ProtoVAE

January 24, 2024

Reviewr : Park Seok Hun

Table of Contents

Table of Contents

Definition of self-explainable model(SEM)

- They proposed the three properties that are prerequisites for SEM.
- Properties : transparent, diverse , trustworthy

Transparent

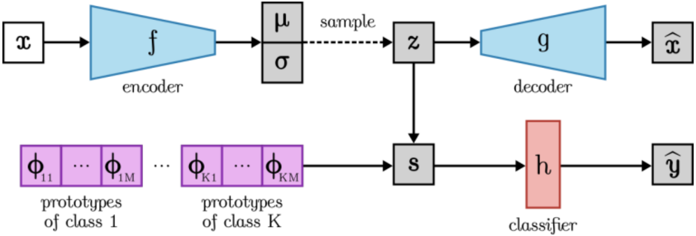
- An SEM is transparent if
  1. its concepts are utilized to perform the downstream task without leveraging a complex black box model
  2. its concepts are visualizable in input space.

Diverse and trustworthy

- An SEM is diverse if
  1. its concepts represent non-overlapping information in the latent space.
- An SEM is trustworthy if
  1. its performance matches to that of the closet black-box counterpart.
  2. the explanations are robust.
  3. the explanations represent the real contribution of the input features to the prediction.

Table of Contents

Model



x
f
encoder
μ
σ
sample
z
g
decoder
x̂
ϕ11 ⋯ ϕ1M
prototypes
of class 1
⋯
ϕK1 ⋯ ϕKM
prototypes
of class K
s
h
classifier
ŷ

Model

- Let $\Phi = \{\phi_{kj}\}_{k=1,\dots,K, j=1,\dots,M}$ be prototypes parameters where K is the number of class and M is the number of prototypes per class.
- $z_i = f(x_i)$ be the latent vector for input x_i where f is the encoder.
- Using the following function to calculate the similarity between z_i and the parameters of the prototypes.

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

  where $0 < \epsilon < 1$.
- $\hat{y}_i = h(s_i)$ where $s_i = (s_i(k,j), k, j)'$ and h is linear classifier.

Loss

- $Loss = L_{pred} + L_{orth} + L_{VAE}$
- $L_{pred} = \frac{1}{n}\sum_{i=1}^{N} CE(h(s_i), y_i)$ where y_i is true label.
- $L_{orth} = \sum_{k=1}^{K} \|\Phi_k^t \Phi_k - I_M\|_F^2$ where $\Phi_k = (\phi_{kj}, j = 1, .., M)'$
- L_{orth} forces the prototypes of vae to be diverse in the class.

Table of Contents

visualization

- The prototypes parameter can decoded via decoder of VAE.

Table of Contents

Experiments

Table 2: Performance results of ProtoVAE compared to other state-of-the-art methods (measured in accuracy (in %)). The reported numbers are means and standard deviations over 4 runs. Best and statistically non-significantly different results are marked in bold. *Results for SITE are taken from the original paper and thus based on more complex architectures.

	Black-box encoder	FLINT [13]	SENN [8]	*SITE [17]	ProtoPNet [9]	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**