ETH zürich

UC Irvine

D-INFK

Unity by Diversity:
Improved Representation Learning in Multimodal VAEs

Thomas M. Sutter
ETH Zurich

medical data science

Collaborators

Yang Meng
UC Irvine

Andrea Agostini
ETH zürich

Daphné Chopard
ETH zürich

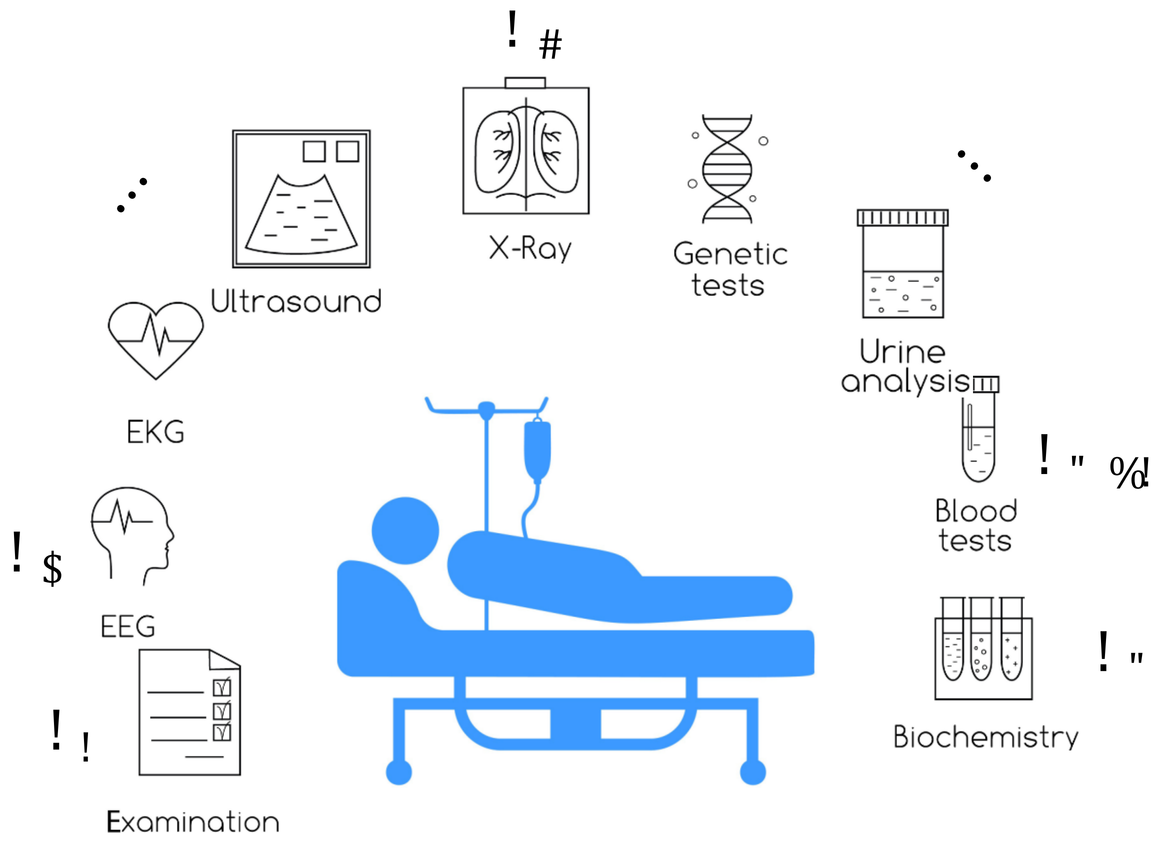
Norbert Fortin
UC Irvine

Julia Vogt
ETH zürich

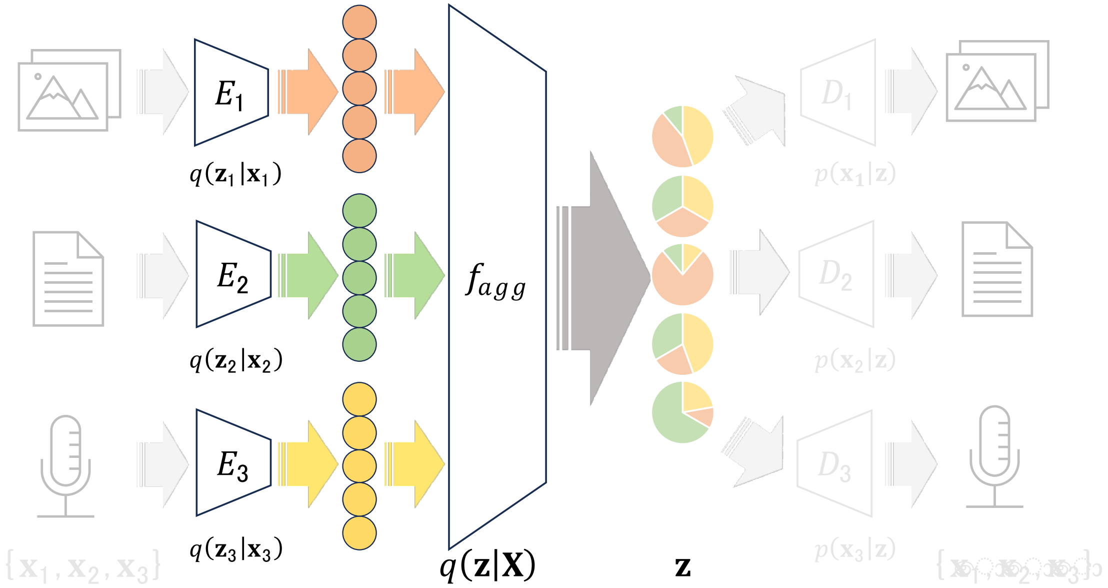
Bahbak Shahbaba
UC Irvine

Stephan Mandt
UC Irvine

Multimodal Learning



Multimodal Variational Autoencoders



$$\mathcal{E}(\boldsymbol{X}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X} \mid \boldsymbol{z}) - \log \frac{q_\phi(\boldsymbol{z} \mid \boldsymbol{X})}{p_\theta(\boldsymbol{z})}\right]$$

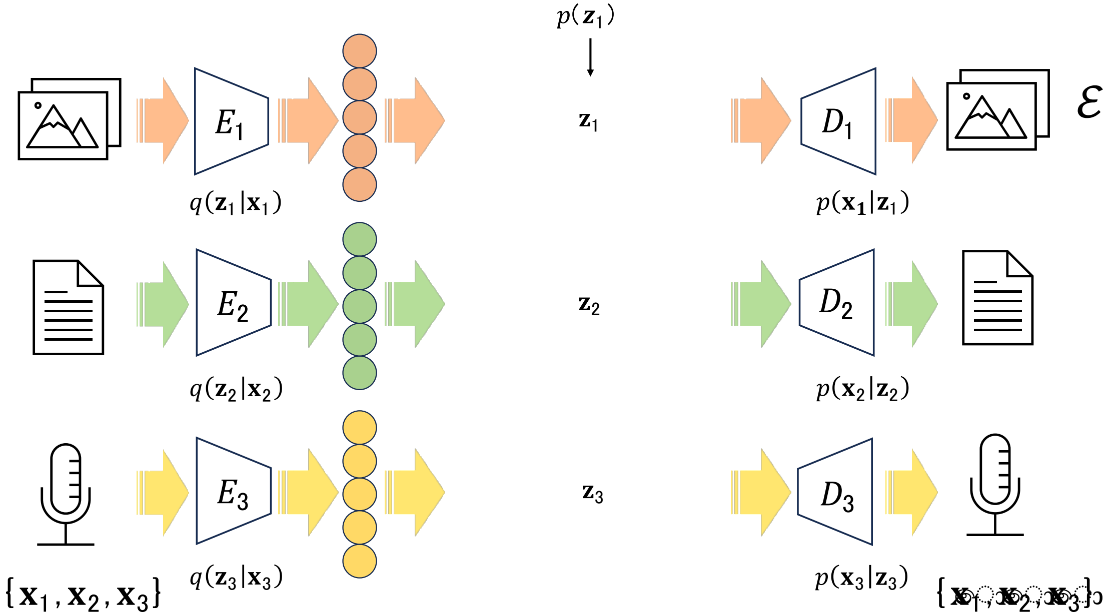
$$\mathcal{E}(\boldsymbol{X}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{X})}\left[\sum_{m=1}^{M} \log p_\theta(\boldsymbol{x}_m \mid \boldsymbol{z}) - \log \frac{q_\phi(\boldsymbol{z} \mid \boldsymbol{X})}{p_\theta(\boldsymbol{z})}\right]$$

- Multimodal sample: $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$
- Encoder (Posterior Approximation): $q_\phi(\boldsymbol{z} \mid \boldsymbol{X})$
- Decoder (Conditional Generation): $p_\theta(\boldsymbol{X} \mid \boldsymbol{z})$

- Multimodal fusion via joint posterior approximation: $q_\phi(\boldsymbol{z} \mid \boldsymbol{X})$
    - Product of Experts [1]
    - Mixture of Experts [2]
    - Mixture of Product of Experts [3]

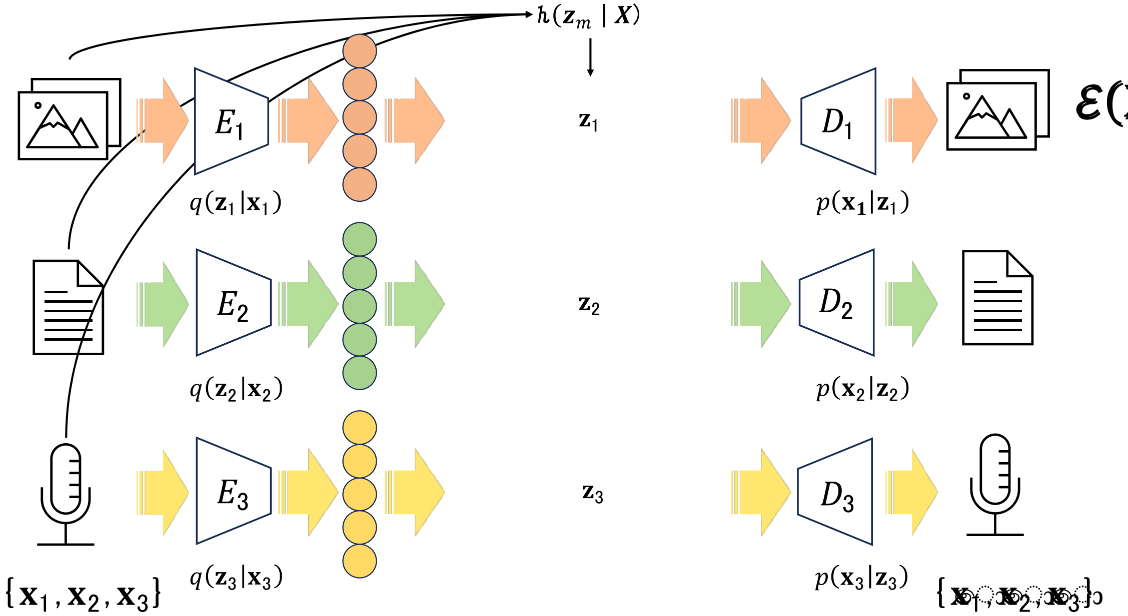
1. Wu and Goodman, "Multimodal Generative Models for Scalable Weakly-Supervised Learning", Neurips 2018
2. Shi et al., "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models", Neurips 2019
3. Sutter et al., "Generalized Multimodal ELBO", ICLR 2021
4. Daunhawer et al., "On the limitations of multimodal VAEs", ICLR 2022

Unimodal VAEs



$$\mathcal{E}(\mathbf{X}) = \sum_{m=1}^{M} \mathrm{E}_{q_\phi(\mathbf{z}_m|\mathbf{x}_m)}\left[\log p_\theta(\mathbf{x}_m \mid \mathbf{z}_m) - \log \frac{q_\phi(\mathbf{z}_m \mid \mathbf{x}_m)}{p_\theta(\mathbf{z}_m)}\right]$$

Multimodal Variational Mixture of Experts Prior



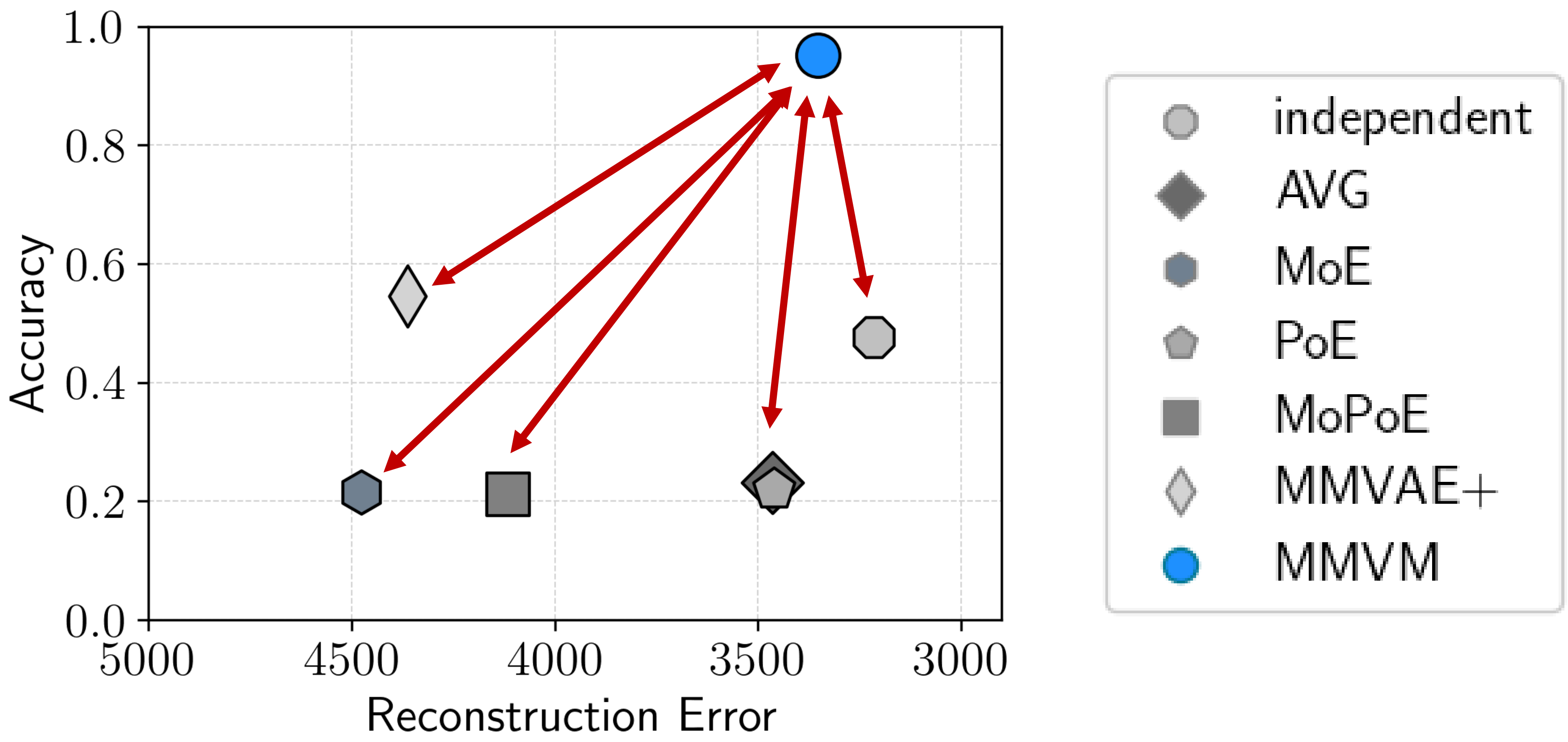
$$\mathcal{E}(X) = \sum_{m=1}^{M} E_{q_\phi(z_m|x_m)}\left[\log p_\theta(x_m \mid \ldots\right]$$

We provide a detailed derivation and analysis of the proposed MMVM in the paper!

...xture of Experts

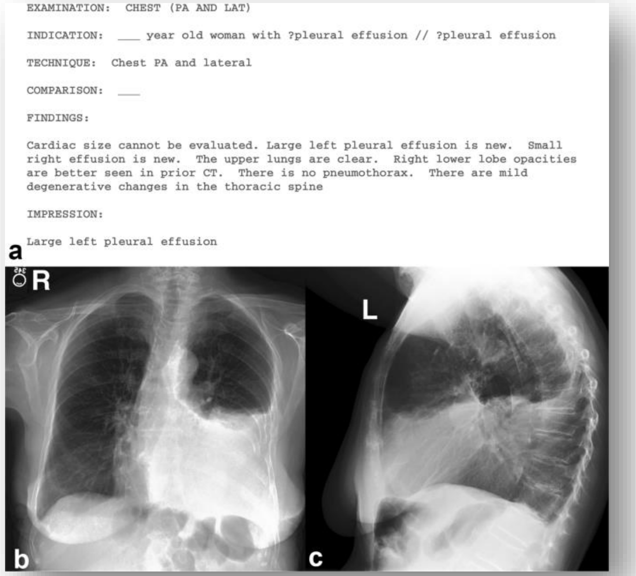
$$h(z_m \mid X) = \frac{1}{M} \sum_{\tilde{m}=1}^{M} q_\phi(z_m \mid x_{\tilde{m}})$$

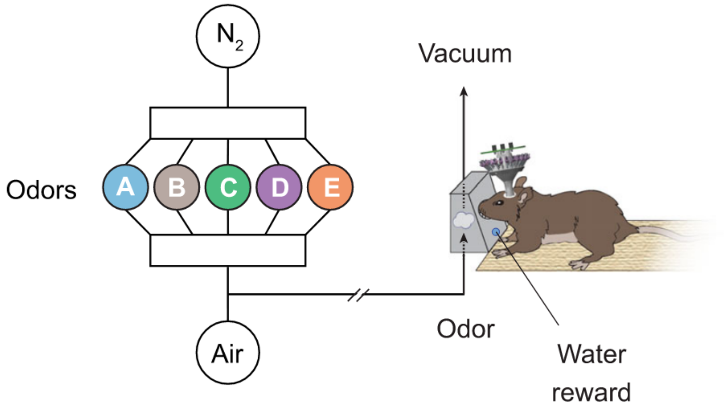
Benchmark Experiments
Translated PolyMNIST



Summary

- Novel Multimodal Learning objective that leverages *soft-sharing* instead of fusion
- Multimodal Variational Mixture of Experts prior (MMVM): we show the optimality of the chosen prior distribution
- Strong results on benchmark experiments and real-world datasets
- To follow: extending the proposed objective to additional architectures learning objectives





Visit us at our poster!

Friday, 13 Dec at 11 am

Thomas M. Sutter
Postdoc
thomas.sutter@inf.ethz.ch

ETH Zürich
Medical Data Science

https://thomassutter.github.io/