Finding better Sparse Neural Networks is hard

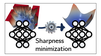
- Sparse neural networks are neural networks with a lot of zero-valued weight parameters.
- They can use up less memory and computation than dense ones, making them increasingly sought after as modern models grow to extreme sizes.

$$\min_{\|x\|_0 \le d} f(x)$$

- Various neural network pruning techniques were developed to find sparse neural networks x with good performance or minimal loss f.
- Still, it **remains a constant challenge to push for higher sparsity with minimal performance degradation**.

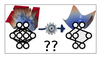
Degraded performance by loss sharpness

- Sharpness of loss has been suggested as a major cause for diminishing the trainability of sparse networks in high sparsity (Lee et al., 2021).
- More generally, its strong correlations with degradation in generalization performance has been well studied (Keskar et al., 2017).



- Various techniques were proposed to explicitly minimize sharpness during training (Foret et al., 2021).
- This has been shown to be effective in improving the generalization performance and robustness of neural networks.

Enforcing flatness to improve sparsification



Despite studies connecting degraded performance by sparsification to loss sharpness, very little work has attempted to develop a general framework to explicitly reduce sharpness while sparsifying neural networks.

SAFE! To fill this gap, we propose SAFE, a **general constrained-optimization-based framework to simultaneously enforce flatness and sparsification while optimizing the neural network**.

SAFE: Sparsification via ADMM with Flatness Enforcement

Sharpness-aware sparsity-constrained optimization problem

$$\min_{\|x\|_0 \le d} \max_{\|\epsilon\|_2 \le \rho} f(x+\epsilon)$$

We first formulate this as an optimization problem, where the goal is to find a sparse solution x^* with at most d non-zero elements that minimizes the objective function in the whole ϵ-neighborhood, *i.e.*, seek flat minima.

Augmented Lagrangian based approach

To solve this, we form the augmented Lagrangian dual problem of the following:

$$\max_{u} \min_{x,z} \left[\mathcal{L}(x,z,u) := \max_{\|\epsilon\|_2 \le \rho} f(x+\epsilon) + I_{\|\cdot\|_0 \le d}(z) - \frac{\lambda}{2}\|u\|_2^2 + \frac{\lambda}{2}\|x - z + u\|_2^2 \right],$$

where we separate the sparsity-constraint satisfaction using variable z so that it can be handled more easily.

Alternating Direction Method of Multipliers

$$x_{k+1} = \arg\min_{x} \max_{\|\epsilon\|_2 \le \rho} f(x+\epsilon) + \frac{\lambda}{2}\|x - z_k + u_k\|_2^2 \quad \text{(x-min)}$$

$$z_{k+1} = \arg\min_{z} I_{\|\cdot\|_0 \le d}(z) + \frac{\lambda}{2}\|x - z + u\|_2^2 \quad \text{(z-min)}$$

$$u_{k+1} = u_k + x_{k+1} - z_{k+1}$$

We apply dual ascent and minimize x and z in an alternating fashion, which gives us this ADMM iterate.

x-minimization: iterative minimization while enforcing flatness

We solve this iteratively using *Sharpness-aware minimization (SAM)*, where we approximately solve for ϵ through first-order Taylor approximation:

$$\epsilon^*(x) \approx \operatorname*{argmax}_{\|\epsilon\|_2 \le \rho} f(x) + \epsilon^\top \nabla f(x) = \rho \frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

Applying this back to the objective and applying gradient descent gives us the following iteration for x-minimization

$$\text{(x-min)} \quad x_k^{(t+1)} = x_k^{(t)} - \eta^{(t)} \left[\nabla f\left(x_k^{(t)} + \rho \frac{\nabla f(x_k^{(t)})}{\|\nabla f(x_k^{(t)})\|_2} \right) + \lambda (x_k^{(t)} - z_k + u_k) \right]$$

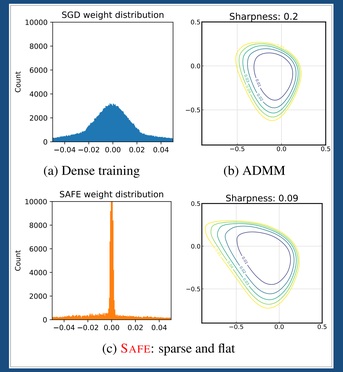
ICML 2025 Spotlight Paper

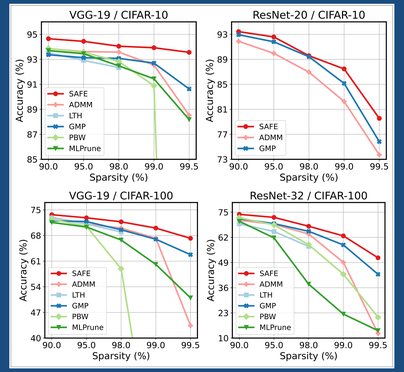
SAFE: Finding Sparse and Flat Minima to Improve Pruning

Dongyeop Lee, Kwanhee Lee, Jinseok Chung, and Namhoon Lee
Pohang University of Science and Technology (POSTECH)

: We propose **SAFE** and **SAFE+**, constrained optimization algorithms to enforce flatness simultaneously with sparsity

: strong/robust performance across various sparsity/settings





Sparsity	Method	LLaMa-2 7B Wikitext/C4	LLaMa-2 13B Wikitext/C4	LLaMa-3 8B Wikitext/C4
0%	Dense	5.47 / 7.26	4.88 / 6.72	6.23 / 9.53
50%	Magnitude	16.03 / 21.33	6.82 / 9.37	134.20 / 273.3
	SparseGPT	6.99±0.03 / 9.20±0.03	6.06±0.03 / 8.20±0.01	9.36±0.11 / 13.98±0.02
	Wanda	6.92±0.01 / 9.23±0.00	5.98±0.01 / 8.28±0.01	9.71±0.03 / 14.88±0.04
	ALPS	6.87±0.01 / 8.98±0.00	5.96±0.02 / 8.09±0.04	9.05±0.12 / 13.40±0.06
	SAFE	6.78±0.01 / 8.93±0.00	5.76±0.01 / 7.85±0.02	9.59±0.06 / 14.60±0.04
	SAFE+	**6.56**±0.01 / **8.71**±0.00	**5.67**±0.01 / **7.74**±0.01	**8.62**±0.06 / **13.26**±0.06
60%	Magnitude	1864 / 2043	11.81 / 14.62	5335 / 7438
	SparseGPT	10.19±0.08 / 12.86±0.05	8.31±0.09 / 10.85±0.09	15.46±0.40 / 21.25±0.18
	Wanda	10.75±0.07 / 13.87±0.01	8.43±0.07 / 11.55±0.01	22.06±0.19 / 32.28±0.37
	ALPS	9.55±0.00 / 11.24±0.03	7.54±0.03 / 9.87±0.05	14.03±0.35 / 18.72±0.15
	SAFE	9.20±0.04 / 11.51±0.04	7.18±0.03 / 9.59±0.03	15.90±0.25 / 22.26±0.16
	SAFE+	**8.30**±0.06 / **10.59**±0.00	**6.78**±0.04 / **9.02**±0.15	**12.18**±0.22 / **17.30**±0.02
4:8	Magnitude	15.91 / 31.61	7.32 / 9.96	212.5 / 336.3
	SparseGPT	8.42±0.05 / 10.73±0.03	7.02±0.06 / 9.33±0.04	12.16±0.20 / 17.36±0.06
	Wanda	8.64±0.03 / 11.35±0.01	7.01±0.02 / 9.70±0.03	13.84±0.04 / 21.14±0.06
	ALPS	8.11±0.09 / 10.21±0.04	6.81±0.07 / 9.33±0.04	11.38±0.17 / 16.10±0.10
	SAFE	8.21±0.01 / 10.61±0.04	6.60±0.02 / 8.95±0.02	12.15±0.14 / 17.90±0.15
	SAFE+	**7.59**±0.03 / **9.88**±0.01	**6.37**±0.03 / **8.61**±0.01	**10.51**±0.13 / **15.67**±0.02
2:4	Magnitude	37.77 / 74.70	8.88 / 11.72	792.8 / 2245
	SparseGPT	11.00±0.20 / 13.54±0.03	8.78±0.09 / 11.26±0.11	15.87±0.32 / 22.45±0.12
	Wanda	12.17±0.02 / 15.60±0.11	9.01±0.04 / 12.40±0.01	23.03±0.38 / 34.91±0.31
	ALPS	9.99±0.19 / 12.04±0.04	8.16±0.17 / 10.35±0.18	14.53±0.33 / 19.74±0.18
	SAFE	10.53±0.13 / 13.20±0.07	7.64±0.05 / 10.10±0.01	17.49±0.27 / 24.45±0.13
	SAFE+	**8.96**±0.07 / **11.34**±0.03	**7.20**±0.04 / **9.52**±0.01	**13.39**±0.23 / **19.03**±0.01

(Left) SAFE successfully finds sparse models at flat minima

(Center) SAFE achieves strong image classification performance

(Right) SAFE and SAFE+ achieves strong LLM pruning performance

Contact: {dongyeop.lee_, kwanhee.lee, jinseokchung, namhoon.lee}@postech.ac.kr

z-minimization: Euclidean projection onto sparsity constraint

$$\text{(z-min)} \quad z_{k+1} = \arg\min_{z} I_{\|\cdot\|_0 \le d}(z) + \frac{\lambda}{2}\|x_{k+1} - z + u_k\|_2^2 = \text{proj}_{\|\cdot\|_0 \le d}(x_{k+1} + u_k).$$

- z-minimization corresponds to projecting $x_{k+1} + u_k$ onto the sparsity constraint in terms of Euclidean distance.
- This leads to the classic hard thresholding operator, where we zero out except d elements with the largest magnitude.

SAFE+: Improving projection through generalized distance

Generalized quadratic distance to improve projection

$$z_{k+1} = \text{proj}^{\mathbf{P}}_{\|\cdot\|_0 \le d}(x_{k+1} + u_k) := \arg\min_{\|z\|_0 \le d} \frac{1}{2}\|z - (x_{k+1} + u_k)\|_{\mathbf{P}}^2 = \arg\min_{\|z\|_0 \le d} \frac{1}{2}(z - (x_{k+1} + u_k))^\top \mathbf{P}(z - (x_{k+1} + u_k)).$$

- However, this magnitude-based projection often yields subpar performance in practice.
- To improve this, we introduce a generalized distance $\frac{1}{2}\|\cdot\|_{\mathbf{P}}^2$ with diagonal positive definite matrix $\mathbf{P}$.

Distance vs. saliencies

Criteria	$\mathbf{P}$
Magnitude	$\mathbf{I}$
OBD	$\text{diag}(H)$
SNIP	$\text{diag}(\nabla f(\nabla f)^\top)$
Wanda	$\text{diag}(\mathbf{A}^\top \mathbf{A})$

- This generalized projection framework allows us to employ various saliency scores within the projection step
- Here we use this primarily for LLM pruning, though it is generally applicable to other domains

SAFE demonstrates strong empirical performance

←1. Successful sparsity and flatness enforcement

- Weights are concentrated near zero → **sparse**
- Wide minima + smaller largest Hessian eigenvalue → **flat**

←2. Strong image classification performance

- SAFE retains strong performance at high sparsities compared to baselines.
- Epoch: 200 (300 for ResNet-20) / Recompute batch statistics after final hard projection step

←3. Strong LLM pruning performance

- Compared to methods specifically designed for LLMs, SAFE performs competitively and SAFE+ outperforms all baselines across all models and sparsities.
- Blockwise SAFE: Following common practice, we sequentially apply SAFE to each transformer block to minimize the reconstruction error as $\min_{\|\mathbf{x}\|_0 \le d} \|\text{Block}(\text{inputs}; \mathbf{x}) - \text{Block}(\text{inputs}; x_{\text{original}})\|$.
- SAFE+: Projection based on diagonal Hessian of the Layerwise reconstruction error is employed, which corresponds to using the Wanda score.
- Epoch: 30 / Dataset: 128 random samples from the first shard of the C4 dataset / seq len: 2048
- (Sidenote) As model scales, the computation of SAFE scales quadratically with the model width. In contrast, methods such as SparseGPT and ALPS scales cubically (see Appendix E).

4. Strong robustness to various data noise

SAFE is robust to noisy label training (left), as well as common image corruptions and adversarial attacks (right) on ResNet-20/CIFAR-10.

Sparsity	Method	Noise ratio 25%	50%	75%
70%	ADMM	77.00±0.91	59.18±0.55	32.62±0.89
	SAFE	**90.58**±0.30	**86.51**±0.16	**67.01**±0.54
80%	ADMM	76.18±0.56	62.67±0.58	32.86±1.12
	SAFE	**91.25**±0.12	**86.55**±0.07	**66.49**±0.56
90%	ADMM	79.40±0.12	66.64±0.13	36.84±0.91
	SAFE	**90.68**±0.21	**86.49**±0.06	**64.72**±0.61
95%	ADMM	77.71±0.52	67.10±1.37	39.68±1.44
	SAFE	**89.86**±0.11	**85.18**±0.15	**64.25**±0.36

Sparsity	Method	Common corruption (avg.) intensity=3	intensity=5	Adversarial l_∞-PGD	l_2-PGD
90%	ADMM	70.06±0.03	52.01±0.38	49.81±1.02	49.71±1.06
	SAFE	**73.98**±0.09	**55.11**±0.27	**56.43**±1.03	**56.36**±1.11
95%	ADMM	68.87±0.25	50.56±0.07	49.84±1.78	49.68±1.79
	SAFE	**72.92**±0.41	**54.86**±0.51	**51.40**±0.89	**51.36**±0.94
98%	ADMM	65.46±0.24	48.65±0.04	43.33±1.59	**43.42**±1.60
	SAFE	**68.20**±0.47	**49.96**±0.83	**43.34**±0.90	43.41±1.03
99%	ADMM	59.21±0.47	43.81±0.44	30.29±0.64	30.32±0.58
	SAFE	**66.02**±0.56	**49.34**±1.03	**43.70**±1.28	**32.70**±1.28
99.5%	ADMM	55.72±0.44	41.55±0.78	23.25±1.92	23.25±1.85
	SAFE	**56.58**±0.36	**42.27**±0.63	**29.48**±0.68	**29.45**±0.74

SAFE converges to stationary point within sparsity constraint

(δ-stationary point) We say a point $\bar{x}$ is a δ-stationary point of the sparsity-constrained optimization problem if $\bar{x} \in \arg\min_{a \in \mathcal{A}} \left\| a - \left(\bar{x} - \delta^{-1}\nabla f(\bar{x})\right) \right\|$,

(Convergence of SAFE) Suppose that f is smooth and weakly convex. Assume further that δ is chosen large enough so that $\delta^{-1}\beta^2 - (\delta - \mu)/2 < 0$. Let $(\bar{x}, \bar{z}, \bar{u})$ be a limit point of SAFE algorithm. Then $\bar{x}$ is a δ-stationary point of the sparsity-constrained optimization problem.