

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

13 August 2017

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 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 learn even with missing data. This last point makes them particularly interesting for research fields where missing data are commonplace: in medical research for example.

In this article, we present MultiVae, an open-source Python library designed to bring together unified implementations of multimodal variational autoencoders models. It has been designed for easy, 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 popular datasets, 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 available modalities.

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 $X = (x_1, x_2, \dots, x_M)$ contain M modalities. In the VAE setting, we suppose that the generative process behind the observed data is the following:

$$z \sim p(z) \quad \forall 1 \leq i \leq M, x_i | z \sim p_\theta(x_i | z) \quad (1)$$

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 Networks. 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 map the multimodal observation to the latent space. $q_\phi(z|X)$ is also parameterized by a neural network. Derived from variational inference theory, the VAE objective writes:

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

A simple interpretation of this objective is to see that the first term is a reconstruction loss and the second term is a regularization term that avoids overfitting. A typical training of a multimodal VAE consists in encoding the data with the encoder, reconstructing each modality with the decoders and take a gradient step to optimize the loss $\mathbb{J}(X)$.

Most multimodal VAEs differ in how they construct the encoder $q_\phi(z|x_1, \dots, x_M)$. In Figure ?? we summarize several approaches: Aggregated models use a mean or a product operation to aggregate the information coming from all modalities, where joint models uses a neural network taking all modalities as input. Finally coordinated models uses different latent spaces but add a constraint term in the loss to force them to be similar.

Aggregated models have a natural way of learning on incomplete datasets: for an incomplete sample X , we use only the available modalities to encode the data and compute the loss $\mathbb{J}(X)$. However, except in MultiVae, there doesn't exist an implementation of these models that can be used on incomplete datasets in a straightforward manner.

Recent extensions of multimodal VAEs include additional terms to the loss, multiple or hierarchical latent spaces to more comprehensively describe the multimodal data.

Another application of these models is data augmentation: from sampling latent codes z and decoding them, new synthetic multimodal samples can be generated. Data augmentation has been proven useful in many deep learning applications. In MultiVae we propose different ways of sampling latent codes z to further explore the generative abilities of these models.

Statement of need

Although multimodal VAEs have interesting applications in different fields, the lack of easy-to-use and verified implementations might hinder applicative research. With MultiVae, we offer unified implementations, designed to be easy to use by non-specialists and even on incomplete data. To this end, we offer online documentation and tutorials. In order to propose reliable implementations

of each method, we tried to reproduce, whenever possible, a key result from the original paper. Some works similar to ours have grouped together model implementations: the Multimodal VAE Comparison Toolkit includes 4 models and the Pixyz library groups 2 multimodal models. The work closest to ours and developed in parallel is the multi-view-ae library, which contains a dozen of models. Nevertheless, we are convinced that our library complements what already exists: our API is quite different, the models implemented are not all the same, and for those we have in common, our implementation offers additional parameterization options. Indeed, for each model, we’ve made sure to offer great flexibility on parameters and to include all implementation details present in the original codes that boost results. What’s more, our library offers numerous additional features: compatibility with incomplete data, which we consider essential for real-life applications, and a range of tools dedicated to the research and development of new algorithms: benchmark datasets, metrics modules and samplers, for testing and analyze models. Our library also supports distributed training and straightforward model sharing via HuggingFace Hub. In this way, our work complements existing work and addresses different needs.

Mathematics

Single dollars (\$) are required for inline mathematics e.g. $f(x) = e^{\pi/x}$

Double dollars make self-standing equations:

$$\Theta(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{else} \end{cases}$$

You can also use plain \LaTeX for equations

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{i\omega x} dx \tag{2}$$

and refer to Equation 2 from text.

Citations

Citations to entries in paper.bib should be in rMarkdown format.

If you want to cite a software repository URL (e.g. something on GitHub without a preferred citation) then you can do it with the example BibTeX entry below for @fidgit.

For a quick reference, the following citation commands can be used: - @author:2001 -> “Author et al. (2001)” - [@author:2001] -> “(Author et al., 2001)” - [@author1:2001; @author2:2001] -> “(Author1 et al., 2001; Author2 et al., 2002)”

Figures

Figures can be included like this: `Caption for example figure.` and referenced from text using `section` .

Figure sizes can be customized by adding an optional second parameter: `Caption for example figure.`

Acknowledgements

We acknowledge contributions from Brigitta Sipocz, Syrtis Major, and Semyeong Oh, and support from Kathryn Johnston during the genesis of this project.

References