Multimodal Generative Learning on the MIMIC-CXR Database

A presentation of my semester project

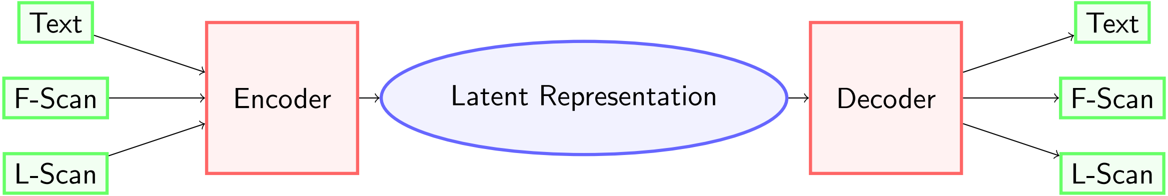
Hendrik Klug

Institute for Electrical Engineering, ETH

2021-02-23

In this work, we applied a method for self-supervised, multimodal and generative training from [3] on the MIMIC-CXR Database [1].

The General Idea



Multimodal, Unsupervised Generative Learning On Medical Data

- No need for labeled data
- Can extract features from multiple modalities
- Can generate *coherent* samples from one input modality

The Mixture-of-Products-of-Experts-VAE

Combination of:

- The Product-of-Experts (PoE) from [4]
- The Mixture-of-Experts (MoE) from [2]

Both differ in their choice of the joint posterior approximation functions.

The PoE-VAE

- Uses a geometric mean: the joint posterior is a product of individual posteriors

$$q_{\Phi}(z|x_{1:M}) = \prod_{m} q_{\Phi_m}(z|x_m) \tag{1}$$

- Results in a good approximation of the joint distribution but struggles in optimizing the individual experts.

The MoE-VAE

- Uses an arithmetic mean

$$q_{\Phi}(z|x_{1:M}) = \sum_{m} \alpha_m \cdot q_{\Phi_m}(z|x_m) \tag{2}$$

- Optimizes individual experts well but is not able to learn a distribution that is sharper than any of its experts.

The Mixture-of-Products-of-Experts-VAE

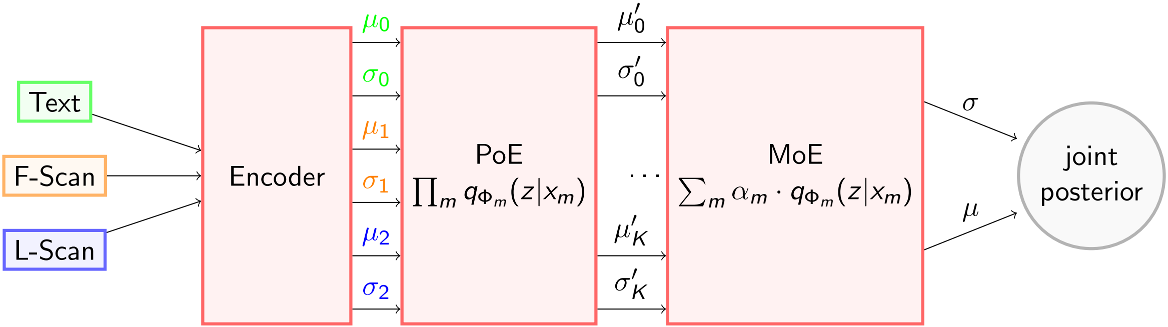
The generalized multimodal ELBO utilizes the PoE to get the posterior approximation of a subset $\mathbb{X}_k \in \mathcal{P}(\mathbb{X})$:

$$\tilde{q}_{\phi}(\mathbf{z}|\mathbb{X}_k) = PoE(\{q_{\phi_j}(\mathbf{z}|\mathbf{x}_j) \forall \mathbf{x}_j \in \mathbb{X}_k\}) \propto \prod_{\mathbf{x}_j \in \mathbb{X}_k} q_{\phi_j}(\mathbf{z}|\mathbf{x}_j) \tag{3}$$

And the MoE to get the joint posterior:

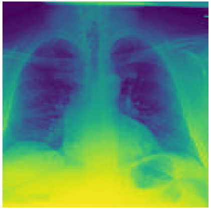
$$q_{\phi}(\mathbf{z}|\mathbb{X}) = \frac{1}{2^3} \sum_{\mathbf{x}_k \in \mathbb{X}} \tilde{q}_{\phi}(\mathbf{z}|\mathbb{X}_k) \tag{4}$$

Frame Title

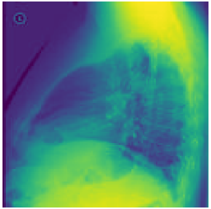

Text
F-Scan
L-Scan
Encoder
μ_0
σ_0
μ_1
σ_1
μ_2
σ_2
PoE
$\prod_m q_{\Phi_m}(z|x_m)$
μ'_0
σ'_0
$\dots$
μ'_K
σ'_K
MoE
$\sum_m \alpha_m \cdot q_{\Phi_m}(z|x_m)$
σ
μ
joint posterior

1. Implemented word encoding
2. tested image size
3. tested beta
4. tested class dim

Lateral view



Frontal view



Text report

Heart size is normal. Aorta is tortuous. Decrease in lung volume. However, the Lungs are clear. There is no pleural effusion or pneumothorax.

[1] Alistair EW Johnson et al. “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs”. In: *arXiv preprint arXiv:1901.07042* (2019).

[2] Yuge Shi et al. “Variational mixture-of-experts autoencoders for multi-modal deep generative models”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 15718–15729.

[3] Thomas M Sutter, Imant Daunhawer and Julia E Vogt. “Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence”. In: *arXiv preprint arXiv:2006.08242* (2020).

[4] Mike Wu and Noah Goodman. “Multimodal generative models for scalable weakly-supervised learning”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 5575–5585.