Improving Sparse Autoencoder with Dynamic Attention

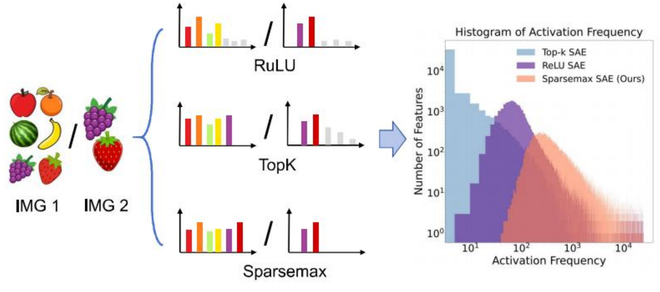
Dongsheng Wang, Jinsen Zhang, Dawei Su, Hui Huang* | Shenzhen University, China

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

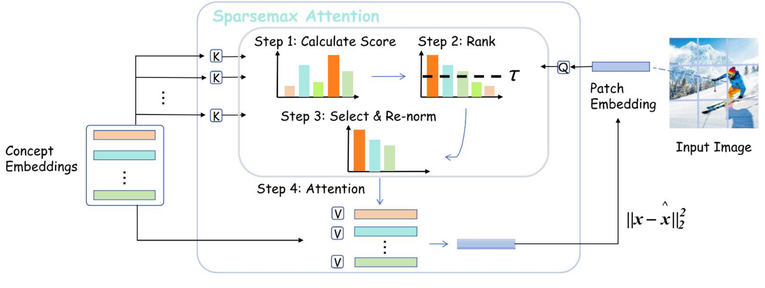
$$z = \sigma(\mathbf{W}_{\text{enc}}(x - b_{\text{enc}})),$$
$$\hat{x} = \mathbf{W}_{\text{dec}} z + b_{\text{dec}},$$

- ReLU-based SAEs need extra sparsity regularization and may suffer feature shrinkage.
- TopK-based SAEs rely on a fixed K, which is not ideal for inputs with different complexity.



Method

Question:Can SAEs identify the optimal number of activations based on the feautre itself?



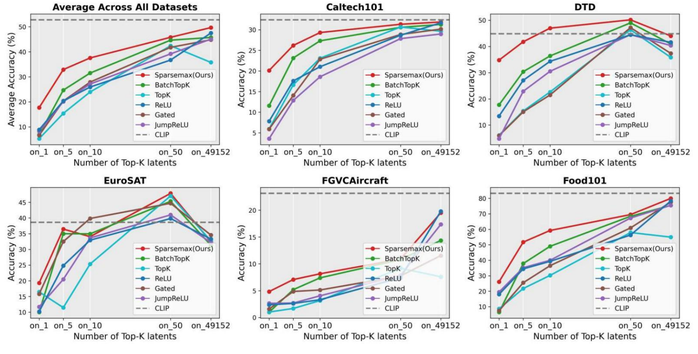
$$\mathbf{Q} = x^T \mathbf{W}_Q, \quad \mathbf{K} = \mathcal{C}^T \mathbf{W}_K, \quad \mathbf{V} = \mathcal{C}^T \mathbf{W}_V,$$
$$\hat{x} = \sigma(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \quad \text{sparsemax}(z) = \arg\min_{p \in \Delta^{K-1}} \|p - z\|^2$$

Algorithm

Algorithm 1 Sparsemax Attention

Input: z
Sort z as $z_{(1)} \geq \cdots \geq z_{(M)}$
Find $k(z) := \max\left\{ r \in [M] \,\middle|\, z_{(r)} + \frac{1-\sum_{i=1}^{r} z_{(i)}}{r} > 0 \right\}$, where $[M] := \{1, \ldots, M\}$
Define $\tau(z) = \frac{\sum_{i=1}^{k} z_{(i)} - 1}{kz}$.
Output: p such that $p_i = max\{0, z_i - \tau(\mathbf{z})\}$

Main results



Comparisions of zero-shot image classfication using top-n concepts on 11 datasets.

Model	L_0 ↑	FVU↓	CS↑	CKNNA↑	DO↓
ReLU	0.928	**0.098**	**0.953**	**0.812**	0.003
TopK	0.966	0.169	0.925	0.701	0.003
BatchTopK	0.814	0.278	0.904	0.750	0.002
Ours	**0.979**	0.129	0.934	0.796	**0.001**

Sparsity metric

Model	MEAN-MS	MAX-MS
ReLU	0.1627	0.9172
TopK	0.0548	0.8751
BatchTopK	0.1243	0.9031
Ours	**0.3484**	**0.9575**

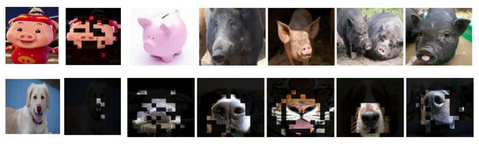
Interpretability metrics

Ablation

	on_1	on_5	on_10	on_50	on_49152
ReLU SAE	3.12	15.83	22.17	34.87	63.67
Transformer + ReLU	3.86	16.85	24.08	36.33	63.94
MLP + Sparsemax	7.91	29.87	39.73	55.32	64.74
Sparsemax SAE (Ours)	10.93	33.47	42.13	59.95	65.09

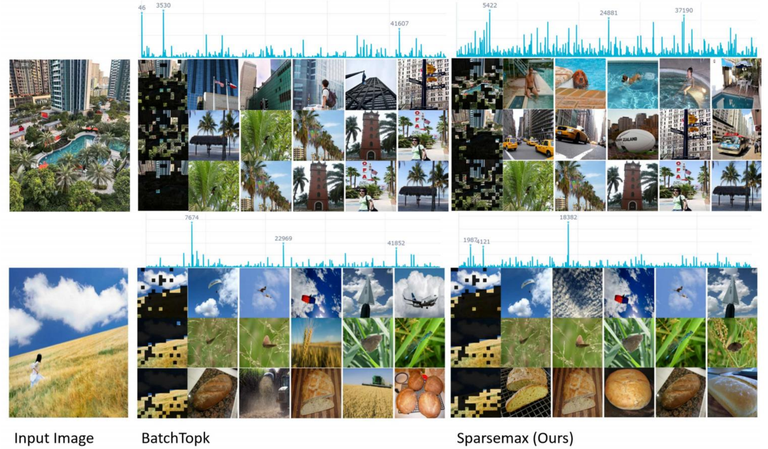
Concept Visualization

found concepts and related images



found concepts and related masked images

Case Study



Input Image BatchTopk Sparsemax (Ours)