Indirectly Parameterized Concrete Autoencoders

Alfred Nilsson*, Klas Wijk*, Sai bharath chandra Gutha, Erik Englesson, Alexandra Hotti, Carlo Saccardi, Oskar Kviman, Jens Lagergren, Ricardo Vinuesa, Hossein Azizpour

We propose an improvement to Concrete Autoencoders (CAEs), a state-of-the-art technique for embedded feature selection in neural networks. By learning an embedding and mapping it to the parameters of the Gumbel-Softmax distribution, our Indirectly Parameterized CAEs (IP-CAEs) improve training stability.

Keywords

- Feature selection
- Gumbel-Softmax
- End-to-end differentiable optimization

Problem

Embedded Feature Selection

- CAEs enable the simultaneous learning of complex models and feature selection, extending beyond classical linear methods.
- Currently state-of-the-art in neural network-based embedded feature selection.

CAE Training Instability

We identify that CAEs often learn *duplicate selections,* and it affects convergence speed and generalization.

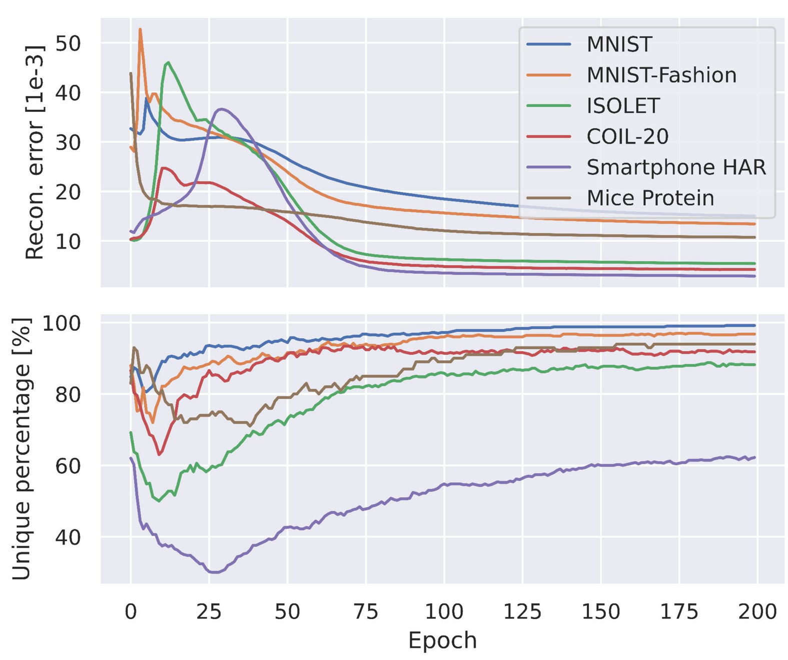


Figure 1: Top) Unstable reconstruction loss. Bottom) The Unique Percentage, a measure of the diversity of feature selections. We observe that the learning of duplicate selections is correlated with training instability.

Method

Concrete Autoencoders and Gumbel-Softmax

CAEs learn features through k stochastic nodes. Each node entails: Drawing a sample $\boldsymbol{m}_j \in \mathbb{R}^D$ from a learned Gumbel-Softmax (GS) distribution

$$\boldsymbol{m}_j = \frac{\exp\{(\log \boldsymbol{\alpha}_j + \boldsymbol{g}_j)/T\}}{\sum_{i=1}^{D} \exp\{(\log \boldsymbol{\alpha}_{j,i} + \boldsymbol{g}_{j,i})/T\}},$$

and multiplying it with the input $\mathbf{x} \in \mathbb{R}^D$. Each GS distribution is parameterized by a learned vector $\log \boldsymbol{\alpha}_j \in \mathbb{R}^D$.

Indirect Parameterization

We propose parameterizing $\log \boldsymbol{\alpha} \in \mathbb{R}^{K \times D}$ with an array of learnable parameters $\Psi \in \mathbb{R}^{K \times P}$ with a linear transformation $(\boldsymbol{W}, \boldsymbol{b})$, where $\boldsymbol{W} \in \mathbb{R}^{D \times P}$ and $\boldsymbol{b} \in \mathbb{R}^D$.

$$\log \boldsymbol{\alpha}_i = \boldsymbol{W}\boldsymbol{\psi}_i + \boldsymbol{b}, \quad i \in [K],$$

Empirically, we observe that this indirect parameterization results in:

- Fewer duplicate selections.
- Increased convergence speed.
- Better performance in classification and reconstruction tasks.

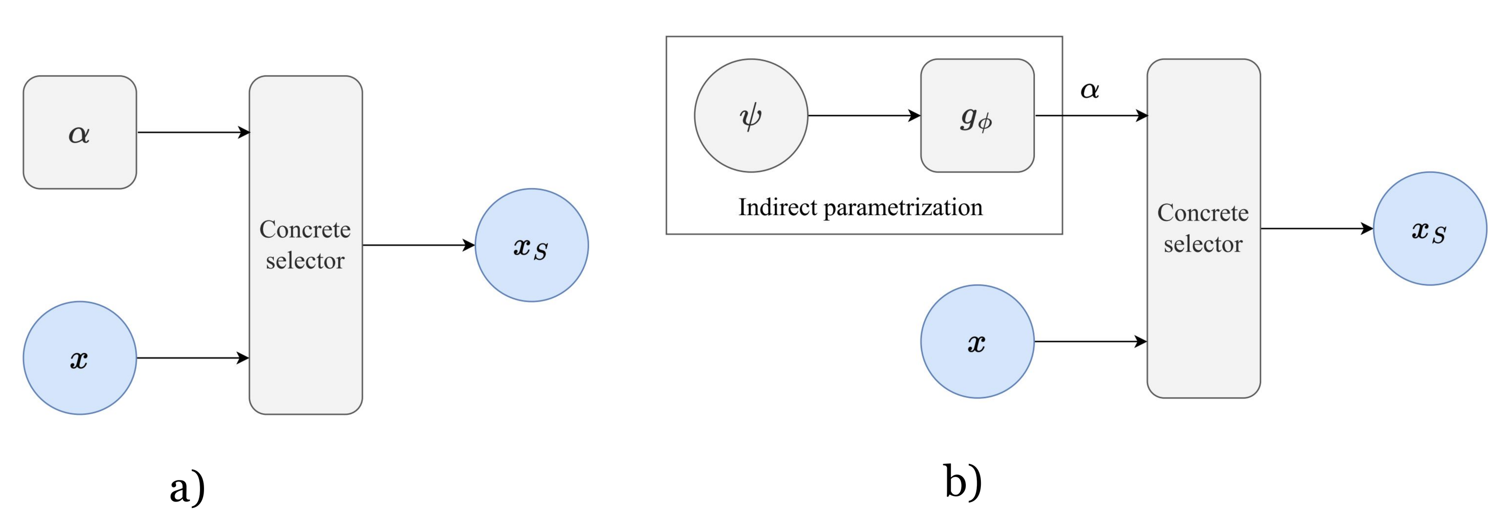


Figure 2: a) CAE parameterization. b) Indirect parameterization.

Results

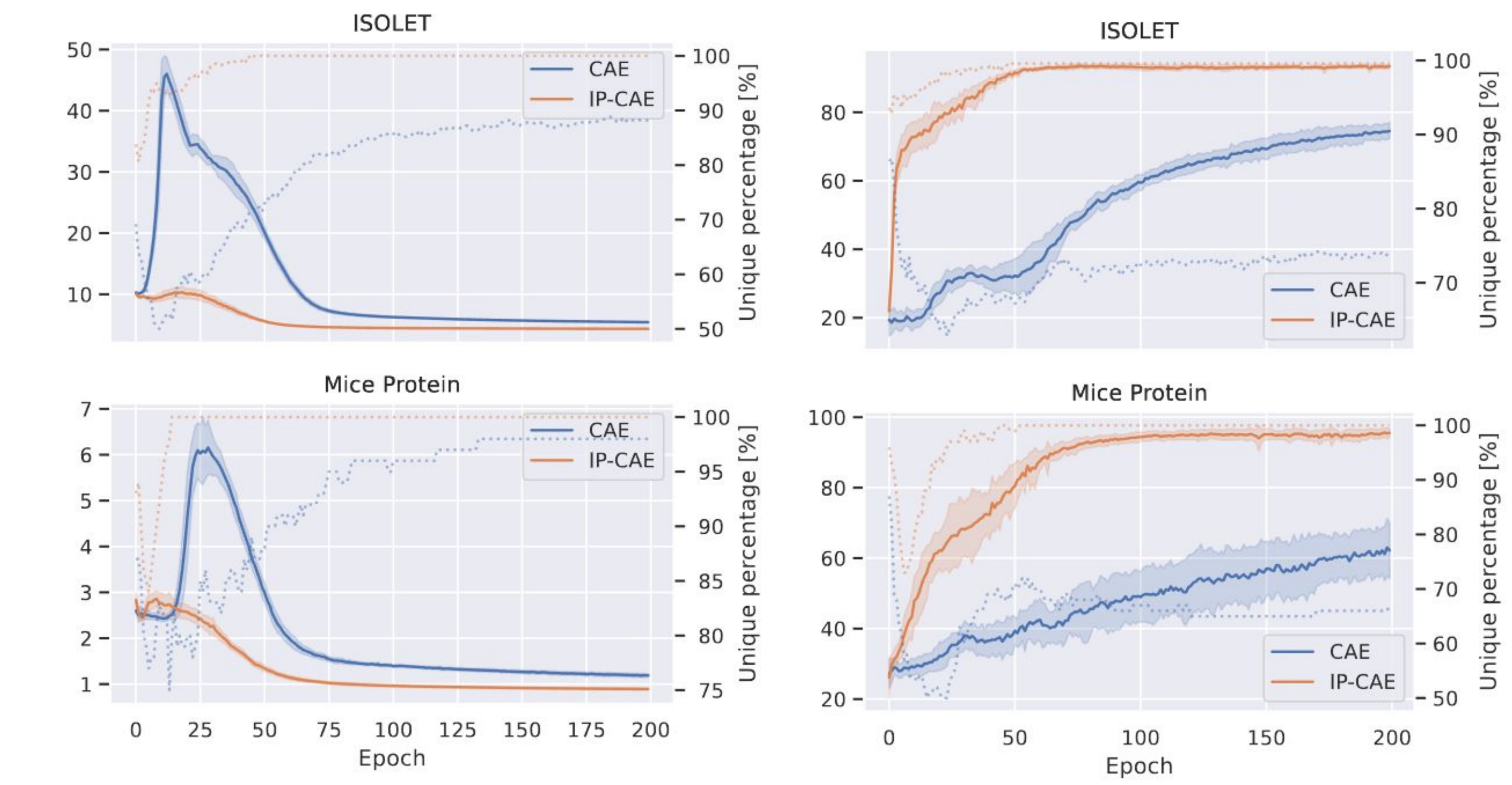


Figure 3: Improved convergence speed, MSE (left), and accuracy (right) for the ISOLET and Mice Protein datasets.

Table: Accuracy. Comparison to related works on feature selection.

Model	MNIST	MNIST-Fashion	ISOLET	COIL-20	Smartphone HAR	Mice Protein
STG	92.29 ±0.30	80.85 ±0.27	84.95 ±0.31	96.80 ±0.25	88.80 ±0.08	68.24 ±1.11
LassoNet	90.06 ±0.33	78.28 ±0.36	84.33 ±0.28	89.37 ±0.47	92.44 ±0.11	77.12 ±0.80
CAE	83.10 ±1.23	73.19 ±0.82	75.82 ±2.31	80.70 ±2.93	82.72 ±0.80	63.10 ±6.51
GJSD	84.38 ±1.50	74.13 ±0.64	77.56 ±0.82	82.10 ±3.72	84.78 ±1.04	68.43 ±7.75
IP-CAE	**94.07** ±0.37	**82.68** ±0.80	**91.85** ±0.55	**97.92** ±0.57	**93.71** ±0.62	**94.26** ± 1.48

Future Work

- Theoretical understanding
- Gumbel-Softmax applications beyond feature selection

Paper

Code

Indirectly Parameterized Concrete Autoencoders

Alfred Nilsson, Klas Wijk, Sai bharath chandra Gutha, Erik Englesson, Alexandra Hotti, Carlo Saccardi, Oskar Kviman, Jens Lagergren, Ricardo Vinuesa, Hossein Azizpour
KTH Royal Institute of Technology

We propose an improvement to Concrete Autoencoders (CAEs), a state-of-the-art technique for embedded feature selection in neural networks. By learning an embedding and mapping it to the parameters of the Gumbel-Softmax distribution, our Indirectly Parameterized CAEs (IP-CAEs) improve training stability.

Embedded feature selection.

- CAEs enable the simultaneous learning of complex models and feature selection, extending beyond classical linear methods.
- Currently state-of-the-art in neural network-based embedded feature selection,

Training instability in CAEs.

We identify that CAEs often learn *duplicate selections,* and it affects convergence speed and generalization.

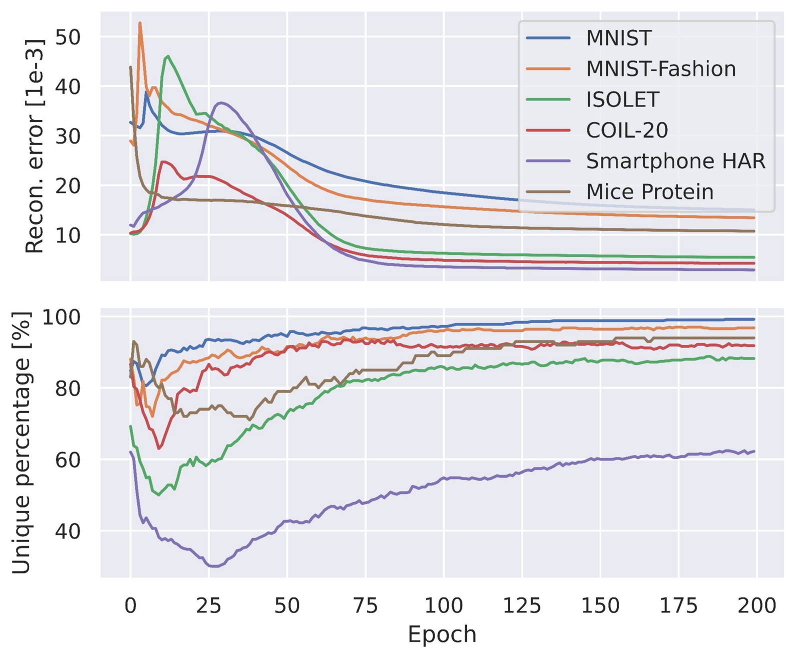


Figure 1: Top) Unstable reconstruction loss. Bottom) The Unique Percentage, a measure of the diversity of feature selections. We observe that the learning of duplicate selections is correlated with training instability.

Concrete Autoencoders and Gumbel-Softmax

CAEs learn features through k stochastic nodes. Each node entails:

- Drawing a sample $\boldsymbol{m}_j \in \mathrm{R}^{\mathrm{D}}$ from a learned Gumbel-Softmax (GS) distribution,

$$\boldsymbol{m}_j = \frac{\exp\{(\log \boldsymbol{\alpha}_j + \boldsymbol{g}_j)/T\}}{\sum_{i=1}^{D} \exp\{(\log \boldsymbol{\alpha}_{j,i} + \boldsymbol{g}_{j,i})/T\}},$$

- Multiplying it with the input $\mathbf{x} \in \mathrm{R}^{\mathrm{D}}$.

Each GS distribution is parameterized with a by a learned vector $\log \boldsymbol{\alpha}_j \in \mathrm{R}^{\mathrm{D}}$.

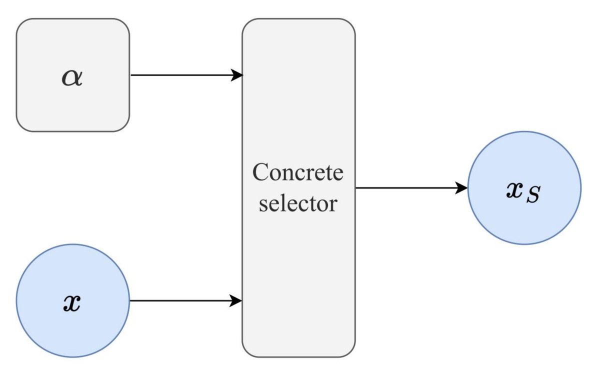
Indirect Parameterization

We propose parameterizing $\log \boldsymbol{\alpha} \in \mathrm{R}^{\mathrm{K \times D}}$ with an array of learnable parameters $\Psi \in \mathrm{R}^{\mathrm{K \times P}}$ with a linear transformation $(\boldsymbol{W}, \boldsymbol{b})$, where $\boldsymbol{W} \in \mathrm{R}^{\mathrm{D \times P}}$ and $\boldsymbol{b} \in \mathrm{R}^{\mathrm{D}}$.

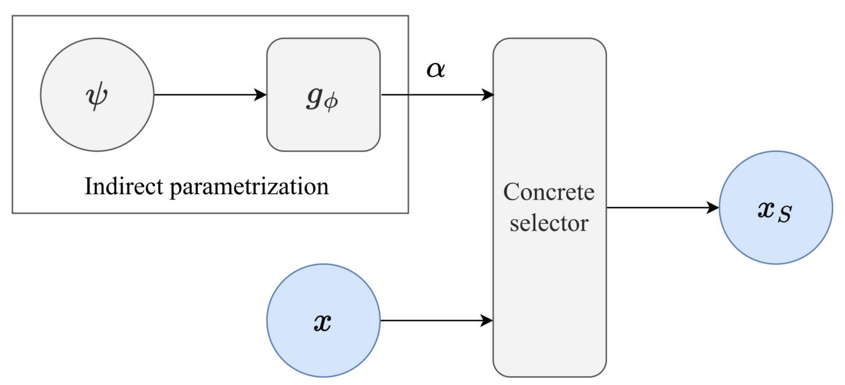
$$\log \boldsymbol{\alpha}_i = \boldsymbol{W} \boldsymbol{\psi}_i + \boldsymbol{b}, \quad i \in [K],$$

Empirically, we observe that this indirect parameterization results in:

- Less duplicate selections.
- Increased convergence speed.
- Better performance in classification and reconstruction tasks.



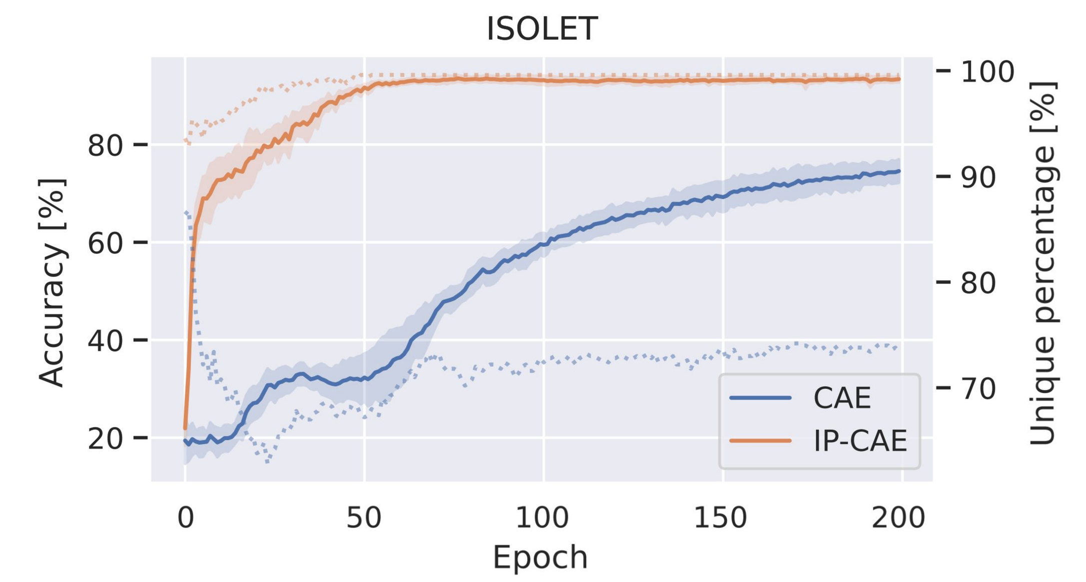
a)



b)

Figure 2: a) CAE parameterization. b) Indirect parameterization.

Results



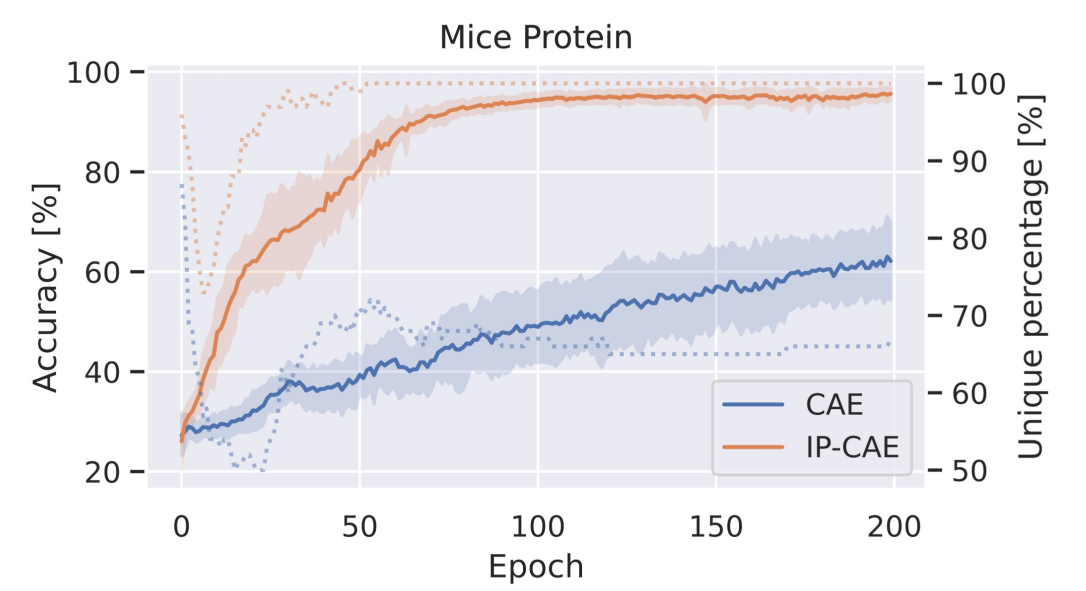


Figure 3: Improved convergence speed and accuracy for the ISOLET and Mice Protein datasets.

Model	MNIST	MNIST-Fashion	ISOLET	COIL-20	Smartphone HAR	Mice Protein
STG	92.29 ±0.30	80.85 ±0.27	84.95 ±0.31	96.80 ±0.25	88.80 ±0.08	68.24 ±1.11
LassoNet	90.06 ±0.33	78.28 ±0.36	84.33 ±0.28	89.37 ±0.47	92.44 ±0.11	77.12 ±0.80
CAE	83.10 ±1.23	73.19 ±0.82	75.82 ±2.31	80.70 ±2.93	82.72 ±0.80	63.10 ±6.51
GJSD	84.38 ±1.50	74.13 ±0.64	77.56 ±0.82	82.10 ±3.72	84.78 ±1.04	68.43 ±7.75
IP-CAE	**94.07** ±0.37	**82.68** ±0.80	**91.85** ±0.55	**97.92** ±0.57	**93.71** ±0.62	**94.26** ± 1.48

Table: Accuracy. Comparison to related works on feature selection.

Assets

Paper Code

Stacking the k $\{\boldsymbol{m}_j\}$ samples in a matrix M, the selected features $\mathbf{x}_S$ can be expressed as:

$$\mathbf{x}_S = M\mathbf{x},$$