DEEP GENERATIVE MODELS FOR LEARNING COHERENT LATENT REPRESENTATIONS FROM MULTI-MODAL DATA

Timo Korthals, Marc Hesse & Ulrich Rckert
Bielefeld University
Cluster of Excellence Cognitive Interaction Technologies
Cognitronics & Sensor Systems
Inspiration 1, 33619 Bielefeld, Germany
{tkorthals,mhesse,rueckert}@cit-ec.uni-bielefeld.de

Jürgen Leitner
Australian Centre for Robotic Vision
Queensland University of Technology
Brisbane, Australia
j.leitner@qut.edu.au

ABSTRACT

The application of multi-modal generative models by means of a *Variational Auto Encoder* (VAE) is an upcoming research topic for sensor fusion and bi-directional modality exchange. This contribution gives insights into the learned joint latent representation and shows that expressiveness and coherence are decisive properties for multi-modal datasets. Furthermore, we propose a multi-modal VAE derived from the full joint marginal log-likelihood that is able to learn the most meaningful representation for ambiguous observations. Since the properties of multi-modal sensor setups are essential for our approach but hardly available, we also propose a technique to generate correlated datasets from uni-modal ones.

1 INTRODUCTION

It is quite common in the multi-modal VAE community to model a bi-modal dataset as follows (Wang et al. (2016); Ngiam et al. (2011); Suzuki et al. (2017); Vedantam et al. (2017)): The first modality a denotes the raw data and b denotes the label (e.g. the digits' images and labels as one-hot vector wrt. the MNIST dataset). This is a rather artificial assumption and only sufficient when the objective is within a semi-supervised training framework. Real multi-modal data does not show this behavior as there are commonly multiple raw data inputs. Unfortunately, only complex multi-modal datasets of heterogeneous sensor setups exist (Ofli et al. (2013); Udacity (2016); Kragh et al. (2017)), which makes a comprehensive evaluation for VAEs futile. On the other hand, creating own multi-modal datasets is exhaustive since training generative models either demand dense sampling or supervised signals to form a consistent latent manifold (Bengio et al. (2012)).

While naïve consolidation of non-coherently datasets does not meet the conditions of data continuity, as discussed later, we propose a consolidation technique by sampling from superimposed latent spaces of various uni-modal trained CVAEs in Sec. 1.1. This approach allows the generation of multi-modal datasets from distinct and disconnected uni-modal sets.

1.1 MNIST-E

Perry et al. (2010) state that Hebbian learning relies on the fact that the same objects are continuously transformed to their nearest neighbor in the observable space. Higgins et al. (2016) adopted this approach to their assumptions, that this notion can be generalized within the latent manifold learning. Further, neither a coherent manifold nor a proper factorization of the latent space can be trained if these assumptions are not fulfilled by the dataset. In summary, this means that observed data has to have the property of continues transformation wrt. to their properties (e.g. position and shape of an object), such that a small deviation of the observations results in proportional deviations in the latent space. We adopt this assumption for multi-modal datasets where observations should correlate if the same quantity is observed, such that a small deviation in the common latent representation between all modalities conducts a proportional impact in all observations. This becomes an actual fundamental requirement for any multi-modal dataset, as correlation and coherence are within the objective of multi-modal sensor fusion. In the following, we propose a technique to generate new multi-modal datasets, given different uni-modal enclosed sets which meet the former conditions.

A valuable property of the VAE's learned posterior distribution is, that it matches the desired prior quite sufficiently if only a single class is observed. This characteristic can be found again in the conditional VAE (CVAE) Kingma et al. (2014); Sohn et al. (2015) as it's training is supported by the ground truth labels of the observations. Thus, it actually builds non-related posterior distribution for each class label, where every distribution matches a given prior. Furthermore, we adopt the idea of β-VAE Higgins et al. (2017) which learns disentangled and factorized latent representations. Combining the properties of both advantages allows the superimposing of latent manifolds from various uni-modal encoders as shown in Fig. 1 (Top-Right). Now, latent samples can be drawn from the posterior to operate all CVAE encoders, with the desired label, to generate continues multi-modal data.

To test the approach we consolidate MNIST (LeCun Yann et al. (1998)) and fashion-MNIST (Xiao et al. (2017)) to an entangled *MNIST* (MNIST-E) set by sampling from the prior (i.e. $z \sim \mathcal{N}(0,\mathbf{I})$) to generate observation tuples from the corresponding encoder networks $p_{\theta_a}(a|z,C)$ and $p_{\theta_b}(b|z,C)$ with class label C. The network architecture is explained in Sec. 2. To avoid artifacts, only samples from within 2σ of the prior are obtained.

Furthermore, we train a bi-modal JMVAE on the newly generated data to depict properties of the different datasets. We are aware of the fact that consolidation of uni-modal datasets cannot be achieved easily since continuity is hardly measurable. Therefore, naïve consolidation results in a mixed dataset (i.e. mixed-MNIST) as shown in Fig. 1. To mimic this behavior and to achieve a fair comparison of the ELBO, we shuffle the generated fashion-MNIST per class label of MNIST-E to generate an equivalent mixed *MNIST-E* (MNIST-ME) set.

As shown in Fig. 1 (bottom), the JMVAE's latent space reveals that for MNIST-M single clusters share the same mean as the best representative of a single label, but the variance of any uni-modal trained encoder remains orthogonal. Thus, the continuity in the observations does not correlate with each other by any means. On the other hand, the MNIST-E set with continues samples shows the desired behavior of multi-modal datasets as the JMVAE trains a coherent distribution for all uni- and multi-modal encoders. These observations show that our proposed approach for generating new entangled datasets meet the formulated requirements of multi-modal datasets.

Acknowledgments

This research was supported by 'CITEC' (EXC 277) at Bielefeld University and the Federal Ministry of Education and Research (57388272). The responsibility for the content of this publication lies with the author.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. (1993):1–30, 2012. ISSN 15324435. doi: 10.1145/1756006.1756025. URL http://arxiv.org/abs/1206.5538.

Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early Visual Concept Learning with Unsupervised Deep Learning. 2016. URL http://arxiv.org/abs/1606.05579.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: Improving Zero-Shot Transfer in Reinforcement Learning. 2017. ISSN 1938-7228. URL http://arxiv.org/abs/1707.08475.

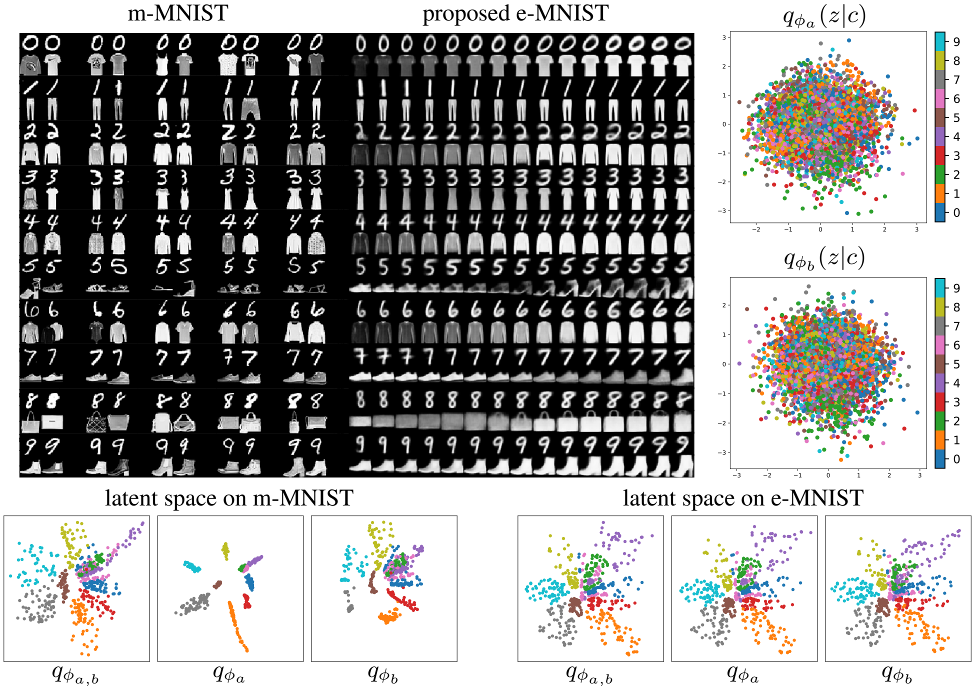

Figure 1: **Top-Left**: Depiction of naïve mixed MNIST (m-MNIST) vs. proposed entangled MNIST (e-MNIST). m-MNIST is pairwise plotted with the closest match of MNIST digits according to the mean-squared-error. The corresponding fashion-MNIST samples show no continuity nor correlation (despite the intended class correlation). e-MNIST shows the desired entanglement for changes of a single latent space factor. **Top-Right**: Latent space of the CVAE for the modalities a (MNIST) and b (fashion-MNIST). **Bottom**: Latent space of a trained JMVAE (c.f. Sec. 1.2.4). m-MNIST shows clear orthogonalization between modalities of the same class and segregation between classes (colorization is wrt. the CVAE legend). e-MNIST shows a coherently learned latent space between the uni- and multi-modal encoders. Thus, the JMVAE learns the correlation inside the dataset sufficiently ($\mathcal{L}_{a,b|\text{me-MNIST}} = -204.48$ vs. $\mathcal{L}_{a,b|\text{e-MNIST}} = -199.23$).

Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised Learning with Deep Generative Models. pp. 1–9, 2014. ISSN 10495258. URL `http://arxiv.org/abs/1406.5298`.

Mikkel Fly Kragh, Peter Christiansen, Morten Stigaard Laursen, Morten Larsen, Kim Arild Steen, Ole Green, Henrik Karstoft, and Rasmus Nyholm Jørgensen. FieldSAFE: Dataset for Obstacle Detection in Agriculture. *Sensors*, 17(11), 2017.

LeCun Yann, Cortes Corinna, and Burges Christopher. THE MNIST DATABASE of handwritten digits. *The Courant Institute of Mathematical Sciences*, 1998.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal Deep Learning. *Proceedings of The 28th International Conference on Machine Learning (ICML)*, 2011. ISSN 9781450306195. doi: 10.1145/2647868.2654931.

Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, Rene Vidal, and Ruzena Bajcsy. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *Proceedings of IEEE Workshop on Applications of Computer Vision*, 2013. ISBN 9781467350532. doi: 10.1109/WACV.2013.6474999.

G. Perry, E. T. Rolls, and S. M. Stringer. Continuous transformation learning of translation invariant representations. *Experimental Brain Research*, 204(2):255–270, 2010. ISSN 00144819. doi: 10.1007/s00221-010-2309-0.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning Structured Output Representation using Deep Conditional Generative Models. In C Cortes, N D Lawrence, D D Lee, M Sugiyama, and R Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 3483–3491. Curran Associates, Inc., 2015.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. pp. 1–12, 2017.

Udacity. Self-Driving Car: Annotated Driving Dataset, 2016. URL `https://github.com/udacity/self-driving-car/tree/master/annotations`.

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative Models of Visually Grounded Imagination. pp. 1–21, 2017. URL `http://arxiv.org/abs/1705.10762`.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep Variational Canonical Correlation Analysis. 1, 2016. URL `http://arxiv.org/abs/1610.03454`.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. pp. 1–6, 2017. URL `http://arxiv.org/abs/1708.07747`.

APPENDIX

1.2 Extension to three Modalities

The proposed, as well as approach by Suzuki et al. (2017), can be extended to multiple modalities $\mathcal{M}=\{a,b,c\}$. The conditional marginal log-likelihood of a can be written as

$$\log p(a|b,c) = \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}} + \mathrm{D}_{\mathrm{KL}}(q(z|\mathcal{M})\|p(z|\mathcal{M})) \geq \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}}. \tag{1}$$

1.2.1 JMVAE for three Modalities

The VI between a set of distributions $\mathcal{M}$ can be written as $-\mathbb{E}_{p(\mathcal{M})}\sum_{m\in\mathcal{M}}\log p(m|\mathcal{M}\setminus m)$, which leads to an expression of maximizing the ELBO of negative VI (c.f. Suzuki et al. (2017)). Following this approach, the log-likelihood $L_{3\mathrm{M}}$ can be expressed by the ELBOs, by utilizing Eq. 1, of their conditionals and KL divergence:

$$\begin{aligned}
L_{3\mathrm{M}} &= \log p(a|b,c) + \log(p(b|a,c)) + \log(p(c|b,c)) && (2)\\
&\geq \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{a}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{b}} + \mathcal{L}_{\widetilde{\mathrm{M}}_\mathrm{c}} && (3)\\
&\geq \mathcal{L}_{\widetilde{\mathrm{J}}} - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c)\|p(z|b,c)) && (4)\\
&\quad - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c)\|p(z|a,c)) - \mathrm{D}_{\mathrm{KL}}(q(z|a,b,c)\|p(z|b,c)) && (5)
\end{aligned}$$

with $\mathcal{L}_{\widetilde{\mathrm{J}}}$ being the joint ELBO of a joint probability $p(\mathcal{M})$ which expression is analog to Eq. **??**.

1.2.2 M^{2}VAE for three Modalities

Applying the proposed scheme to the joint log-likelihood of three modalities results in the following expression:

$$\begin{aligned}
&L_{3\mathrm{M}^2} && (6)\\
&= {}^{3}\!/\!{}_{3}\log p(a,b,c) = {}^{1}\!/\!{}_{3}\log p(a,b,c)^3 && (7)\\
&= {}^{1}\!/\!{}_{3}\log p(a,b,c)p(a,b,c)p(a,b,c) && (8)\\
&= {}^{1}\!/\!{}_{3}\log p(a,b)p(b,c)p(a,c)p(a|b,c)p(b|a,c)p(c|a,b) && (9)\\
&= {}^{1}\!/\!{}_{3}(\log(p(a,b)) + \log(p(b,c)) + \log(p(a,c)) && (10)\\
&\quad + \log p(a|b,c) + \log p(b|a,c) + \log p(c|a,b)) && (11)\\
&= {}^{1}\!/\!{}_{3}({}^{2}\!/\!{}_{2}(\log p(a,b) + \log p(b,c) + \log p(a,c)) + L_{3\mathrm{M}}) && (12)\\
&= {}^{1}\!/\!{}_{6}\big(\log p(a,b)^2 + \log p(b,c)^2 + \log p(a,c)^2\big) + {}^{L_{3\mathrm{M}}}\!/\!{}_{3} && (13)\\
&= {}^{1}\!/\!{}_{6}\big(L_{\mathrm{M}^2{}_{\mathrm{ab}}} + L_{\mathrm{M}^2{}_{\mathrm{bc}}} + L_{\mathrm{M}^2{}_{\mathrm{ac}}}\big) + {}^{1}\!/\!{}_{3}L_{3\mathrm{M}} && (14)
\end{aligned}$$

From here on, one can substitute all log-likelihoods given the expressions in Sec. **??** and **??**, to derive the ELBO $\mathcal{L}_{3\text{M}^2}$.

1.2.3 M²VAE Derivation

$$L_{\text{M}^2_{\mathcal{M}}} = \log p(\mathcal{M}) \overset{\text{mul. 1}}{=} {}^{|\mathcal{M}|}/_{|\mathcal{M}|} \log p(\mathcal{M}) \overset{\text{log. mul.}}{=} {}^{1}/_{|\mathcal{M}|} \log p(\mathcal{M})^{|\mathcal{M}|} \tag{15}$$

$$\overset{\text{Bayes}}{=} {}^{1}/_{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m) p(m|\mathcal{M} \setminus m) \tag{16}$$

$$\overset{\text{log. add}}{=} {}^{1}/_{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m) + \log p(m|\mathcal{M} \setminus m) \tag{17}$$

The expression $\sum_{m \in \mathcal{M}} \log p(m|\mathcal{M} \setminus m)$ is the general form of the marginal log-likelihood for the *variation of information* (VI), as introduced by Suzuki et al. (2017) for the JMVAE, for any set $\mathcal{M}$. Thus, it can be directly substituted with $L_{\text{M}_{\mathcal{M}}}$. The expression $\sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m)$ is the combination of all joint log-likelihoods of the subsets of $\mathcal{M}$ which have one less element. Therefore, this term can be rewritten as

$$\sum_{m \in \mathcal{M}} \log p(\mathcal{M} \setminus m) = \sum_{\widetilde{m} \in \widetilde{\mathcal{M}}} \log p(\widetilde{m}) \tag{18}$$

with $\widetilde{\mathcal{M}} = \{m | m \in \mathcal{P}(\mathcal{M}), |m| = |\mathcal{M}| - 1\}$ Finally, $\log p(\widetilde{m})$ can be substituted by $L_{\text{M}^2_{\widetilde{m}}}$ without loss of generality. However, it is worth noticing that substitution stops at the end of recursion and therefore, all final expressions $\log p(\widetilde{m}) \; \forall \; |\widetilde{m}| \equiv 1$ remain. □

1.2.4 Network Architecture

We designed all VAEs such that the latent space prior is given by a Gaussian with unit variance. Furthermore, all VAEs sample from a Gaussian variational distribution that is parametrized by the encoder networks. A summary of all architectures used in this paper can be seen in Tbl. 1. The reconstruction loss for calculating the evidence lower bound was performed by *binary cross-entropy* (BCE) for the e-MNIST and *root-mean-squared error* (RMS) for the MoG experiment.

Table 1: Various VAE architectures and optimizers for the e-MNIST and MoG experiments. um/mm stand for uni- and multi-modal while fc refers to fully-connected layers.

Issue	**VAE**	**Optimizer**		**VAE architecture**
e-MNIST	JMVAE-Z.	adam	encoder	fc 2x784-2x128-2x64-concat-64-2 (ReLU)
			decoder	fc 2x64-2x128-2x786 (tanh)
e-MNIST	tVAE	adam	um enc.	fc 784-128-64-2 (ReLU)
			mm enc.	fc 2x784-2x128-2x64-concat-64-2 (ReLU)
			decoder	fc 2x64-2x128-2x786 (tanh)
e-MNIST	M²VAE	adam	um enc.	fc 784-128-64-2 (ReLU)
			mm enc.	fc 2x784-2x128-2x64-concat-64-2 (ReLU)
			decoder	fc 2x64-2x128-2x786 (tanh)
MoG	JMVAE-Z.	rmsprop	encoder	fc 2x2-2x128-concat-64-2 (ReLU)
			decoder	fc 2x128-2x2 (tanh)
MoG	tVAE	rmsprop	um enc.	fc 2x2-2x128-2x2 (ReLU)
			mm enc.	fc 2x2-2x128-concat-64-2 (ReLU)
			decoder	fc 2x128-2x2 (tanh)
MoG	M²VAE	rmsprop	um enc.	fc 2-128-2 (ReLU)
			mm enc.	fc 2x2-2x128-concat-64-2 (ReLU)
			decoder	fc 2x128-2x2 (tanh)

Furthermore, the CVAE for training the e-MNIST dataset is designed as depicted in Tbl. 2.

Table 2: CVAE architecture for each dataset MNIST and fashion-MNIST. The label as one-hot-vector is concatenated after the convolution layers and fed into the fully-connected (fc) layers. For convolutional architectures the numbers in parenthesis indicate strides, while padding is always *same*.

	CVAE architecture
encoder	conv 1x2x2-64x2x2 (2)-64x3x3-64x3x3-concat label C-fc 128-2
decoder	concat label C-fc 128-deconv reverse of encoder (ReLU)