MultiVae: A Python package for Multimodal Variational Autoencoders on Partial Datasets.

Agathe Senellart[1,2,3¶]**, Clément Chadebec**[1,2,3]**, and Stéphanie Allassonnière**[1,2,3]

1 Université de Paris-Cité **2** Inria **3** Inserm ¶ Corresponding author

Summary

In recent years, there has been a major boom in the development of multimodal machine learning models. Among open topics, representation (fusion) and generation of multimodal data are very active fields of research. Recently, Multimodal Variational Autoencoders (VAEs) have been attracting growing interest for both tasks, thanks to their versatility, scalability, and interpretability as probabilistic latent variable models. They are also particularly interesting models in the *partially observed* setting, as most of them can be trained even with missing data. This last point makes them particularly interesting for the medical field, where available datasets are often incomplete (Antelmi et al., 2019; Lawry Aguila et al., 2023).

We present MultiVae, an open-source Python library for bringing together unified implementations of multimodal VAEs. It has been designed for easy and customizable use of these models on fully or partially observed data. This library also facilitates the development and benchmarking of new algorithms by integrating several benchmark datasets, a variety of evaluation metrics and tools for monitoring and sharing models.

Multimodal Variational Autoencoders

In Multimodal Machine Learning, two goals are generally targeted: (1) Learn a shared representation from multiple modalities; (2) Learn to generate one missing modality given the ones that are available.

Multimodal Variational Autoencoders aim at solving both issues at the same time. These models learn a latent representation z of all modalities in a lower dimensional common space and learn to *decode* z to generate any modality.

Let $X = (x_1, x_2, ...x_M)$ a sample with M modalities. In the VAE setting, we suppose that the generative process behind the observed data is the following:

$$z \sim p(z) \quad \text{and} \quad \forall 1 \le i \le M \quad x_i | z \sim p_\theta(x_i | z),$$ (1)

where $p(z)$ is a prior distribution that is often fixed, and $p_\theta(x_i | z)$ are called *decoders* and are parameterized by neural network. Typically, $p_\theta(x_i | z) = \mathcal{N}(x_i; \mu_\theta(z), \sigma_\theta(z))$ where $\mu_\theta, \sigma_\theta$ are neural networks. We aim to learn these *decoders* that translate z into the high dimensional data x_i. At the same time, we aim to learn an *encoder* $q_\phi(z|X)$ that maps observations to the latent space. $q_\phi(z|X)$ is also parameterized by a neural network. Derived from variational inference (Kingma & Welling, 2014), the VAE objective writes:

$$\mathcal{L}(X) = \mathbb{E}_{q_\phi(z|X)} \left(\sum_i \ln(p_\theta(x_i|z)) \right) - \text{KL}(q_\phi(z|X)|p(z)).$$

The first term is a reconstruction loss and the second term can be seen as a regularization term. A typical training step of a multimodal VAE consists in encoding a batch of samples

³⁷ with the encoder, reconstructing each modality with the decoders and taking a gradient step
³⁸ to optimize the loss $\mathcal{L}(X)$.

³⁹ A key differentiator of multimodal VAEs relies in the choice of the encoder $q_\phi(z|X)$. As
⁴⁰ illustrated in the figure below, they can be categorized into three main groups: *Aggregated*
⁴¹ *models* (Shi et al., 2019; Sutter et al., 2021; Wu & Goodman, 2018) use a mean or a
⁴² product operation to aggregate the information coming from all modalities, where *Joint models*
⁴³ (Senellart et al., 2023; Suzuki et al., 2016; Vedantam et al., 2018) use a neural network taking
⁴⁴ all modalities as input. Finally *coordinated models* (Tian & Engel, 2019; Wang et al., 2017)
⁴⁵ use different latent spaces but add a constraint term in the loss to force them to be similar.

Aggregated Models   Joint Models   Coordinated Models

⟷ : Similarity constraint

⁴⁶
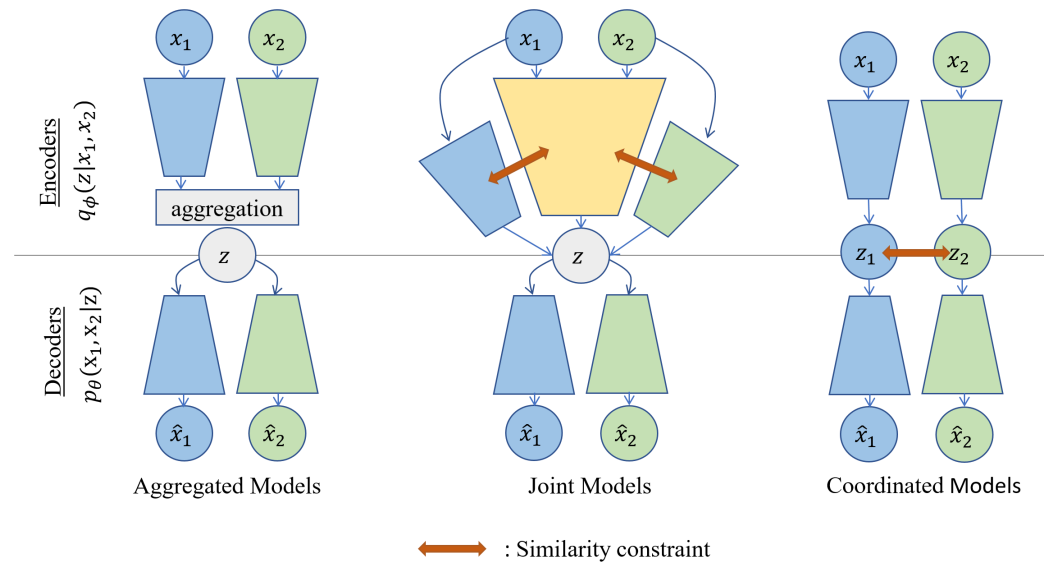
⁴⁷ The design of our library MultiVae was driven by the desire to implement all the approaches in
⁴⁸ a unified yet modular way.

⁴⁹ Notably, aggregated models offer a natural way of learning on incomplete datasets: for an
⁵⁰ incomplete sample X, the encoding z and the loss can be computed using only available
⁵¹ modalities. However, except in our library MultiVae, there does not exist an implementation
⁵² of these models that can be used on incomplete datasets in a straightforward manner. We
⁵³ propose a convenient way to handle missing modalities using *masks* in the loss computation of
⁵⁴ each aggregated model.

Data Augmentation

⁵⁶ Another application of VAEs is Data Augmentation (DA): from sampling new latent codes z and
⁵⁷ decoding them with trained models, *fully synthetic multimodal* samples can be generated to
⁵⁸ augment a dataset. This approach has been successfully used with unimodal VAEs to augment
⁵⁹ datasets for data-intensive deep learning applications (Chadebec et al., 2023). However, the
⁶⁰ use of similar sampling techniques with multimodal VAEs remains largely unexplored. In our
⁶¹ library, we provide a module `multivae.samplers` with popular sampling strategies to further
⁶² explore the generative abilities of these models.

Statement of Need

⁶⁴ Although multimodal VAEs have interesting applications in different fields, the lack of easy-to-
⁶⁵ use and verified implementations might hinder applicative research. With MultiVae, we offer
⁶⁶ unified implementations, designed to be accessible even for non-specialists. In order to propose
⁶⁷ reliable implementations, we reproduced, whenever possible, a key result from the original paper.

Some works similar to ours have grouped together model implementations: the Multimodal VAE Comparison Toolkit (Sejnova et al., 2024) includes 4 models and the Pixyz(Masahiro Suzuki & Matsuo, 2023) library contains 2 multimodal models. The closest work to ours and released while we were developing our library is `multi-view-ae` (Aguila et al., 2023), which contains a dozen of models. We compare in a summarizing table below, the different features of each work. Our library complements existing software: our API is quite different compared to previous work, the models implemented are not all the same, and for those we have in common, our implementation offers additional options. Indeed, for each model, we made sure to offer great flexibility on parameters' choices and to include all implementation details present in the original codes. Our library also offers additional features: **compatibility with incomplete data**, which we consider essential for real-life applications, **samplers** to boost the generative abilities of models, and a range of tools dedicated to research and development such **benchmark datasets** and **metrics**. We implement the most commonly used metrics in a modular way to easily evaluate any model.

List of Models and Features

In the Table below, we list available models and features, and compare to previous work. This symbol (✓*) indicates that the implementation includes additional options.

Models/ Features	Ours	(Aguila et al., 2023)	(Sejnova et al., 2024)
JMVAE(Suzuki et al., 2016)	✓*	✓	
MVAE(Wu & Goodman, 2018)	✓*	✓	✓
MMVAE(Shi et al., 2019)	✓*	✓	✓
MoPoE(Sutter et al., 2021)	✓*	✓	✓
DMVAE(Lee & Pavlovic, 2021)	✓	✓*	✓
MVTCAE(Hwang et al., 2021)	✓	✓	
MMVAE+(Palumbo et al., 2023)	✓*	✓	
CMVAE(Palumbo et al., 2024)	✓		
Nexus(Vasco et al., 2022)	✓		
CVAE(Kingma & Welling, 2014)	✓		
MHVAE(Dorent et al., 2023)	✓		
TELBO(Vedantam et al., 2018)	✓		
JNF(Senellart et al., 2023)	✓		
CRMVAE(Suzuki & Matsuo, 2023)	✓		
MCVAE(Antelmi et al., 2019)		✓	
mAAE		✓	
DVCCA(Wang et al., 2017)		✓	
DCCAE(Wang et al., 2015)		✓	
mWAE		✓	
mmJSD(Sutter et al., 2020)		✓	
gPoE(Lawry Aguila et al., 2023)		✓	
Support of Incomplete datasets	✓		
GMM Sampler	✓		
MAF Sampler, IAF Sampler	✓		
Metrics: {Likelihood, Coherences, FIDs, Reconstruction, Clustering}	✓		
Benchmark Datasets	✓		✓
Model sharing via Hugging Face	✓		

Code Quality and Documentation

Our code is available on Github (https://github.com/AgatheSenellart/MultiVae) and Pypi and we provide a full online documentation at (https://multivae.readthedocs.io/). Our code is unit-tested with a code coverage of 94%. The main features are illustrated through **tutorials** made available either as notebooks or scripts allowing users to get started easily. To further showcase how to use our library for research applications, we provide detailed *case studies* in the documentation.

Acknowledgements

References

Aguila, A. L., Jayme, A., Montaña-Brown, N., Heuveline, V., & Altmann, A. (2023). Multi-view-AE: A python package for multi-view autoencoder models. *Journal of Open Source Software*, *8*(85), 5093. https://doi.org/10.21105/joss.05093

Antelmi, L., Ayache, N., Robert, P., & Lorenzi, M. (2019). *Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data*. *97*, 302–311. https://proceedings.mlr.press/v97/antelmi19a.html

Chadebec, C., Thibeau-Sutre, E., Burgos, N., & Allassonnière, S. (2023). Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 2879–2896. https://doi.org/10.1109/TPAMI.2022.3185773

Dorent, R., Haouchine, N., Kogl, F., Joutard, S., Juvekar, P., Torio, E., Golby, A. J., Ourselin, S., Frisken, S., Vercauteren, T., Kapur, T., & Wells, W. M. (2023). Unified brain MR-ultrasound synthesis using multi-modal hierarchical representations. In *Medical image computing and computer assisted intervention − MICCAI 2023* (pp. 448–458). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43999-5_43

Hwang, H., Kim, G.-H., Hong, S., & Kim, K.-E. (2021). Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, *34*, 12194–12207.

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes*. arXiv. http://arxiv.org/abs/1312.6114

Lawry Aguila, A., Chapman, J., & Altmann, A. (2023). *Multi-modal variational autoencoders for normative modelling across multiple imaging modalities*. 425–434. https://doi.org/10.1007/978-3-031-43907-0_41

Lee, M., & Pavlovic, V. (2021). *Private-shared disentangled multimodal VAE for learning of latent representations*. 1692–1700. https://doi.org/10.1109/CVPRW53098.2021.00185

Masahiro Suzuki, T. K., & Matsuo, Y. (2023). Pixyz: A python library for developing deep generative models. In *Advanced Robotics* (No. 0; Vol. 0, pp. 1–16). Taylor & Francis. https://doi.org/10.1080/01691864.2023.2244568

Palumbo, E., Daunhawer, I., & Vogt, J. E. (2023). MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=sdQGxouELX

129 Palumbo, E., Manduchi, L., Laguna, S., Chopard, D., & Vogt, J. E. (2024). *Deep generative*
130 *clustering with multimodal diffusion variational autoencoders*. https://openreview.net/
131 forum?id=k5THrhXDV3

132 Sejnova, G., Vavrecka, M., Stepanova, K., & Taniguchi, T. (2024). *Benchmarking multimodal*
133 *variational autoencoders: CdSprites+ dataset and toolkit*. https://arxiv.org/abs/2209.
134 03048

135 Senellart, A., Chadebec, C., & Allassonnière, S. (2023). Improving multimodal joint varia-
136 tional autoencoders through normalizing flows and correlation analysis. *arXiv Preprint*
137 *arXiv:2305.11832*.

138 Shi, Y., Siddharth, N., Paige, B., & Torr, P. H. S. (2019). Variational Mixture-of-Experts
139 Autoencoders for Multi-Modal Deep Generative Models. *arXiv:1911.03393 [Cs, Stat]*.
140 http://arxiv.org/abs/1911.03393

141 Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2020). Multimodal generative learning utilizing
142 jensen-shannon-divergence. *CoRR*, *abs/2006.08242*. https://arxiv.org/abs/2006.08242

143 Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2021). Generalized Multimodal ELBO. *ICLR*.

144 Suzuki, M., & Matsuo, Y. (2023). *Mitigating the limitations of multimodal VAEs with*
145 *coordination-based approach*. https://openreview.net/forum?id=Rn8u4MYgeNJ

146 Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint Multimodal Learning with Deep
147 Generative Models. *arXiv:1611.01891 [Cs, Stat]*. http://arxiv.org/abs/1611.01891

148 Tian, Y., & Engel, J. (2019). Latent Translation: Crossing Modalities by Bridging Generative
149 Models. *ArXiv*.

150 Vasco, M., Yin, H., Melo, F. S., & Paiva, A. (2022). Leveraging hierarchy in multimodal
151 generative models for effective cross-modality inference. *Neural Networks*, *146*, 238–255.

152 Vedantam, R., Fischer, I., Huang, J., & Murphy, K. (2018). Generative Models of Visually
153 Grounded Imagination. *arXiv:1705.10762 [Cs, Stat]*. http://arxiv.org/abs/1705.10762

154 Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation
155 learning. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on*
156 *machine learning* (Vol. 37, pp. 1083–1092). PMLR. https://proceedings.mlr.press/v37/
157 wangb15.html

158 Wang, W., Yan, X., Lee, H., & Livescu, K. (2017). *Deep Variational Canonical Correlation*
159 *Analysis*. https://doi.org/10.48550/arXiv.1610.03454

160 Wu, M., & Goodman, N. (2018). Multimodal Generative Models for Scalable Weakly-Supervised
161 Learning. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.
162 neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html