Increasing Expressivity of a Hyperspherical VAE

Davidson, T.R.; Tomczak, J.M.; Gavves, E.

DOI
10.48550/arXiv.1910.02912
Publication date
2019
Document Version
Final published version

Increasing Expressivity of a Hyperspherical VAE

Tim R. Davidson
Aiconic
tim.davidson@aiconic.com

Jakub M. Tomczak
University of Amsterdam
jmk.tomczak@gmail.com

Efstratios Gavves
University of Amsterdam
egavves@uva.nl

Abstract

Learning suitable latent representations for observed, high-dimensional data is an important research topic underlying many recent advances in machine learning. While traditionally the Gaussian normal distribution has been the go-to latent parameterization, recently a variety of works have successfully proposed the use of manifold-valued latents. In one such work [4], the authors empirically show the potential benefits of using a hyperspherical von Mises-Fisher (vMF) distribution in low dimensionality. However, due to the unique distributional form of the vMF, expressivity in higher dimensional space is limited as a result of its scalar concentration parameter leading to a 'hyperspherical bottleneck'. In this work we propose to extend the usability of hyperspherical parameterizations to higher dimensions using a product-space instead, showing improved results on a selection of image datasets.

1 Introduction

Following the manifold hypothesis, in unsupervised generative learning we strive to recover a distribution on a (low-dimensional) latent manifold, capable of explaining observed, high-dimensional data, e.g. images. One of the most popular frameworks to achieve this goal is the Variational Auto-Encoder (VAE) [13, 22], a latent variable model which combines variational inference and auto-encoding to directly optimize the parameters of some latent distribution. While originally restricted to 'flat' space using the classic Gaussian normal distribution, there has recently been a surge in research extending the VAE to distributions defined on manifolds with non-trivial topologies [4, 6, 16, 17, 21, 8, 7]. This is fruitful, as most data is not best represented by distributions on flat space, which can lead to undesired 'manifold mismatch' behavior.

In [4], the authors propose a hyperspherical parameterization of the VAE using a von Mises-Fisher distribution, demonstrating the improved results over the especially bad pairing of the 'blob-like' Gaussian distribution and hyperspherical data. Surprisingly, they further show that these positive results extend to datasets *without* a clear hyperspherical interpretation, which they mostly attribute to the restricted surface area and the absence of a 'mean-biased' prior in the vMF as the Uniform distribution is feasible in the compact, hyperspherical space. However, as dimensionality increases performance begins to decrease. This could possibly be explained by taking a closer look at the vMF's functional form

$$q(\mathbf{z}|\mu,\kappa) = \mathcal{C}_m(\kappa)\exp\left(\kappa\mu^T\mathbf{z}\right), \tag{1}$$

$$\mathcal{C}_m(\kappa) = \frac{\kappa^{m/2-1}}{(2\pi)^{m/2}\mathcal{I}_{m/2-1}(\kappa)}, \tag{2}$$

where $||\mu||^2 = 1$, κ a scalar, $\mathcal{C}_m(\kappa)$ is the normalizing constant, and $\mathcal{I}_v$ denotes the modified Bessel function of the first kind at order v. Note that the scalar concentration parameter κ is fixed in all dimensions, severely limiting the distribution's expressiveness as dimensionality increases.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

2 Method: A Hyperspherical Product-Space

To improve on the vMF's per-dimension concentration flexibility limitation we propose a simple idea: breaking up the single latent hyperspherical assumption, into a concatenation of multiple independent hyperspherical distributions. Such a compositional construction increases flexibility through the addition of a new concentration parameter for each hypersphere, as well as providing the possibility of sub-structure forming . Given a hyperspherical random variable $\mathbf{z} \in \mathcal{S}_M$, we want to choose $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_k$ in respectively $\mathcal{S}_{M_0}, \mathcal{S}_{M_1}, \cdots, \mathcal{S}_{M_k}$ s.t. $\sum_{i=0}^{k} M_i = M$, and $\mathbf{z} = \mathbf{z}_0 \frown \mathbf{z}_1 \frown \cdots \frown \mathbf{z}_k$, where $(\frown)$ denotes concatenation. The probabilistic model becomes:

$$p(\mathbf{z}) = p(\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_k) \overset{*}{=} \prod_{i=0}^{k} p(\mathbf{z}_i), \tag{3}$$

which factorizes in (*) if we assume independence between the new sub-structures. Assuming conditional independence of the approximate posterior as well, i.e. $q(\mathbf{z}|\mathbf{x}) = q(\mathbf{z}_0|\mathbf{x})q(\mathbf{z}_1|\mathbf{x}) \cdots q(\mathbf{z}_k|\mathbf{x})$, it can be shown[1] that the Kullback-Leibler divergence simplifies as

$$KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \int_{\mathcal{S}_M} q(\mathbf{z}|\mathbf{x}) \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{z} = \sum_i KL(q(\mathbf{z}_i|\mathbf{x})||p(\mathbf{z}_i)) \tag{4}$$

Flexibility Trade-Off Given a single hypersphere and keeping the ambient space fixed, for each additional 'break', a degree of freedom is exchanged for a concentration parameter. In the base case of $\mathcal{S}_{k+1}$, we can potentially support $k+1$ 'independent' feature dimensions, that must share a single concentration parameter κ, and hence are globally restricted in their flexibility per dimension. On the other hand, the moment we break $\mathcal{S}_{k+1}$ up in the Cartesian cross-product of $\mathcal{S}_{k/2} \times \mathcal{S}_{k/2}$, we 'lose' an independent dimensions (or degree of freedom), but in exchange the two resulting sub-hyperspheres have to share their concentration parameters κ_1, κ_2 over fewer dimensions increasing flexibility[2].

The reason a vMF is uniquely suited for such a decomposition as opposed to a Gaussian, is that assuming a factorized variance the Gaussian distribution is already equipped with a concentration parameter for each dimension. However, in the case of the vMF, which has only a single concentration parameter κ for *all* dimensions, we gain flexibility. This is an important distinction: while all dimensions are implicitly connected through the shared loss objective in both cases, in the case of the vMF this connection is amplified through the *direct connection* of the shared concentration parameter.

Related Work The work closest to our model is that of [20], where a Cartesian product of Gaussian Mixture Models (GMMMs) is proposed, with hyperpriors on all separate components to create a fully data-inferred model. They use results from [10, 12] on structured VAEs, and extend the work on VAEs with single GMMs of [18, 5, 11]. Partially following similar motivations to our work, the authors hypothesize and empirically show the structured compositionality encourages disentanglement. By working with GMMs instead of single Gaussians, they circumvent the factorized single Gaussian break-up limitation described before. Another recent work proposing to break up a large, single latent representation into a composition of sub-structures in the context of Bayesian optimization is [19].

3 Experiments and Discussion

To test the ability of a hyperspherical product-space model to improve performance over its single-shell counterpart, we perform product-space interpolations breaking up a single shell into an increasing amount of independent components.

Experimental Setup We conduct experiments on Static MNIST, Omniglot [14], and Caltech 101 Silhouettes [15] mostly following the experimental setup of [4], using a simple MLP encoder-decoder architecture with `ReLU()` activations between layers. We train for 300 epochs using early-stopping with a look-ahead of 50 epochs, and a linear *warm-up* scheme of 100 epochs as per [2, 23], during which the KL divergence is annealed from 0 to β [9, 1]. Marginal log-likelihood is estimated using importance sampling with 500 sample points per [3], reporting the mean over three random seeds.

[1] See Appendix A for derivation.

[2] In the most extreme case, this will lead to a latent space of $[\mathcal{S}_1]_{\times(k+1)/2}$ - which is equal to the n-Torus.

Table 1: Overview of best results of various $\mathcal{S}_{40}$ product-space ambient dimensionality interpolations compared to best single $\mathcal{S}_m$-VAE ($m \leq 40$) indicated (*). LL represents the negative log-likelihood, $\mathcal{L}|q|$ the ELBO, a indicates the ambient space dimensionality, κ the number of concentration parameters, i.e. breaks, and $[\mathcal{S}_k]$ the product-space composition.

			Static MNIST			
a	κ	$[\mathcal{S}_k]$	LL	$\mathcal{L}\|q\|$	LL*	$\mathcal{L}\|q\|$*
41	4	$\mathcal{S}_{10} \times [\mathcal{S}_9]_{\times 3}$	-92.65	-98.23	-93.37	-98.88
41	4	$\mathcal{S}_{20\times 10\times 6\times 1}$	-92.59	-98.27		
41	6	$\mathcal{S}_{15\times 10\times 4\times 3\times 2\times 1}$	-92.25	-98.10		
41	6	$[\mathcal{S}_6]_{\times 5} \times \mathcal{S}_5$	-92.71	-98.46		
			Caltech			
41	4	$\mathcal{S}_{10} \times [\mathcal{S}_9]_{\times 3}$	-139.30	-151.67	-143.49	-152.25
41	4	$\mathcal{S}_{20\times 10\times 6\times 1}$	-140.64	-153.05		
			Omniglot			
41	4	$\mathcal{S}_{20\times 10\times 6\times 1}$	-112.79	-119.17	-113.83	-120.48
41	6	$[\mathcal{S}_6]_{\times 5} \times \mathcal{S}_5$	-112.58	-118.49		
41	10	$\mathcal{S}_4 \times [\mathcal{S}_3]_{\times 9}$	-112.64	-118.67		

Keeping in mind the *flexibility trade-off* consideration, we analyze both the effects of keeping the total degrees of freedom fixed (increasing ambient space dimensionality), as well as the case of keeping the ambient space fixed (decreasing the degrees of freedom). We break up $\mathcal{S}_{40}$ respectively into 2, 4, 6, 10, and 40 sub-spaces. In each break-up, we try a balanced, leveled, and unbalanced hyperspherical composition.

Results A summary of best results for fixed ambient space is shown in Table 1, with a summary of best results for fixed degrees of freedom and complete interpolations in Appendix B. Initial inspection shows that partially breaking up a single $\mathcal{S}_{40}$ hypersphere into a hyperspherical product-space indeed allows us to improve performance for all examined datasets. Diving deeper into the results, we do find that both the number of breaks as well as the dimensional composition of these breaks strongly inform performance and learning stability.

A high number of breaks appears to negatively influence both performance and learning stability. Indeed, for most datasets the 'Torus' setting, i.e. full factorization in $\mathcal{S}_1$ components proved too unstable to train to convergence. One explanation for this result could be found in the fact that we omit the REINFORCE part of the vMF reparameterization during training[3]. While only of very limited influence on a single hyperspherical distribution, the accumulated bias across many shells might lead to a non-trivial effect. On the other hand, adding as few as four breaks extends the model's expressivity enough to outperform a single shell consistently.

Balance of the subspace composition plays a key role as well. We find that when the subspaces are too unbalanced, e.g. $\mathcal{S}_{37}$ v. $[\mathcal{S}_1]_{\times 3}$, the network starts to 'choose' between subspace channels. Effectively, it will for example start encoding all information in the $\mathcal{S}_1$ shells and *completely ignore* the $\mathcal{S}_{37}$ shell, leading to an effective latent space of $\mathcal{S}_3$ degrees of freedom[4], see for example Fig. 2(a). On the contrary, better balanced compositions appear capable of cleanly separating semantically meaningful features across shells as displayed in Fig. 2(b).

Conclusion and Future Work In summary, breaking up a single hypersphere into multiple components effectively increases concentration expressiveness leading to more stable training and improved results. In future work we'd like to investigate the possibility of *learning* an optimal break-up as opposed to fixing it a-priori, as well as mixing sub-spaces with different topologies.

[3] See [4], Appendix D for more details.

[4] For a more extended discussion on the interplay between balance and the KL divergence see Appendix B.

Acknowledgments

We would like to thank Luca Falorsi and Nicola De Cao for insightful discussions during the development of this work.

References

[1] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *ICML*, pages 159–168, 2018.

[2] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21, 2016.

[3] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *ICLR*, 2016.

[4] Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical Variational Auto-Encoders. *UAI*, 2018.

[5] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[6] Luca Falorsi, Pim de Haan, Tim R Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S Cohen. Explorations in homeomorphic variational auto-encoding. *ICML Workshop*, 2018.

[7] Luca Falorsi, Pim de Haan, Tim R Davidson, and Patrick Forré. Reparameterizing distributions on lie groups. *AISTATS*, 2019.

[8] Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *NeurIPS*, pages 439–450. Curran Associates, Inc., 2018.

[9] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

[10] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[11] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *IJCAI*, pages 1965—-1972, 2017.

[12] Matthew Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *NIPS*, pages 2946–2954, 2016.

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, 2013.

[14] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[15] Benjamin Marlin, Kevin Swersky, Bo Chen, and Nando Freitas. Inductive principles for restricted boltzmann machine learning. In *AISTATS*, pages 509–516, 2010.

[16] Emile Mathieu, Charline Le Lan, Chris J Maddison, Ryota Tomioka, and Yee Whye Teh. Hierarchical representations with poincar\'e variational auto-encoders. *NeurIPS*, 2019.

[17] Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A differentiable gaussian-like distribution on hyperbolic space for gradient-based learning. *arXiv preprint arXiv:1902.02992*, 2019.

[18] Eric Nalisnick, Lars Hertel, and Padhraic Smyth. Approximate inference for deep latent gaussian mixtures. In *NIPS Workshop on Bayesian Deep Learning*, volume 2, 2016.

[19] Changyong Oh, Jakub M. Tomczak, Efstratios Gavves, and Max Welling. Combinatorial bayesian optimization using graph representations. *ICML Workshop*, 2019.

[20] Ulrich Paquet, Sumedh K Ghaisas, and Olivier Tieleman. A factorial mixture prior for compositional deep generative models. *arXiv preprint arXiv:1812.07480*, 2018.

[21] Luis A. Pérez Rey, Vlado Menkovski, and Jacobus W. Portegies. Diffusion variational autoencoders. *arXiv preprint arXiv:1901.08991*, 2019.

[22] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, pages 1278–1286, 2014.

[23] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *NIPS*, pages 3738–3746, 2016.

A Dimensionality Decomposition

Given a latent variable $\mathbf{z} \in \mathbb{R}^M$, we choose $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_k$ in respectively $\mathbb{R}^{M_0}, \mathbb{R}^{M_1}, \cdots, \mathbb{R}^{M_k}$ s.t. $\sum_{i=0}^{k} M_i = M$, and $\mathbf{z} = \mathbf{z}_0 \frown \mathbf{z}_1 \frown \cdots \frown \mathbf{z}_k$, where $(\frown)$ denotes concatenation. The probabilistic model becomes:

$$p(\mathbf{z}) = p(\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_k) \overset{*}{=} \prod_{i=0}^{k} p(\mathbf{z}_i), \tag{5}$$

which factorizes in (*) if we assume independence. Assuming conditional independence of the approximate posterior as well, i.e. $q(\mathbf{z}|\mathbf{x}) = q(\mathbf{z}_0|\mathbf{x})q(\mathbf{z}_1|\mathbf{x}) \cdots q(\mathbf{z}_k|\mathbf{x})$, the Kullback-Leibler divergence simplifies as

$$\begin{aligned}
&KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \int_{\mathbb{R}^M} q(\mathbf{z}|\mathbf{x}) \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{z} \\
&= \int_{\mathbb{R}^{M_0} \times \cdots \times \mathbb{R}^{M_k}} q(\mathbf{z}_0|\mathbf{x})q(\mathbf{z}_1|\mathbf{x}) \cdots q(\mathbf{z}_k|\mathbf{x}) \log \frac{q(\mathbf{z}_0|\mathbf{x})q(\mathbf{z}_1|\mathbf{x}) \cdots q(\mathbf{z}_k|\mathbf{x})}{p(\mathbf{z}_0)p(\mathbf{z}_1) \cdots p(\mathbf{z}_k)} d\mathbf{z}_0 \cdots d\mathbf{z}_k \\
&= \int_{R^{M_0}} q(\mathbf{z}_0|\mathbf{x}) \log \frac{q(\mathbf{z}_0)}{p(\mathbf{z})} d\mathbf{z}_0 + \cdots + \int_{\mathbb{R}^{M_k}} q(\mathbf{z}_k|\mathbf{x}) \log \frac{q(\mathbf{z}_k|\mathbf{x})}{p(\mathbf{z}_k)} d\mathbf{z}_k \\
&= \sum_i KL(q(\mathbf{z}_i|\mathbf{x})||p(\mathbf{z}_i)),
\end{aligned} \tag{6}$$

B Supplementary Tables and Figures

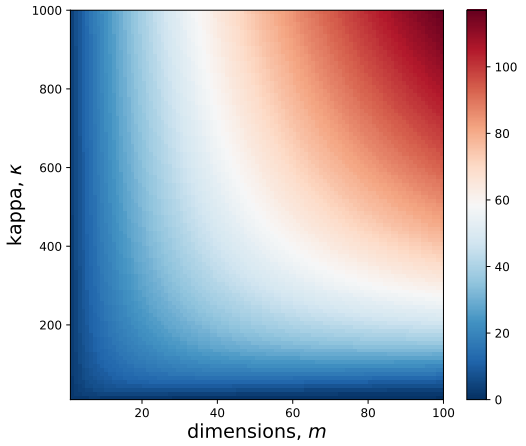


Figure B.1: Value of the von Mises-Fisher Kullback-Leibler divergence varying the concentration parameter κ on the y-axis, and the dimensionality m on the x-axis. (Best viewed in color)

Another way of understanding the importance of balance is by examining the KL divergence form of the vMF and its influence in the loss objective: In order to achieve high quality reconstruction performance, it is necessary for the concentration parameter κ to concentrate, i.e. take on a high value. Given the Uniform prior setting in which $\kappa = 0$, this logically leads to an increase in the KL-divergence. The crucial observation here however, is that the strength of the KL-divergence is also strongly dependent on the dimensionality as can be observed in Fig. B.1. Hence during learning over a product-space containing several lower dimensionality components and a single high dimensionality component, if the reconstruction error can be made sufficiently low using the lower dimensionality components, the optimal loss minimization strategy would be to set the concentration parameter of the largest component to 0, effectively ignoring it. A possible strategy to prevent this from happening could be to set separate β parameters for each hyperspherical component, however we fear that this will quickly blow up the hyperparameter search-space.

B.1 Fixed Ambient Space

Table 2: Summary of results of $\mathcal{S}_{40}$ ambient interpolations for unsupervised model on Static MNIST. RE and KL correspond respectively to the reconstruction and the KL part of the ELBO.

m	κ	$[\mathcal{S}_k]$	LL	$\mathcal{L}\|q\|$	RE	KL
41	2	$\mathcal{S}_{20} \times \mathcal{S}_{19}$	-93.18	-98.72	69.78	28.94
41	2	$\mathcal{S}_{38} \times \mathcal{S}_{1}$	-95.69	-103.67	71.67	32.01
41	4	$\mathcal{S}_{10} \times [\mathcal{S}_9]_{\times 3}$	-92.65	-98.23	70.55	27.68
41	4	$\mathcal{S}_{20\times 10\times 6\times 1}$	-92.59	-98.27	71.33	26.94
41	4	$\mathcal{S}_{34} \times [\mathcal{S}_1]_{\times 3}$	-108.42	-116.86	99.62	17.23
41	6	$\mathcal{S}_{15\times 10\times 4\times 3\times 2\times 1}$	-92.25	-98.10	69.78	28.32
41	6	$\mathcal{S}_{30} \times [\mathcal{S}_1]_{\times 5}$	-93.86	-100.99	69.46	31.54
41	6	$[\mathcal{S}_6]_{\times 5} \times \mathcal{S}_5$	-92.71	-98.46	70.97	27.48
41	10	$\mathcal{S}_{10\times 5\times 4\times 3} \times [\mathcal{S}_2]_{\times 3} \times [\mathcal{S}_1]_{\times 3}$	-92.93	-99.07	70.67	28.41
41	10	$\mathcal{S}_{22} \times [\mathcal{S}_1]_{\times 9}$	-93.45	-100.29	68.75	31.54
41	10	$\mathcal{S}_{4} \times [\mathcal{S}_3]_{\times 9}$	-93.36	-99.40	71.93	27.47

Table 3: Summary of results of $\mathcal{S}_{40}$ ambient interpolations for unsupervised model on Caltech.

m	κ	$[\mathcal{S}_k]$	LL	$\mathcal{L}\|q\|$	RE	KL
41	2	$\mathcal{S}_{20} \times \mathcal{S}_{19}$	-142.43	-155.24	123.35	31.89
41	2	$\mathcal{S}_{38} \times \mathcal{S}_{1}$	-147.41	-166.64	130.41	36.22
41	4	$\mathcal{S}_{10} \times [\mathcal{S}_9]_{\times 3}$	-139.30	-151.67	120.82	30.85
41	4	$\mathcal{S}_{20\times 10\times 6\times 1}$	-140.64	-153.05	123.23	29.82
41	4	$\mathcal{S}_{34} \times [\mathcal{S}_1]_{\times 3}$	-168.25	-186.47	170.44	16.03
41	6	$\mathcal{S}_{15\times 10\times 4\times 3\times 2\times 1}$	-142.84	-156.59	126.59	30.00
41	6	$\mathcal{S}_{30} \times [\mathcal{S}_1]_{\times 5}$	-169.15	-177.23	161.68	15.55
41	6	$[\mathcal{S}_6]_{\times 5} \times \mathcal{S}_5$	-139.99	-152.68	121.91	30.77
41	10	$\mathcal{S}_{10\times 5\times 4\times 3} \times [\mathcal{S}_2]_{\times 3} \times [\mathcal{S}_1]_{\times 3}$	-144.73	-159.27	126.14	33.13
41	10	$\mathcal{S}_{22} \times [\mathcal{S}_1]_{\times 9}$	-154.91	-164.90	140.06	24.83
41	10	$\mathcal{S}_{4} \times [\mathcal{S}_3]_{\times 9}$	-144.72	-160.13	126.34	33.79

Table 4: Summary of results of $\mathcal{S}_{40}$ ambient interpolations for unsupervised model on Omniglot.

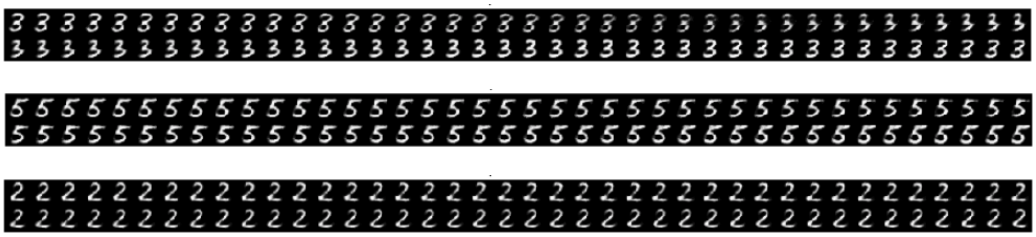
m	κ	$[\mathcal{S}_k]$	LL	$\mathcal{L}\|q\|$	RE	KL
41	2	$\mathcal{S}_{20} \times \mathcal{S}_{19}$	-114.32	-120.72	92.10	28.62
41	2	$\mathcal{S}_{38} \times \mathcal{S}_{1}$	-115.19	-122.30	91.82	30.48
41	4	$\mathcal{S}_{10} \times [\mathcal{S}_9]_{\times 3}$	-113.29	-118.97	88.93	30.05
41	4	$\mathcal{S}_{20\times 10\times 6\times 1}$	-112.79	-119.17	87.94	31.23
41	4	$\mathcal{S}_{34} \times [\mathcal{S}_1]_{\times 3}$	-136.39	-142.03	132.75	9.28
41	6	$\mathcal{S}_{15\times 10\times 4\times 3\times 2\times 1}$	-114.07	-119.99	91.26	28.72
41	6	$\mathcal{S}_{30} \times [\mathcal{S}_1]_{\times 5}$	-131.55	-137.29	124.66	12.62
41	6	$[\mathcal{S}_6]_{\times 5} \times \mathcal{S}_5$	-112.58	-118.49	88.27	30.23
41	10	$\mathcal{S}_{10\times 5\times 4\times 3} \times [\mathcal{S}_2]_{\times 3} \times [\mathcal{S}_1]_{\times 3}$	-113.53	-119.83	90.00	29.83
41	10	$\mathcal{S}_{22} \times [\mathcal{S}_1]_{\times 9}$	-114.95	-121.42	92.42	29.00
41	10	$\mathcal{S}_{4} \times [\mathcal{S}_3]_{\times 9}$	-112.64	-118.67	88.98	29.68

B.2 Fixed Degrees of Freedom

Table 5: Overview of best results (mean over 3 runs) of $\mathcal{S}_{40}$ product-space interpolations compared to best single $\mathcal{S}_m$-VAE ($m \leq 40$) indicated (*). Here a indicates the ambient space dimensionality, κ the number of concentration parameters, i.e. breaks, and $[\mathcal{S}_k]$ the product-space composition.

			Static MNIST			
a	κ	$[\mathcal{S}_k]$	LL	$\mathcal{L}\|q\|$	LL*	$\mathcal{L}\|q\|$*
44	4	$[\mathcal{S}_{10}]_{\times 4}$	-92.62	-98.26	-93.38	-98.88
46	6	$[\mathcal{S}_7]_{\times 5} \times \mathcal{S}_5$	-92.59	-98.46		
46	6	$\mathcal{S}_{15\times10\times5\times4\times3\times3}$	-92.50	-98.28		
50	10	$\mathcal{S}_{10\times7\times6\times5\times4\times3\times2} \times [\mathcal{S}_1]_{\times 3}$	-92.57	-98.81		
				Caltech		
44	4	$[\mathcal{S}_{10}]_{\times 4}$	-137.95	-150.86	-143.49	-152.25
46	6	$[\mathcal{S}_7]_{\times 5} \times \mathcal{S}_5$	-139.84	-152.92		
				Omniglot		
44	4	$[\mathcal{S}_{10}]_{\times 4}$	-112.28	-118.21	-113.83	-120.48
46	6	$[\mathcal{S}_7]_{\times 5} \times \mathcal{S}_5$	-112.78	-118.84		
50	10	$[\mathcal{S}_4]_{\times 10}$	-112.61	-118.70		

B.3 Ignored and Disentangled Shells



(a) Ignored Sub-space

(b) Thick to Thin

Figure B.2: $\mathcal{S}_1$ interpolations of broken up $\mathcal{S}_9$. On top an example of an 'ignored' sub-space, leading to little to no semantic change when decoded. Bottom a semantically meaningful sub-space that consistently changes the stroke thickness.