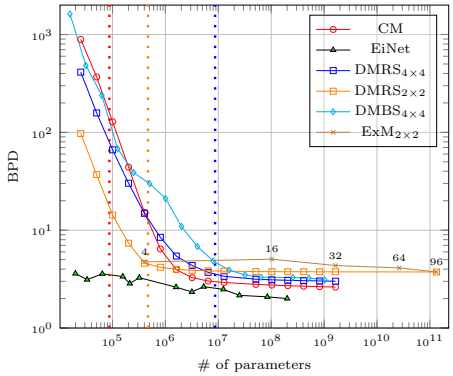

Figure 1: *BPD performance vs model size for different tractable models.* Lower is better. DMs are denoted with hollow squares, while EiNets have filled ones, as they are fully trained and non-distilled models. Dotted vertical lines indicate the sizes of the source VQ-VAEs used for distillation. For the CM, the red dotted line shows the decoder size. The numbers above the cross markers express the VQ-VAE codebook sizes, K, for the exact models. All results are averaged over 5 runs with different seeds.

its size. We use the class-conditioned stochastic BS [12], to discover latent variables for distillation, resulting in the DM via Beam Search (DMBS). DMBS requires $\mathcal{O}(NHWK)$ operations, i.e., dramatically fewer than $\mathcal{O}(K^{HW})$ needed for the exhaustive enumeration. This trade-off manifests in reduced expressivity, but, as we demonstrate in Section 3, its effects are modest. We use uniform mixture weights to turn the skewed prior into a balanced distribution with equal component contributions, consistent with Monte Carlo random sampling and improving the effectiveness of mode-seeking beam search.

3 Experimental results

We demonstrate the tractability of our DMs on two core probabilistic-inference tasks: density estimation and image inpainting. The goal is to exhibit that the DMs can accurately answer complex probabilistic queries, yielding answers that approach state-of-the-art models. We choose two tractable probabilistic models as baselines: continuous mixtures (CMs) [6] and Einsum networks (EiNets) [13]. Evaluations are conducted on MNIST [14], modeling the pixels by the Gaussian distribution. All settings used to reproduce the experiments reported in this paper are available at `https://github.com/ahadzic7/vqvae-distillation`

Density estimation. Figure 1 compares all the aforementioned models in terms of bits per dimension (BPD). We can see that the exact model $\text{ExM}_{2\times2}$, which uses a VQ-VAE with a latent grid of shape 2×2, sets a performance reference which is quickly approached by the $\text{DMRS}_{2\times2}$, indicating that even the randomized distillation captures critical information encoded by the VQ-VAE. Importantly, we can see that the $\text{DMBS}_{4\times4}$ closely follows the $\text{DMRS}_{4\times4}$, which demonstrates that the BS has the ability to select important parts of the latent space without the exhaustive enumeration necessary for the random sampling. The CM model outperforms all the DM models; however, as shown in Figure 2,

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Figure 2: *Image inpainting by tractable probabilistic models.* The unobserved $\mathbf{x}_\mathrm{u}$ and observed $\mathbf{x}_\mathrm{o}$ parts are highlighted by the red and blue colors, respectively.

its sample quality is lower. We conjecture that this is attributed to the smaller size of the CM's decoder (the red dashed line), from which the model is distilled. The performance of all the DMs plateaus, which shows that distilling more of the same or similar latent variables does not bring additional information into the resulting DMs. Interestingly, the EiNet model outperforms all the other models for all parameter counts; however, its sample quality seems lower (Figure 2).

Tractable inference. Our DMs support diverse inference tasks, including marginalization, expectation, and maximum a posteriori estimation. We demonstrate the tractability of the DMs on the image inpainting, which corresponds to the conditional inference task $p(\mathbf{x}_\mathrm{u}|\mathbf{x}_\mathrm{o})$, where $\mathbf{x}_\mathrm{u}$ and $\mathbf{x}_\mathrm{o}$ are unobserved and observed image parts, respectively. Figure 2 shows that all models successfully infill missing parts to form correct digits. Importantly, the image quality of the reconstructions done by our DMs is visually sharper and more coherent.

4 Conclusions

We have proposed a novel framework for distilling tractable mixtures from intractable VQ-VAEs. Our DMs are able to answer a broad range of probabilistic inference tasks, while retaining the expressive power of VQ-VAEs. We have investigated two strategies for identifying informative latent space regions: the random sampling and the beam search. Though the random sampling has proven efficient, its key disadvantage is the full enumeration of the latent space, making it impractical for scaling to state-of-the-art VQ-VAEs. Importantly, the beam search delivers almost identical performance, while avoiding the exhaustive enumeration. In future work, we plan to design data-driven exploration strategies of the latent space that will allow us to improve the performance of our DMs.

References

[1] Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities. In *Advances in Neural Information Processing Systems*, 2024.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[2] Luong-Ha Nguyen and James-A. Goulet. Analytically tractable inference in deep neural networks, 2021.

[3] Gonçalo Correia, Vlad Niculae, Wilker Aziz, and André Martins. Efficient marginalization of discrete and structured latent variables via sparsity. In *Advances in Neural Information Processing Systems*, 2020.

[4] Robert Peharz, Robert Gens, Franz Pernkopf, and Pedro Domingos. On the latent variable interpretation in sum-product networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2030–2044, 2016.

[5] YooJung Choi, Antonio Vergari, and Guy Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. October 2020.

[6] Alvaro H.C. Correia, Gennaro Gala, Erik Quaeghebeur, Cassio De Campos, and Robert Peharz. Continuous mixtures of tractable probabilistic models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[7] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[9] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with PixelCNN decoders, 2016.

[10] Haohan Guo, Fenglong Xie, Dongchao Yang, Hui Lu, Xixin Wu, and Helen Meng. Addressing index collapse of large-codebook speech tokenizer with dual-decoding product-quantized variational auto-encoder, 2024.

[11] Yuehua Xu, Alan Fern, and Sungwook Yoon. Learning linear ranking functions for beam search with application to planning. *Journal of Machine Learning Research*, 10(7), 2009.

[12] Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models, 2017.

[13] Robert Peharz, Steven Lang, Antonio Vergari, Karl Stelzner, Alejandro Molina, Martin Trapp, Guy Van Den Broeck, Kristian Kersting, and Zoubin Ghahramani. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.