ProtoVAE: A Trustworthy Self-Explainable Prototypical Variational Model

Srishti Gautam

Ahcene Boubekki

Stine Hansen

Suaiba Amina Salahuddin

Robert Jenssen

Marina M.-C. Höhne

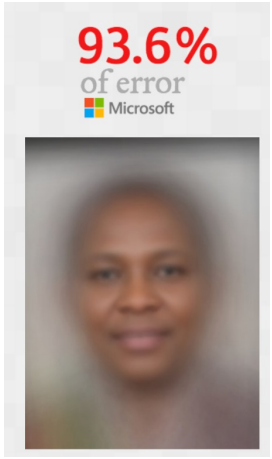
Michael Kampffmeyer

Why is Explainable AI important?

Capturing unwanted bias in the data

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Joy Buolamwini, Timnit Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”



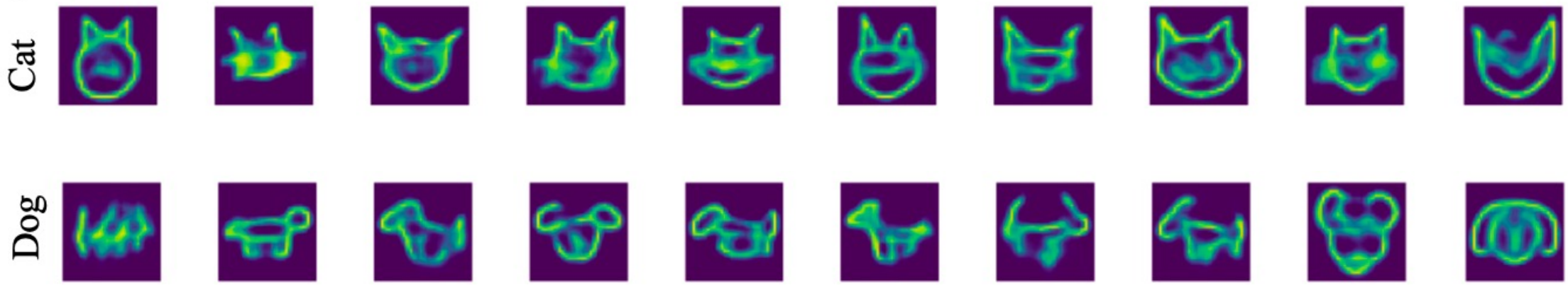
93.6% of faces misgendered by Microsoft were those of darker subjects

Concept/Prototypical Self-Explainable Models

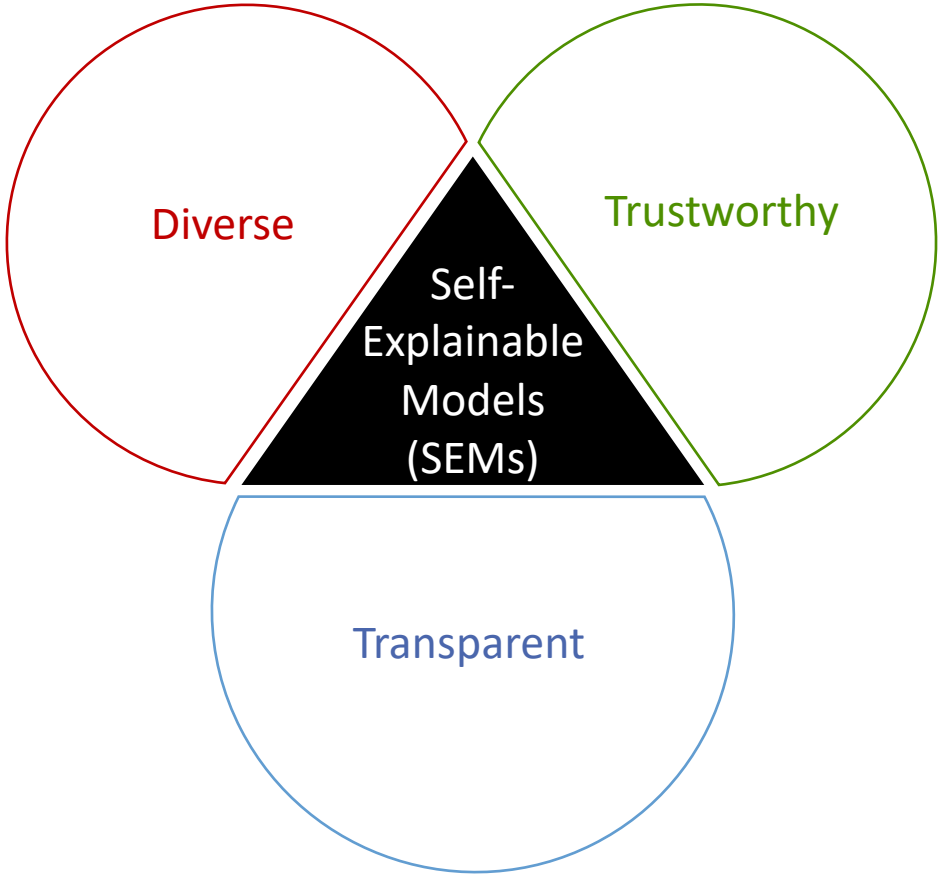
Self-Explainable Models: Provides explanations and labels at the same time.

Prototypical Self-Explainable Models: Learns representatives of the class

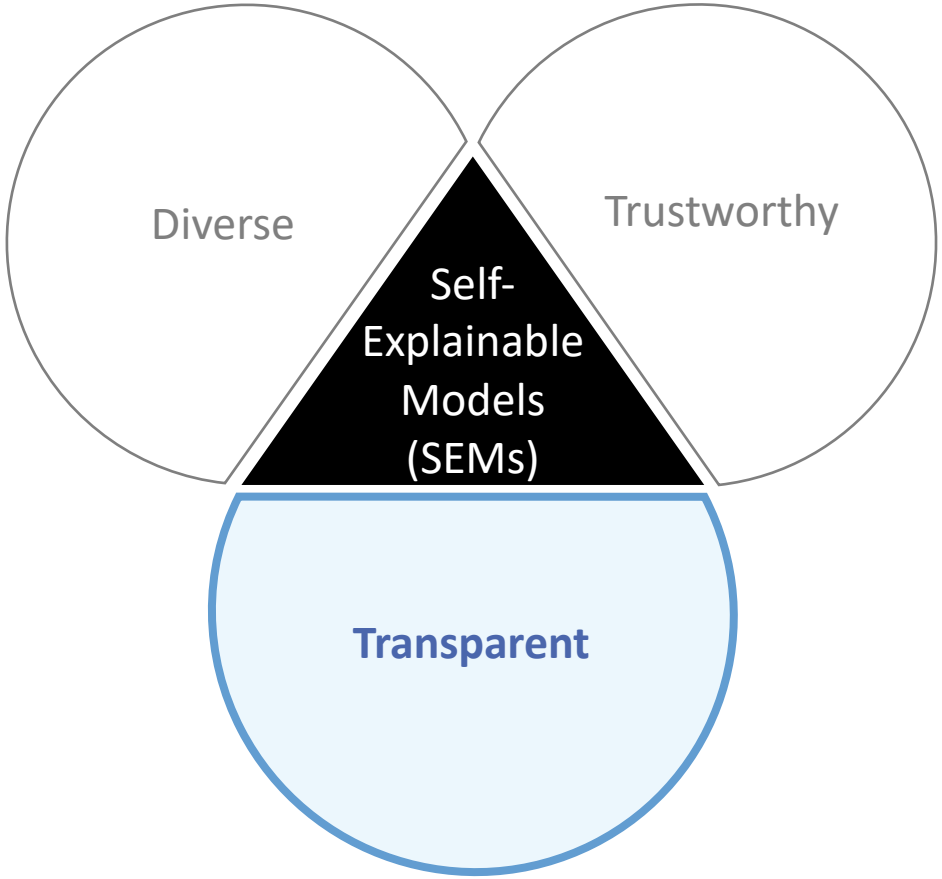
QuickDraw



Predicates for a self-explainable model



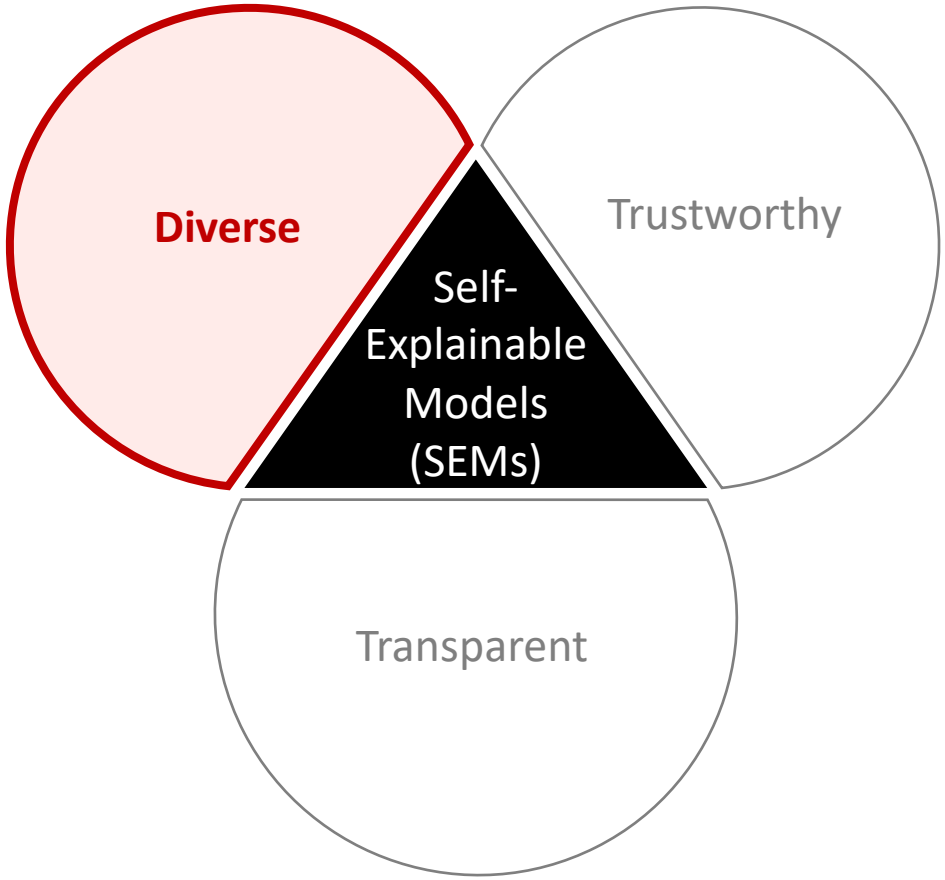
Predicates for a self-explainable model



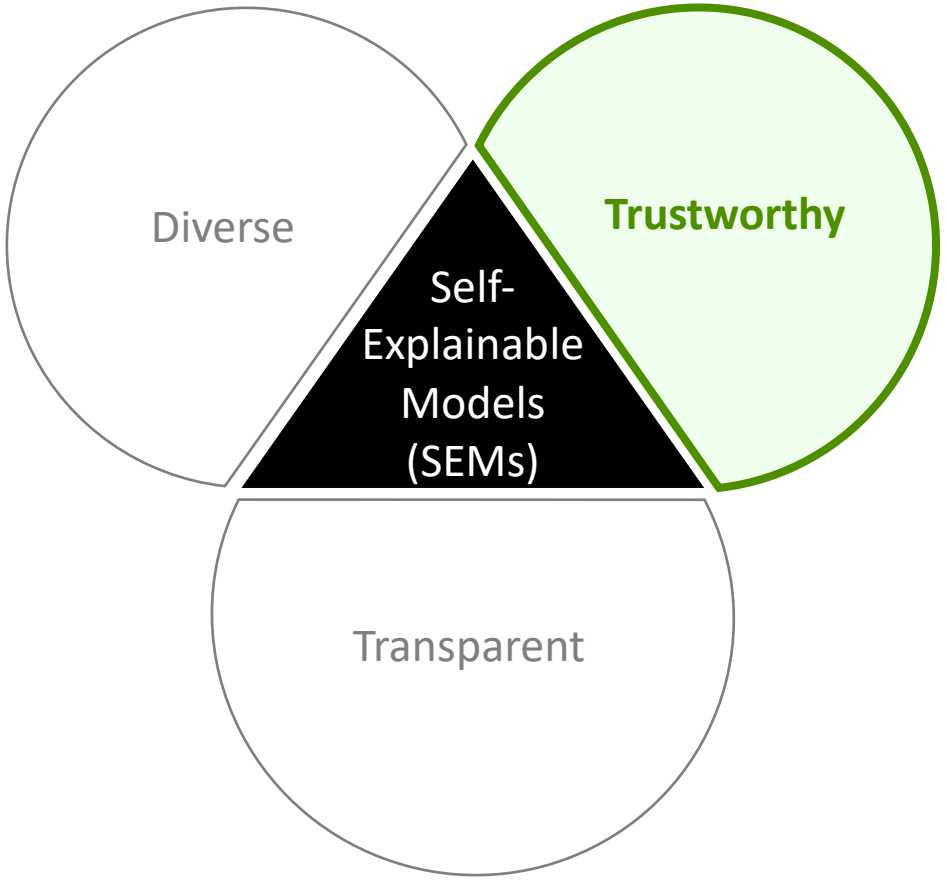
Definition 1 *An SEM is* ***transparent*** *if:*
(i) its ***concepts*** *are utilized to perform the* ***downstream tasks*** *without leveraging a complex black-box model;*
(ii) its ***concepts*** *are* ***visualizable*** *in input space*

Predicates for a self-explainable model

Definition 2 *An SEM is **diverse** if: its **concepts** represent **non-overlapping information** in the input space.*



Predicates for a self-explainable model



Definition 3 *An SEM is* ***trustworthy*** *if:*

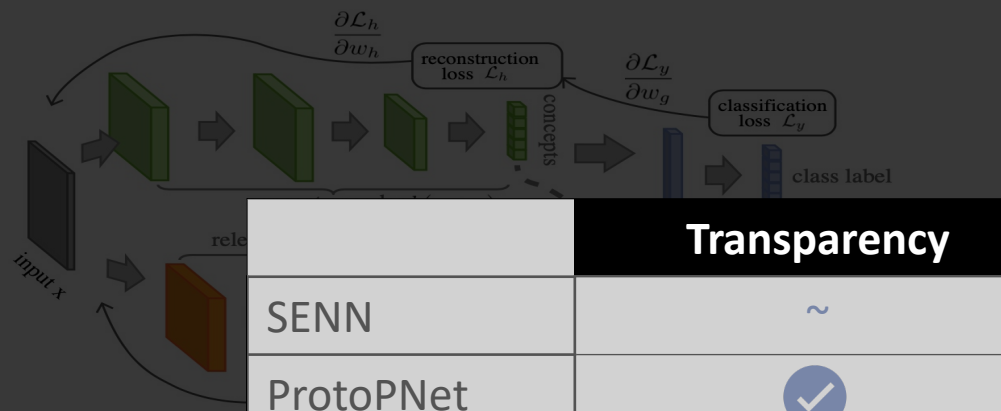
(i) the ***performance*** *matches to that of the closest black-box counterpart;*

(ii) the ***explanations*** *are* ***robust*** *i.e., similar images yield similar explanations;*

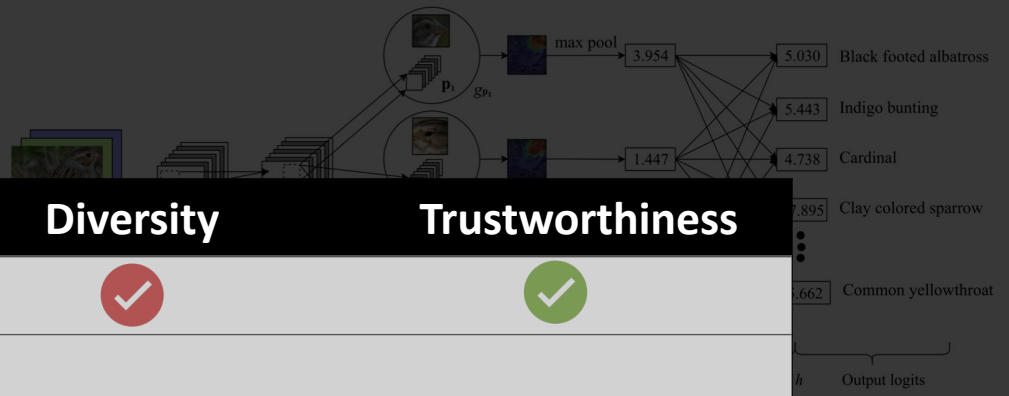
(iii) the ***explanations*** *represent the* ***real contribution*** *of the input features to the prediction.*

Prior Self-Explainable Models

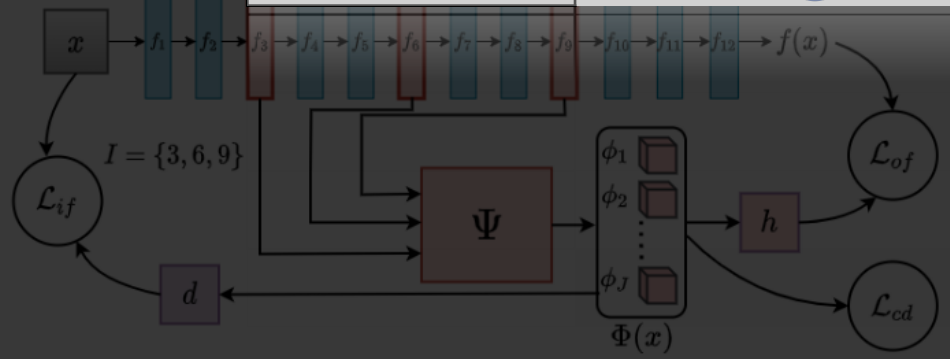
	Transparency	Diversity	Trustworthiness
SENN	~	✓	✓
ProtoPNet	✓		
SITE		✓	✓
FLINT	~	✓	
ProtoVAE	✓	✓	✓



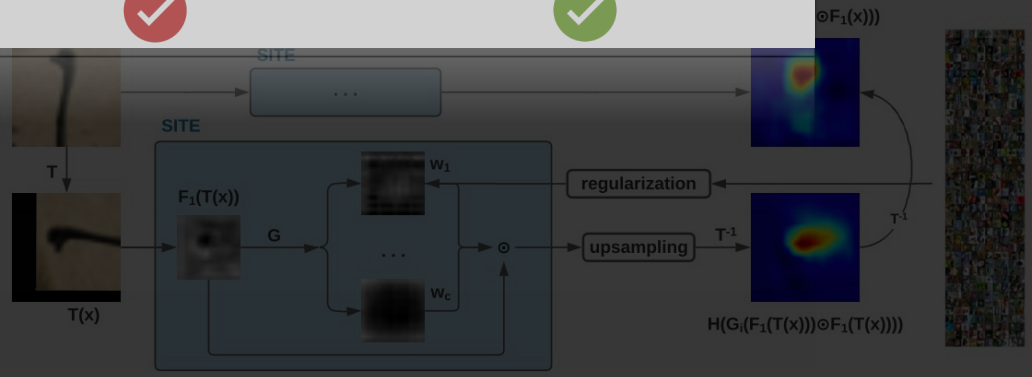
SENN



19)



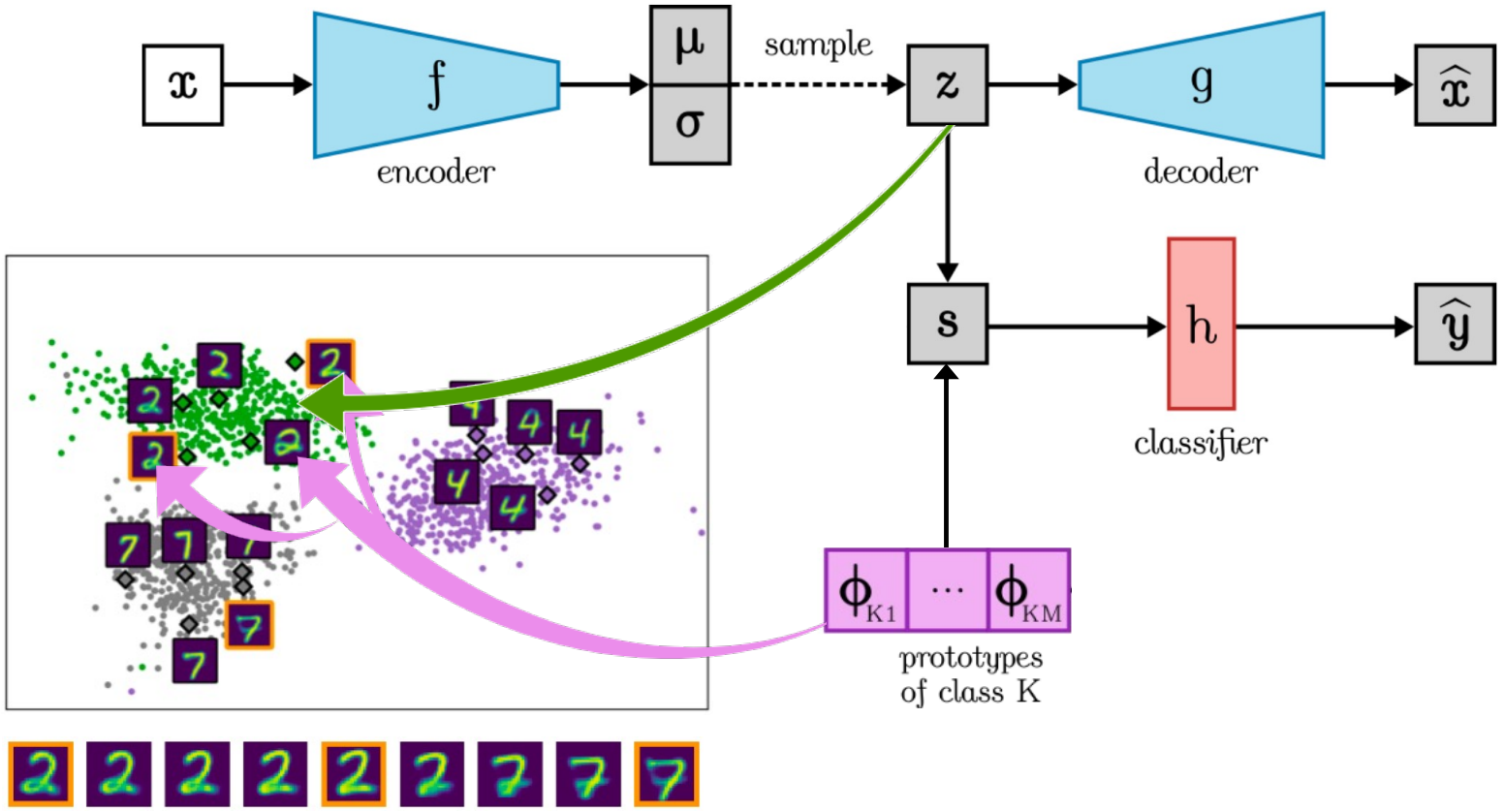
FLINT (Parekh et al. NeurIPS 2021)



SITE (Wang et al. NeurIPS 2021)

ProtoVAE

Transparent architecture



The input image $\boldsymbol{x}$ is encoded by f into a tuple $(\boldsymbol{\mu}, \boldsymbol{\sigma})$. A vector $\boldsymbol{z}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ which, on one side, is decoded by g into the reconstructed input $\hat{\boldsymbol{x}}$ and, on the other side, is compared to the prototypes ϕ_{kj} resulting in the similarity scores s. The latter are passed through the classifier h to get the final prediction $\hat{\boldsymbol{y}}$.

ProtoVAE

Diversity and trustworthiness through loss

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

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

Inter-class diversity

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

Intra-class diversity

$$\mathcal{L}_{\text{VAE}} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||^2 + \sum_{k=1}^{K}\sum_{j=1}^{M} \boldsymbol{y}_i(k) \frac{\boldsymbol{s}_i(k,j)}{\sum_{l=1}^{M} \boldsymbol{s}_i(k,l)} D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) || \mathcal{N}(\boldsymbol{\phi}_{kj}, \mathbf{I}_d)\big)$$

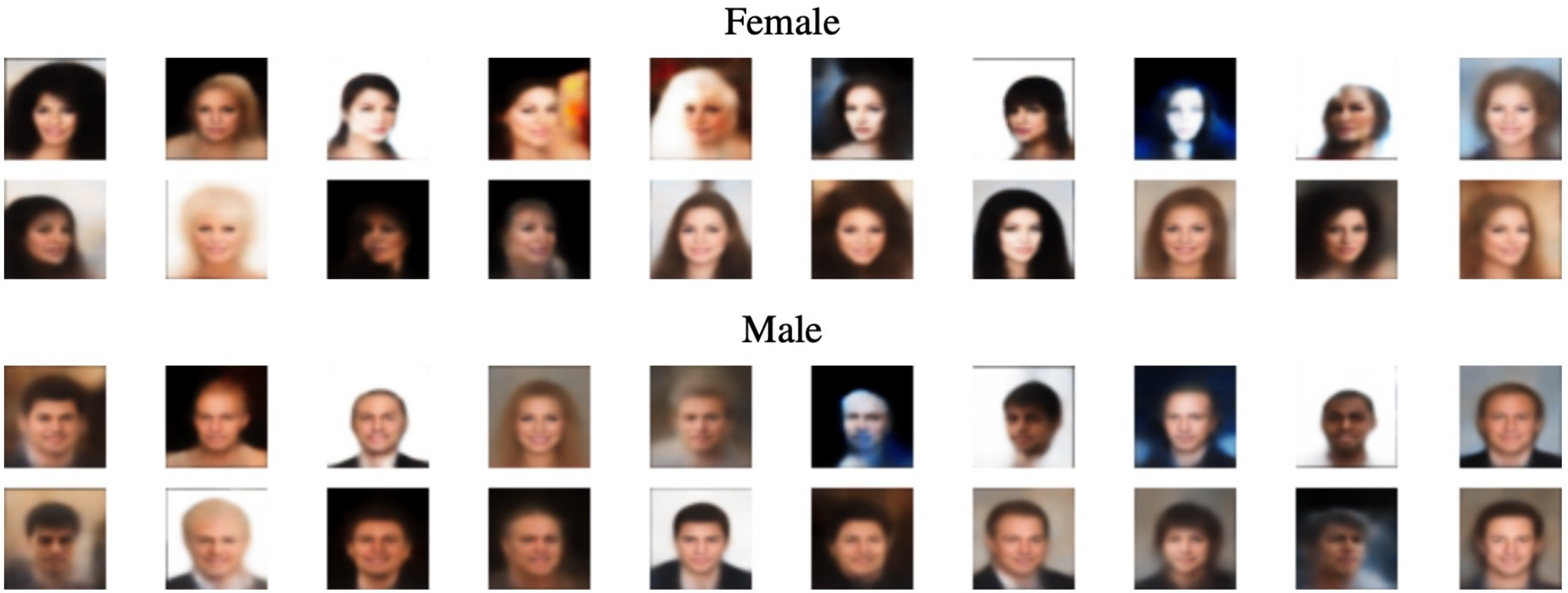
Robust classification and reconstruction

Predictive performance

	Black-box encoder	FLINT	SENN	*SITE	ProtoPNet	ProtoVAE
MNIST	99.2±0.1	**99.4±0.1**	98.8±0.7	98.8	94.7±0.6	**99.4±0.1**
fMNIST	91.5±0.2	91.5±0.2	88.3±0.3	-	85.4±0.6	**91.9±0.2**
CIFAR-10	83.9±0.1	79.6±0.6	76.3±0.2	84.0	67.8±0.9	**84.6±0.1**
QuickDraw	86.7±0.4	82.6±1.4	79.3±0.3	-	58.7±0.0	**87.5±0.1**
SVHN	**92.3±0.3**	90.8±0.4	91.5±0.4	-	88.6±0.3	**92.2±0.3**

Results for accuracy (in %) for ProtoVAE and comparison with other state-of-the-art methods. *Results for SITE are taken from the original paper and thus based on more complex architectures.

Global explanations

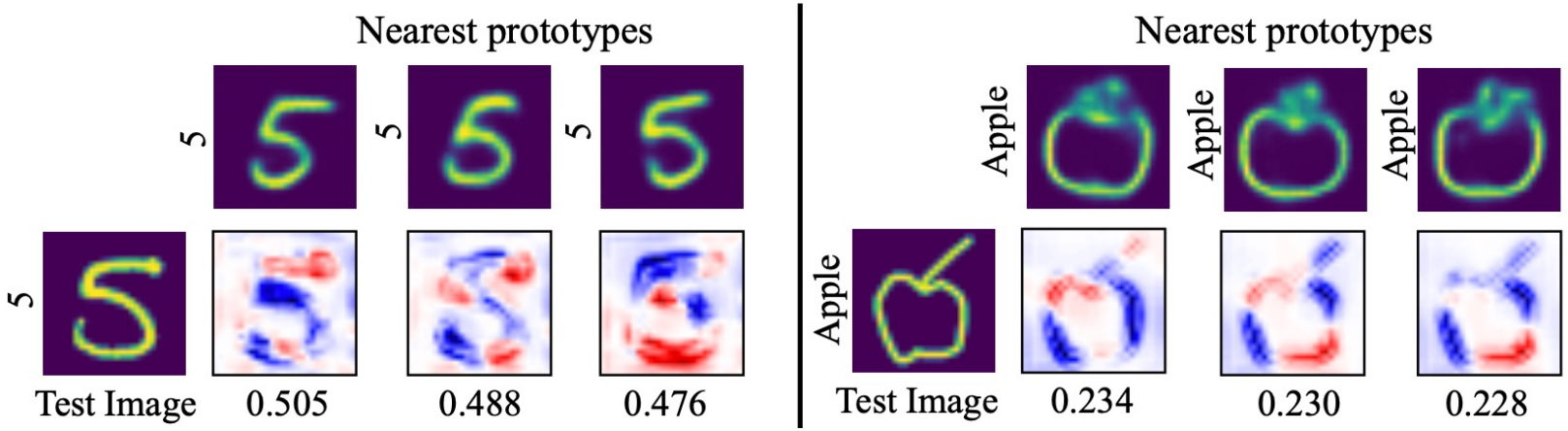


Prototypes learned for CelebA dataset with ProtoVAE

Local explanations

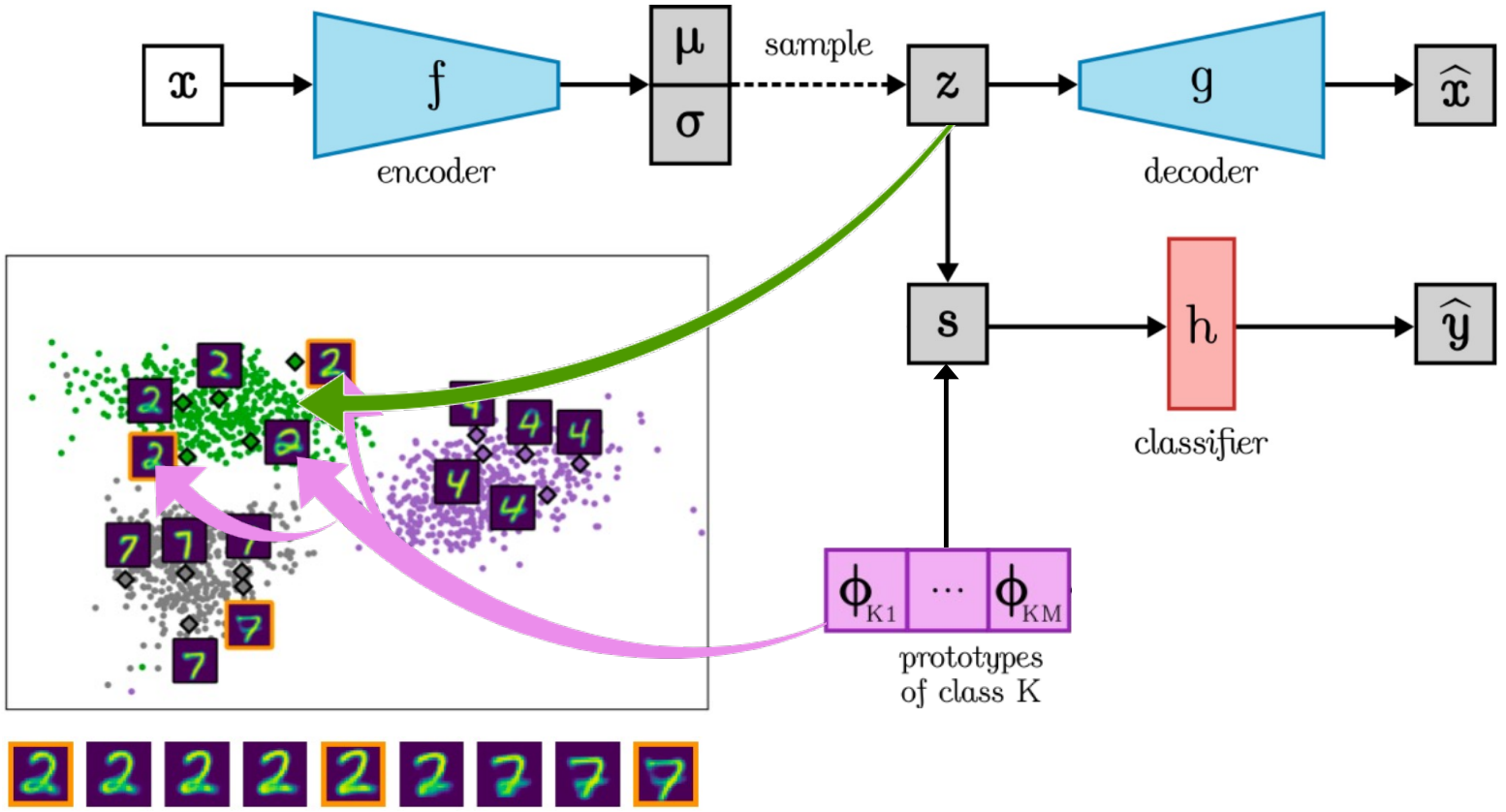
Prototypical Relevance Propagation (PRP)

[S. Gautam, M. Höhne, S. Hansen, R. Jenssen, M. Kampffmeyer
"This looks more like that: Enhancing Self-Explaining Models by Prototypical Relevance Propagation", 2021 arXiv.]



Three maximally activated prototypes, the corresponding prototypical activations, and corresponding similarity scores for a test image of class 5 (for MNIST) and apple (for QuickDraw).

ProtoVAE: A Trustworthy Self-Explainable Prototypical Variational Model



Snapshot

- A *transparent* probabilistic self-explainable model.
- Generates *diverse* and *trustworthy* prototypical explanations.
- Performs on-par or better than existing self-explainable models as well as black-box counterparts.

NEURAL INFORMATION PROCESSING SYSTEMS

github.com/SrishtiGautam/ProtoVAE

srishti.gautam@uit.no