Coarsening the Granularity: Towards Structurally Sparse Lottery Tickets

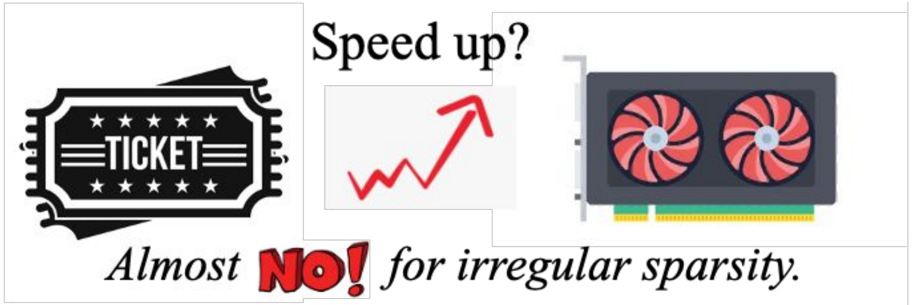
[ICML 2022] Tianlong Chen[1], Xuxi Chen[1], Xiaolong Ma[2], Yanzhi Wang[2], Zhangyang Wang[1]
[1]University of Texas at Austin, [2]Northeastern University

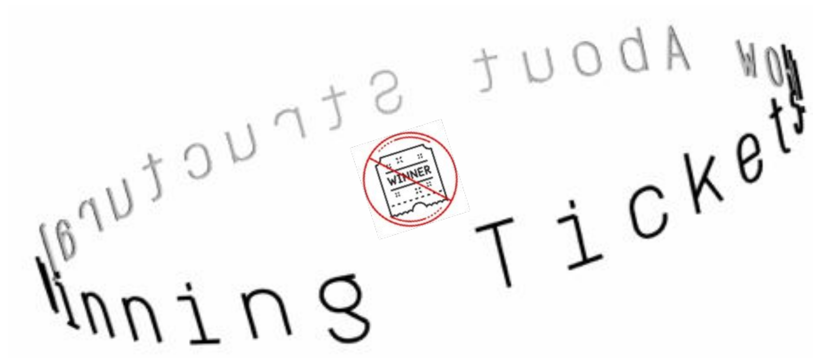
The University of Texas at Austin
Electrical and Computer Engineering

Agenda

- ➢ The Current Limitation of (Unstructured) Lottery Tickets
- ➢ Insightful Findings
- ➢ Our Solutions
- ➢ Our Main Experimental Results

The Current Limitation of (Unstructured) Lottery Tickets





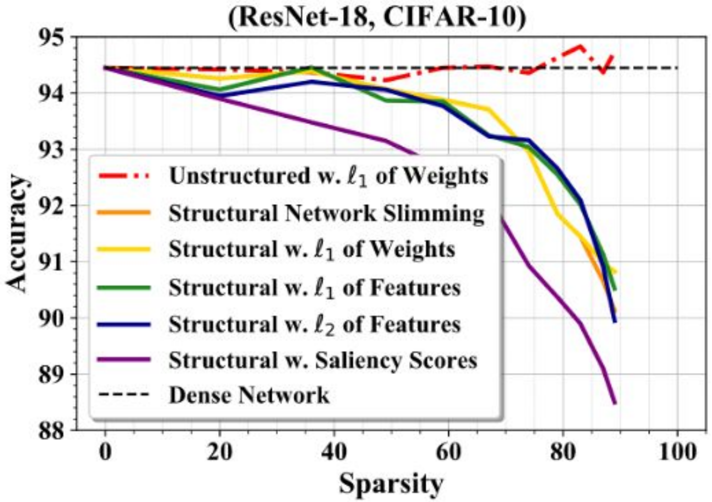


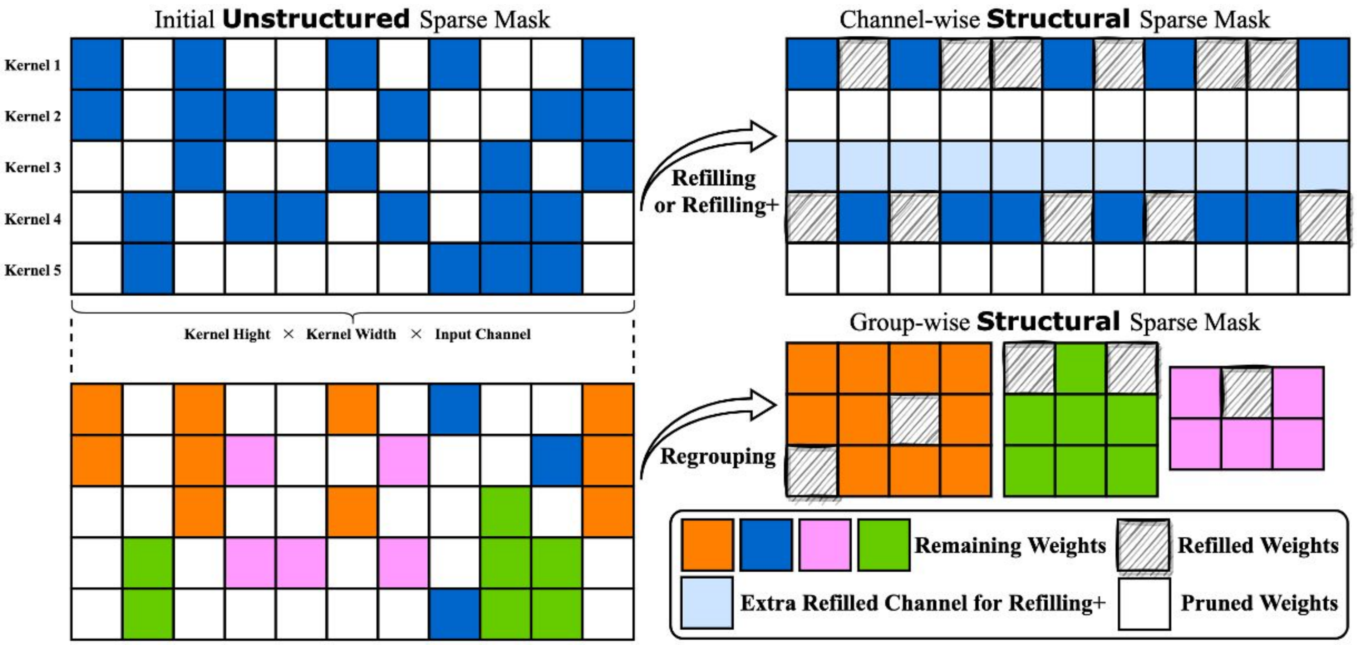
Figure 1. Achieved test accuracy over different sparsity levels of diverse unstructured and structural subnetworks. Sparse models from classical channel-wise structural pruning algorithms (He et al., 2017; Liu et al., 2017; Bartoldson et al., 2019; Molchanov et al., 2019) can not match the full accuracy of the dense model.

Insightful Findings

- To our best knowledge, we are the first to demonstrate the existence of structurally sparse winning tickets at non-trivial sparsity levels (*i.e.*, > 30%), and with both channel-wise and group-wise sparse patterns.

- Extensive experiments validate our proposal on diverse datasets (*i.e.*, CIFAR-10/100, Tiny-ImageNet, and ImageNet) across multiple network architectures, including ResNets, VGG, and MobileNet. Specifically, our structural winning tickets achieve 53.75%~64.93% GPU running time savings at 45%~80% channel- and group-wise sparsity.

Our Solutions



Algorithm 2 `IMP-Refill(+)`

Input: $f(x; m \odot \theta_i)$ with unstructured sparsity s (Algo. 1)
Output: $f(x; m \odot \theta_i)$ with channel-wise structural mask m at sparsity $\tilde{s}$

```
Calculate importance scores of each channel according to certain criterion
Pick top-k channels in m, refill back their 0 (pruned) elements with 1 (trainable) and update m, maintaining s̃ ∼ s
Pick and refill back extra channels in m with s̃+ < s
  # Optional for Refill+
```

Algorithm 3 `IMP-Regroup`

Input: $f(x; m \odot \theta_i)$ with unstructured sparsity s from Algorithm 1, hyperparameters t_1, t_2, b_1, and b_2
Output: $f(x; m) \odot \theta_i$ with group-wise structural mask m at sparsity s^*

```
while dense block can be found do
  Divide the rows of the sparse pruning mask m into t1 groups using hypergraph partitioning (hMETIS)^a
  for group c_i ∈ {c_1, c_2, ..., c_t1} do
    if c_i has ≥ b1 rows then
      Select columns in c_i that has no less than t2 non-zero items
      if ≥ b2 columns are selected then
        Group and Refill the selected columns as well as rows to a dense block, and update m
      end if
    end if
  end for
end while
Set other elements out of dense blocks to 0
```

Our Main Experimental Results

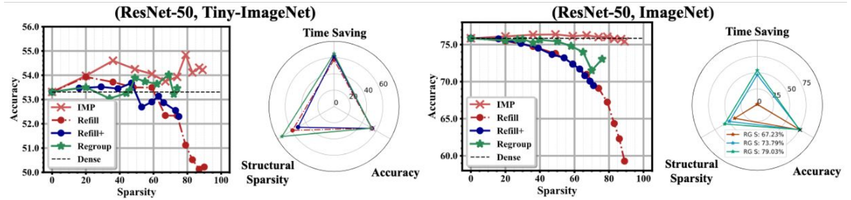


Figure 3. (*Curve plots*) Testing accuracy (%) over network sparsity (%) on Tiny-ImageNet and ImageNet datasets with ResNet-50 (25.56 M). (*Radar plots*) The end-to-end inference time saving of extreme structural winning tickets. Unstructured subnetworks or dense models do not have structural sparsity, and thus they are plotted as dots in the axes of accuracy in the corresponding radar plot. The rightmost plot includes three extreme regroup tickets with accuracy drop $< 1\%$, where "RG S: $x\%$" indicates unstructured sparsity before regrouping.

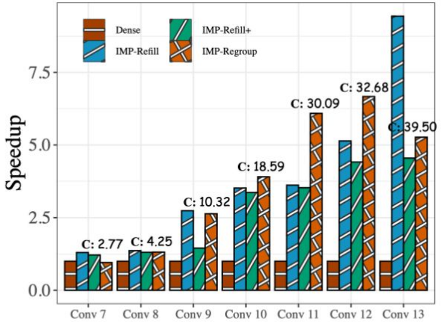


Figure 7. The layer-wise performance of convolution operations in extreme structural winning tickets of (VGG-16, C10). The first six conv. operations are omitted since there is no meaningful speedup, coincided with Rumi et al. (2020). Marks like "C: 2.77" indicate the layer-wise compression ratio of `IMP-Regroup`.

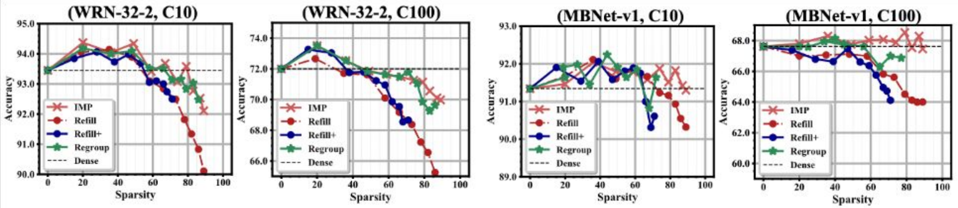


Figure 4. Testing accuracy (%) over sparsity (%) on CIFAR-10/100 with Wide-ResNet-32-2 (1.86 M) and MobileNet-v1 (3.21 M).

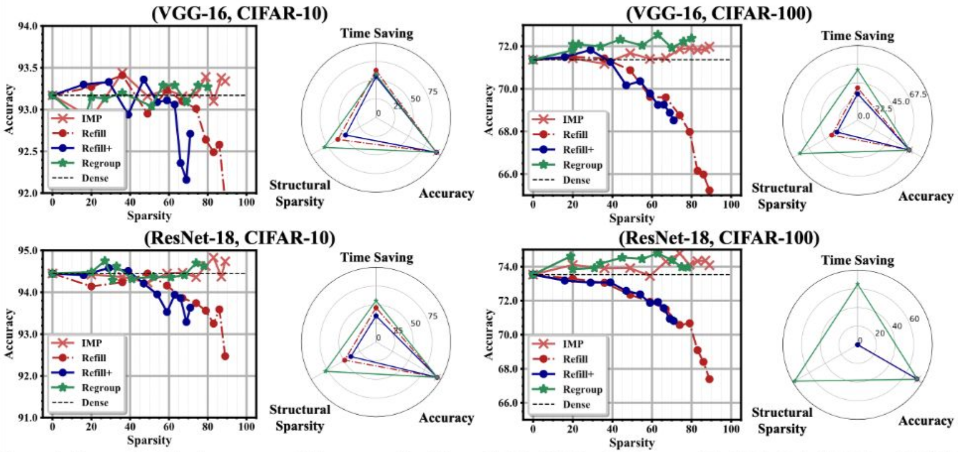


Figure 5. (*Curve plots*) Testing accuracy (%) over sparsity (%) on CIFAR-10/100 with large models VGG-16 (14.72 M) and RN-18 (11.22 M). (*Radar plots*) The end-to-end inference time saving of extreme structural winning tickets. Note that unstructured subnetworks or dense models do not have structural sparsity, and thus they are plotted as dots in the axes of accuracy in the corresponding radar plot.

Coarsening the Granularity: Towards Structurally Sparse Lottery Tickets

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering
Northeastern University

VITA GitHub

Tianlong Chen[1], Xuxi Chen[1], Xiaolong Ma[2], Yanzhi Wang[2], Zhangyang Wang[1]
[1]University of Texas at Austin, [2]Northeastern University

ICML International Conference On Machine Learning

➢ The Current Limitation of (Unstructured) Lottery Tickets

*Almost NO! for irregular sparsity**

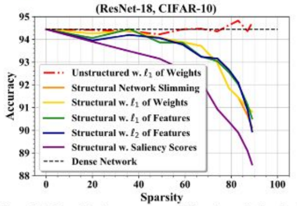


Figure 1. Achieved test accuracy over different sparsity levels of diverse unstructured and structural subnetworks. Sparse models from classical channel-wise structural pruning algorithms (He et al., 2017; Liu et al., 2017; Bartoldson et al., 2019; Molchanov et al., 2019) can not match the full accuracy of the dense model.

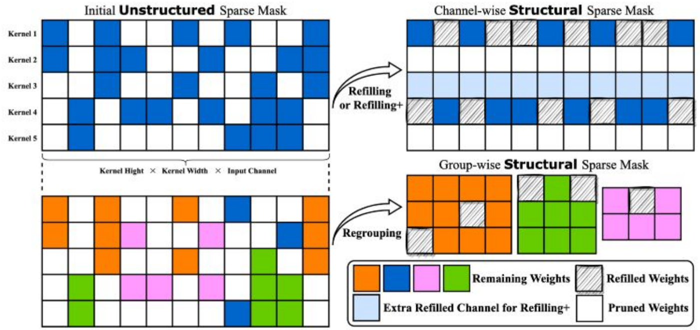
➢ Insightful Findings

- ❑ To our best knowledge, we are the first to demonstrate the existence of structurally sparse winning tickets at non-trivial sparsity levels (*i.e.*, > 30%), and with both channel-wise and group-wise sparse patterns.
- ❑ Extensive experiments validate our proposal on diverse datasets (i.e., CIFAR-10/100, Tiny-ImageNet, and ImageNet) across multiple network architectures, including ResNets, VGG, and MobileNet. Specifically, our structural winning tickets achieve 53.75% ~ 64.93% GPU running time savings at 45% ~ 80% channel- and group-wise sparsity.

➢ Our Solutions

- ❖ Refilling
- ❖ Refilling+
- ❖ Regrouping

* R.K. Some packages like XNNPACK can accelerate unstructured sparse neural networks on certain hardware platforms like smartphone processors.



➢ Our Main Experimental Results

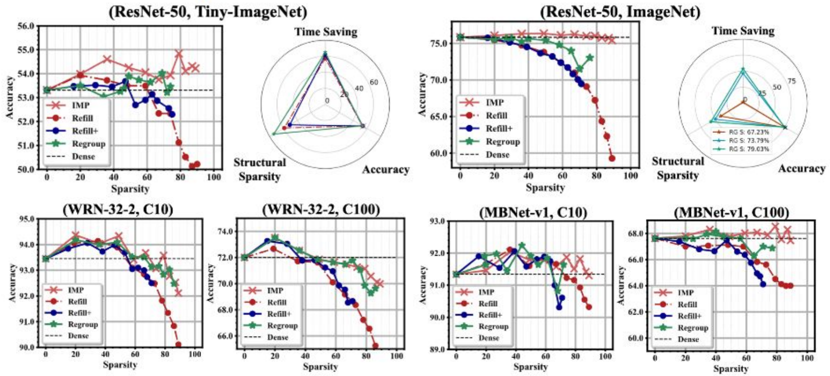


Figure 4. Testing accuracy (%) over sparsity (%) on CIFAR-10/100 with Wide-ResNet-32-2 (1.86 M) and MobileNet-v1 (3.21 M).

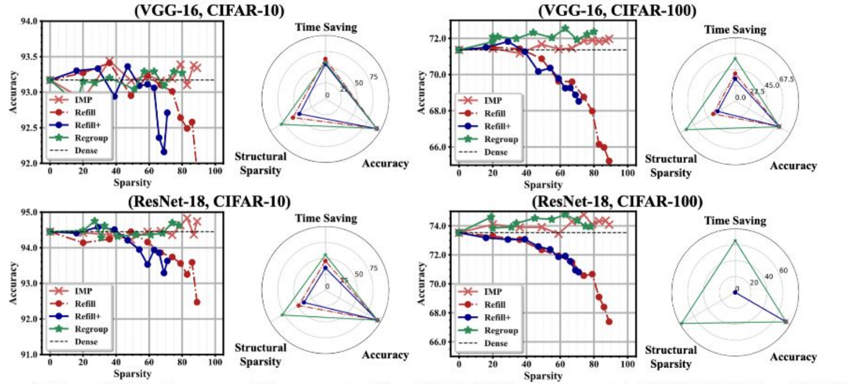


Figure 5. (*Curve plots*) Testing accuracy (%) over sparsity (%) on CIFAR-10/100 with large models VGG-16 (14.72 M) and RN-18 (11.22 M). (*Radar plots*) The end-to-end inference time saving of extreme structural winning tickets. Note that unstructured subnetworks or dense models do not have structural sparsity, and thus they are plotted as dots in the axes of accuracy in the corresponding radar plot.

Contact: {tianlong.chen, xxchen, atlaswang}@utexas.edu, {ma.xiaol, yanz.wang}@northeastern.edu

Q&A