Evaluating Neural Network Pruning Techniques on Vision Transformers (ViTs)

Sarah Chen [1] Victor Kolev [1] Kaien Yang [1] Jonathan Frankle [2]

[1]Computer Science, Stanford [2]Mosaic ML

Overview

We provide a comprehensive benchmark for vision transformer (ViT) pruning by evaluating a spectrum of magnitude-based train-then-sparsify pruning methods on a small ViT applied to CIFAR-10.

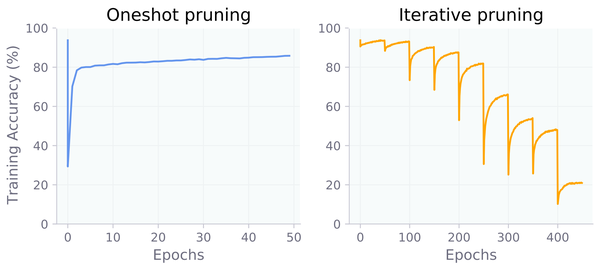
Background

- Modern Transformers often comprise billions of parameters, incurring large computational and environmental costs.
- Pruning enables gains in efficiency via sparsification and model compression.
- However, standard methods (shown to be effective on CNNs, NLP Transformers, etc.) have not yet been thoroughly evaluated on ViTs.

Methods

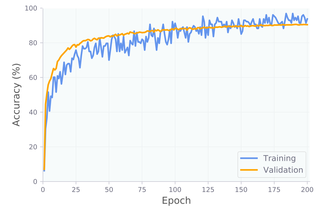
We apply all pruning strategies after standard training is run to convergence.

- Pruning methods
  - Unstructured: remove individual weights
  - Structured per-row/column: remove weight matrix rows/columns
- Pruning distributions
  - Global: remove low-magnitude weights, regardless of layer
  - Layerwise: remove a uniform ratio of low-magnitude weights per layer
  - Random: shuffle masks, maintaining per-layer sparsity ratios
- Pruning schedules
  - One-shot with fine-tuning: prune once and then fine-tune with learning rate rewinding (25% of original training time)
  - Iterative pruning: take small pruning steps to mitigate accuracy drops

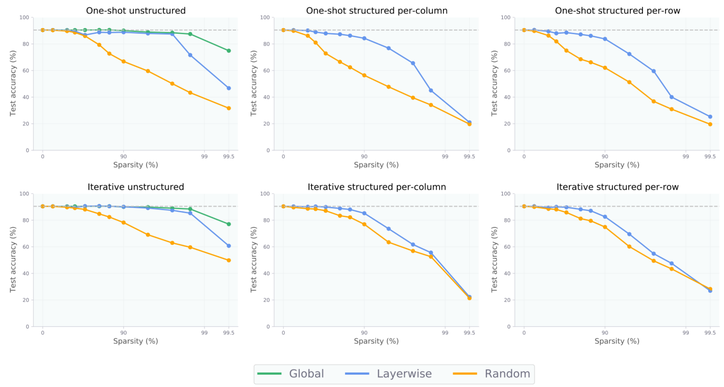


Results

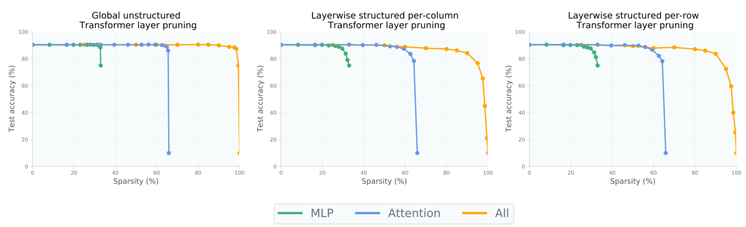
- The ViT is a 6.3-million-parameter network with 7 layers, 12 attention heads, and 384 embedding dimensions. It achieves 90% validation accuracy on CIFAR-10 after 200 epochs.



- Generally, magnitude-based methods outperform random baselines and iterative methods outperform their one-shot counterparts, but all methods show high resilience to pruning.

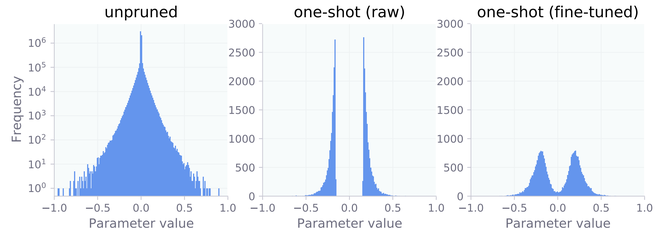


- Model performance when one-shot pruning is applied to MLP layers only, attention layers only, and all layers. Even when pruning all MLP layers, the network is able to maintain reasonable accuracy via residual connections; this highlights the key role of attention units.

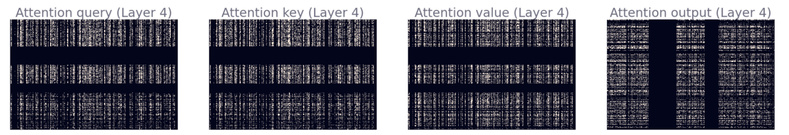


Visualisation

- Weight magnitudes for global unstructured pruning at 99.5% sparsity. Fine-tuning recovers from 8.15% to 74.89% validation accuracy.



- Visualization of global unstructured pruning masks on layer 4 attention matrices at 95% sparsity. We can observe structural patterns that implicitly arise, and result in whole attention heads being pruned.



Conclusion

1. Compared to random baselines, ViTs can sustain high (>95%) sparsity levels without significant accuracy loss, especially for global unstructured pruning methods.
2. There is a notable difference between the prunability of different layers, suggesting that information is unevenly distributed across layers. Specifically, feedforward layers are more prunable than attention layers.
3. Fine-tuning mitigates performance degradation even with sub-optimal pruning strategies or one-shot pruning.
4. Implicit structured trends emerge in unstructured pruning approaches, suggesting that weight matrices of ViTs are highly structured.
5. Magnitude-based structured pruning approaches do not significantly outperform random pruning — a potential direction for future work.