SCIENCE ▪ PASSION ▪ TECHNOLOGY

Institute for Technical Informatics < Inffeldgasse 16/I, 8010 Graz, Austria

GRAIL: Post-hoc Compensation by Linear Reconstruction for Compressed Networks

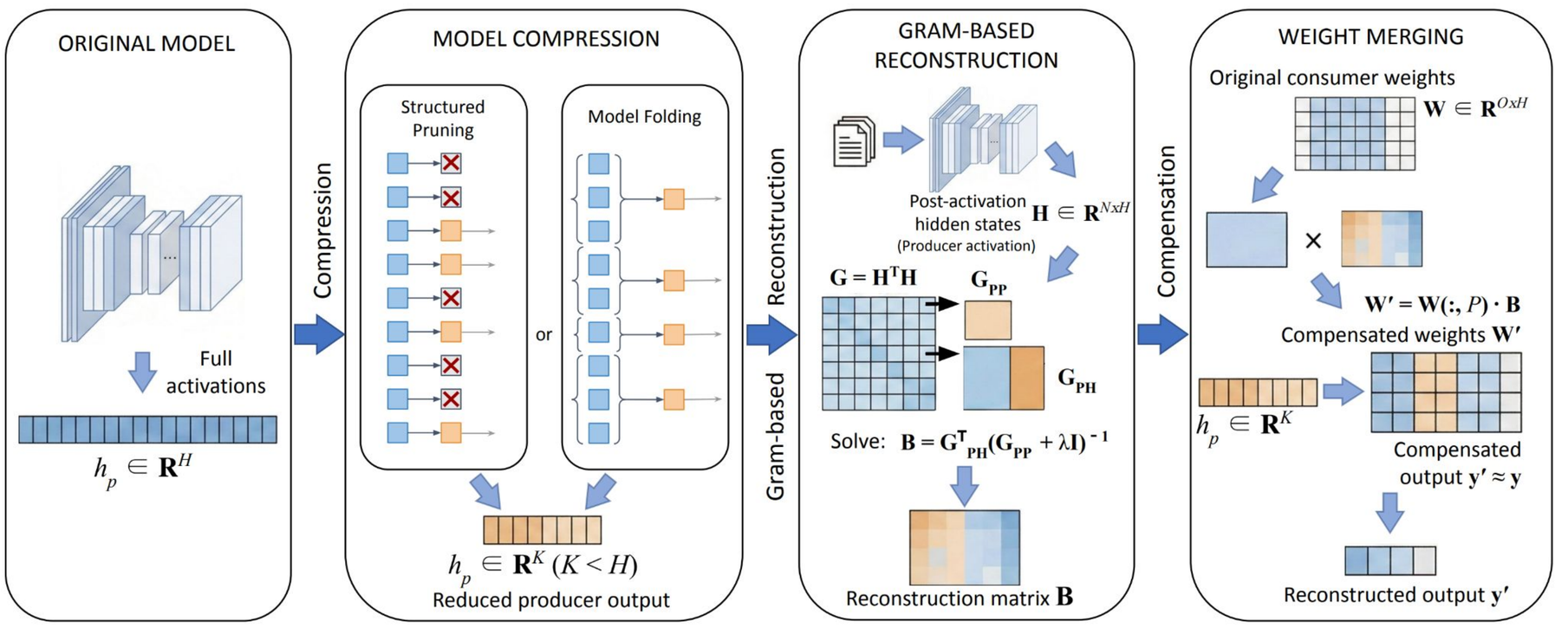
Wenwu Tang[1], Dong Wang[1], Lothar Thiele[3], Olga Saukh[12]

[1] Institute for Technical Informatics, Graz University of Technology, Austria
[2] Complexity Science Hub Vienna, Austria
[3] ETH Zurich, Switzerland
{wenwu.tang, dong.wang, saukh} @tugraz.at, thiele@tik.ee.ethz.ch

Motivation

- Model compression is necessary for the deployment of deep learning model in resource-constrained setting.
- Model compression suffers from notable **performance degradation** under aggressive compression.

Model Compression




```
X
|
▼
W_fc → prune
Layer(l)
|
▼
H_p
|
▼
W_proj → compensate
Layer(l+1)
|
▼
Y
```

$$\min_{\mathbf{B}} \left\|\mathbf{H} - \mathbf{H}_P\mathbf{B}^\top\right\|_F^2 + \lambda\|\mathbf{B}\|_F^2,$$

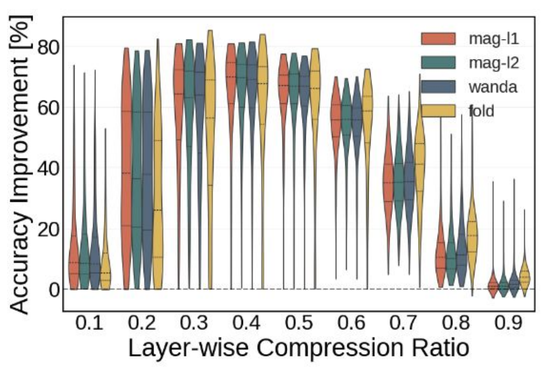
- *Model Compression*
  - GRAIL is **Method- and Model-Agnostic**, it can be applied after Pruning (Magnitude, Wanda, etc.) or Model Folding to both Vision Model (ResNets, ViTs, etc.) and LLMs
- *Gram-Based reconstruction*
  - Collect Gram Matrices ***G*** through a few forward passes without gradients or labels and calculate Reconstruction matrix ***B***:
  - Merge Reconstruction matrix ***B*** to consumer weights

Results

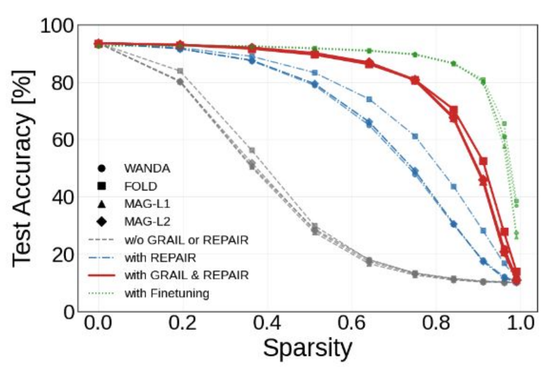
- Performance Improvement
  - At **65% sparsity** using L1-magnitude pruning, accuracy of ResNet-18 collapses to 17.6%, whereas GRAIL restores it to 84.8%, a remarkable **67.2%** improvement
- Data and resource efficiency
  - As few as **128 sequences** are sufficient for LLaMA-2-7B, and approximately **100 images** suffice for ResNet-18 to recover most of the lost accuracy.
  - For LLaMA-2-7B, the memory footprint is only **3 GB**.

Method	10%	20%	30%	40%	50%	60%	70%
ZipLM	5.84	6.36	7.64	nan	12.65	21.11	45.74
Wanda	6.18	7.45	9.18	15.16	171.29	272.47	1839.20
Wanda + GRAIL	**5.75**	6.44	7.45	9.98	18.85	39.59	408.68
Wanda++	5.80	6.56	7.59	10.18	23.29	44.00	128.00
Wanda++ + GRAIL	**5.75**	6.45	7.44	10.44	16.14	35.17	288.79
SlimGPT	7.69	9.81	nan	62.56	590.75	1220.71	16764.33
SlimGPT + GRAIL	5.81	**6.32**	**7.34**	**9.71**	16.30	23.45	**43.14**
FLAP	6.01	7.16	8.85	11.49	16.67	31.80	490.85
FLAP + GRAIL	5.88	6.80	8.08	10.18	**13.45**	**20.46**	71.63

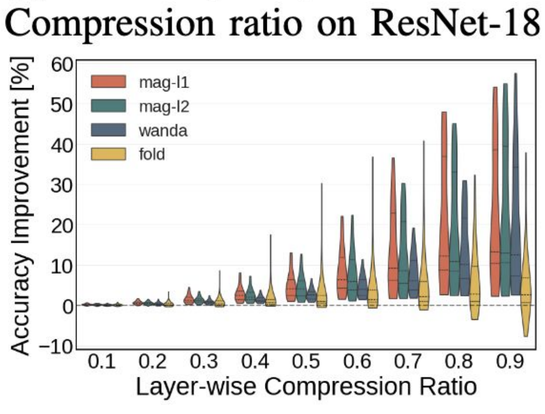
Perplexity (↓) on WikiText-2 for LLaMA-2-7B under differer sparsity levels, evaluated with sequence length 2048 and 128 calibration samples.



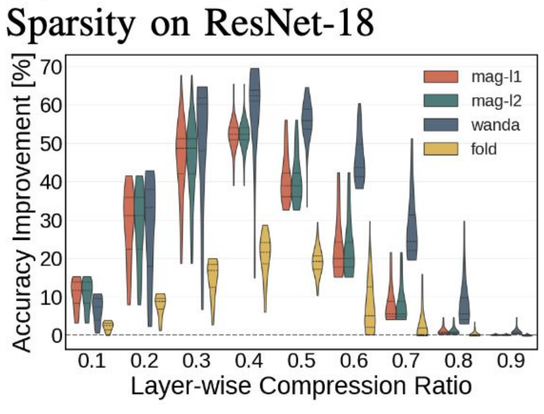
(a) Accuracy improvement vs. Compression ratio on ResNet-18



(b) Accuracy comparison vs. Sparsity on ResNet-18



(c) Accuracy improvement vs. Compression ratio on ViT



(d) Accuracy improvement vs. Compression ratio on CLIP

ANT
Funded by the European Union
ELSS