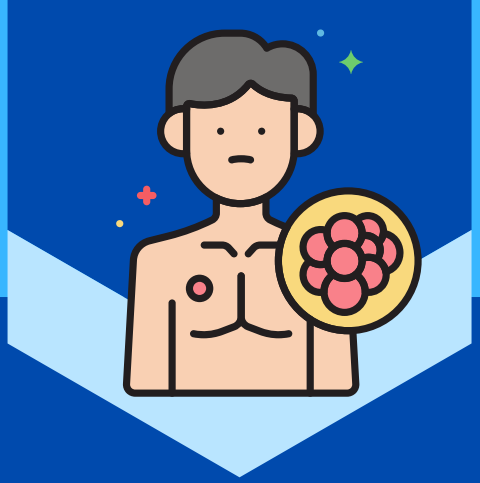


Benchmarking VAE latent features in downstream tasks for cancer related predictions

1. Background

Variational Autoencoder (VAE)

- Lower dimensional latent space with good generative properties
- Variants: VanillaVEA, β-VAE, β-TCVAE, Categorical VAE, NoVAE

Datasets (The Cancer Genome Atlas)

- Gene expression - RNA Sequence
- Curated clinical data

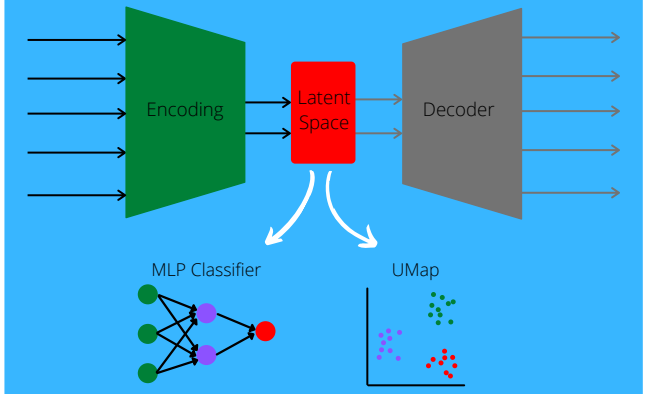
2. Research question

Benchmarking VAE latent features in downstream tasks for cancer related predictions.

- Prediction accuracies using the VAEs' latent features
- Analyzing VAE learning capabilities

3. Method

- Normalizing input data and taking the 5000 most variable features for the most common cancer types
- Using a 70/30% split for training the VAE and encoding both 70% and 30% once the model is trained to obtain latent features



MLP Classifier

- Encoded 70% used for training
- Encoded 30% used for predictions
- Prediction accuracy scores:
  - Mean of 10 runs
  - Standard deviation of 10 runs

Tasks:

- Cancer stages
- Survival time
  - Divided into 3 classes
- Cancer types

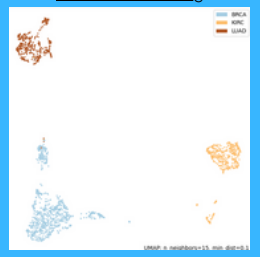
4. Results

- Predictive model is mostly guessing
  - Most accuracies lie in the 35-40% for predictions with 3 classes

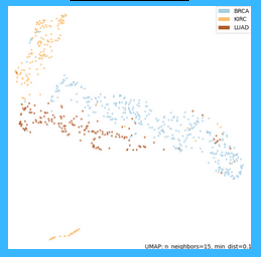
	NoVAE	VanillaVAE	β-VAE	β-TCVAE	CatVAE
Mean	40.05	35.94	36.63	37.79	39.87
SD	7.85	1.45	2.07	3.15	2.20

- VAE models unable to distinguish between classed when faced with new data

VanillaVAE Training



VanillaVAE Test



- Results seen are for the cancer types task, but are similar for the other tasks
- VAEs seem to be acting like regular autoencoder
  - Attempting to just recreate the data instead of learning underlyning probability distribution
  - Changing β term has almost no effect outcome
  - The higher the learning rate, the more the classes are intertwined in UMaps

Responsible professor: Marcel Reinders

Other Supervisors: Stavros Makrodimitris, Mostafa elTager, Tamim Abdelaal, Mohammed Charrout

TU Delft

Boris van Groeningen
29-06-2021
b.vangroeningen@student.tudelft.nl