ETH zürich

medical data science

Multimodal Generative Learning on the MIMIC-CXR Database

Hendrik Klug[1], Thomas M. Sutter[2], Julia E. Vogt[2]
[1]**Department of Electrical Engineering, ETH Zurich**; [2] **Department of Computer Science, ETH Zurich**
klugh@ethz.ch, {thomas.sutter, julia.vogt}@inf.ethz.ch

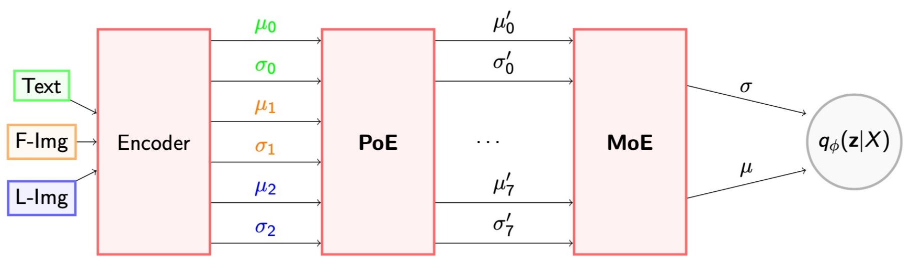
1. Introduction

Goals:

- Applying and evaluating a method for multimodal, unsupervised and generative learning on challenging medical data from the MIMIC-CXR database
- Learning a joint embedding of multiple data types
- Handling of missing data

2. Method Overview

Merging embeddings of multiple data types into one joint embedding is still an open problem. We use the MoPoE method from Sutter et al. [1], which is a combination of the PoE from Wu & Goodman [2] and the MoE from Shi et al. [3].



$$\log\ q_\theta(X) \geq E_{q_\phi(z|X)}[log\ q_\theta(X|z)] - KLD(q_\phi(z|X)|q_\theta(z))$$

With: $q_\phi(z|X) = \boldsymbol{MoE}(\{\widetilde{q_\phi} \forall \mathbb{X}_k \in \mathcal{P}(\mathbb{X})\}) = \frac{1}{2^3} \sum_{\mathbb{X}_k \in \mathcal{P}(\mathbb{X})} \widetilde{q_\phi}(z|\mathbb{X}_k)$

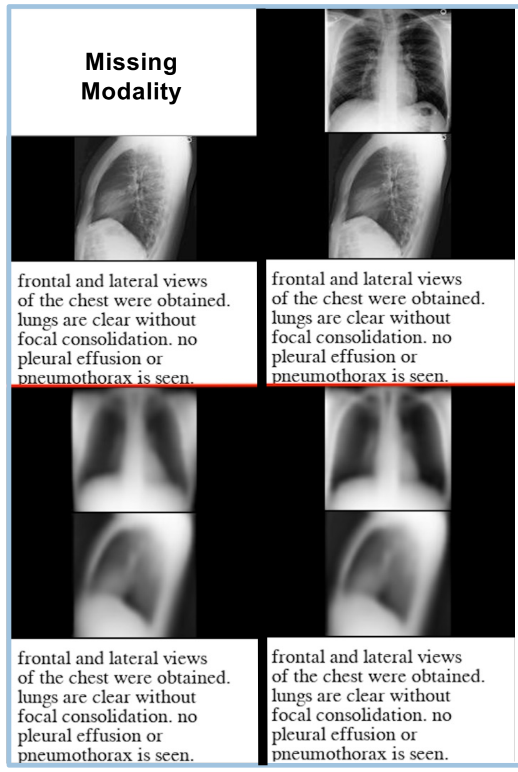
and: $\widetilde{q_\phi}(z|\mathbb{X}_k) = \boldsymbol{PoE}\left(\{q_{\phi_j} \forall x_j \in \mathbb{X}_k\}\right) = \prod_{x_j \in \mathbb{X}_k} q_{\phi_j}(z|x_j)$

3. Evaluation of Latent Representation Quality

We evaluate the quality of the latent representation for each subset of modalities by verifying if a linear classifier can separate between encoded samples with or without any pathology. We report the *mean average precision* over the test set for each subset (**F**: frontal image, **L**: lateral image, **T**: text report).

MODEL	F	L	T	L,F	F,T	L,T	L,F,T
MoPoE	0.467	0.460	0.473	0.476	0.493	0.475	**0.494**
Random				0.235			

4. Conditioned Generation



Examples of generated samples.
On the left, the L and T modality are given to the model as input.
On the right, all modalities (F, L and T) are given as input. The samples above the red line are the input samples and those below are generated.

5. Method Details

- We create a binary label "Finding", which indicates if a sample presents any pathology in the MIMIC-CXR database. This gives 14529 positive and 47218 negative samples.
- We use ResNet type architectures for all encoders and decoders.
- We use a word encoding for the text:

$$\text{"Heart size is normal."} \rightarrow [0,1,2,3] \rightarrow \text{MoPoE} \rightarrow \begin{bmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{bmatrix} \rightarrow [0,1,2,3] \rightarrow \text{"Heart size is normal."}$$

6. Results and Discussion

We provide a useful baseline for multimodal, unsupervised and generative methods on challenging medical data for real world applications.

We highlight challenges that can be addressed in future work:

- Features that are needed to classify for pathologies are lost due to the blurriness of the generated samples.
- The separability of the latent representation could be leveraged in a better way by using more advanced methods than linear classification.
- We use basic encoder and decoder architectures. The usage of more ad hoc architectures could further improve the results.

References

1. Sutter, Thomas M, Imant Daunhawer, and Julia E Vogt (2020). "Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence". arXiv preprint arXiv:2006.08242
2. Wu, Mike and Noah Goodman (2018). "Multimodal generative models for scalable weakly-supervised learning". Advances in Neural Information Processing Systems, pp. 5575–5585.
3. Shi, Yuge et al. (2019). "Variational mixture-of-experts autoencoders for multi-modal deep generative models". Advances in Neural Information Processing Systems, pp. 15718–15729.