Provable concept learning for interpretable predictions

Armeen Taeb[1], Nicolo Ruggeri[2,3], Carina Schnuck[2], Fanny Yang[2]
[1]Seminar for Statistics, ETH Zurich; [2]Department of Computer Science, ETH Zurich; [3]Max-Planck-Institute Tubingen

Interpretable predictions for science

One goal of machine learning for science:

- a model producing interpretable predictions that uses
- learned concepts close to the 'true' mechanism

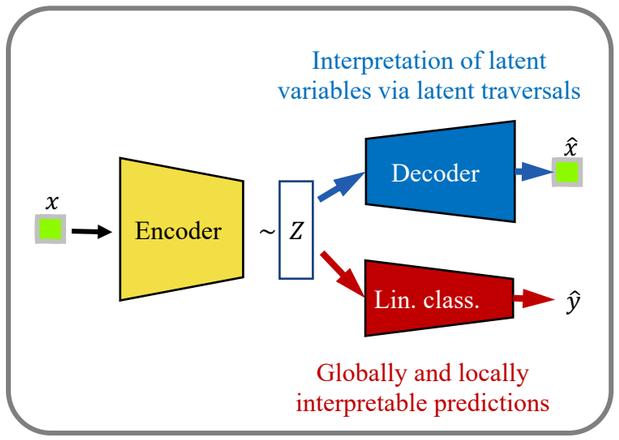
Our contribution: CLAP, a VAE-based classifier

- that provably learns unlabeled ground-truth concepts
- uses these learned features to obtain optimal predictions

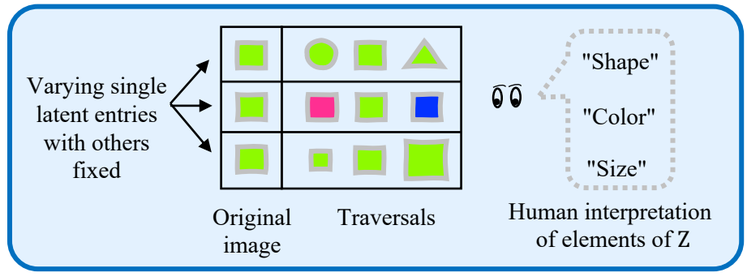
Comparison with related work

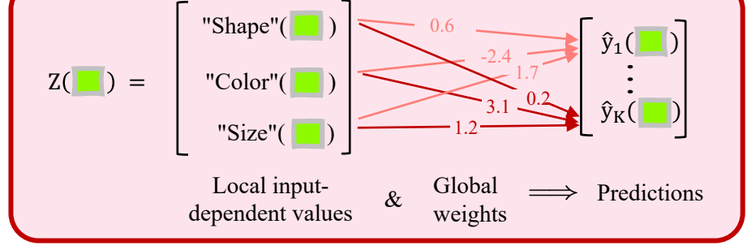
Properties	Post-hoc explanations			Inherently interpretable	
	pixel attribution+ counterfactuals	pre-defined concepts	StyleGANs	existing VAEs/ autoencoders	*CLAP*
Learning visually distinct features	×	×	✓	✓*	✓
Global importance of predictive features	×	✓	×	×	✓
Guarantees: concept learning+prediction	×	×	×	×	✓

C(oncept) L(earning) a(nd) P(rediction)

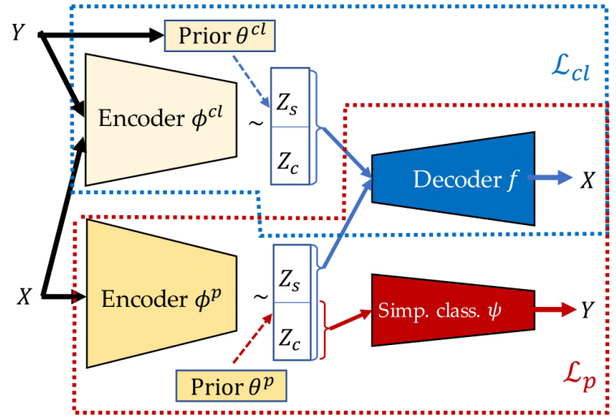


Trained predictive VAE



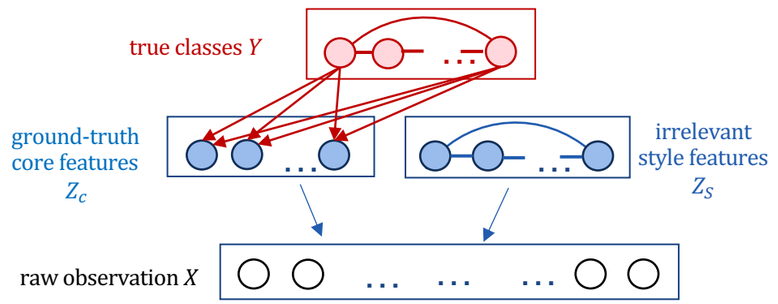


Training CLAP



$$\max \hat{\mathbb{E}}[\underbrace{\mathcal{L}_{CL}(\phi^{cl}, \theta^{cl}, f; X, Y)}_{\text{ELBO for concept learning VAE}} + \underbrace{\mathcal{L}_{P}(\phi^{p}, \theta^{p}, f, \psi; X, Y)}_{\text{ELBO for prediction VAE (used during test time)}}] - \underbrace{\lambda\rho(f, \psi)}_{\text{sparsity regularizer}}$$

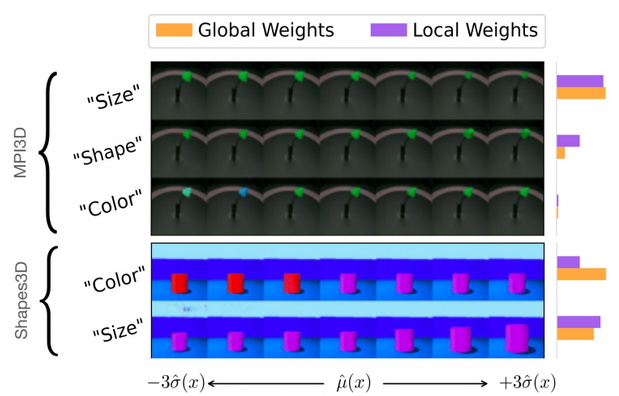
Theoretical guarantees (informal)



Under above data-generating and further regularity and heterogeneity assumptions, in the infinite data limit:

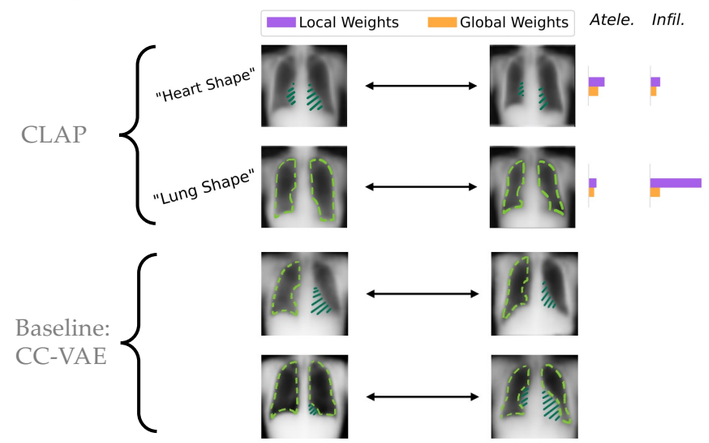
- CLAP identifies Z_c up to permutation and scaling
- CLAP learns the optimal classifier that uses Z_c

Example I: CLAP on toy datasets



Recovers ground-truth concepts + 99% class. accuracy!

Example II: CLAP on the Chest–Xray dataset



At similar accuracy 90%, CLAP achieves better disentanglement & indicates global and local feature importance

Work in progress

- Using GANs for sharper reconstructions
- Aggregate data across multiple sources
- Incorporating dependence between style features and Y