Find A Winning Sign: Sign Is All We Need to Win the Lottery

Junghun Oh[1] Sungyong Baik[3] Kyoung Mu Lee[1,2]

[1]Dept. of ECE&ASRI, [2]IPAI, Seoul National University

[3]Dept. of Artificial Intelligence, [4]Dept. of Data Science, Hanyang University

{dh6dh, kyoungmu}@snu.ac.kr, dsybaik@hanyang.ac.kr

Introduction - Lottery Ticket Hypothesis (LTH)

- Is over-parameterization required for strong generalization?
  - Many redundant parameters emerge **after training**
  - Can we prune a network **before training**? -> often leads to a sparse network with degraded generalization

Introduction - Lottery Ticket Hypothesis (LTH)

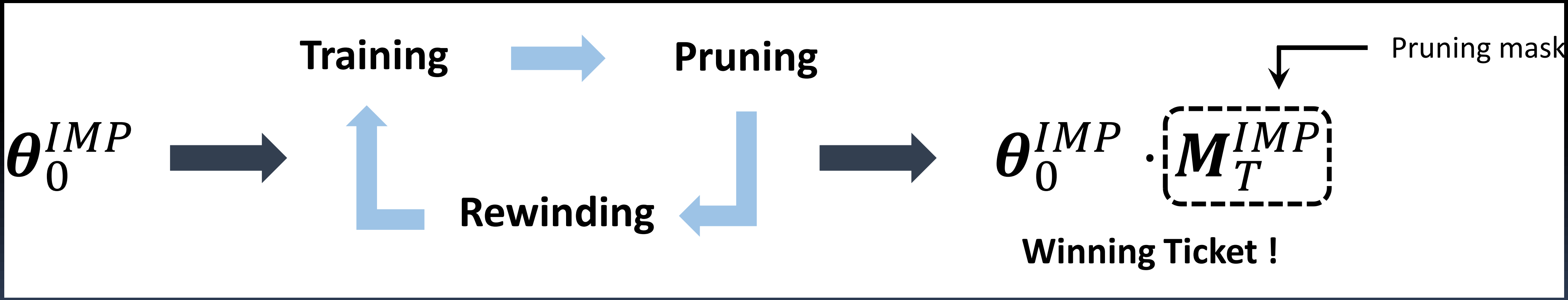
- Is over-parameterization required for strong generalization?
  - Many redundant parameters emerge **after training**
  - Can we prune a network **before training**? -> often leads to a sparse network with degraded generalization

Lottery Ticket Hypothesis (LTH)

There is a sparse network (a.k.a. a winning lottery ticket) that generalizes comparably to its over-parameterized counterpart when trained from scratch

Introduction - Iterative Magnitude Pruning (IMP)

- Iterative Magnitude Pruning (IMP)
  - Find a winning ticket through iterating three phases
    - Training
    - Pruning using parameter magnitudes
    - Rewinding to initialization

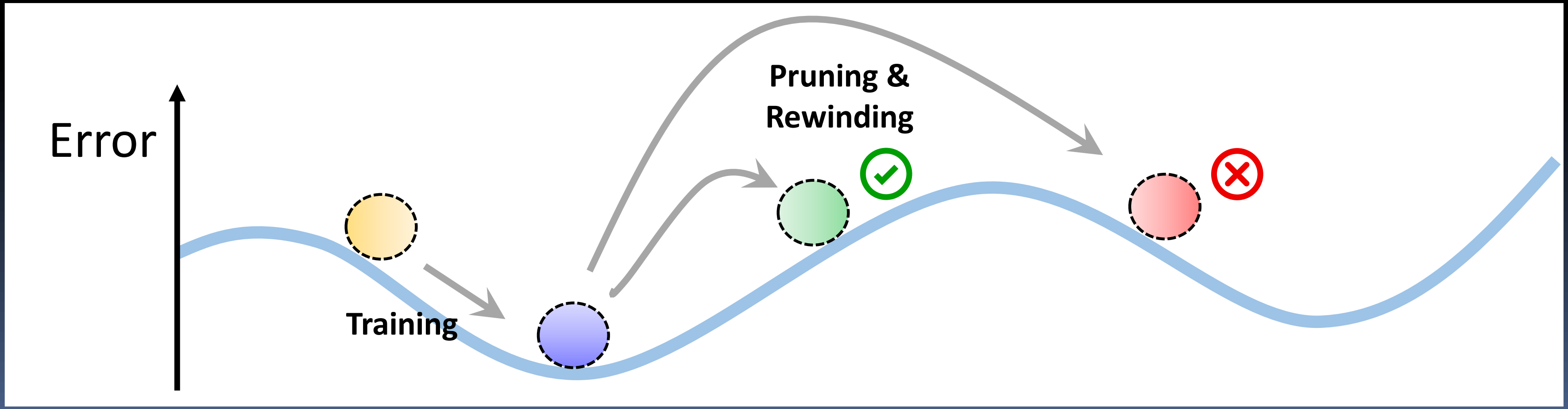


Introduction - Iterative Magnitude Pruning (IMP)

- For IMP to succeed,
  - Network should maintain **stability against SGD noise** after pruning and rewinding, ensuring it can still converge to a solution that remains **linearly mode-connected** to the original solution
  - A network is considered **stable against SGD noise** if it converges to a set of solutions **without high error barriers along the linear path** between them (linearly mode-connected), despite different SGD randomness

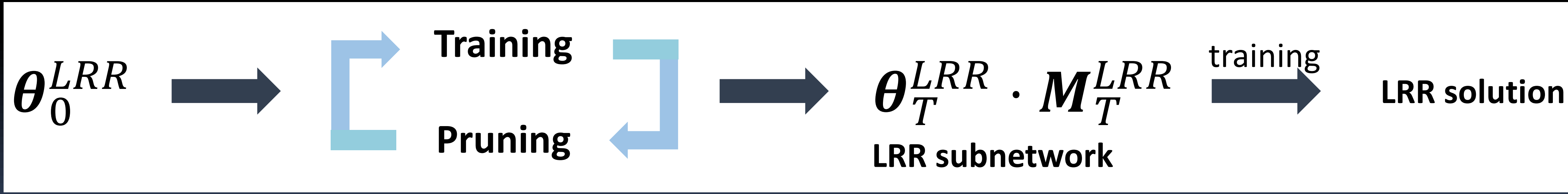
Introduction - Iterative Magnitude Pruning (IMP)

- For IMP to succeed,
  - Network should maintain **stability against SGD noise** after pruning and rewinding, ensuring it can still converge to a solution that remains **linearly mode-connected** to the original solution
  - A network is considered **stable against SGD noise** if it converges to a set of solutions **without high error barriers along the linear path** between them (linearly mode-connected), despite different SGD randomness



Introduction - Learning Rate Rewinding (LRR)

- Learning Rate Rewinding (LRR)
  - Bypass the challenges to satisfy the condition by eliminating parameter rewinding phase
  - Learning **effective parameter signs** is key to find a high-performing sparse network [1]



[1] Gadhikar et al, Masks, signs, and learning rate rewinding. In ICLR, 2024.

Motivation

- Learning Rate Rewinding (LRR)
  - Bypass the challenges to satisfy the condition by eliminating parameter rewinding phase
  - Learning **effective parameter signs** is key to find a high-performing sparse network [1]

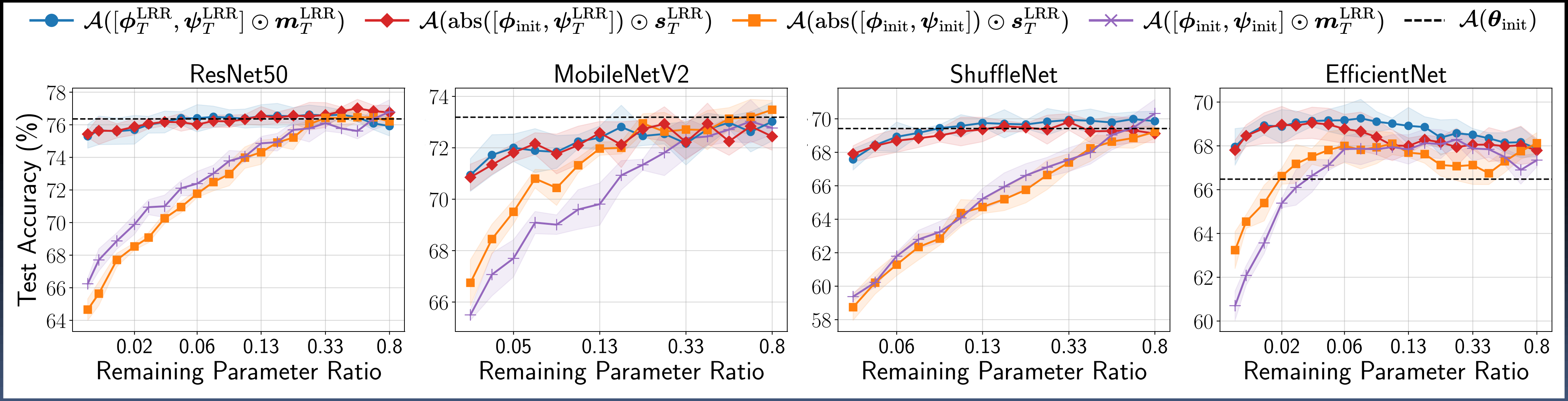
Motivation

Does the parameter sign information of the LRR subnetwork help in finding an effective sparse network at initialization?

[1] Gadhikar et al, Masks, signs, and learning rate rewinding. In ICLR, 2024.

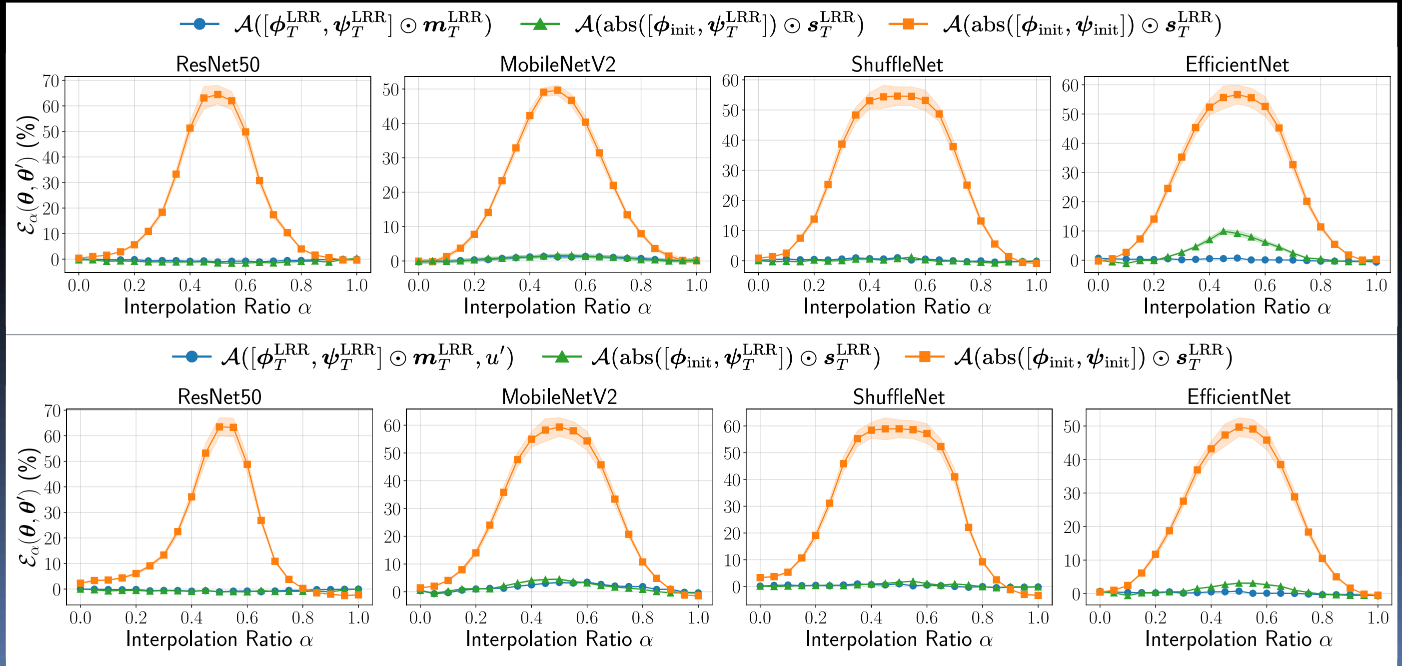
Motivation

- Experiments
  - For LRR subnetwork, randomly initializing all parameters while preserving their signs results in performance worse than LRR solution and comparable to when signs are not preserved
  - Preserving signs and normalization parameters yields performance on par with LRR solution



Motivation

- Experiments
  - Random initialization while preserving signs fails to maintain stability against SGD noise
  - Preserving signs and normalization parameters makes the resulting network **stable against to SGD noise** and **converge to a solution with linear mode connectivity with LRR**



SGD noise stability

Linear mode connectivity

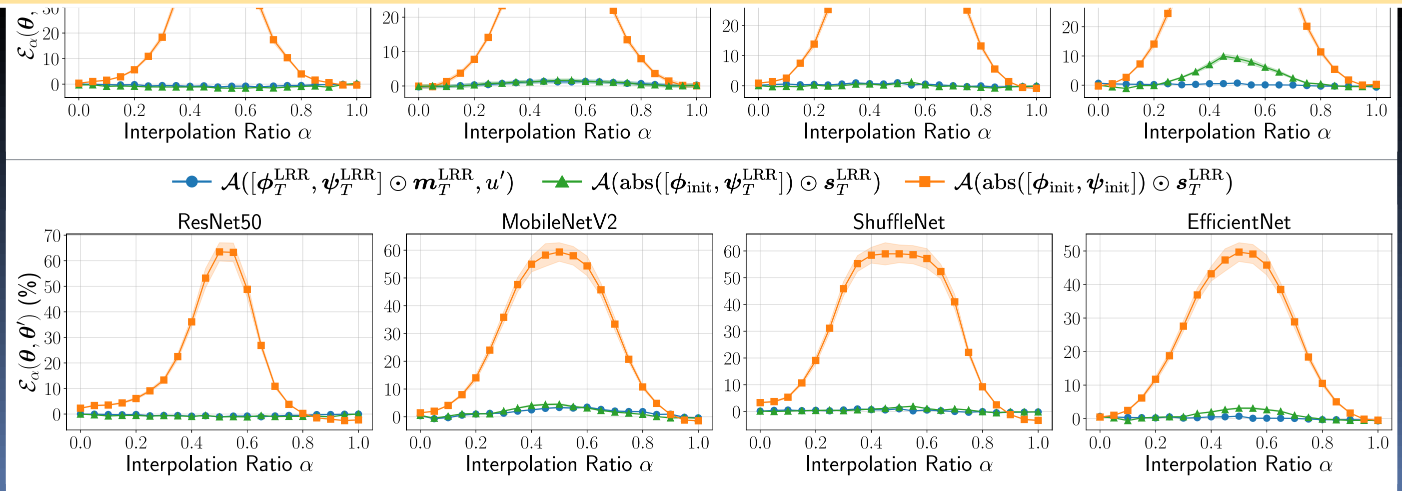
Motivation

- Experiments
  - Random initialization while preserving signs fails to maintain stability against SGD noise
  - Preserving signs and normalization parameters makes the resulting network ... ear

Takeaway

Any randomly initialized network can inherit the generalization potential of the LRR subnetwork through its **signed mask and the normalization layer parameters**

... stability



Linear mode connectivity

Proposed Method

- AWS: Find A Winning Sign
  - Mitigate the impact of normalization parameters by **preventing high error barriers between AWS subnetwork and its counterpart with initialized normalization parameters**
  - Randomly and linearly interpolate between normalization parameters and their initialization
  - Use the interpolated parameters for a network forward pass during training

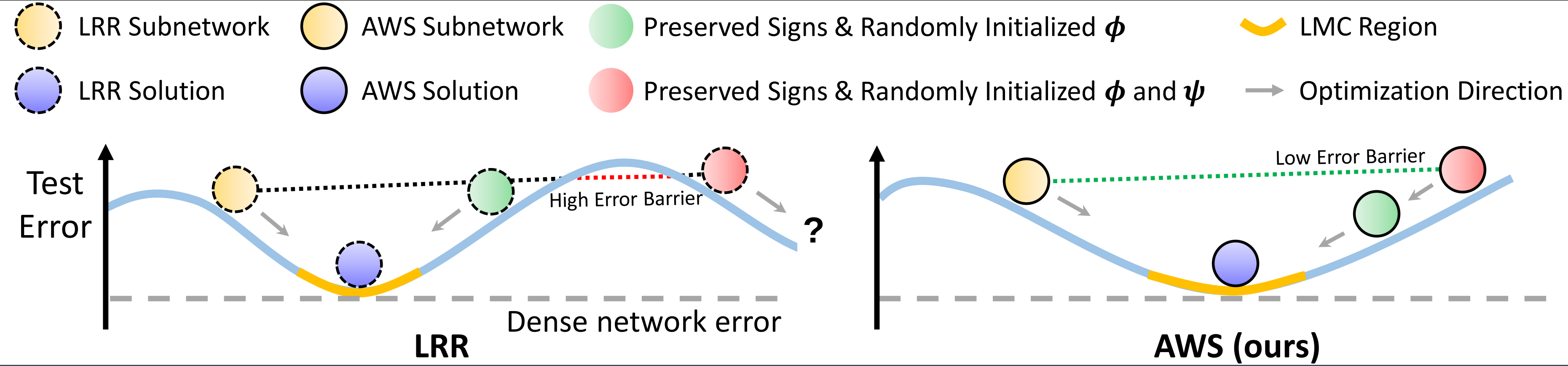
$$(\boldsymbol{\psi}_t^{AWS}, \boldsymbol{\psi}_{init})_\alpha = \alpha \cdot \boldsymbol{\psi}_t^{AWS} + (1-\alpha) \cdot \boldsymbol{\psi}_{init}$$

  - Apply the resulting signed mask to any random initialization and train from scratch

$$\boldsymbol{\theta}_{init} \cdot \mathrm{sign}(\boldsymbol{\theta}_T^{AWS} \cdot \boldsymbol{M}_T^{AWS})$$

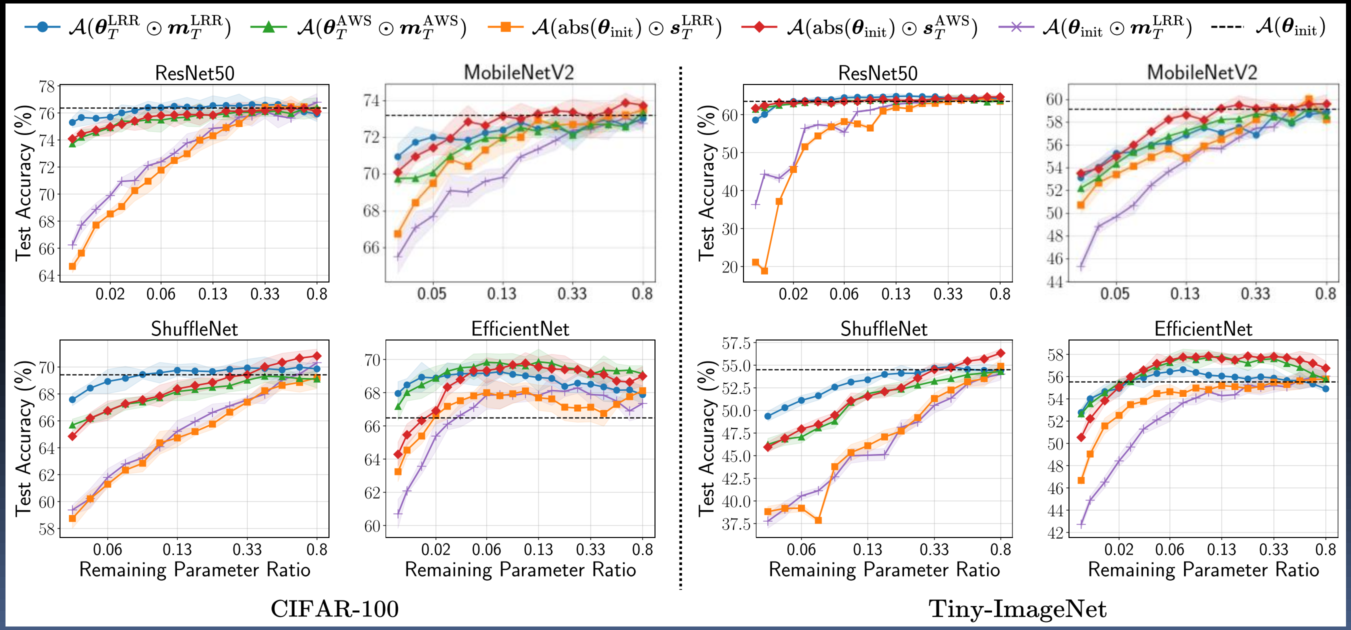
Proposed Method

- AWS: Find A Winning Sign
  - LRR subnetwork can preserve its basin of attraction through the signed mask and the normalization parameters
  - Any random initialization can inherit the basin of attraction of the AWS subnetwork through **its signed mask**



Experiments

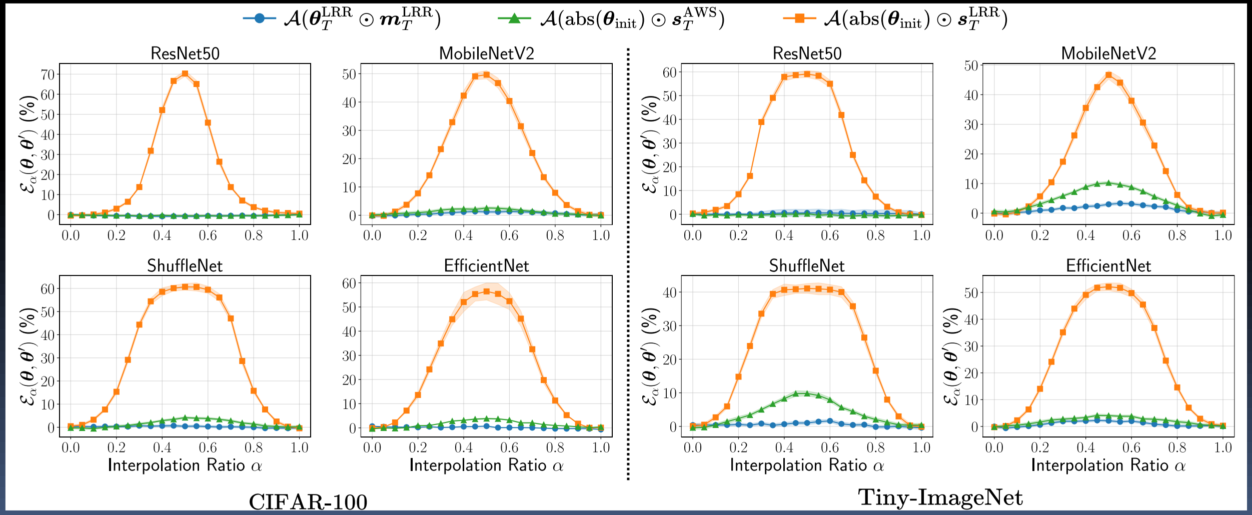
- Performance
  - Initialized networks with LRR signed masks perform worse than LRR solution and comparably to when sign information is not used
  - Initialized networks with AWS signed masks perform comparably to AWS solution



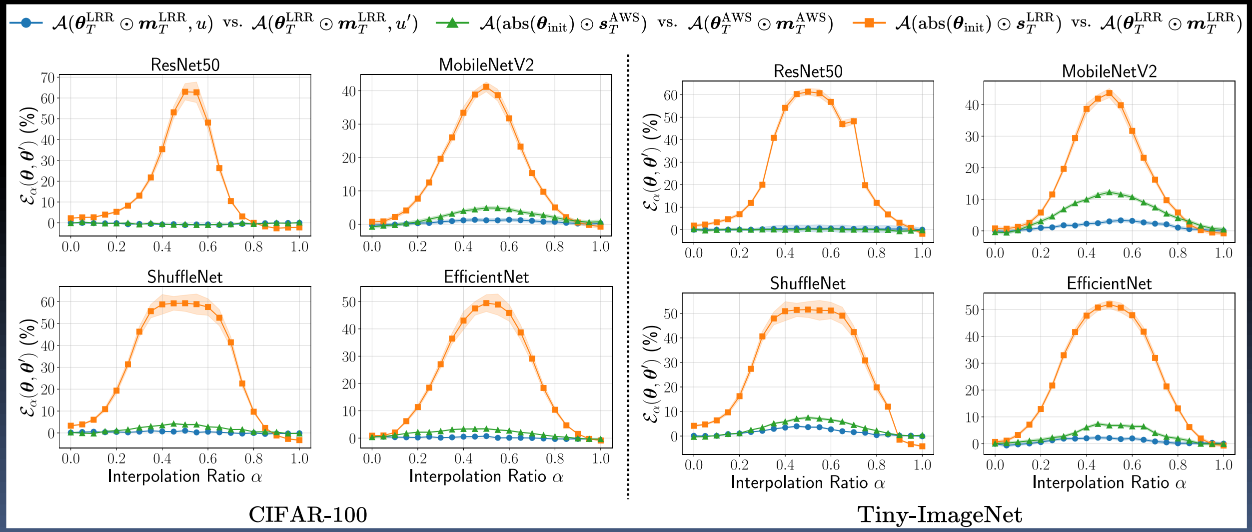
Experiments

- Analysis of SGD noise stability and linear mode connectivity
  - Initialized networks with AWS signed masks are stable against SGD noise and converge to a solution with linearly mode connectivity to AWS solution
  - But, Initialized networks with LRR signed masks are not stable against SGD noise

SGD noise stability



Linear mode connectivity



Conclusion

- Investigating the role of sign in finding a winning ticket
- LRR signed mask is not effective unless it is used with normalization parameters
- AWS can make any randomly initialized network generalize comparably to dense network by transferring AWS signed mask

Thank you