Appendix for
Pantypes: Diverse Representatives for Self-Explainable Models

Rune Kjærsgaard, Ahcène Boubekki, Line Clemmensen

Appendix A

In this appendix we include the training details and the hyperparameters used for our experiments. The experiments for MNIST, FMNIST, QuickDraw and UTK Face have been performed on an Intel 6 core i7, UHD 630 CPU laptop. The experiments for CelebA have been performed on a GPU HPC cluster. For the MNIST, FMNIST and QuickDraw datasets we train the networks on images with the original dimensions from the published datasets. For CelebA we rescale the images from a dimension of 178 × 218 to a dimension of 224 x 224. For UTK Face we use the aligned and cropped version of the data and rescale the images from a dimension of 200 x 200 to 32 x 32. The UTK Face dataset contains images of all ages from 0-116. We filter the dataset to include any individuals over the age of 18. The UTK Face dataset is trained on 20 prototypes per class, which causes the initial volume of the randomly initialized prototypes to be excessively large, leading to computational precision issues. To resolve this, we downscale the volume loss kernel $\boldsymbol{G}_k$ with a multiplicative factor of $c = 0.1$ before computing the volume. This results in a volume loss of:

$$\mathcal{L}_{\text{vol}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|c \cdot \boldsymbol{G}_k|^{\frac{1}{2}}}. \tag{1}$$

For the MNIST, FMNIST and QuickDraw datasets we used the standard encoder and decoder structures reported for these datasets in the ProtoVAE paper (Gautam et al. 2022). For the UTK Face dataset we use the CIFAR-10 structure reported in the ProtoVAE paper. For CelebA we use a ResNet-34 encoder and the usual decoder designed to output 224 x 224 images.

The hyperparameters used for the experiments are reported in Tabs. 2 and 3. We tested a range of values in interval [0.01-1000] for the loss scaling parameters reported in Tab. 3. The final parameters were chosen to balance accuracy, DB scores and decoded prototype appearance.

Appendix B

In this appendix we include additional illustrations of the concepts learned by PanVAE. All UMAP (McInnes, Healy, and Melville 2018) illustrations in this appendix and the paper throughout have been created using the following UMAP parameters: Minimum distance of 0.99, learning rate

Table 1: Overview of the datasets used for our experiments. K is the number of classes.

Dataset	N_{train} / N_{test}	Input Size	K
MNIST	60,000 / 10,000	28 x 28	10
FMNIST	60,000 / 10,000	28 x 28	10
QuickDraw	80,000 / 20,000	32 x 32	10
CelebA	162,770 / 39,829	224 x 224	2
UTK Face	16,000 / 3,210	32 x 32	2

Table 2: Overview of hyperparameters used for our experiments. LR indicates the learning rate and z Dim is the dimensionality of the latent space.

Dataset	LR	Epochs	Batch Size	z Dim
MNIST	$1e^{-3}$	100	128	256
FMNIST	$1e^{-3}$	100	128	256
QuickDraw	$1e^{-3}$	100	128	512
CelebA	$1e^{-3}$	50	128	512
UTK Face	$1e^{-3}$	50	128	512

Table 3: Overview of number of prototypes pr. class (M) and loss term scalings used for our experiments. $\mathcal{L}_{\text{div}}$ indicates the scaling on the respective diversity inducing loss in ProtoVAE and PanVAE (orthonormalization or volumetric loss). $\mathcal{L}_{\text{rec}}$ is the reconstruction loss term in $\mathcal{L}_{\text{VAE}}$ and $\mathcal{L}_{\text{kl}}$ is the KL-divergence loss term in $\mathcal{L}_{\text{VAE}}$.

Dataset	M	$[\mathcal{L}_{\text{pred}}, \mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{kl}}, \mathcal{L}_{\text{div}}]$
MNIST	5	[1,1,1,1]
FMNIST	5	[1,1,1,1]
QuickDraw	10	[1,1,1,1]
CelebA (Proto)	10	[1,0.1,100,10]
CelebA (Pan)	10	[1,0.1,100,100]
UTK Face (Proto)	20	[1,1,1000,1]
UTK Face (Pan)	20	[1,1,1,0.1]

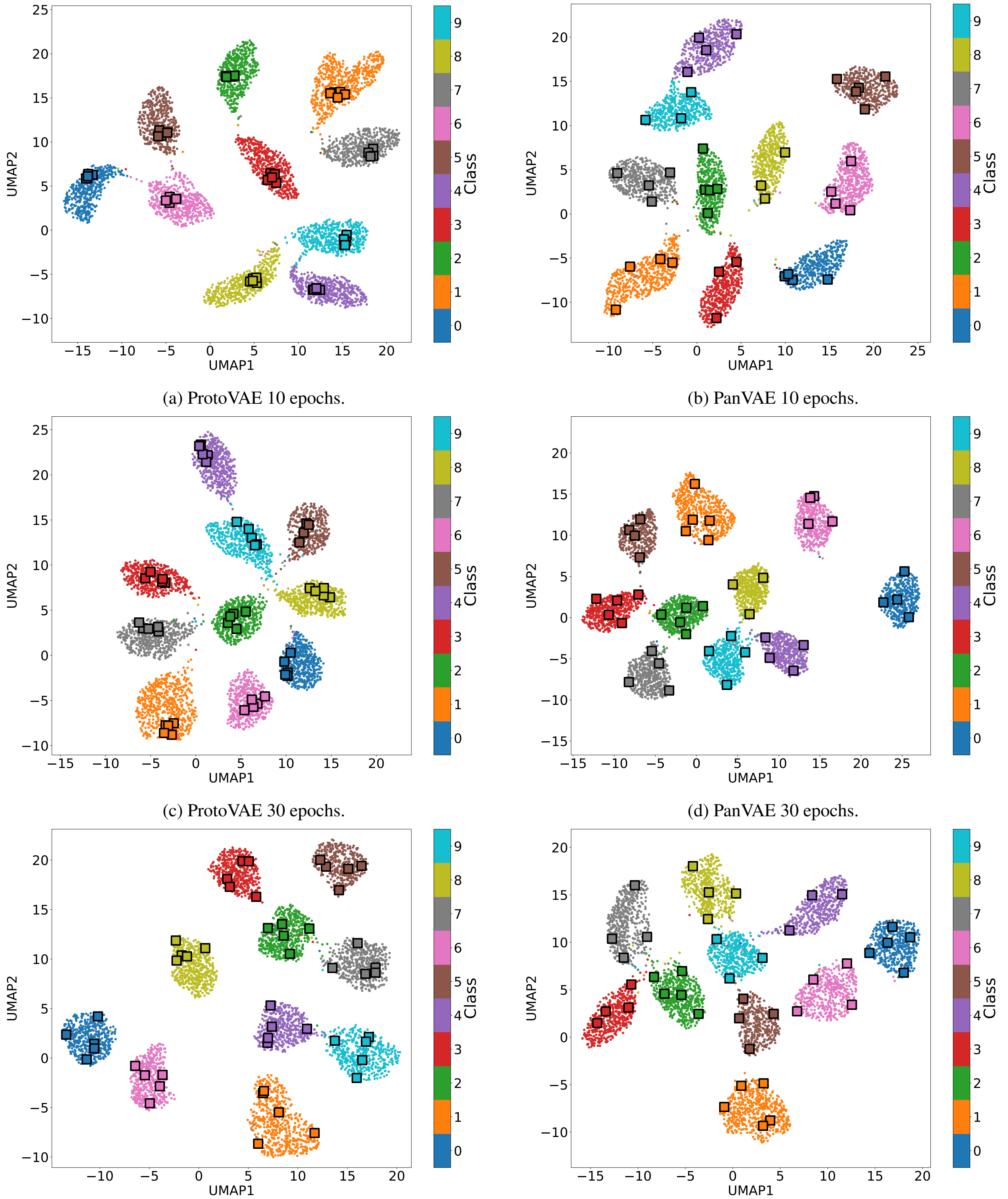
of 1.0, local connectivity of 1 and the number of neighbors at 25.

Fig. 1 shows the evolution of the latent space of MNIST during training. Here it is evident that PanVAE achieves high prototype diversity early in the training phase (as soon as 10 epochs), while the orthonormalization loss in ProtoVAE does not cause significant prototype diversity before 50 epochs of training. This finding is mirrored in Fig. **??**, which shows that ProtoVAE uses significantly more training time to achieve good separation between the prototypes and that at separation convergence PanVAE achieves the best representation. Fig. 2 shows the final decoded prototypes learned on the QuickDraw dataset. Figs. 3, 4 and 5 show prototype data coverage on the FMNIST dataset in UMAP and PCA space.

References

Gautam, S.; Boubekki, A.; Hansen, S.; Salahuddin, S.; Jenssen, R.; Höhne, M.; and Kampffmeyer, M. 2022. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35: 17940–17952.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.



(a) ProtoVAE 10 epochs.

(b) PanVAE 10 epochs.

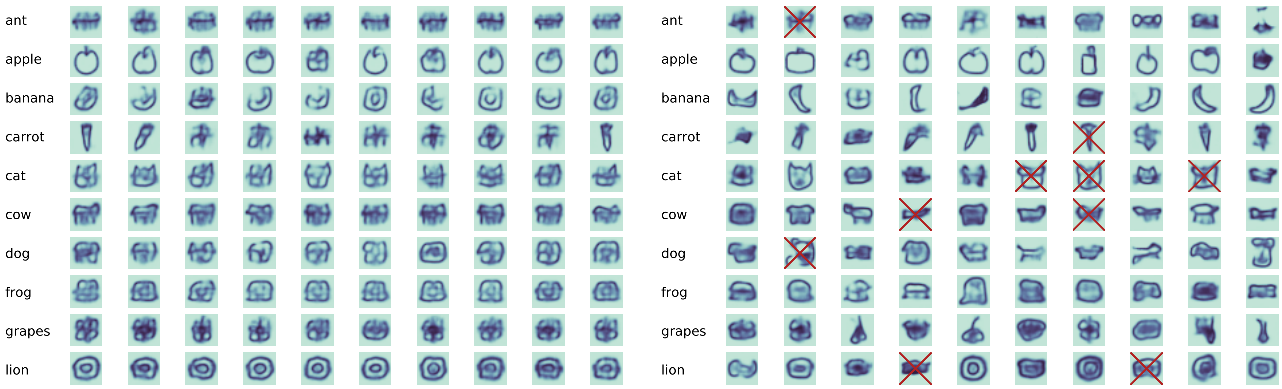
(c) ProtoVAE 30 epochs.

(d) PanVAE 30 epochs.

(e) ProtoVAE 50 epochs.

(f) PanVAE 50 epochs.

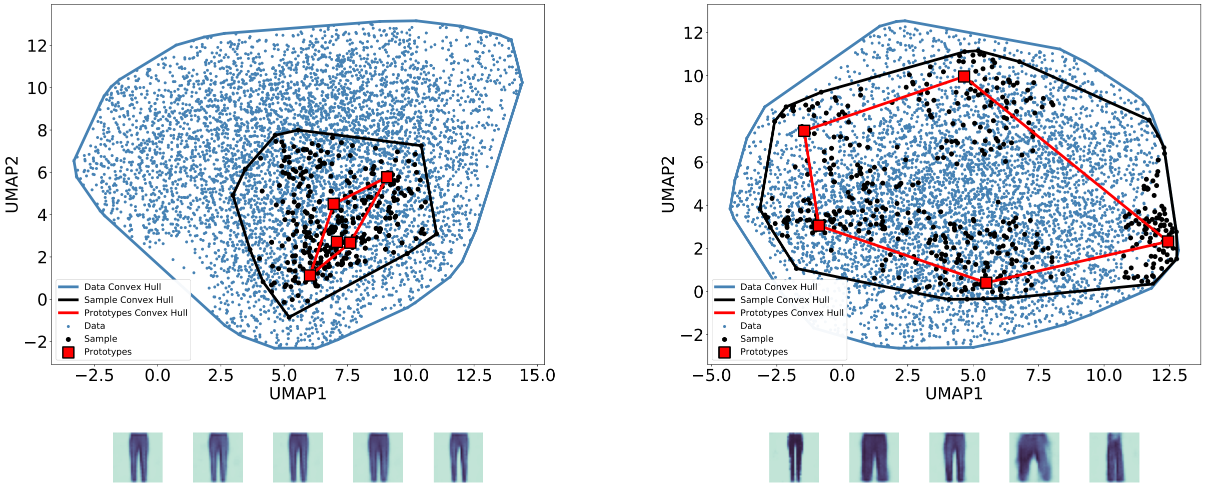
Figure 1: UMAP representations for the latent space of the MNIST data with overlaid latent representations of the prototypes (squares) for ProtoVAE and PanVAE respectively. The figures show the evolution of prototype latent location with training time. Both models are initiated with 50 total prototypes, but PanVAE is using prototype elimination and has eliminated 8 prototypes converging at 42 total prototypes.



(a) ProtoVAE

(b) PanVAE.

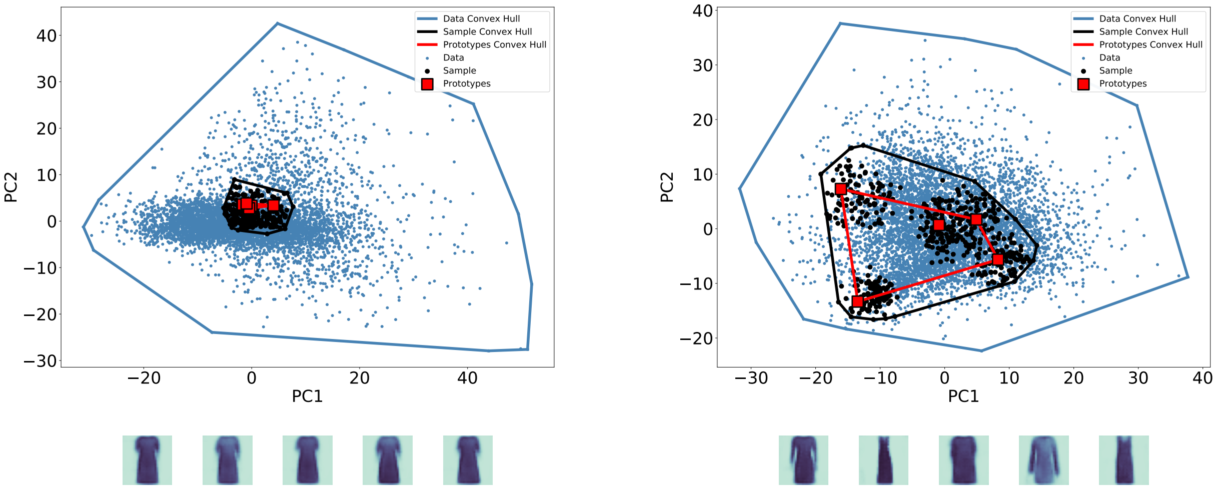
Figure 2: QuickDraw prototypes from ProtoVAE and PanVAE after 100 epochs of training with 10 prototypes per class. For PanVAE the pantypes that do not have the maximal similarity score with any training image have been marked with red crosses. In ProtoVAE all prototypes have maximal similarity score with at least one training image.



(a) Trouser ProtoVAE

(b) Trouser PanVAE.

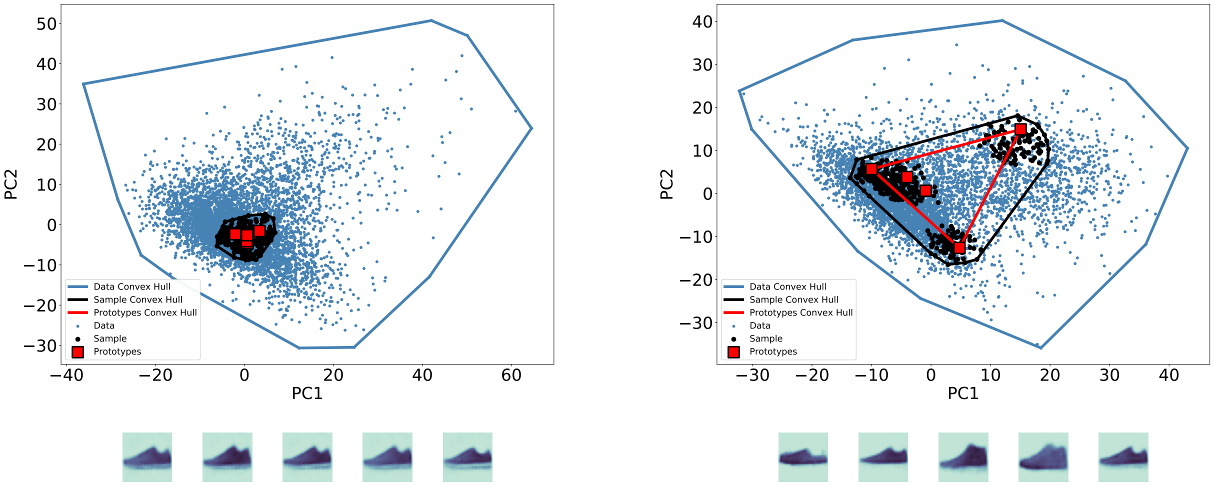
Figure 3: Prototype coverage from 20 epochs of training on FMNIST with 5 prototypes for the trouser class. The PanVAE sample convex hull covers 73% of the volume of the full class convex hull, whereas the ProtoVAE sample convex hull covers 25%.



(a) Dress ProtoVAE.

(b) Dress PanVAE.

Figure 4: PCA coverage from 20 epochs of training on FMNIST with 5 prototypes for the dress class. The principal components account for 54% and 52 % of the variation in the Prototype and PanType networks respectively. The ProtoVAE sample convex hull covers 3% of the volume of the full class convex hull, whereas the PanVAE sample convex hull covers 24%.



(a) Sneaker ProtoVAE

(b) Sneaker PanVAE.

Figure 5: PCA coverage from 20 epochs of training on FMNIST with 5 prototypes for the sneaker class. The principal components account for 65% and 68% of the variation in the ProtoVAE and PanVAE networks respectively. The ProtoVAE sample convex hull covers 2% of the volume of the full class convex hull, whereas the PanVAE sample convex hull covers 19%.