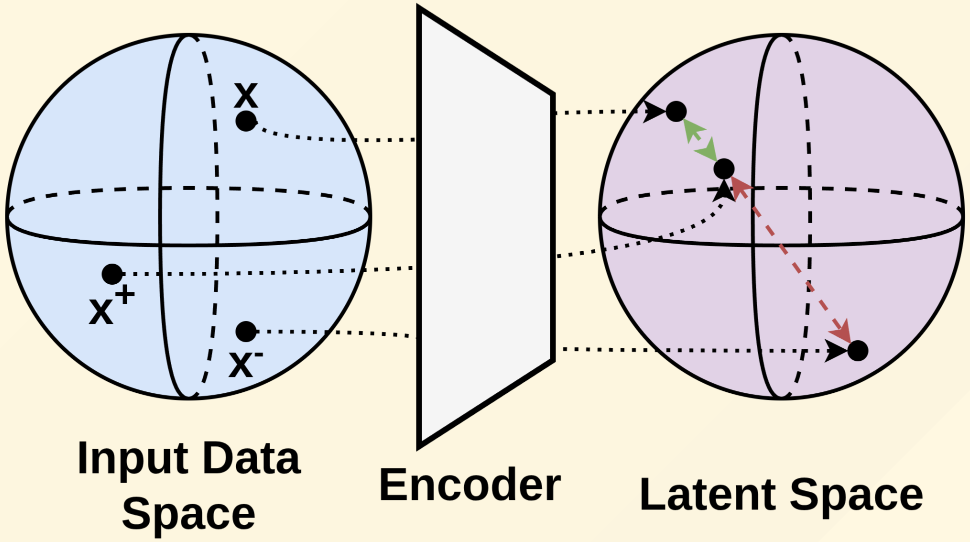
A Contrastive Approach to Weight Space Learning

DJ Swanevelder

Supervised by: Ruan van der Merwe (Bytefuse)

Context

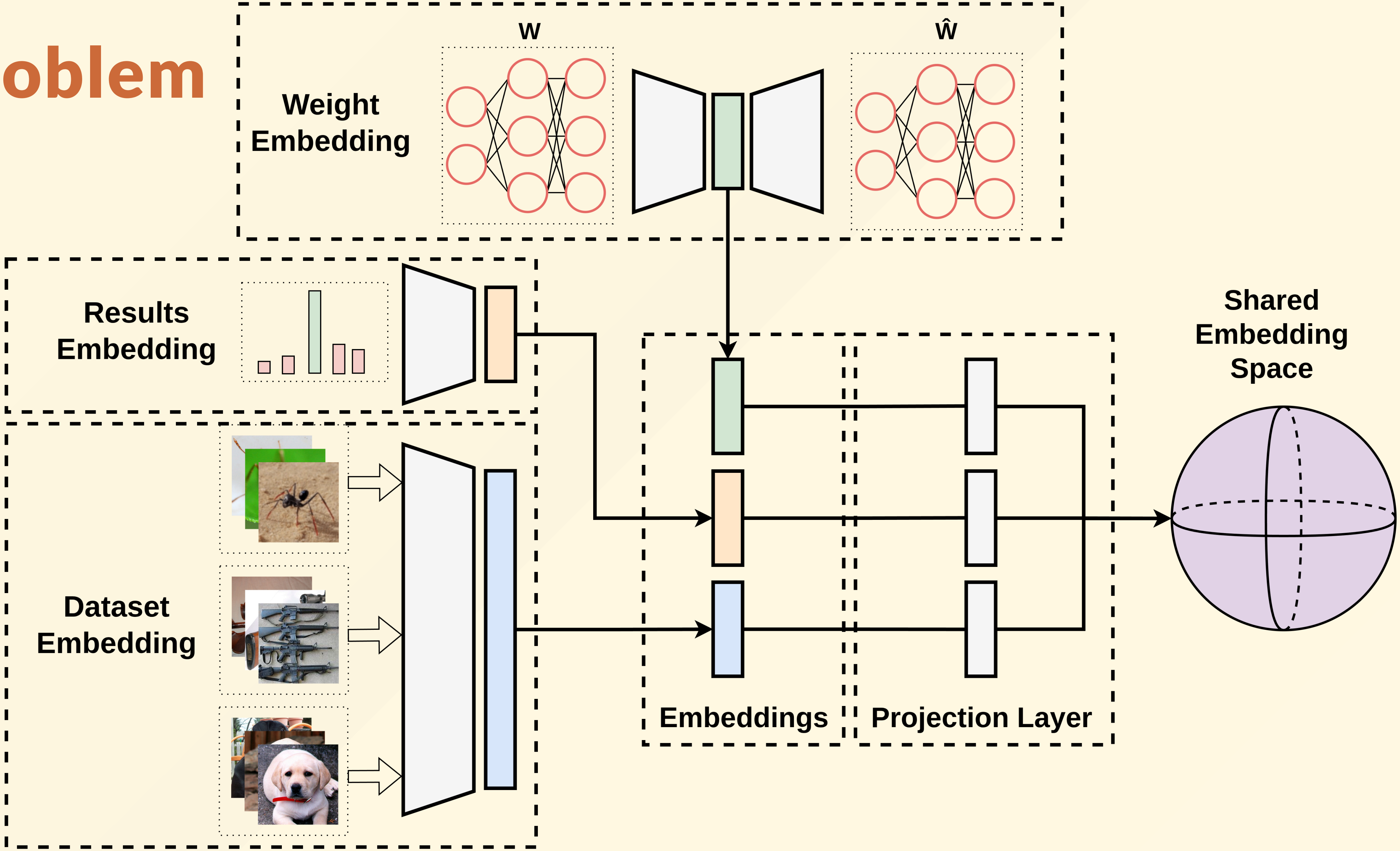
Contrastive Learning



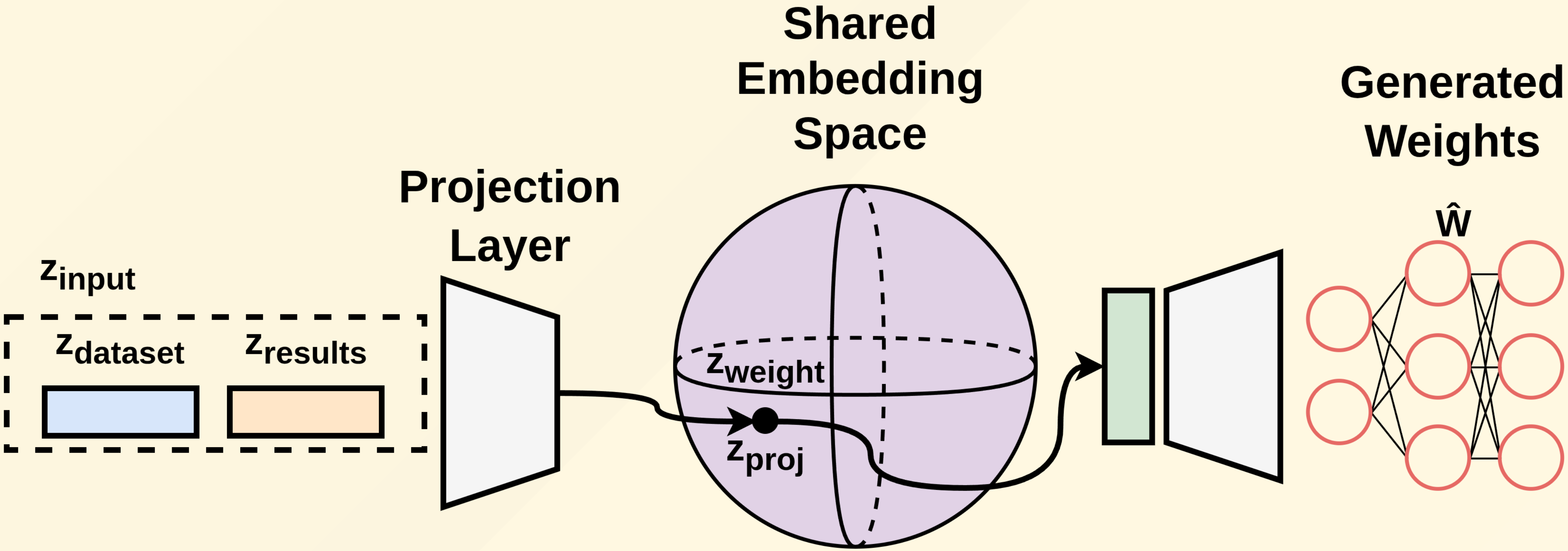
Weight Space Learning

- 1 point → 1 model
- Discriminative
  - Predict model properties
- Generative
  - High-performing Model Weights

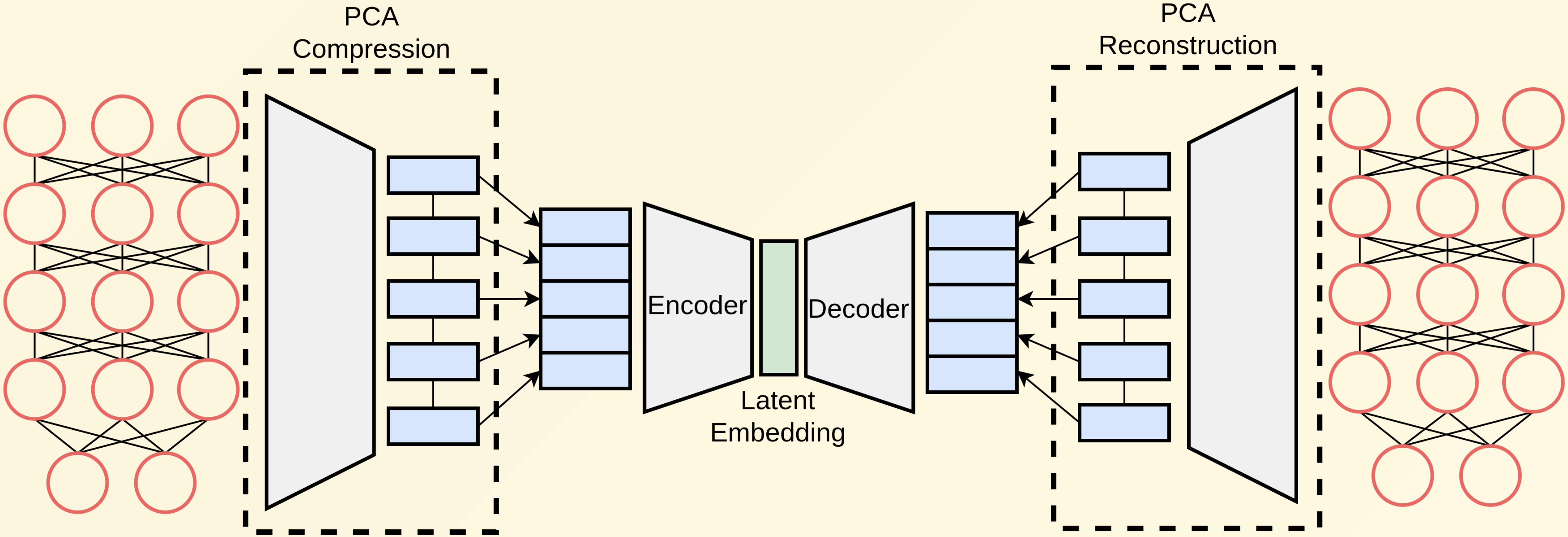
Problem


Weight Embedding
W
Ŵ
Results Embedding
Dataset Embedding
Embeddings
Projection Layer
Shared Embedding Space

Conditional Model Sampling

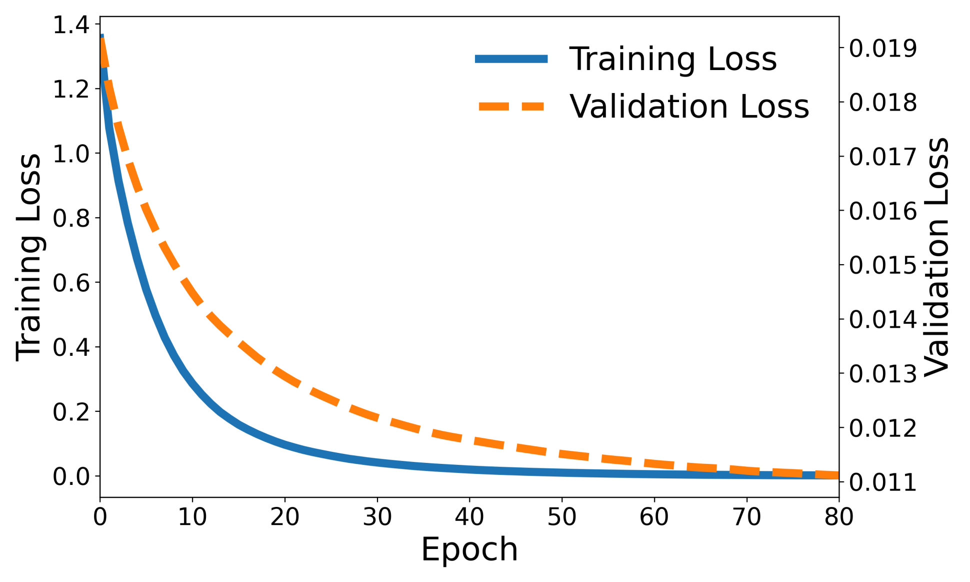


PCA + Weight Autoencoder

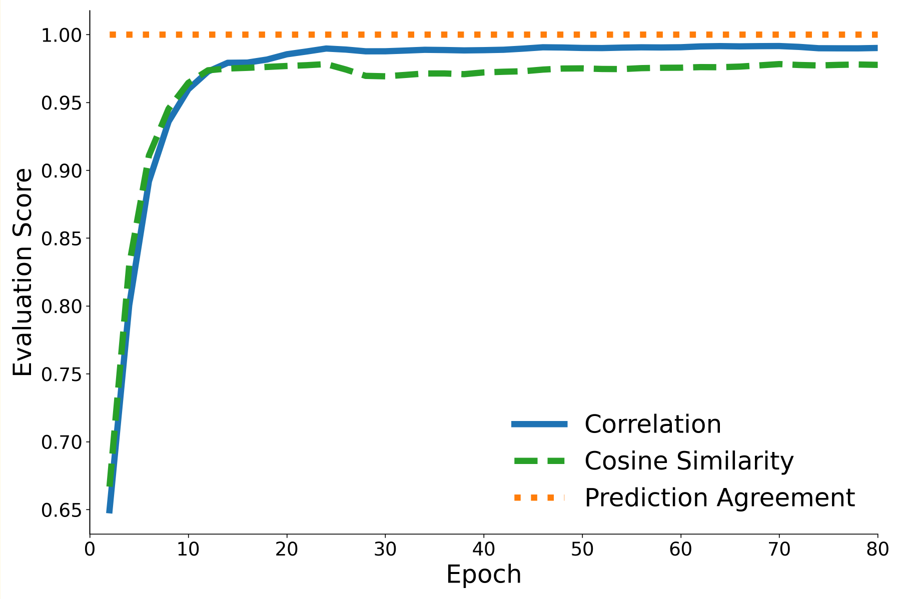


Results

Weight Autoencoder



(a) Loss



(b) Output Comparison (random input)

Results

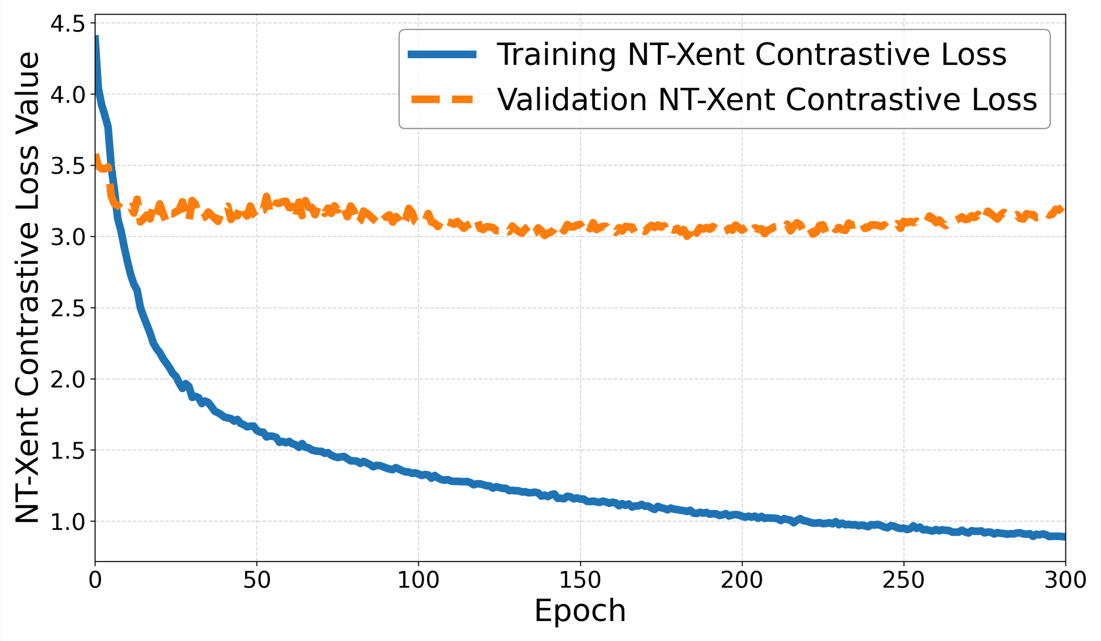
Weight Autoencoder

Output comparison (real data)

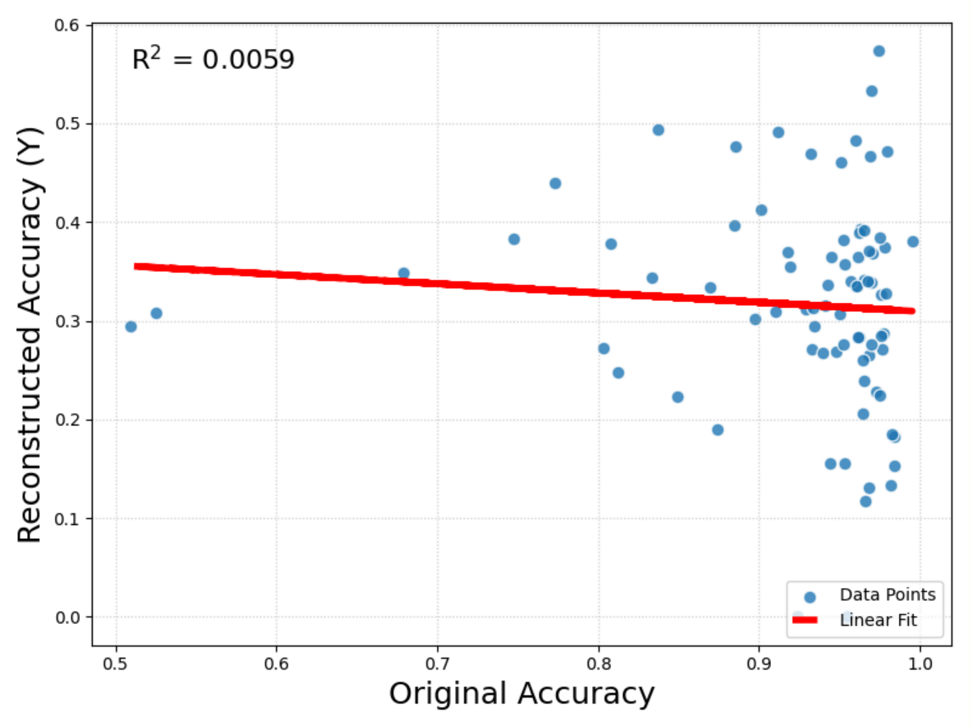
Metric	Mean Value	Range
Cosine Similarity	-0.082	[−0.547, 0.526]
Correlation	0.016	[−0.329, 0.344]
Prediction Agreement	31.21%	[2.06%, 64.88%]

Results

Shared Encoder



(a) Loss



(b) Input vs Output Accuracy

Conclusion

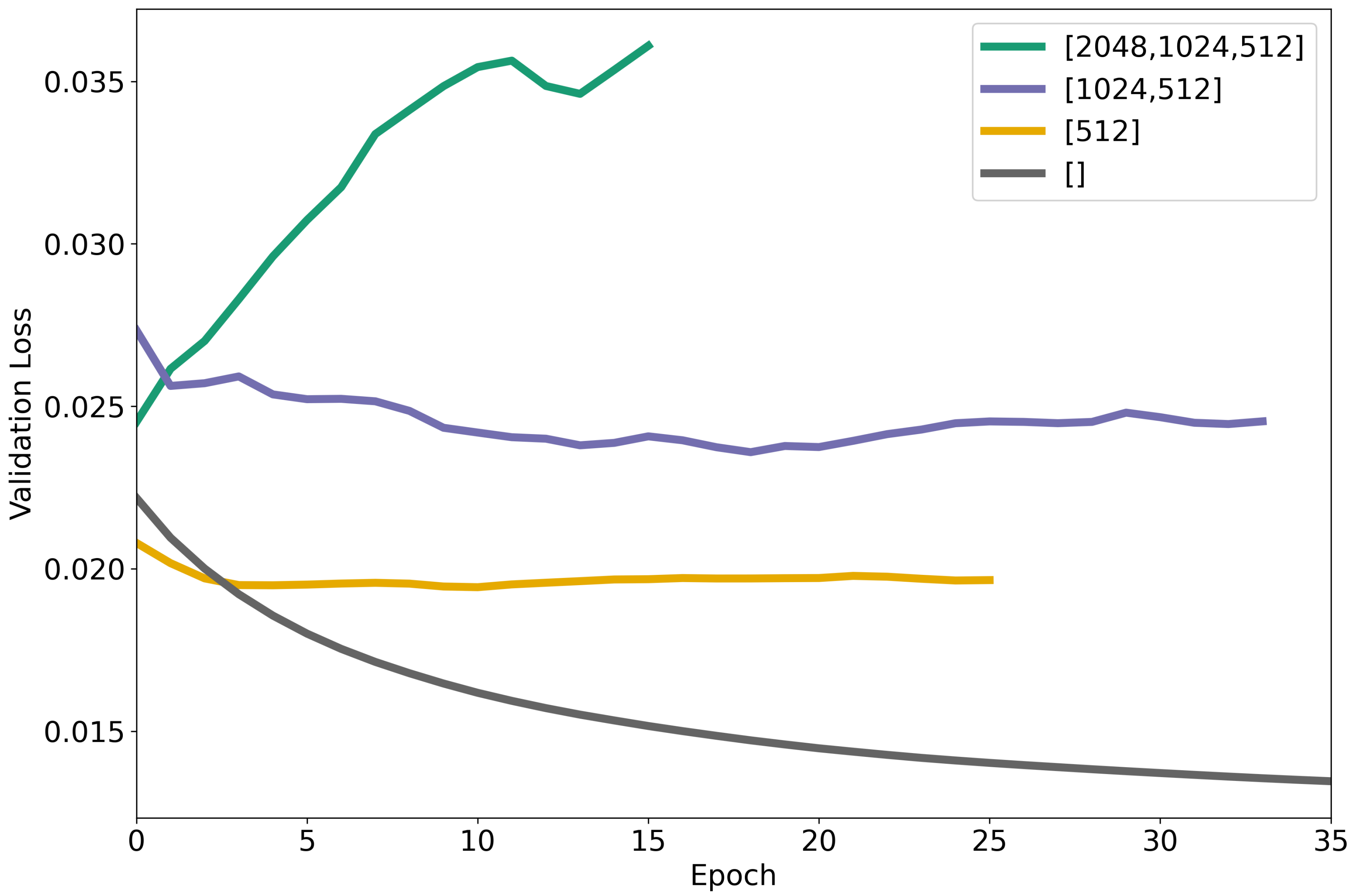
Contribution

- First joint modeling of $P(W|D,R)$
- Demonstrates viability of contrastive alignment for heterogeneous modalities

Future Work

- Improve weight encoder
  - Reverse PCA
  - PCA + Non-linear Stages
- Sequential Autoencoder for Neural Embeddings
- Dataset distillation
- Variational Autoencoder

Thank you


[2048,1024,512]
[1024,512]
[512]
[]
Validation Loss
0.015
0.020
0.025
0.030
0.035
Epoch
0
5
10
15
20
25
30
35