Pruning via Sparsity-indexed ODE: A Continuous Sparsity Viewpoint

Zhanfeng Mo and Haosen Shi
and Sinno Jialin Pan

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

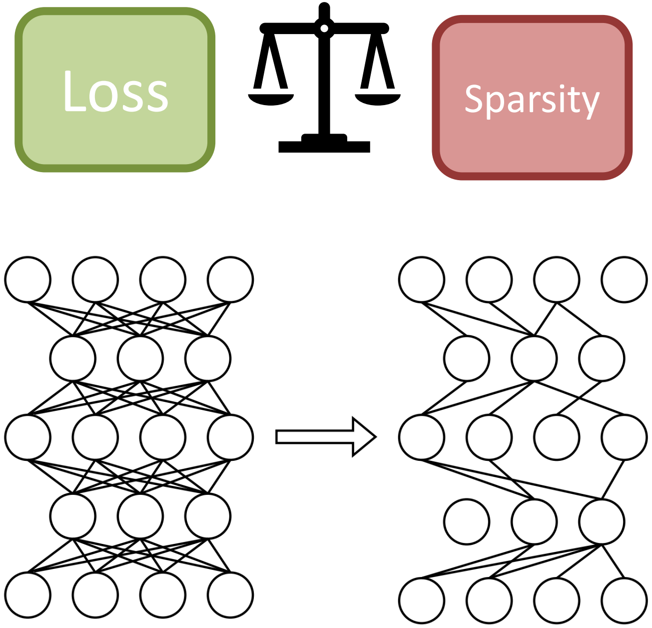
ICML 2023

Neural Pruning

- **Goals**: compress large neural network to a target sparsity level by pruning model weights, with a minimal performance drop.

- **How**: minimize model loss, with a sparsity constraint.

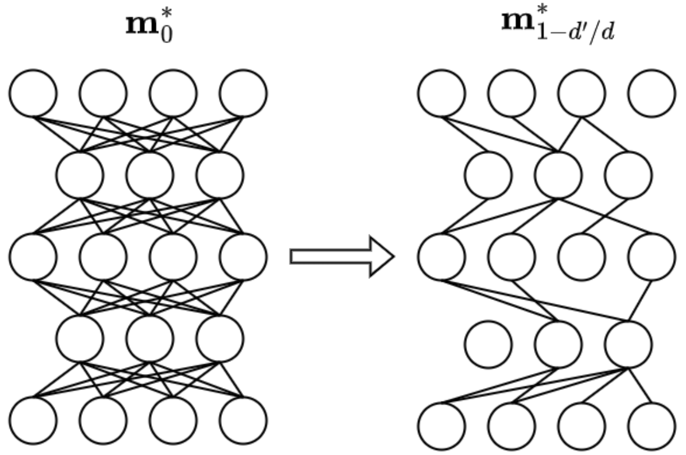
$$\min_{\mathbf{m}\in\{0,1\}^d} \boxed{\mathcal{E}(\mathbf{m}\odot\boldsymbol{\theta}^*)}, \ \text{s.t.} \boxed{\|\mathbf{m}\|_0 = d'}$$



Neural Pruning

- $\boldsymbol{\theta}^* \in \mathbf{R}^d$: reference (pretrained) model.
- $\boldsymbol{m} \in \{0,1\}^d$: 0-1 mask.
- $d' < d$: target parameter budget.
- **Sparsity**: $1 - |\boldsymbol{m}|_0/d$.
- $\mathcal{E}(\cdot) : \mathbf{R}^d \mapsto \mathbf{R}_+$, **model loss**.
- $\boldsymbol{m}_t$: the **optimal mask** of sparsity t.

$$\min_{\mathbf{m}\in\{0,1\}^d} \mathcal{E}(\mathbf{m} \odot \boldsymbol{\theta}^*), \text{ s.t. } \|\mathbf{m}\|_0 = d'$$



Limitations of Existing Methods

$$\min_{\mathbf{m}\in\{0,1\}^d} \mathcal{E}(\mathbf{m} \odot \boldsymbol{\theta}^*), \text{ s.t. } \|\mathbf{m}\|_0 = d'$$

Method	Pros	Cons
Score-based	cheap & fast	**biased** due to locality
Regularization-based	differentiable	numerically **instable**
Sparse-training	better performance	**biased** & slow

- Finding $\boldsymbol{m}^*_{1-d'/d}$ **directly** is hard due to the sparsity and irregularity!

Ease Neural Pruning with a Hint

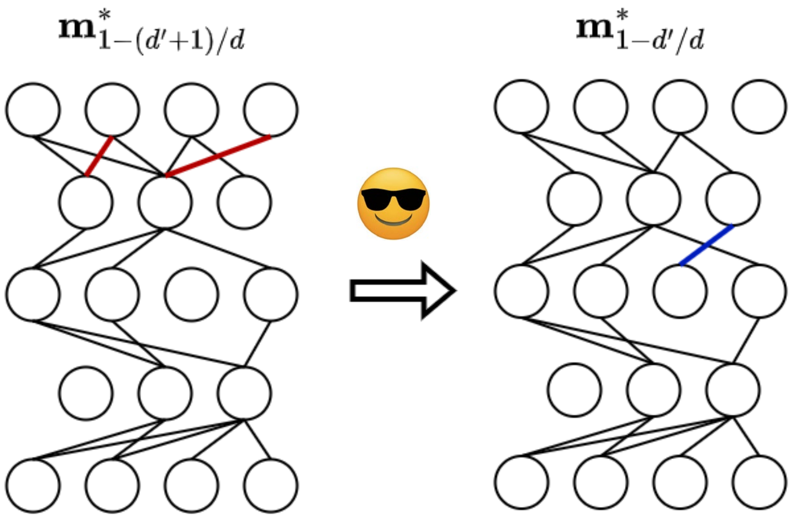
- Finding $\boldsymbol{m}^*_{1-d'/d}$ **directly** is hard due to the sparsity and irregularity!

- Q: what if we know a **hint**?

- A: travel from a $1/d$ denser optimal mask may help!

- The denser optimal mask enable us to **preserve optimality** with **a few alteration.**

$$\mathbf{m}_{t+\Delta t} - \mathbf{m}_t = F(\mathbf{m}_t)\Delta t$$



One-step Evolution of Optimal Mask

- Q: what is the dynamic of $t \mapsto \boldsymbol{m}_t$?

- Q: how $\boldsymbol{m}_t$ evolves for an **infinitesimal increase** in sparsity?

1. Sparsity t is not continuous.

2. What is the one-step evolution $F(\cdot)$?

$$\mathbf{m}_{t+\Delta t} - \mathbf{m}_t \stackrel{?}{=} F(\mathbf{m}_t)\Delta t$$

$$\Big\downarrow \ \Delta t \to 0$$

$$\mathrm{d}\mathbf{m}_t = F(\mathbf{m}_t)\mathrm{d}t$$

Polarized Soft Neural Pruning

- Q: how $\boldsymbol{m}_t$ evolves for an **infinitesimal increase** in sparsity?

1. Sparsity t is not continuous.

2. What is the one-step evolution $F(\cdot)$?

$$\min_{\mathbf{m}\in\{0,1\}^d} \mathcal{E}(\mathbf{m} \odot \boldsymbol{\theta}^*)$$
$$\text{s.t. } \|\mathbf{m}\|_0 = d'$$

$$\min_{\mathbf{m}\in\mathbb{R}^d} \mathcal{E}_\varepsilon(\mathbf{m}) \triangleq \mathcal{E}(\mathcal{P}_\varepsilon(\mathbf{m}) \odot \boldsymbol{\theta}^*)$$
$$\text{s.t. } G(\mathbf{m}) = t,$$

Solution:

- Generalize mask $\boldsymbol{m}$ and sparsity t to be **continuous-valued** by a soft sparsity measure $G(\cdot)$: $\mathbf{R}^d \mapsto [0,1]$.
- Polarize the soft mask $\boldsymbol{m}$ to be **nearly-binary** via a polarizer $P_\varepsilon(\cdot)$: $\mathbf{R}^d \mapsto ([0,1]\backslash(\varepsilon, 1-\varepsilon))^d$.

One-step Evolution of Optimal Mask

- Q: how $\boldsymbol{m}_t$ evolves for an **infinitesimal increase** in sparsity?

1. Sparsity t is not continuous.

2. <u>What is the one-step evolution $F(\cdot)$?</u>

$$\min_{\mathbf{m}\in\mathbb{R}^d} \mathcal{E}_\varepsilon(\mathbf{m}) \triangleq \mathcal{E}(\mathcal{P}_\varepsilon(\mathbf{m}) \odot \boldsymbol{\theta}^*)$$
$$\text{s.t. } G(\mathbf{m}) = t,$$

Localization trick

$$\min_{\boldsymbol{\delta}\in\mathbb{R}^d} \nabla\mathcal{E}_\varepsilon(\mathbf{m}_{t+\Delta t})^\top \boldsymbol{\delta},$$
$$\text{s.t. } \nabla G(\mathbf{m}_{t+\Delta t})^\top \boldsymbol{\delta} = -\Delta t$$

$$\boxed{\|\boldsymbol{\delta}\| \leqslant r_t \Delta t}$$

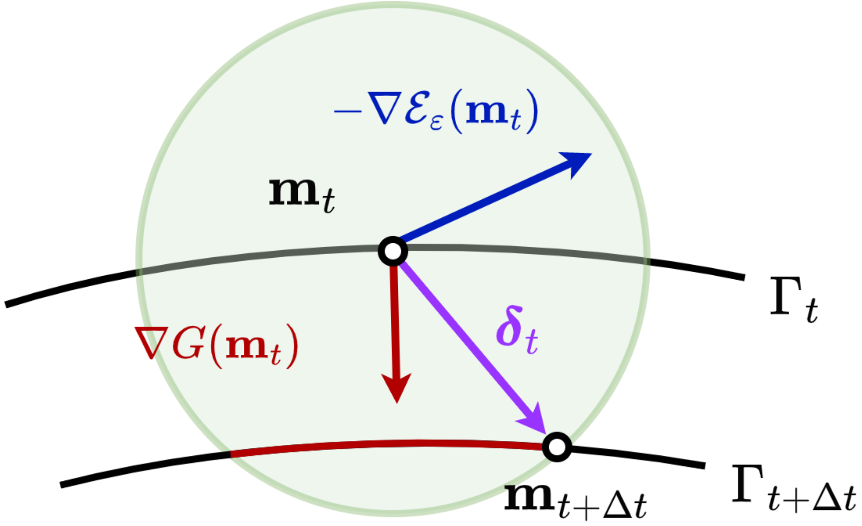
Localization radius

<u>Solution:</u>

- **Localize** around the hint $\boldsymbol{m}_t$ and solve $\boldsymbol{\delta}_t \coloneqq \boldsymbol{m}_{t+\Delta t} - \boldsymbol{m}_t$.
- Solve $\boldsymbol{\delta}_t$ from the localized problem with an **explicit solution**.

Explicit One-step Evolution of Optimal Mask

$$\mathbf{m}_{t+\Delta t} - \mathbf{m}_t = F(\mathbf{m}_t)\Delta t$$



$$\min_{\boldsymbol{\delta}\in\mathbb{R}^d} \nabla\mathcal{E}_\varepsilon(\mathbf{m}_{t+\Delta t})^\top \boldsymbol{\delta},$$

$$\text{s.t. } \nabla G(\mathbf{m}_{t+\Delta t})^\top \boldsymbol{\delta} = -\Delta t$$

$$\boxed{\|\boldsymbol{\delta}\| \leqslant r_t \Delta t}$$

$$\boldsymbol{\delta}_t = F(\mathbf{m}_{t+\Delta t})\Delta t$$

$$\boxed{\begin{aligned} F(\mathbf{m}) &\triangleq \begin{cases} \mathbf{g}/\|\mathbf{g}\|^2, \text{ if } \|\mathbf{g}\|\|\mathbf{e}\| = |\mathbf{g}^\top \mathbf{e}|^2, \\ x\mathbf{e} + y\mathbf{g}, \text{ else}, \end{cases} \\ x &\triangleq \sqrt{((r^2-1)/((\|\mathbf{g}\|\|\mathbf{e}\|)^2 - (\mathbf{g}^\top\mathbf{e})^2))}, \\ y &\triangleq (1 - \mathbf{g}^\top \mathbf{e} x)/\|\mathbf{g}\|^2, \end{aligned}}$$

Pruning via Sparsity-indexed ODE (SpODE)

- **In intuition**: SpODE guides us towards the **sparsity increase direction** with a **minimal performance drop**.

- **In theory**: solving $\boldsymbol{m}_{1-d'/d} \Leftrightarrow$ evaluating the **optimal mask dynamic** $t \mapsto \boldsymbol{m}_t$ at $t = 1 - d'/d$.

- Now we can evaluate $\boldsymbol{m}_{1-d'/d}$ via SpODE discretization!

$$\mathbf{m}_{t+\Delta t} - \mathbf{m}_t = F(\mathbf{m}_t)\Delta t$$

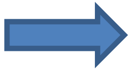
$$\Big\Downarrow \; \Delta t \to 0$$

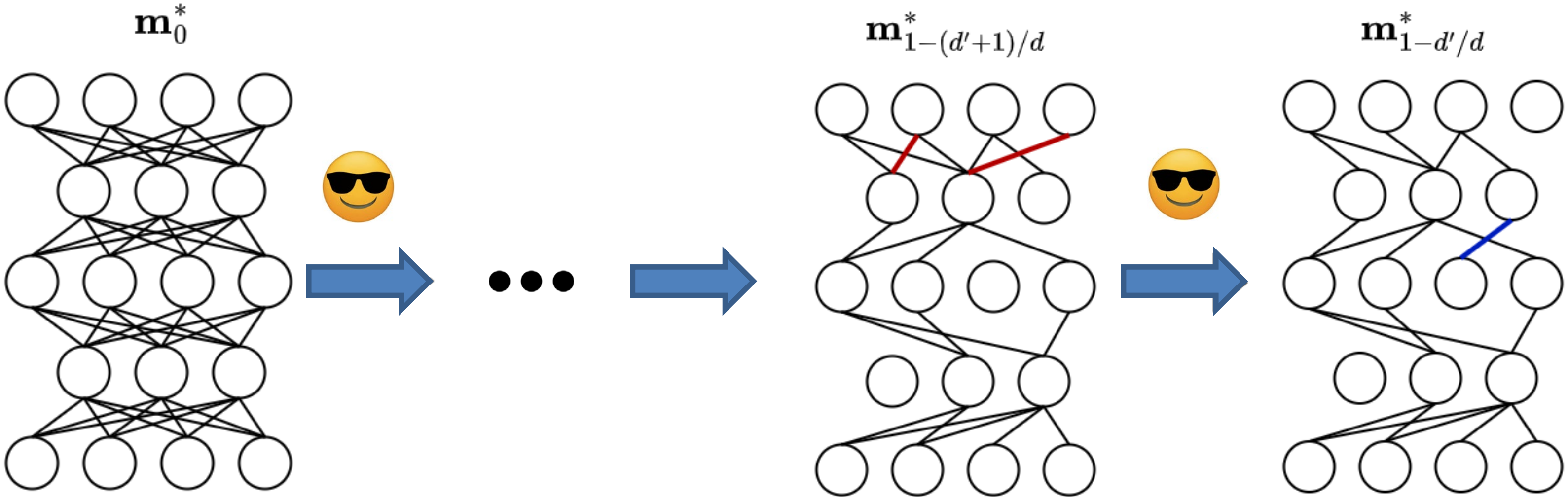
$$\boxed{\begin{aligned} &\mathrm{d}\mathbf{m}_t = F(\mathbf{m}_t)\mathrm{d}t,\ t \in [0, 1 - d'/d] \\ &\mathbf{m}_0 \triangleq \mathbf{1} \end{aligned}}$$

$\mathrm{d}t$-sparsity increase

Minimal loss increase

How Sparsity-indexed ODE works?

 : One-step discretization of SpODE.



One-shot performance on CIFAR-10 / 100

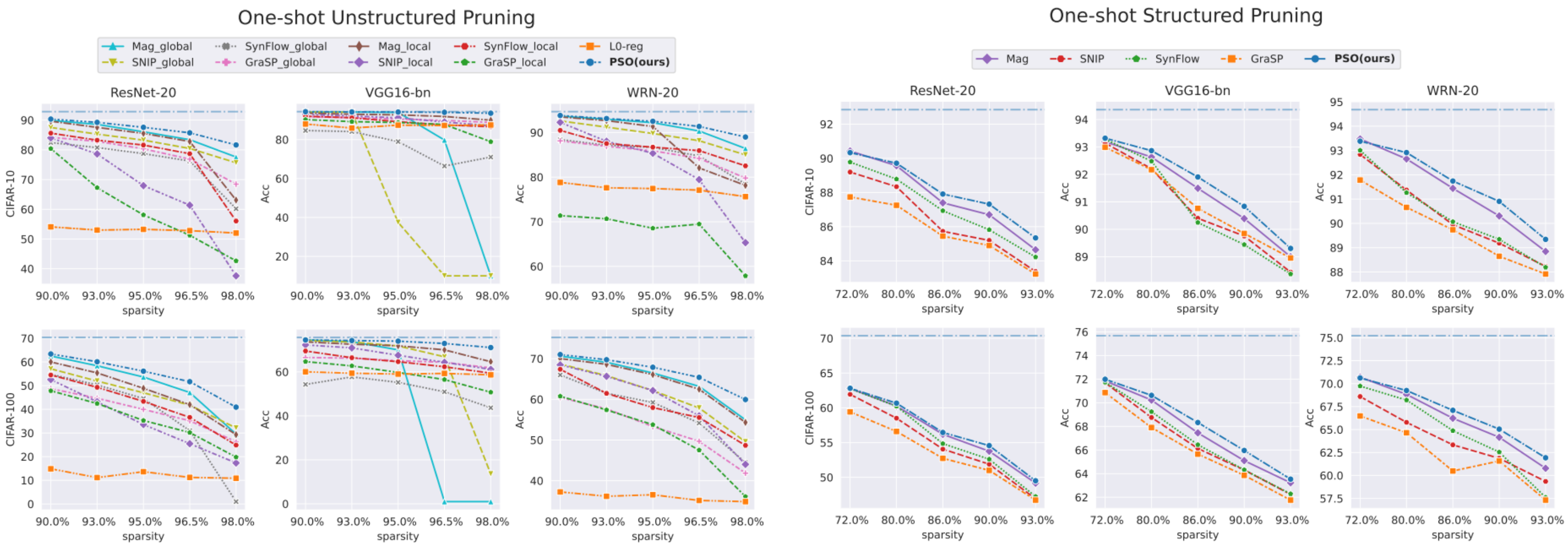


Figure 3. Results of one-shot pruning on CIFAR-10/100 dataset. The x-axis is the sparsity and the y-axis is the top-1 accuracy of the tuned sparse model. The horizontal dash line represents the performance of the unpruned model $\boldsymbol{\theta}^*$.

One-shot performance on Tiny-ImageNet

	Sparsity	Mag	SNIP	SynFlow	GraSP	**PSO (ours)**
ResNet-50 (67.06)	90% / 72%	60.51 / 63.40	57.62 / 64.26	56.09 / 63.78	52.78 / 59.14	**63.47 / 63.60**
	93% / 80%	57.18 / 62.84	53.91 / 63.03	54.79 / 63.07	49.16 / 57.21	**59.37 / 63.37**
	95% / 86%	56.64 / 61.36	54.33 / 61.77	53.85 / 61.37	43.95 / 57.00	**61.20 / 62.73**
	96.5% / 90%	53.42 / 58.98	53.87 / 56.16	50.33 / 58.45	32.42 / 54.76	**57.61 / 61.63**
	98% / 93%	51.90 / 56.81	52.94 / 58.07	41.00 / 57.23	10.56 / 53.99	**53.82 / 59.62**
VGG19-bn (63.47)	90% / 72%	62.32 / 59.58	61.66 / 59.27	0.50 / 0.50	43.73 / 56.84	**62.67 / 60.37**
	93% / 80%	61.65 / 58.33	60.22 / 58.01	0.50 / 0.50	43.52 / 55.51	**62.05 / 59.08**
	95% / 86%	**61.74 / 57.34**	56.08 / 55.75	0.50 / 0.50	42.86 / 50.90	61.54 / 57.08
	96.5% / 90%	60.46 / 54.99	46.54 / 52.38	0.50 / 0.50	42.37 / 46.16	**60.60 / 55.22**
	98% / 93%	53.26 / 49.82	23.33 / 46.75	0.50 / 0.50	39.42 / 41.06	**57.63 / 50.65**
WRN-34 (64.74)	90% / 72%	61.59 / 60.23	61.43 / 60.32	57.87 / 60.44	53.37 / 59.85	**62.36 / 61.11**
	93% / 80%	61.11 / 59.35	60.16 / 59.28	57.57 / 59.61	52.75 / 57.13	**61.17 / 59.91**
	95% / 86%	59.97 / 58.30	51.31 / 58.14	50.11 / 58.20	49.67 / 55.30	**60.14 / 58.38**
	96.5% / 90%	**58.98** / 57.25	56.45 / 55.65	54.32 / 56.23	49.67 / 53.11	58.75 / **57.28**
	98% / 93%	56.39 / 55.40	54.60 / 54.12	50.42 / 52.03	47.17 / 51.62	**57.54 / 57.03**

Table 2. Comparison results of one-shot Unstructured / Structured Pruning on Tiny-ImageNet. The numbers in the parentheses are the performance of the unpruned model.

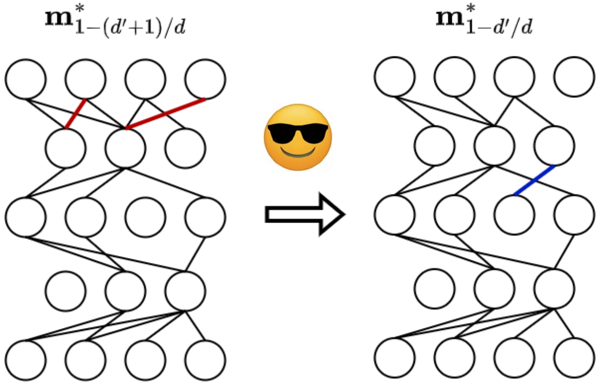
Iterative Performance on ImageNet

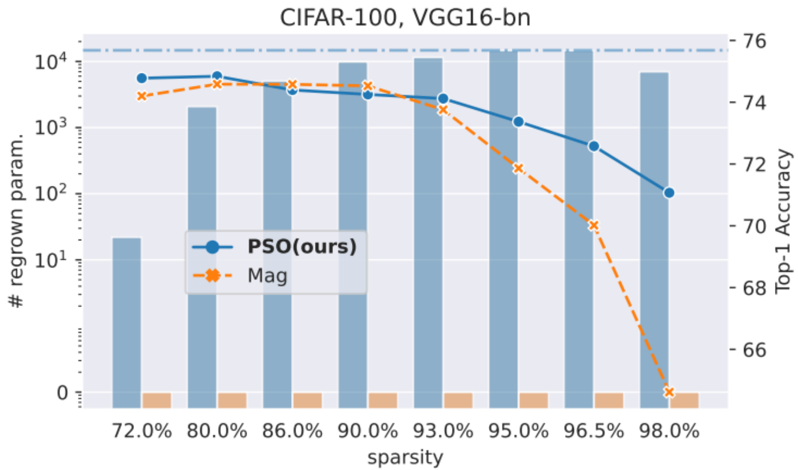
ResNet-50	Unpruned Acc.	Sparisty	Final Acc.
STR (Kusupati et al., 2020)	77.01%	87.70%	74.73%
WoodFisher (Singh & Alistarh, 2020)	77.01%	90.00%	75.21%
Powerprop (Schwarz et al., 2021)	76.80%	90.00%	74.40%
ProbMask (Zhou et al., 2021)	77.01%	90.00%	74.68%
PSO (ours)	77.01%	90.00%	**75.10%**

Table 5. Iterative PSO achieves either better or comparable performance than the state-of-the-art baselines on ImageNet. The experiment follows the same settings of (Kusupati et al., 2020).

Exhibits implicit mask regrowing

- Score-based methods are **biased** since they **can NOT regrow**.
- # regrowing $(t, t+\Delta t)$ = # masks exist at $t+\Delta t$ but absent at t.





Conclusions

- Sparsity-indexed ODE (SpODE) illuminates the **evolution of optimal masks** as the sparsity level increases continuously.

- SpODE enables effective pruning by **traveling along the path of optimal masks**.

- Pruning via SpODE achieves the state-of-the-art performance on various pruning settings and datasets.

- SpODE allows for **implicit mask regrowing**, making it more robust in high sparsity regimes.

Thank you!

Q & A