# SPDY: Accurate Pruning with Speedup Guarantees

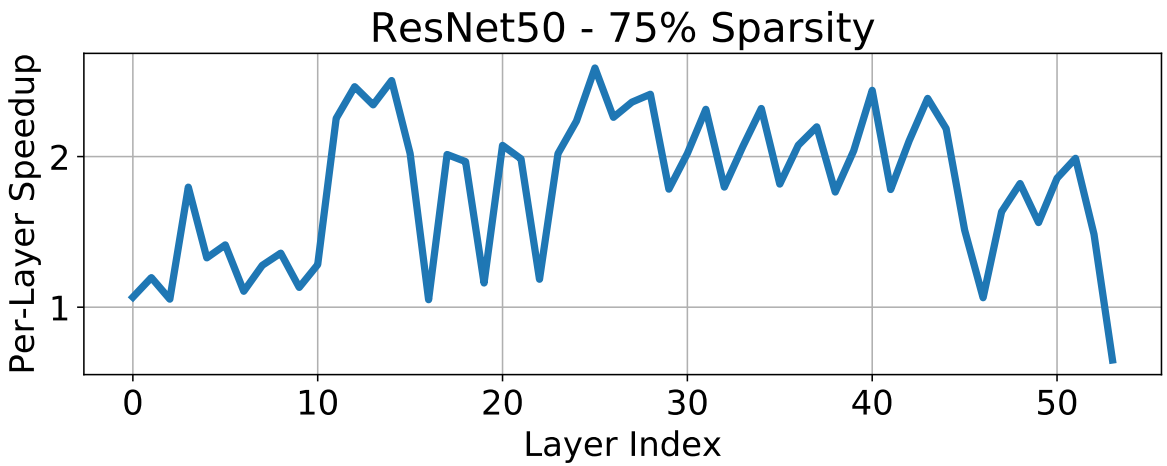
## ICML 2022

Elias Frantar, Dan Alistarh


July 15, 2022

# Motivation

- Pruning methods (e.g.: RIGL [4], WoodFisher [13], AC/DC [12]) produce accurate *unstructured* sparse models:
  - Reduced compute → used to be difficult to utilize in practice
- Increasingly advanced software & hardware acceleration techniques (e.g.: DeepSparse [11], Sputnik [6] & others [3, 1]) achieve better and better *real* speedups
  - Complex relationship between sparsity & speedup




Existing pruning algorithms do not take this into account.

# Methods

**Goal:** find per-layer target sparsities $s_\ell$ that

- Maximize model accuracy
- While achieving a certain real inference speedup

→ Existing techniques not reliable enough for unstructured sparsity

**SPDY:** *Sparsity Profiles via Dynamic Programming search*

- Highly efficient dynamic programming algorithm for solving constrained layer-wise optimization problem
- Automatic search process to inject global cross-layer information into layer-wise problem
- Enhancements to the AdaPrune [9] methods for fast & accurate one-shot pruning

# Experiments: CPU Inference with DeepSparse [11]

- Run SPDY *once* in the beginning to find sparsity profile
- Then apply state-of-the-art gradual pruning methods: AC/DC [12], M-FAC [5], gradual magnitude [14]
- ResNet50 [7], MobileNetV1 [8], YOLOv5 [10], BERT [2]

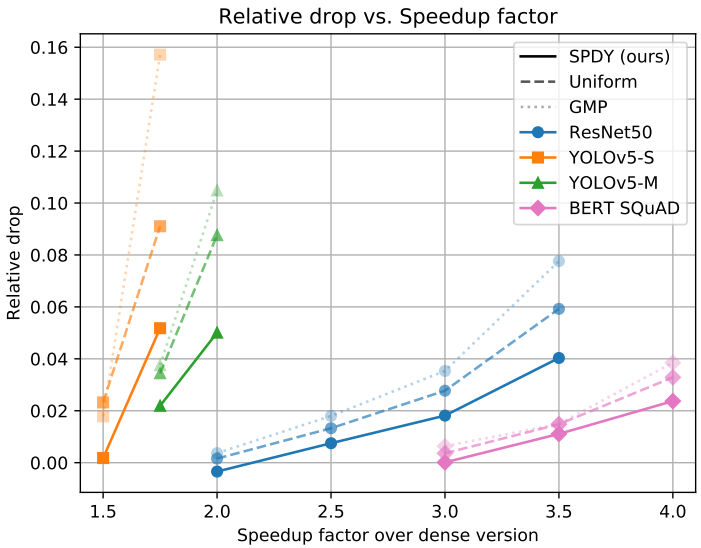



Figure: Uniform and GMP.

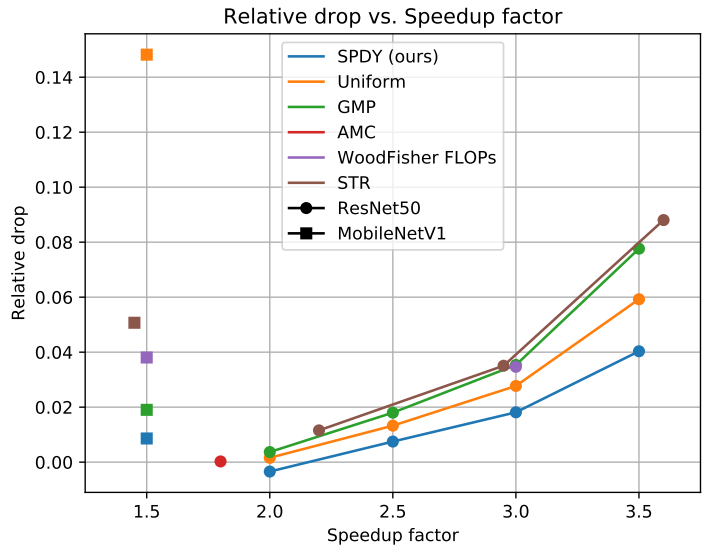



Figure: Other pruning methods.

**Code:** `https://github.com/IST-DASLab/spdy`

# Bibliography I


[1] Shail Dave, Riyadh Baghdadi, Tony Nowatzki, Sasikanth Avancha, Aviral Shrivastava, and Baoxin Li.
Hardware acceleration of sparse and irregular tensor computations of ML models: A survey and insights.
*Proceedings of the IEEE*, 109(10):1706–1752, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: Pre-training of deep bidirectional transformers for language understanding.
In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

# Bibliography II


[3] Erich Elsen, Marat Dukhan, Trevor Gale, and Karen Simonyan.

Fast sparse convnets.
In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen.
Rigging the lottery: Making all tickets winners.
In *International Conference on Machine Learning (ICML)*, 2020.

[5] Elias Frantar, Eldar Kurtic, and Dan Alistarh.
M-FAC: Efficient matrix-free approximations of second-order information.
In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

# Bibliography III


[6] Trevor Gale, Matei Zaharia, Cliff Young, and Erich Elsen.
Sparse GPU kernels for deep learning.
In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition.
In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam.
MobileNets: Efficient convolutional neural networks for mobile vision applications.
*arXiv preprint arXiv:1704.04861*, 2017.

# Bibliography IV


[9] Itay Hubara, Brian Chmiel, Moshe Island, Ron Banner, Seffi Naor, and Daniel Soudry.
Accelerated sparse neural training: A provable and efficient method to find N:M transposable masks.
In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[10] Glenn Jocher.
YOLOv5.
https://github.com/ultralytics/yolov5, 2022.

[11] Mark Kurtz, Justin Kopinsky, Rati Gelashvili, Alexander Matveev, John Carr, Michael Goin, William Leiserson, Sage Moore, Bill Nell, Nir Shavit, and Dan Alistarh.
Inducing and exploiting activation sparsity for fast inference on deep neural networks.

# Bibliography V

In *International Conference on Machine Learning (ICML)*, 2020.

[12] Alexandra Peste, Eugenia Iofinova, Adrian Vladu, and Dan Alistarh.
AC/DC: Alternating compressed/decompressed training of deep neural networks.
In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[13] Sidak Pal Singh and Dan Alistarh.
Woodfisher: Efficient second-order approximation for neural network compression.
In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

# Bibliography VI

[14] Michael Zhu and Suyog Gupta.
To prune, or not to prune: exploring the efficacy of pruning for model compression.
*arXiv preprint arXiv:1710.01878*, 2017.