Pantypes: Diverse Representatives for Self-Explainable Models

Rune Kjærsgaard [1], **Ahcène Boubekki** [2], **Line Clemmensen** [1]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[2] Machine Learning and Uncertainty, Physikalisch-Technische Bundesanstalt, Germany
rdokj@dtu.dk

Abstract

Prototypical self-explainable classifiers have emerged to meet the growing demand for interpretable AI systems. These classifiers are designed to incorporate high transparency in their decisions by basing inference on similarity with learned prototypical objects. While these models are designed with diversity in mind, the learned prototypes often do not sufficiently represent all aspects of the input distribution, particularly those in low density regions. Such lack of sufficient data representation, known as representation bias, has been associated with various detrimental properties related to machine learning diversity and fairness. In light of this, we introduce *pantypes*, a new family of prototypical objects designed to capture the full diversity of the input distribution through a sparse set of objects. We show that pantypes can empower prototypical self-explainable models by occupying divergent regions of the latent space and thus fostering high diversity, interpretability and fairness.

Introduction

Machine learning (ML) systems are increasingly affecting individuals across various societal domains. This has put into question the black-box nature of these systems, and fostered the field of explainable AI (XAI), wherein model inference is corroborated with justifications and explanations in an effort to increase transparency and trustworthiness. In this line of research two approaches have arisen; that of ad-hoc black-box model explanations (Selvaraju et al. 2017; Yosinski et al. 2015), and that of self-explainable models (SEMs) (Chen et al. 2019a; Alvarez Melis and Jaakkola 2018). A popular approach for SEMs substitutes traditional black-box networks with glass-box counterparts, where class representative prototypes are generated and used in the decision process (Chen et al. 2019a) leading to increased trustworthiness and interpretability.

The various initiatives emerging in the literature share the same overarching goals, but there is still a lack of consensus on the exact properties a SEMs should display (Gautam et al. 2023). We adopt three prerequisites of a SEM outlined in (Gautam et al. 2022), namely *transparency*, *trustworthiness* and *diversity*.

Transparency may be defined by two properties; (i) the learned concepts are used in the decision making process without the use of a black-box model and (ii) the learned concepts can be visualized in the input space.

Trustworthiness may be defined by three properties; (i) the predictive performance of the model matches its closest black-box counterpart, (ii) explanations are robust and (iii) the explanations directly represent the contribution of the input features to the model predictions.

Diversity may be defined by one property; (i) the concepts learned by the SEM are represented by non-overlapping information in the latent space.

While significant work has been put forth in the literature to cement the transparency and trustworthiness axis of SEMs, only limited effort using qualitative measures exists for the diversity axis. Similarly, the relation between the diversity axis and appropriate inference remains largely unexplored. Diversity is typically ensured by introducing model regularization towards learning non-overlapping concepts (Vilone and Longo 2020). However, this condition may not be strong enough, as non-overlapping concepts can still be learned in a small region of the input space, causing a lack of representativity for the full data distribution, known as representation bias (Shahbazi et al. 2022). Representation bias can cause smaller sub-populations to remain hidden in low-density regions and ultimately cause biased inference (Jin et al. 2020). To provide sufficient coverage and to mitigate the impact of data bias during model inference, it is critical to capture the full diversity of the data, and to have this diversity be represented in the prototypes learned by the SEM. To this end, we introduce pantypes, a new family of prototypical objects designed to empower SEMs by sufficiently covering the dataspace. Pantype generation is promoted using a novel volumetric loss inspired by a probability distribution known as a Determinantal Point Process (DPP) (Kulesza, Taskar et al. 2012). This loss induces higher prototype diversity, enables more fine-grained diversity control, and at the same time allows prototype pruning wherein the number of prototypes is determined dynamically dependent on the diversity expressed within each class. Prototype pruning enables the capacity to learn additional prototypes for complex classes and to grasp simple classes through a sparser set of objects, improving the interpretability of the class representatives.

Our contributions can be summarized as follows:

- Introduction of a volumetric loss, which promotes the generation of pantypes, a highly diverse set of prototypes.
- Quantitative measures for prototype representativity and diversity in SEMs.
- Dynamic class-specific prototype selection.

PanVAE

The modeling task at hand involves a classification setting on visual image data, where the SEM learns to classify $K > 0$ classes from a training set $\mathbf{X} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^{N}$,where $\boldsymbol{x_i} \in \mathbb{R}^P$ is the i^{th} image and $\boldsymbol{y_i} \in \{0,1\}^K$ is a one-hot label vector. We implement the pantypes[1] on the foundation of a well-tested variational autoencoder based SEM, known as ProtoVAE (Gautam et al. 2022). This model uses an encoder function $f : \mathbb{R}^p \rightarrow \mathbb{R}^d \times \mathbb{R}^d$, to transform the input images into a posterior distribution $(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})$. A latent representation $\boldsymbol{z_i}$ of the i^{th} image is then sampled from the distribution $\mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})$ and passed as input to a decoder function $g : \mathbb{R}^d \rightarrow \mathbb{R}^p$ to generate the reconstructed image $g(\boldsymbol{z_i}) = \hat{\boldsymbol{x_i}}$. To enable transparent predictions, the model does not directly use the feature vector $\boldsymbol{z_i}$ during inference, but rather compares this vector to M prototypes per class $\boldsymbol{\Phi} = \{\phi_{kj}\}_{j=1,,M}^{k=1..K}$ via a similarity function : $R^d \rightarrow R^M$. The resulting similarity vector $\boldsymbol{s}_i \in \mathbb{R}^{K \times M}$ is then used in a glass-box linear classifier $h : \mathbb{R}^{\mathbb{M}} \rightarrow [0,1]^K$ to generate the class prediction $h(\boldsymbol{s_i}) = \hat{\boldsymbol{y}}_i$. The similarity function (Chen et al. 2019b) is given by:

$$\boldsymbol{s}_i(k,j) = \text{sim}(\boldsymbol{z}_i, \boldsymbol{\phi_{kj}}) = \log\left(\frac{||\boldsymbol{z}_i - \boldsymbol{\phi_{kj}}||^2 + 1}{||\boldsymbol{z}_i - \boldsymbol{\phi_{kj}}||^2 + \epsilon}\right), \tag{1}$$

where $0 < \epsilon < 1$. This construction allows the similarity vector to not only capture the distances to the prototypes, but to also reflect the influence of each prototype on the final prediction.

Loss Terms

To further enforce the properties of a SEM, we adopt the same prediction and VAE loss term structure as ProtoVAE:

$$\mathcal{L}_{\text{ProtoVAE}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{orth}}, \tag{2}$$

where

$$\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{CE}(h(\boldsymbol{s}_i); \boldsymbol{y}_i) \tag{3}$$

is a cross-entropy (CE) prediction loss term ensuring inter-class diversity in the prototypes and

$$\mathcal{L}_{\text{VAE}} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||^2 + \sum_{k=1}^{K}\sum_{j=1}^{M} \boldsymbol{y}_i(k)\frac{\boldsymbol{s}_i(k,j)}{\sum_{l=1}^{M}\boldsymbol{s}_i(k,j)} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})||\mathcal{N}(\boldsymbol{\phi_{kj}}, \boldsymbol{I}_d)) \tag{4}$$

[1]Our code and training details are publicly available on GitHub at https://github.com/RuneDK93/pantypes

is the loss for a mixture of VAEs using the same network each with a Gaussian prior distribution centered on one of the prototypes (Gautam et al. 2022). Here $\boldsymbol{I}_d$ is a $d \times d$ identity matrix. Finally, an orthonormality loss term is used:

$$\mathcal{L}_{\text{orth}} = \sum_{k=1}^{K} ||\bar{\boldsymbol{\Phi}}_k^T \bar{\boldsymbol{\Phi}}_k - \boldsymbol{I}_M||_F^2, \tag{5}$$

where $\bar{\boldsymbol{\Phi}}_k$ is the mean subtracted prototype vector for all prototypes of class k and $\boldsymbol{I}_M$ is an $M \times M$ identity matrix.

The orthonormality loss is included to foster intra-class prototype diversity and to uphold the diversity property of a SEM by inducing the learning of non-overlapping concepts in the latent space and thus avoiding prototype collapse (Wang et al. 2021; Jing et al. 2021). While this loss causes the prototypes to be orthogonal, it does not explicitly prevent the prototypes from occupying and representing a small region (volume) of the full data-space. Moreover, prototype orthonormality is typically achieved early during training, and further scaling of the orthonormality loss does not significantly alter the diversity of the prototypes (see results section).

Poor or skewed data representation, known as representation bias, has been associated with various detrimental properties related to ML fairness, where underrepresented minority groups are negatively affected during inference (Phillips et al. 2011). To mitigate these issues it is essential to achieve sufficient coverage of the full diversity represented in the data (Suresh and Guttag 2019). We draw on this idea to empower the ProtoVAE model by exchanging its class-wise orthonormality diversity loss with a volumetric diversity loss, which causes the model to learn prototypical objects with various improved qualities, including an improved coverage of the embedding space. We call these learned objects *pantypes*. The loss term structure of our model is:

$$\mathcal{L}_{\text{PanVAE}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{vol}}, \tag{6}$$

where $\mathcal{L}_{\text{vol}}$ is the volumetric prototype loss, which not only prevents prevents prototype collapse, but causes higher prototype diversity, enables more fine-grained diversity control, and at the same time allows prototype pruning wherein the number of prototypes is determined dynamically dependent on the diversity expressed within each class.

Pantypes

Pantypes are prototypical objects learned in an end-to-end manner during model training. They are inspired by a probability distribution known as a Determinantal Point Process (DPP) (Kulesza, Taskar et al. 2012), which can be used to sample from a population while ensuring high diversity. DPPs have recently garnered attention in the ML community, and have been used to draw diverse sets in a range of ML applications including data from videos, images, documents, sensors and recommendation systems (Gong et al. 2014; Kulesza, Taskar et al. 2012; Lin and Bilmes 2012; Zhou et al. 2010; Krause, Singh, and Guestrin 2008). DPPs describe a distribution over subsets, such that the sampling probability of a subset is proportional to the determinant

of an associated sub-matrix (a minor) of a positive semi-definite kernel matrix. The kernel matrix expresses similarity between feature vectors of observations through a kernel function $\boldsymbol{G}_{ij} = g(\boldsymbol{v}_i, \boldsymbol{v}_j)$. This global measure of similarity is then used to sample such that similar items are unlikely to co-occur. The kernel can be constructed in various ways including the radial basis function (RBF) kernel $\boldsymbol{G}_{ij} = e^{-\gamma||\boldsymbol{v}_i - \boldsymbol{v}_j||^2}$ or the linear kernel, leading to a similarity function of inner products known as the Gram matrix $\boldsymbol{G}_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$. When using the Gram matrix, a DPP is equivalent to sampling with probability proportional to the volume of the paralellotope formed by the feature vectors of the sampled items. We utilize the linear kernel to construct a volumetric loss on the prototypes in the following way:

$$\mathcal{L}_{\text{vol}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\boldsymbol{G}_k|^{\frac{1}{2}}}, \tag{7}$$

where $\boldsymbol{G}_k \in \mathbb{R}^{M \times M}$ is the Gram matrix given by $\boldsymbol{G}_k = \boldsymbol{\Phi}_k^T \boldsymbol{\Phi}_k$ with Φ_{kj} as column vectors in $\boldsymbol{\Phi}_k$ and $|\boldsymbol{G}_k|$ is the Gramian (Gram determinant). $|\boldsymbol{G}_k|^{\frac{1}{2}}$ measures the M-dimensional volume of the parallelotope formed by the M columns of $\boldsymbol{\Phi}_k$ embedded in d-dimensional space. In other words, it expresses the diversity of the M prototypes of class k through the volume spanned by their feature vectors. This loss not only prevents prototype collapse by causing the feature vectors to diverge, but also directly encourages the pantypes to occupy different sectors of the data domain to express a large volume.

Prototype elimination Increasing the scaling on the volume loss punishes pantypes that express a low volume and thus directly alters the diversity of the learned objects. With sufficient scaling, the volumetric loss forces pantypes out-of-distribution (OOD) if they are not necessary to represent the observed diversity of a class. This allows natural pruning, wherein the number of pantypes can be dynamically tuned by elimination of OOD pantypes. This is similar to the discipline of hyperspectral endmember unmixing, where a number of endmembers (prototypes) are disentangled from a hyperspectral image and linear combinations of the endmembers are used to reconstruct the input images. Following training, the learned endmembers can be associated with purity scores (Berman et al. 2004), which express the quality of their explanations. These scores describe the maximal responsibility proportion of endmembers for reconstructing the original images. In other words, a high purity score indicates that an endmember shares a high similarity with individual input images, while a low purity score indicates that an endmember is capturing noise and should be pruned. Such purity scores can be constructed from the similarity scores used in the linear classifier in our SEM. Thus, as proposed by (Berman et al. 2004), we can initiate the model with a sufficiently large number of pantypes, and use the similarity scores to prune individual OOD pantypes. We propose a heuristic for pruning, where a pantype can be pruned if it does not have the maximal similarity score for any of the training images (i.e. it does not individually represents any of the training images more than the other pantypes).

DATASET	PROTOPNET	PROTOVAE	PANVAE
MNIST	98.8 ± 0.1	$\mathbf{99.3 \pm 0.1}$	$\mathbf{99.4 \pm 0.1}$
FMNIST	89.9 ± 0.5	91.6 ± 0.1	$\mathbf{92.2 \pm 0.1}$
QDRAW	58.7 ± 0.0	$\mathbf{85.6 \pm 0.1}$	$\mathbf{85.5 \pm 0.1}$
CELEBA	98.2 ± 0.1	$\mathbf{98.6 \pm 0.0}$	$\mathbf{98.6 \pm 0.0}$

Table 1: Predictive performance (accuracy) of PanVAE ProtoVAE and ProtoPNet on MNIST, FMNIST, QuickDraw and CelebA. The values are the mean and standard deviation of three runs.

Results

We perform experiments across various real-world datasets to monitor the transparency, diversity and trustworthiness of PanVAE. These datasets are FashionMNIST (FMNIST) (Xiao, Rasul, and Vollgraf 2017), MNIST (LeCun et al. 1998), QuickDraw (QDraw) (Ha and Eck 2017) and CelebA (Liu et al. 2015). We demonstrate the trustworthiness of PanVAE by evaluating the predictive performance of the overall model and asses the diversity and transparency using qualitative assessments from visualizations of the input space, as well as quantitative measures of prototype quality and coverage. We compare PanVAE to the performance of ProtoVAE and ProtoPNet.

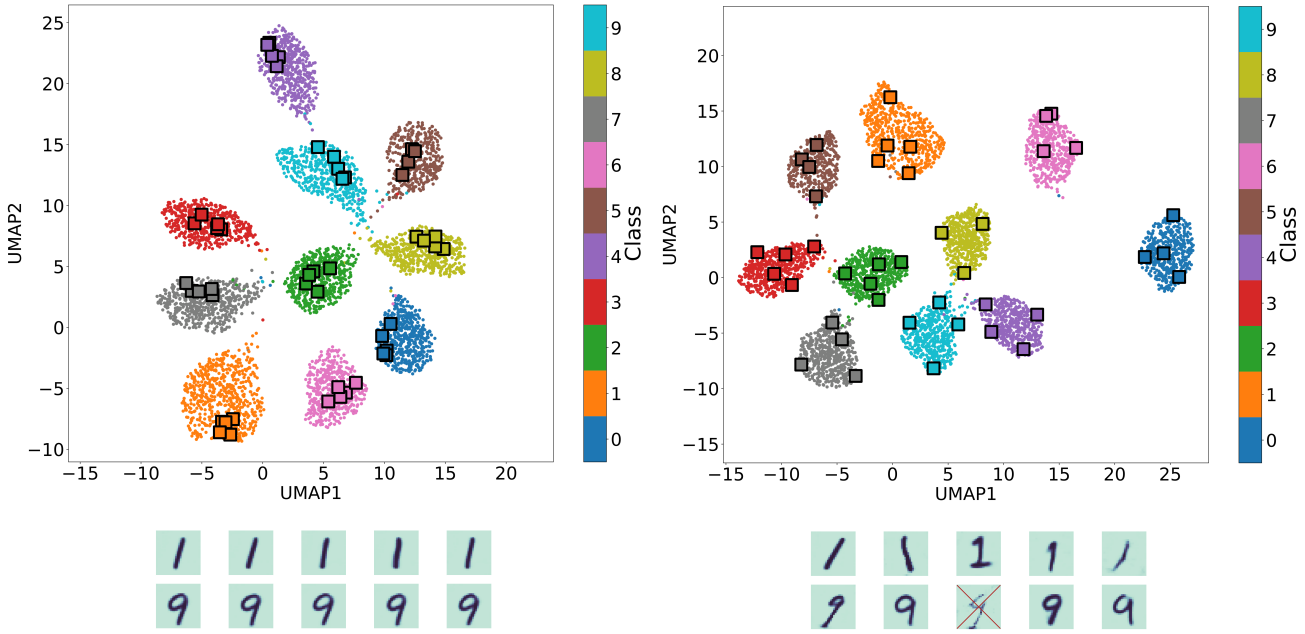
Predictive Performance

The results for the predictive performance are shown in Tab. 1, which demonstrates that PanVAE, like ProtoVAE, achieves higher predictive performance than ProtoPNET on the four datasets. There is no significant predictive performance gap between PanVAE and ProtoVAE on the datasets. This underlines the trustworthiness of PanVAE.

Prototype Representation Quality

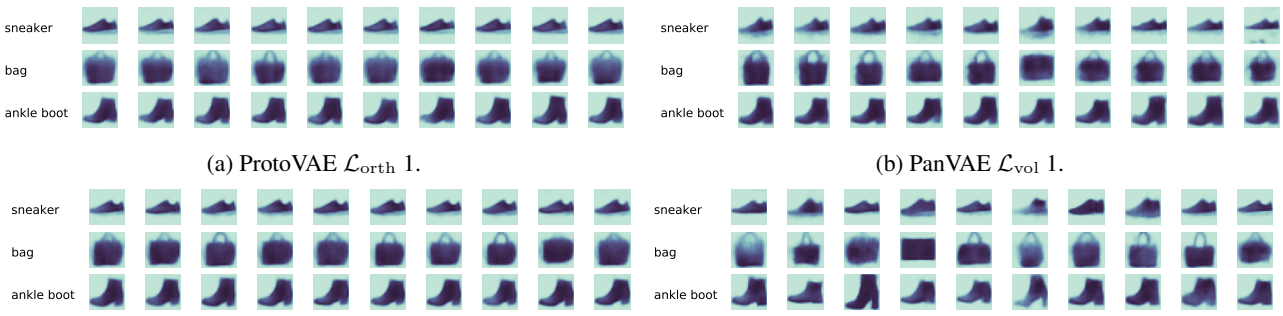
Firstly, we asses prototype representation quality using visual inspection of the learned prototypes and the associated latent space. This can be seen for the MNIST dataset on Fig. 1, where the prototypes for ProtoVAE and PanVAE are shown. The diversity of PanVAE is higher than ProtoVAE. The prototypes from ProtoVAE are mostly orthogonal in latent space, but only occupy a small region of the space. Contrarily, the volume loss in PanVAE has pushed the pantypes away from each other allowing them to occupy and represent diverse regions of the dataspace. This is reflected in the decoded prototypes, which show high diversity by representing various archetypical ways of drawing digits. For instance, the pantypes capture variations between left-handed and right-handed digits of "1" as well as the archetypical "1" with a horizontal base. Moreover, PanVAE has found that the digits of "9" express less diversity and has thus pushed one of the pantypes OOD (indicated by a red cross in the figure). This form of prototype pruning by PanVAE allows the model to asses and represent the individual diversity expressed by each class.

Fig. 2 demonstrates the diversity control enabled by PanVAE by illustrating learned prototypes on the FMNIST datasets with different diversity loss scalings. The objective



(a) ProtoVAE. (b) PanVAE.

Figure 1: ProtoVAE (a) and PanVAE (b) visualizations of the latent space and decoded prototypes learned on MNIST after 30 epochs of training. Top: UMAP representations of the latent space with learned prototypes overlaid as squares. Bottom: Decoded prototypes of class '1' and '9'. One of the prototypes from PanVAE does not have the maximal similarity for any training image, indicated by a red cross. PanVAE has captured variations in the digit '1' pertaining to right-handedness (first '1' from the left), left-handedness (second '1' from the left) and a traditional writing style (third '1' from the left).



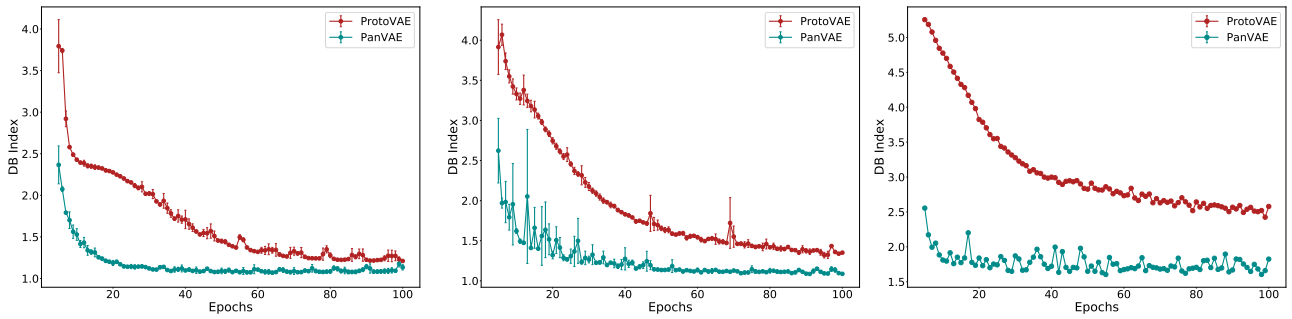
(a) ProtoVAE $\mathcal{L}_{\text{orth}}$ 1. (b) PanVAE $\mathcal{L}_{\text{vol}}$ 1.

(c) ProtoVAE $\mathcal{L}_{\text{orth}}$ 100. (d) PanVAE $\mathcal{L}_{\text{vol}}$ 100.

Figure 2: Diversity control enabled by ProtoVAE and PanVAE. The figure shows the change in decoded prototype appearance as the respective diversity inducing losses are increased. The prototypes are shown for the FMNIST data of classes "sneaker", "bag" and "ankle boot" after 10 epochs of training. Figs. 2a and 2c show the difference between ProtoVAE prototypes with scale factor of 1 and 100 on the diversity loss $\mathcal{L}_{\text{orth}}$. Figs. 2b and 2d show the difference between PanVAE pantypes with scale factor of 1 and 100 on the diversity loss $\mathcal{L}_{\text{vol}}$.

of the orthonormalization loss in ProtoVAE is to enforce intra-class diversity, and hence that the prototypes capture different concepts. While the loss ensures this, it only does so after sufficient training time. Fig. 2 shows that scaling the orthonormalization loss in ProtoVAE does not significantly alter the diversity of the representation. On the other hand, the volumetric loss in PanVAE allows direct control over the diversity of the representation.

Previous work in the literature on prototype based self-explainable classifiers often only qualitative asses the prototype diversity axis (Gautam et al. 2022) (i.e. visual inspection of the diversity prerequisite of non-overlapping prototypes). We propose that self-explainable classifiers should not only be assessed with quantitative measures on the trustworthiness axis, but should also be evaluated by quantitative measures on the diversity axis. This includes thorough evaluations of how well the prototypes represent the dataspace. In order to do this we make use of measures of prototype



(a) MNIST. (b) FMNIST. (c) QuickDraw.

Figure 3: Evolution of prototype DB scores for PanVAE and ProtoVAE on MNIST, FMNIST and QuickDraw. Data points indicate mean values and associated standard deviations over three runs.

quality and representativity by firstly measuring the prototype quality using the Davies-Bouldin (DB) index (Davies and Bouldin 1979) and secondly evaluating the diversity of the class representatives by assessing their data coverage.

Davies-Bouldin Index The DB index is a measure of cluster quality defined by the average similarity between cluster C_i for $i = 1, ..., k$ and its most similar cluster C_j. The similarity measure R_{ij} quantifies a balance between inter- and intra-cluster distances. We adopt this measure and consider the prototypes in a SEM as cluster representatives and assign observations to their closest prototype in latent space according to maximal similarity scores. The intra-cluster size s_i is then measured as the average distance between prototype i and each data point belonging to the prototype, while the the inter-cluster distance d_{ij} is measured by the distance between prototypes i and j. From this the cluster similarity measure R_{ij} can be constructed such that it is non-negative and symmetric by:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}. \tag{8}$$

With these definitions in place the DB index may be defined by:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i=j} R_{ij}, \tag{9}$$

where a lower DB scores equates to a better representation of the underlying data. The DB scores for the different models can be seen in Tab. 2. PanVAE achieves the best DB scores in all cases, demonstrating the ability of the pantypes to represent the underlying dataspaces.

In addition to achieving higher final DB scores, PanVAE also does so using less training time. This is illustrated in Fig. 3, where the DB score evolution is shown for ProtoVAE and PanVAE over 100 epochs of training. PanVAE converges on a lower DB score much quicker than ProtoVAE.

Data Coverage The DB index provides a measure of prototype quality in terms of prototype representation quality, but does not sufficiently asses how well the prototypes cover the diversity in the dataspace. Sufficient coverage of various aspects in the dataspace has been found critical in obtaining unbiased ML algorithms (Jin et al. 2020).

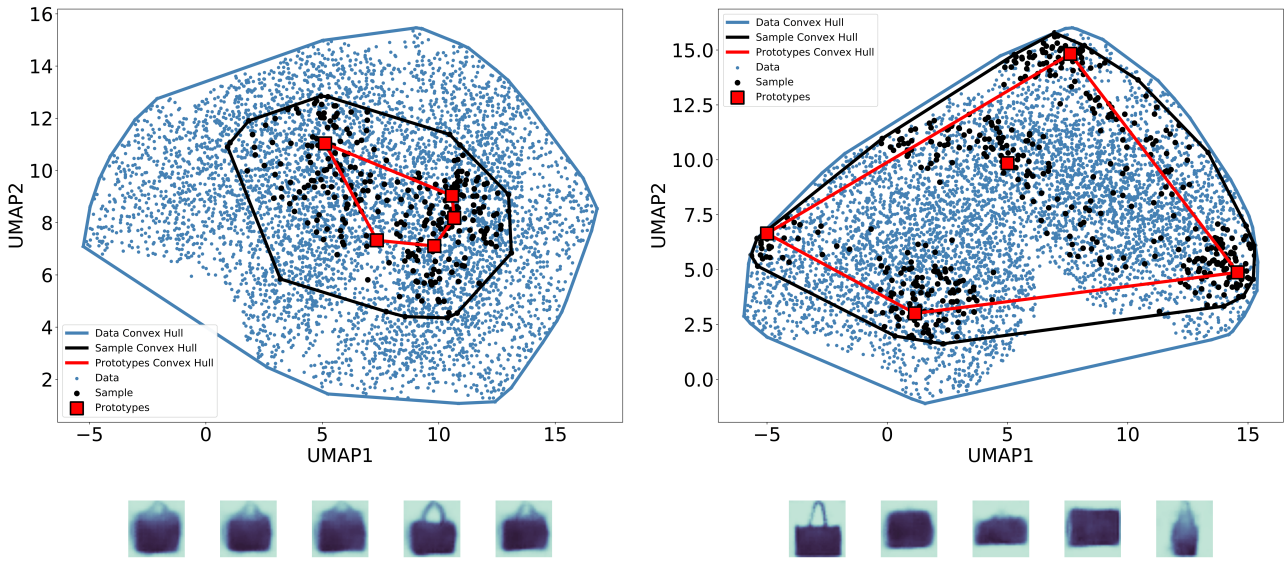
DATASET	PROTOPNET	PROTOVAE	PANVAE
MNIST	2.20 ± 0.18	1.21 ± 0.00	$\mathbf{1.13 \pm 0.03}$
FMNIST	3.43 ± 1.15	1.35 ± 0.01	$\mathbf{1.09 \pm 0.01}$
QDRAW	2.52 ± 0.62	2.57 ± 0.01	$\mathbf{1.82 \pm 0.01}$
CELEBA	27.09 ± 27.23	1.58 ± 0.15	$\mathbf{1.37 \pm 0.01}$

Table 2: Davies-Bouldin scores of prototypes from the different models on the datasets used for our experiments. The values are the mean and standard deviation over three runs.

In order to asses prototype data coverage, we compare the volume spanned by observations represented by the prototypes to the volume of the full data distribution. Ideally, the prototypes are diverse enough, that they sufficiently cover a large volume of data they seek to represent. The coverage may be assessed through the volume of the convex hull of the data. We evaluate our pantypes on this premise by sampling the 100 nearest observations to each pantype. The proximity is measured in the full latent space in terms of the similarity score (Eq. 1). We then compute the volume spanned by the represented observations from their convex hull, and compare this to the volume of the original data. We illustrate the results of this procedure in Fig. 4 using a 2D UMAP projection of the 256 dimensional latent space for the "Bag" class in FMNIST. The increased diversity of the pantypes allow them to occupy and represent a larger region of the dataspace.

Demographic Diversity Sufficient representation of demographic groups has been found critical in ensuring ML fairness (Jin et al. 2020). Image data used to train facial recognition algorithms have historically been biased towards White individuals, which are overrepresented in the training data, resulting in biased inference (Buolamwini and Gebru 2018). The largest disparity is found between white skinned and dark skinned individuals.

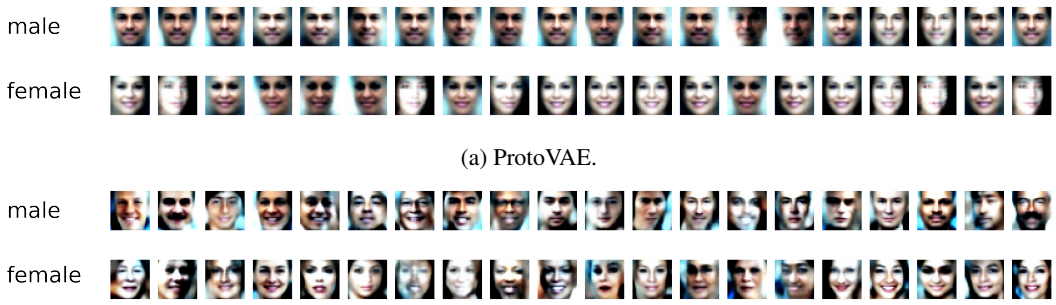
Demographic diversity may be quantified using a measure of combinatorial diversity, also known as diversity index (Simpson 1949). The combinatorial diversity is defined



(a) Bag prototypes from ProtoVAE.

(b) Bag pantypes from PanVAE.

Figure 4: Prototype coverage in UMAP space from 20 epochs of training on FMNIST with 5 prototypes for the "bag" class for ProtoVAE (a) and PanVAE (b). Top: UMAP representations of the latent space with learned prototypes overlaid as red squares. The prototype convex hull in UMAP space is shown as a red outline around the prototypes and the full class dataspace convex hull is shown as a blue outline around the data. A sample of the 100 closest observations to each prototype is shown as black datapoints. The convex hull of the sampled observations is shown as a black outline. The PanVAE sample convex hull covers 77% of the volume of the full class convex hull, whereas the ProtoVAE sample convex hull covers 33%. Bottom: Decoded prototypes.



(a) ProtoVAE.

(b) PanVAE.

Figure 5: Face prototypes learned on the UTK Face dataset. The learned prototypes are shown for ProtoVAE in (a) and for PanVAE in (b). PanVAE has captured variations in race as well as other unseen features such as facial hair in males. The ProtoVAE males all have somewhat neutral expressions with shut mouths while most of the females have slight smiles. The PanVAE males and females all exhibit large variations in expression from full smiles with visible teeth to neutral expressions without visible teeth.

as the information entropy of the distribution (Celis et al. 2016):

$$H = -\sum_{i=1}^{k} p_i \log p_i, \tag{10}$$

where the combinatorial diversity measure H is the entropy, p_i is the probability of event i and $\sum$ is the sum over the possible outcomes k. This measure quantifies the information entropy of the demographic distribution over k demographic groups. A high entropy equates to a more diverse (fair) representation, which is not particularly biased towards any demographic group.

We evaluate how the volumetric loss may aid in mitigating demographic data bias and enhance group level diversity. To do this we train PanVAE on the UTK Face dataset (Zhang, Song, and Qi 2017), which contain images of about 20,000 individuals with associated sex and race labels. The decoded facial prototypes from training on the UTK Face dataset can

METRIC	PROTOVAE	PANVAE
ACC ALL	$\mathbf{95.08 \pm 0.11}$	$\mathbf{95.42 \pm 0.37}$
ACC WHITE MALE	$\mathbf{96.35 \pm 0.31}$	95.21 ± 0.33
ACC BLACK FEMALE	91.67 ± 0.53	$\mathbf{94.90 \pm 0.39}$
ACC GAP	4.69 ± 0.24	$\mathbf{0.32 \pm 0.15}$
DIVERSITY	1.26 ± 0.06	$\mathbf{1.43 \pm 0.07}$

Table 3: UTK results. The values are the mean and standard deviation of three runs. The overall accuracy is reported along with the individual accuracy and accuracy gap between White males and Black females. A positive gap value indicates that the mean accuracy is higher on White males compared to Black females. Diversity is the information entropy (demographic diversity) of the distribution of races represented by the prototypes. The represented races are determined by the nearest test image to each prototype.

be seen in Fig. 5. To evaluate the demographic diversity, we asses the race of the nearest test image to each prototype and use this to compute the combinatorial diversity of the race distribution. The overall accuracy and diversity results are reported in Tab. 3. We also report the accuracy gap between White males and Black females. This accuracy gap has been identified as a ubiquitous problem in facial recognition algorithms. White males account for 23 percent of the individuals in the UTK Face data, while Black females account for 9 percent. PanVAE achieves a lower accuracy gap between these demographics due to a better accuracy on Black females. However, this comes at the expense of a lower accuracy on the majority sub-population of White males as compared to ProtoVAE.

Discussion

The volumetric loss in PanVAE promotes the generation of diverse prototypes, which capture the underlying dataspace and represent distinct archetypical patterns in the data. This leads to increased representation quality and data coverage and can mitigate data bias. However, pantypes are most useful when the diversity expressed by the input data aligns with the diversity a study aims to enforce. This is closely related to the concepts of geometric and combinatorial diversity (Celis et al. 2016), where geometric diversity expresses the volume spanned by a number of high-dimensional feature vectors and combinatorial diversity is related to information entropy of discrete variables. This means that geometric diversity is useful for ensuring what humans perceive as high *visual* diversity, while combinatorial diversity is useful for ensuring high *demographic* diversity (or fairness) of human understandable sensitive variables that take on a small number of discrete values (such as race). The volumetric loss in PanVAE exclusively ensures a large geometric diversity of the learned pantypes and as such only enforces visually diversity. This may not necessarily align with the diversity in unseen protected attributes such as race in facial image data. This misalignment can occur if features like background color and pose in the facial images exhibit larger visual variation than features related to demographic diversity such as skin tone. To enforce high demographic diversity, the images would either have to be pose aligned and background removed (or at at least background noise reduced) or the sensitive features would have to be incorporated directly into the model, if possible. We have trained PanVAE on the cropped and aligned version of the UTK Face dataset to demonstrate that geometric and combinatorial diversity can be obtained simultaneously in noise reduced data with the volumetric loss. More balanced demographic representation can lead to better predictive performance for minority sub-populations in the data and consequently less disparate predictive performance between subpopulations. However, this usually comes at the expense of a reduction in performance for the majority group. Thus, the choice of representation should be carefully considered in coherence with the aim and target population of the trained model.

Conclusion

We have introduced pantypes, a new family of prototypical objects used in a SEM to capture the full diversity of the dataspace. Pantypes emerge by virtue of a volumetric loss and are easily integrated into existing prototypical self-explainable classifier frameworks. The volumetric loss causes the pantypes to diverge early in the training process and to capture various archetypical patterns through a sparse set of objects leading to increased interpretability and representation quality without sacrificing accuracy.

References

Alvarez Melis, D.; and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Berman, M.; Kiiveri, H.; Lagerstrom, R.; Ernst, A.; Dunne, R.; and Huntington, J. F. 2004. ICE: A statistical approach to identifying endmembers in hyperspectral images. *IEEE transactions on Geoscience and Remote Sensing*, 42(10): 2085–2095.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Celis, L. E.; Deshpande, A.; Kathuria, T.; and Vishnoi, N. K. 2016. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019a. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019b. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Davies, D. L.; and Bouldin, D. W. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 224–227.

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.

Gautam, S.; Höhne, M. M.-C.; Hansen, S.; Jenssen, R.; and Kampffmeyer, M. 2023. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. *Pattern Recognition*, 136: 109172.

Gong, B.; Chao, W.-L.; Grauman, K.; and Sha, F. 2014. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27.

Ha, D.; and Eck, D. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

Jin, Z.; Xu, M.; Sun, C.; Asudeh, A.; and Jagadish, H. 2020. Mithracoverage: a system for investigating population bias for intersectional fairness. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2721–2724.

Jing, L.; Vincent, P.; LeCun, Y.; and Tian, Y. 2021. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*.

Krause, A.; Singh, A.; and Guestrin, C. 2008. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2).

Kulesza, A.; Taskar, B.; et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3): 123–286.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lin, H.; and Bilmes, J. A. 2012. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Phillips, P. J.; Jiang, F.; Narvekar, A.; Ayyad, J.; and O'Toole, A. J. 2011. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2): 1–11.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shahbazi, N.; Lin, Y.; Asudeh, A.; and Jagadish, H. 2022. A Survey on Techniques for Identifying and Resolving Representation Bias in Data. *arXiv preprint arXiv:2203.11852*.

Simpson, E. H. 1949. Measurement of diversity. *nature*, 163(4148): 688–688.

Suresh, H.; and Guttag, J. V. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8).

Vilone, G.; and Longo, L. 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Wang, J.; Liu, H.; Wang, X.; and Jing, L. 2021. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 895–904.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; and Lipson, H. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5810–5818.

Zhou, T.; Kuscsik, Z.; Liu, J.-G.; Medo, M.; Wakeling, J. R.; and Zhang, Y.-C. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10): 4511–4515.

Acknowledgments

We would like to acknowledge the authors of the well-tested ProtoVAE. We have used the public code for this model as the foundation of PanVAE.