Improving Sparse Autoencoder with Dynamic Attention

Supplementary Material

6. Proof of Prop. 1

1. Problem formulation. Recall that sparsemax is the Euclidean projection of $\boldsymbol{z} \in \mathbb{R}^M$ onto the probability simplex

$$\Delta^{M-1} = \{\boldsymbol{p} \in \mathbb{R}^M \mid \mathbf{1}^\top \boldsymbol{p} = 1,\ \boldsymbol{p} \geq \mathbf{0}\}, \tag{8}$$

i.e.

$$\text{sparsemax}(\boldsymbol{z}) = \arg\min_{\boldsymbol{p} \in \Delta^{M-1}} \tfrac{1}{2}\|\boldsymbol{p} - \boldsymbol{z}\|_2^2. \tag{9}$$

Equivalently, we solve the constrained quadratic program

$$\min_{\boldsymbol{p} \in \mathbb{R}^M} \frac{1}{2}\sum_{i=1}^{M}(p_i - z_i)^2 \quad \text{subject to} \quad \sum_{i=1}^{M} p_i = 1, \quad p_i \geq 0\ \forall i. \tag{10}$$

2. Lagrangian and KKT conditions. Introduce the Lagrange multiplier $\tau \in \mathbb{R}$ for the equality constraint $\sum_i p_i = 1$ and multipliers $\mu_i \geq 0$ for the inequality constraints $p_i \geq 0$. The (augmented) Lagrangian is

$$\mathcal{L}(\boldsymbol{p}, \tau, \boldsymbol{\mu}) = \frac{1}{2}\sum_{i=1}^{M}(p_i - z_i)^2 + \tau\Big(\sum_{i=1}^{M} p_i - 1\Big) - \sum_{i=1}^{M}\mu_i p_i. \tag{11}$$

Karush–Kuhn–Tucker (KKT) conditions for optimality are:
(i) *Primal feasibility*: $\sum_{i=1}^{M} p_i = 1$, $p_i \geq 0$ for all i.
(ii) *Dual feasibility*: $\mu_i \geq 0$ for all i.
(iii) *Stationarity*:

$$\frac{\partial \mathcal{L}}{\partial p_i} = p_i - z_i + \tau - \mu_i = 0 \quad \Longrightarrow \quad p_i = z_i - \tau + \mu_i, \qquad \forall i. \tag{12}$$

(iv) *Complementary slackness*:

$$\mu_i p_i = 0, \qquad \forall i. \tag{13}$$

Given Eq. 12 and Eq. 13, we have:

- If $p_i > 0$, complementary slackness forces $\mu_i = 0$. Hence

$$p_i = z_i - \tau \qquad \text{and thus} \qquad z_i - \tau > 0.$$

- If $p_i = 0$, complementary slackness allows $\mu_i \geq 0$ and stationarity gives

$$0 = z_i - \tau + \mu_i \quad \Longrightarrow \quad \mu_i = \tau - z_i.$$

Since $\mu_i \geq 0$, we conclude $\tau - z_i \geq 0$, i.e. $z_i \leq \tau$.

Combining both cases yields the compact expression (Eq. 4 in the main paper)

$$p_i = \max(z_i - \tau, 0), \qquad \forall i, \tag{14}$$

3. Determining τ and the active set. Let the active set (support) be

$$S = \{i \mid p_i > 0\} = \{i \mid z_i > \tau\}, \qquad k := |S|. \tag{15}$$

Summing $p_i = z_i - \tau$ over $i \in S$ and using $\sum_{i=1}^{M} p_i = 1$ gives

$$\sum_{i \in S}(z_i - \tau) = 1 \quad \Longrightarrow \quad \sum_{i \in S} z_i - k\tau = 1. \tag{16}$$

Thus

$$\tau = \frac{\sum_{i \in S} z_i - 1}{k}. \tag{17}$$

To find S efficiently, sort the coordinates in descending order:

$$z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(M)}. \tag{18}$$

If the optimal support corresponds to the top k indices (this is always the case: if some $j \in S$ were not among the top k, there would be an index in the top k not in S with a larger z, contradicting $z_j > \tau$), then

$$\tau_k = \frac{\sum_{i=1}^{k} z_{(i)} - 1}{k}. \tag{19}$$

The correct k is the largest integer for which the k-th sorted element exceeds this threshold:

$$z_{(k)} > \tau_k \quad \Longleftrightarrow \quad z_{(k)} + \frac{1 - \sum_{i=1}^{k} z_{(i)}}{k} > 0. \tag{20}$$

Therefore set

$$k = \max\left\{r \in \{1, \ldots, M\} \;\middle|\; z_{(r)} + \frac{1 - \sum_{i=1}^{r} z_{(i)}}{r} > 0\right\}, \tag{21}$$

and then take $\tau = \tau_k$, which completes the proof.

7. More Visualization Results

Analysis of Top 3 Concepts. Fig. 6- 8 show the visualizations of the top three concepts of the test image. Overall, we find that our Sparsemax SAE successfully captures the key visual patterns. For example, the mixture of fruit, wooden background, and apples in Fig. 6. Moreover, our approach is able to understand visual features from different perspectives. For example, the first concept in Fig. 7 corresponds to a pig (the main object of the input image), the second concept is related to cartoon characters, and the third concept is about the dressing.

Table 5. Sparsity metric values.

Model	$L_0 \uparrow$	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**

Table 6. Interpretability metrics.

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**

Failure Cases. We also provide several failure cases of our Sparsemax SAE in Fig. 9- 11. On one hand, we find that the learned concepts sometimes contain unclear visual patterns (The third concept in Fig. 9 and the first concept in Fig. 11). This may stem from feature absorption issues, where the learned concepts fail to decompose into into their subconcepts. On the other hand, the learned concepts occasionally share the similar masking concent of the input image. We attribute this to the fine-grained feature of the learned concepts, where concepts capture the similar visual patterns while focusing on distinct dimensions. For example, the concepts of baby, cute girl, and playing girl in Fig. 10.

Sparsity and Interpretability Analysis To further evaluate the quality of the learned concepts, we report sparsity metric values in Table. 5 and interpretability metrics in Table. 6. Following prior work [30, 38], we compute metrics including L_0 (higher is better), FVU (fraction of variance unexplained, lower is better), CS (cosine similarity, higher is better), CKNNA (higher is better), and DO (dead concepts, lower is better). For interpretability, we follow the evaluation protocol in [47] and report both the mean and maximum monosemantic scores of the learned concepts. Our Sparsemax SAE achieves the best performance under most metrics, demonstrating that the dynamic attention mechanism not only produces sparse representations but also yields more interpretable and semantically coherent concepts.

8. Results of Zero-shot Image Classification

We report the detailed zero-shot image classification results in Table. 7- 17.

Table 7. Zero-shot classification results on the Caltech101 dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	32.403	7.816	17.591	21.033	28.671	31.672
JumpReLU	32.403	3.578	12.840	18.594	27.845	28.974
Gated	32.403	5.894	14.054	22.847	28.847	30.158
TOPK	32.403	5.94	16.632	23.096	30.654	29.645
BatchTopK	32.403	11.567	23.144	27.314	30.582	31.284
Saprsemax SAE (Ours)	32.403	20.121	26.195	29.322	31.325	31.875

Table 8. Zero-shot classification results on the DTD dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	44.840	13.457	27.074	34.326	44.468	41.596
JumpReLU	44.840	4.825	22.884	30.520	44.495	40.367
Gated	44.840	5.989	15.048	21.495	47.187	37.365
TopK	44.840	6.117	15.372	22.624	46.365	35.851
BatchTopK	44.840	17.730	30.372	36.383	49.007	40.957
Saprsemax SAE (Ours)	44.840	34.804	41.791	46.986	50.16	44.025

Table 9. Zero-shot classification results on the EuroSAT dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	38.605	10.200	24.826	32.939	39.832	33.315
JumpReLU	38.605	11.680	20.475	33.648	40.987	32.158
Gated	38.605	15.815	32.487	39.846	44.682	34.577
TopK	38.605	16.46	11.468	25.344	47.041	32.741
BatchTopK	38.605	10.0	34.990	34.957	45.317	31.338
Saprsemax SAE (Ours)	38.605	19.301	36.463	34.056	47.839	32.421

Table 10. Zero-shot classification results on the FGVC dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	23.137	2.371	2.617	3.314	7.431	19.803
JumpReLU	23.137	2.647	2.689	4.047	7.846	17.358
Gated	23.137	1.574	4.855	5.128	7.748	11.547
TopK	23.137	1.02	1.683	3.156	9.18	7.59
BatchTopK	23.136	0.990	5.191	7.390	10.805	14.362
Saprsemax SAE (Ours)	23.137	4.8346	7.069	8.14871	10.894	19.5187

Table 11. Zero-shot classification results on the Food101 dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	83.195	17.991	34.420	39.286	56.2719	78.285
JumpReLU	83.195	19.475	35.187	39.954	67.257	75.481
Gated	83.195	7.458	25.486	36.487	60.875	77.584
TopK	83.195	8.644	21.735	30.207	57.888	54.956
BatchTopK	83.194	6.459	37.864	48.996	68.365	75.440
Saprsemax SAE (Ours)	83.195	26.1134	51.705	59.23	69.49	79.954

Table 12. Zero-shot classification results on the ImageNet dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	67.294	3.116	15.829	22.173	34.872	63.670
JumpReLU	67.294	3.548	17.558	34.975	40.876	62.957
Gated	67.294	5.568	12.875	35.495	50.159	62.547
TopK	67.294	3.421	8.2501	32.646	56.135	47.956
BatchTopK	67.294	1.58	15.562	27.782	58.472	61.072
Saprsemax SAE (Ours)	67.294	10.929	33.469	42.127	59.947	65.087

Table 13. Zero-shot classification results on the ImageNet-Sketch dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	47.851	2.798	10.283	13.451	23.479	31.009
JumpReLU	47.851	3.847	12.657	15.984	26.581	30.257
Gated	47.851	1.782	9.257	13.576	38.712	32.157
TopK	47.851	0.957	8.1269	14.712	36.942	30.991
BatchTopK	47.851	0.3509	7.349	16.103	39.128	34.178
Saprsemax SAE (Ours)	47.851	12.460	25.321	29.678	39.328	42.314

Table 14. Zero-shot classification results on the OxfordPets dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	83.477	26.964	49.611	61.350	71.127	79.061
JumpReLU	83.477	26.981	50.258	63.204	73.980	78.041
Gated	83.477	18.547	53.492	64.581	74.012	76.251
TopK	83.477	6.973	44.886	52.492	79.621	65.274
BatchTopK	83.477	17.081	56.985	67.100	76.918	77.828
Saprsemax SAE (Ours)	83.477	25.186	58.135	67.925	77.351	80.986

Table 15. Zero-shot classification results on the StandfordCars dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	31.976	0.828	3.307	5.383	9.778	24.602
JumpReLU	31.976	1.568	3.964	5.421	9.157	20.367
Gated	31.976	0.540	3.541	6.034	10.652	19.068
TopK	31.976	0.610	2.735	3.263	6.897	8.7586
BatchTopK	31.976	1.248	3.811	6.786	10.1140	18.987
Saprsemax SAE (Ours)	31.976	3.421	8.25	9.337	13.155	27.245

Table 16. Zero-shot classification results on the SUN397 dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	66.257	4.543	19.045	25.902	44.859	60.704
JumpReLU	66.257	5.962	22.084	25.035	48.258	58.367
Gated	66.257	6.058	27.643	31.247	52.796	58.947
TopK	66.257	5.2819	20.468	29.9	54.277	45.027
BatchTopK	66.257	5.117	29.824	40.5	57.203	60.431
Saprsemax SAE (Ours)	66.257	15.946	39.904	48.4794	56.973	63.788

Table 17. Zero-shot classification results on the UCF101 dataset. $K = 32$ in TopK and BatchTopK.

	no_sae	on_1	on_5	on_10	on_50	on_49152
ReLU	61.322	7.607	19.600	26.056	42.945	58.626
TopK	61.322	3.039	18.003	26.157	41.396	34.647
BatchTopK	61.322	3.591	26.474	33.346	45.194	56.986
JumpReLU	61.322	6.257	20.587	26.971	43.579	51.278
Gated	61.322	4.068	24.947	30.189	43.840	52.497
Saprsemax SAE (Ours)	61.322	21.712	33.575	37.779	47.214	59.064

Figure 6. This picture successfully activates three different concepts: fruit, background and apple.

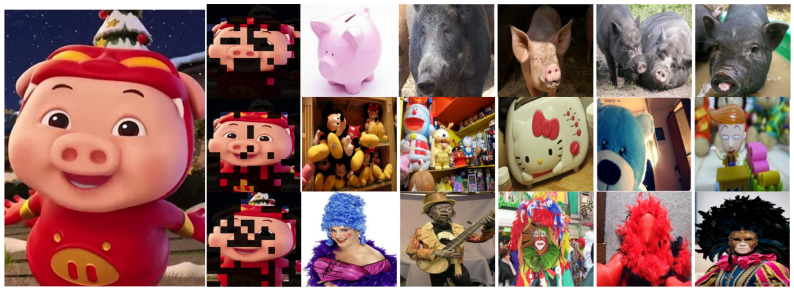

Figure 7. This picture is from the cartoon task image in the animation. It can be seen that it has the characteristics of pigs. Therefore, the features of pig are activated through sae, and other relevant features are activated according to its shape.

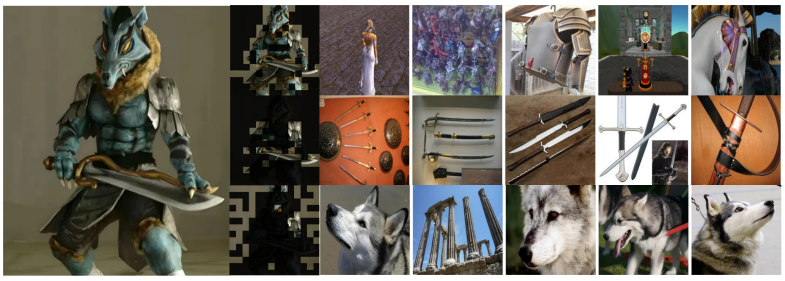

Figure 8. This picture is from the monster in the film and television image. Its wolf's head features activate the wolf's features through sae. In addition, his armor and weapons also activate the corresponding features.

Figure 9. Messi's picture in this example activated the football. In addition, in this experiment, sae activated the "crowd" feature through the background of the picture,but unfortunately there is a phenomenon of feature absorption in the third feature, where images of short sleeves and similar color structures are considered to be the same concept

Figure 10. In this sample, pictures activate children and children's eating characteristics through sae.However,the concept of Apple was not recognized and the concept on the right is too similar.

Figure 11. This picture is from a football match. The picture activates the football feature, field , and the crowd .But in the first concept, there were pictures unrelated to football