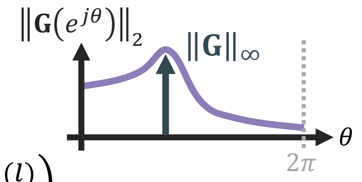
Layer-Adaptive State Pruning for Deep State Space Models

Background and Objective

State space models are efficient alternatives to attention models, offering strong representational capacity for long sequences. The objective of this research is **to optimize trained state space models** by removing insignificant system parameters with minimal accuracy loss. The proposed method performs multi-system approximation, further enhancing the efficiency of overparameterized state space models.

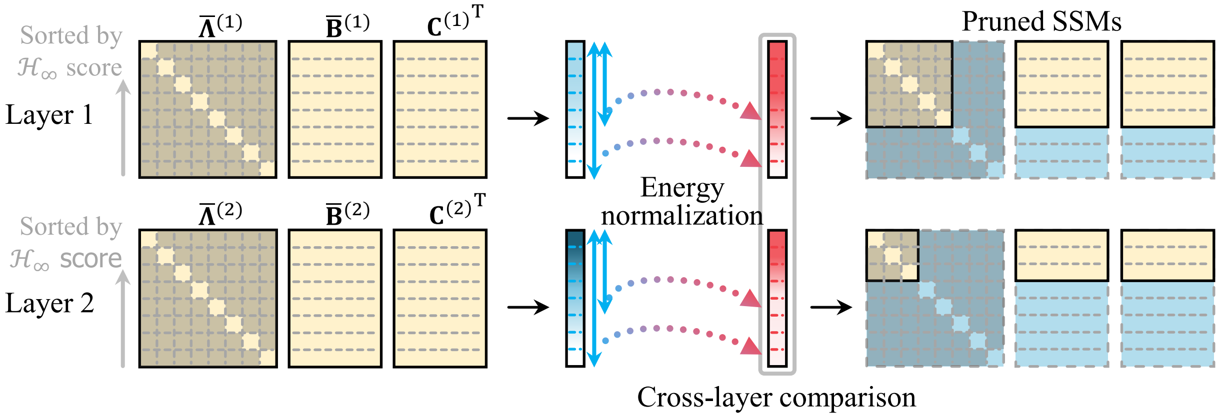
Methods

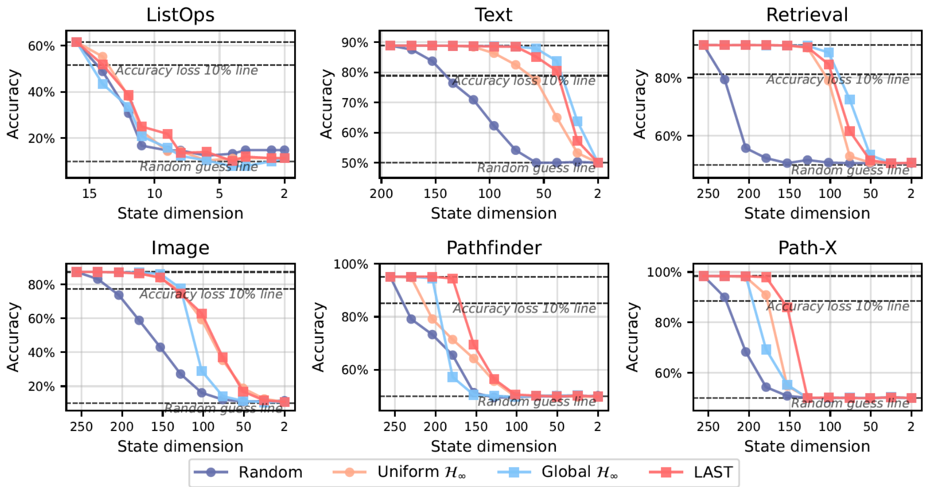
- Structured pruning with the state pruning granularity, layer-adaptive pruning ratios, and H_∞ norm-based pruning criteria



$$\underset{P^{(l)}\subset S^{(l)}}{\text{minimize}} \left\|f_\sigma\left(\mathrm{u}^{(1)};\Sigma^{(1:L)}\right)-f_\sigma\left(\mathrm{u}^{(1)};\hat{\Sigma}^{(1:L)}\right)\right\|_2^2 \Rightarrow \frac{\mathcal{H}_\infty\left(x_i^{(l)};\Sigma^{(l)}\right)}{\sum_{j\le i}\mathcal{H}_\infty\left(x_j^{(l)};\Sigma^{(l)}\right)}$$

By minimizing output energy distortion caused by removing a state, we can derive an importance of each state based on the maximum gain of its corresponding subsystem.





Results

1. Smaller model size (-33% params)
2. Strong insignificant state identification performance (<1% accuracy loss)
3. Faster inference (x1.7, max) and lower memory usage (x0.6, max)

Method	Average accuracy loss ↓
Random	29.53 (32.82)
Uniform magnitude	22.03 (24.48)
Global magnitude	17.49 (19.43)
LAMP	18.07 (20.07)
Uniform $\mathcal{H}_\infty$	4.32 (4.80)
Global $\mathcal{H}_\infty$	7.51 (8.35)
LAST	**0.52 (0.58)**

Minseon Gwak