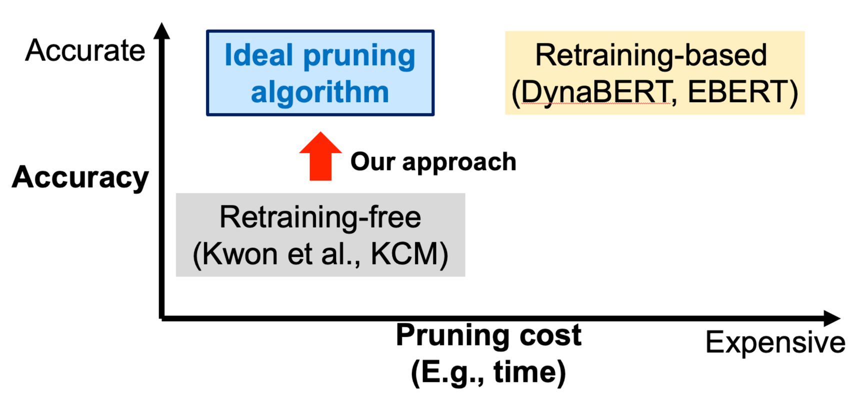
Accurate Retraining-free Pruning for Pretrained Encoder-based Language Models

Seungcheol Park — Seoul National University
Hojun Choi — KAIST
U Kang — Seoul National University

Paper GitHub

Summary

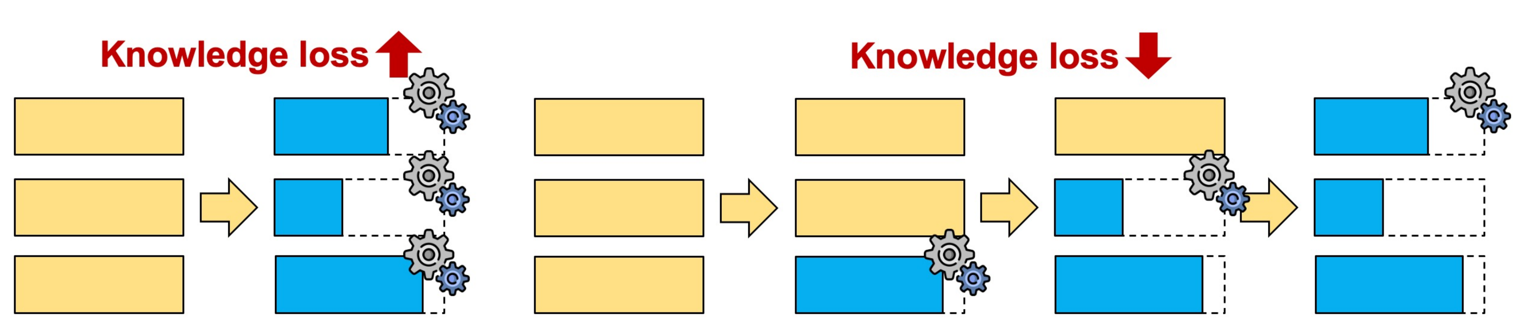
- **Problem: Retraining-free Structured Pruning of PLMs**
  - Given a pre-trained language model (PLM), how can we accurately prune it without retraining?
  - We focus on pruning attention heads and neurons
- **Previous structured pruning algorithms for PLMs**
  - Retraining-based algorithms
    - Accurate, but too expensive
  - Retraining-free algorithms
    - Cheap, but too inaccurate



- **Proposed Method: K-prune**
  - Improving the accuracy of retraining-free pruning algorithms by preserving PLM's knowledge by iterative pruning process
- **Experimental results**
  - Up to 58%p more accurate than existing retraining-free pruning algorithms with similar pruning costs
  - Up to 422× lower pruning cost than existing retraining-based pruning algorithms with similar accuracy

Intuition

- Previous retraining-free algorithms (a) lose PLM's useful knowledge because of its aggressive one-shot pruning process
- An iterative pruning process (b) with an efficient knowledge recovery process the loss of PLM's useful knowledge

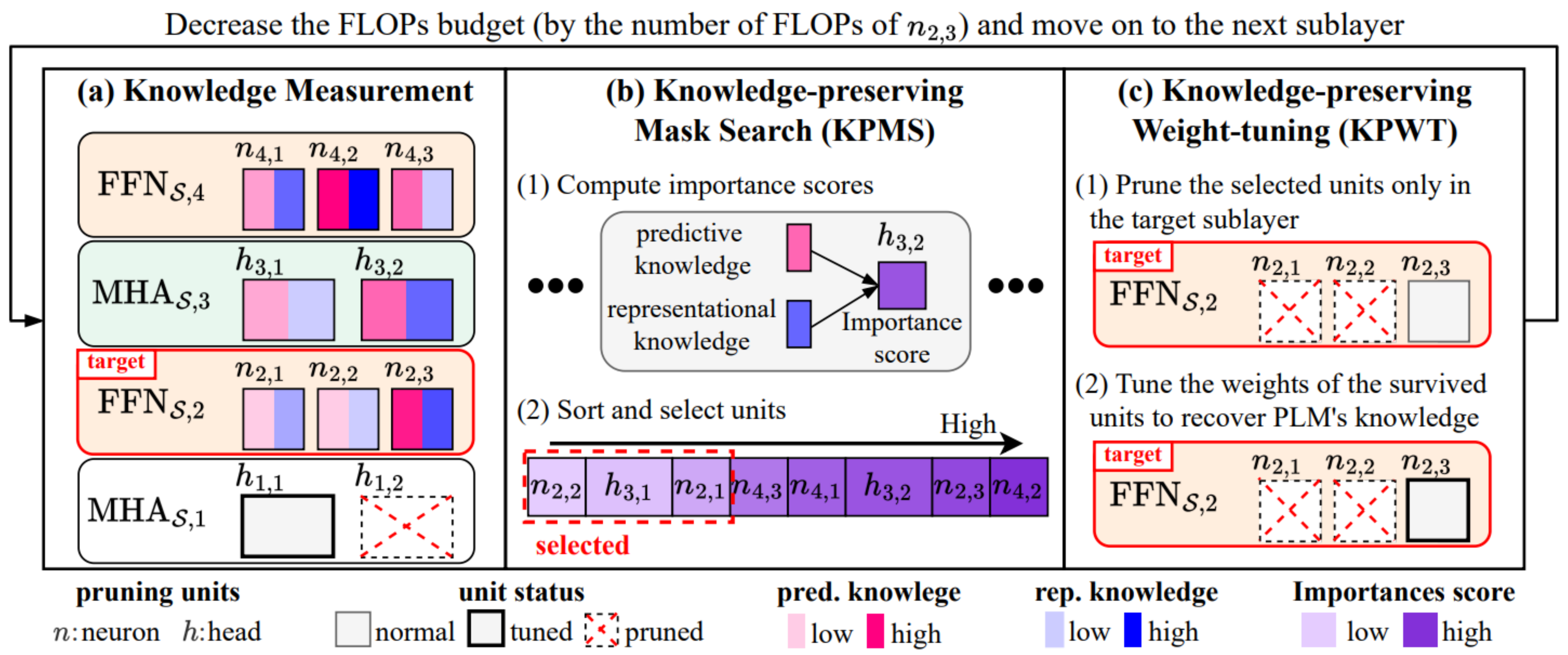


(a) One-shot pruning (b) Iterative pruning

Proposed Method

Knowledge-preserving pruning (K-prune)

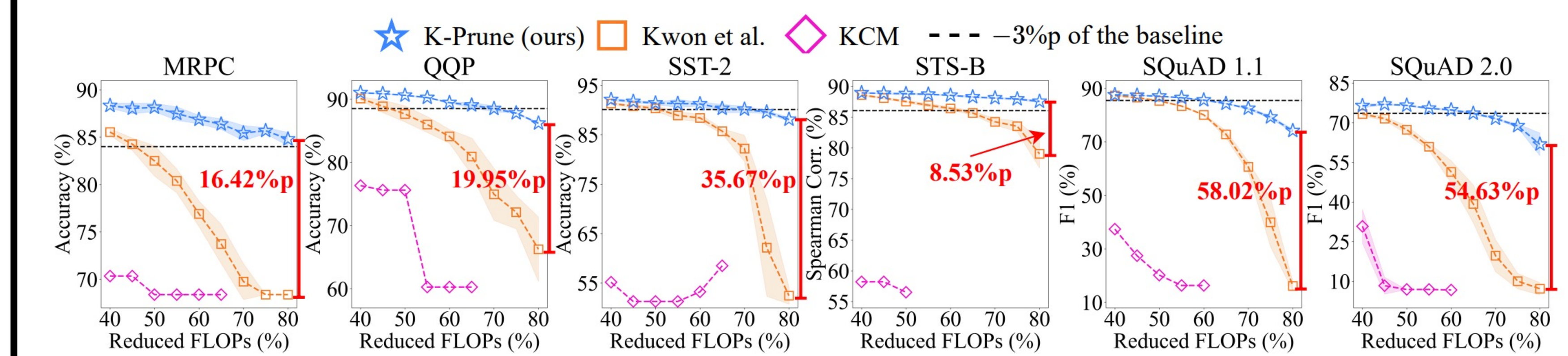
- An accurate retraining-free structured pruning algorithm for pretrained language models
- Focusing on preserving the useful knowledge of pretrained models through a carefully designed sublayer-wise iterative process includes an efficient knowledge recovery process



- **Main ideas**
  - **(a) Knowledge measurement**
    - We measure the amount of inherent knowledge in each attention head and neuron to exploit it as an importance criterion
  - **(b) Knowledge-preserving mask search (KPMS)**
    - We estimate importance scores that reflect the amount of their inherent knowledge considering knowledge types and unit types
    - Selecting uninformative units with the least importance scores
  - **(c) Knowledge-preserving Weight-tuning (KPWT)**
    - We prune the redundant components selected in (b) and perform a short weight-tuning process to reconstruct the knowledge of PLMs
    - Extremely efficient and performs in a second for each sublayer

Experiments

- **Accuracy of the compressed models**
  - K-prune shows up to 58%p higher f1 score than competitors

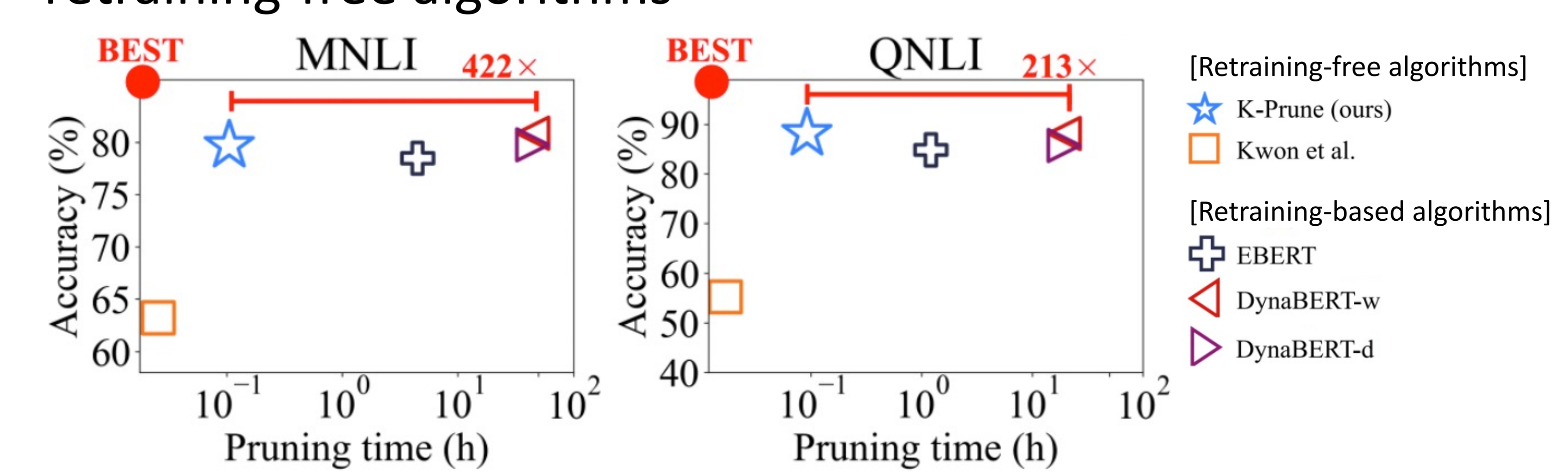


- **Inference speedup** (on 1080Ti without customized kernels)
  - Compare the best inference speedup within a 3%p accuracy drop
  - K-prune shows the largest speedup than competitors

Method	MRPC	STS-B	$\text{SQuAD}_{1.1}$	$\text{SQuAD}_{2.0}$	Avg.*
KCM (Nova et al., 2023)	1.08×	1.23×	1.20×	1.08×	1.15×
Kwon et al. (2022b)	1.59×	2.10×	2.09×	1.75×	1.87×
K-prune (ours)	2.66×	2.43×	2.60×	2.93×	2.65×

∗ Geometric mean

- **Pruning efficiency**
  - Accuracy of the compressed models vs. pruning time under a compression ratio of 75%
  - K-prune shows the best trade-off without losing the efficiency of retraining-free algorithms



- **Pruning of LLMs**
  - Perplexities of OPT models on WikiText2 dataset after pruning with K-prune

OPT-1.3B					
Pruning rate	0%	5%	10%	15%	20%
Perplexity	14.67	14.41	13.96	14.67	15.74
Difference	-	-1.77%	-4.84%	0.00%	7.29%

OPT-2.7B					
Pruning rate	0%	5%	10%	15%	20%
Perplexity	12.46	12.23	11.94	12.01	12.51
Difference	-	-1.85%	-4.17%	-3.61%	0.40%