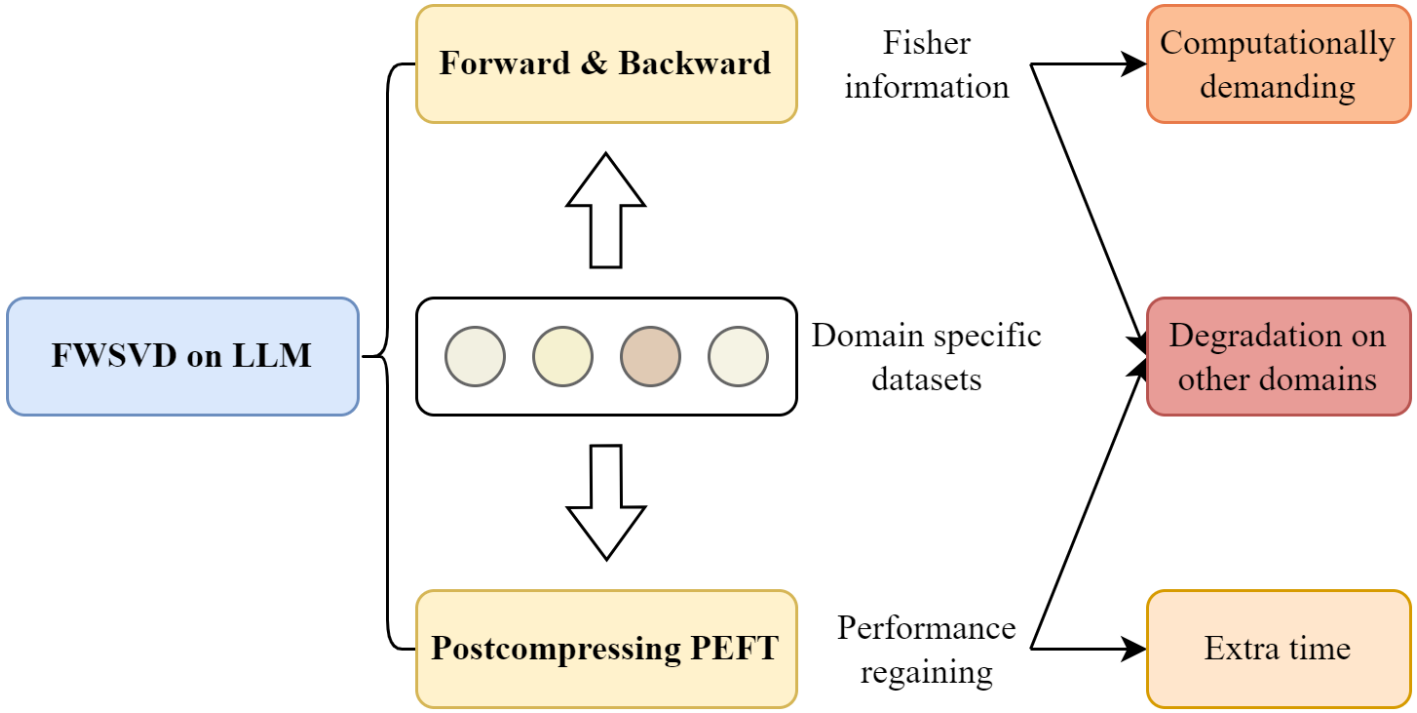
Motivation

Compression based on SVD

- FWSVD
  - Preliminary study: competitive with structure pruning on language modeling
  - Compatible with quantization

- To be improved:
  - Computationally demanding
  - Need peft for calibration



Language model compression with weighted low-rank factorization (ICLR 2022)
Numerical Optimizations for Weighted Low-rank Estimation on Language Model (ACL 2022)

Two Direction

LoRA Weighted SVD

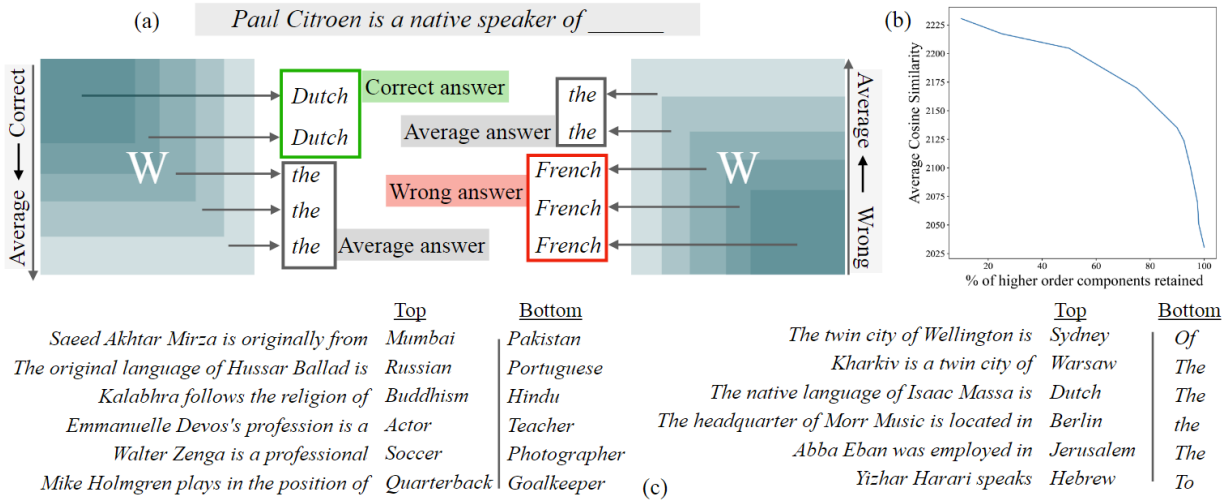
- Degradation on other domains
  - Retuning-free base model with competitive LM ability
- Memory footprint
  - Utilize LoRA to avoid computing full params' gradient, saving 25% GPU mem
  - QLoRA is also compatible
- Further improvement
  - Layer-wise compressing strategy
  - Kernel for efficient SVDLinear OP

Two Direction

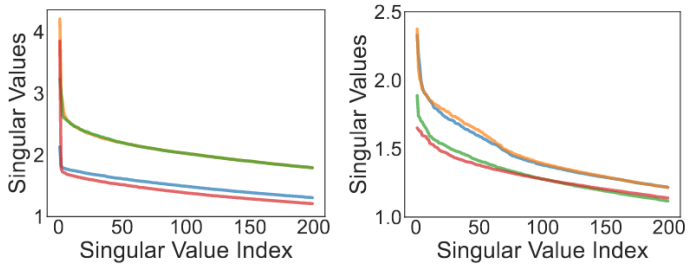
Theoretical Research

- Current trend: low-rank + quant/sparse
  - LQ-LoRA: low-bit weights + LoRA
  - LoSparse: SP weight matrix + LoRA
  - **La**yer-**Se**lective **R**ank Reduction: components with lower singular values may introduce “noise”

- Topic:
  - Low order/coherent parts的作用



values. Therefore, the coherent parts of neurons can be well approximated by the low-rank matrix computed by singular value thresholding.

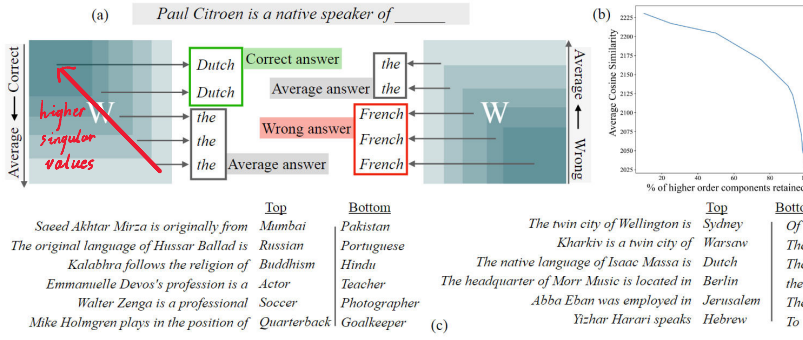


(a) BART-large (b) DeBERTaV3-large

Figure 3. Singular values in language models. (a) Singular values of weight matrices of the 10th decoder layer in BART-large; (b) Singular values of weight matrices of the 14th encoder layer in DeBERTaV3-large.

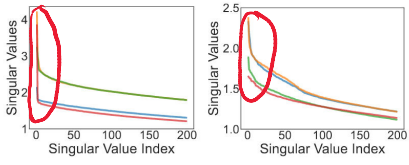
LQ-LORA (MIT)
LoSparse (ICML 2023)
Layer-Selective Rank Reduction (MIT)

日期: /



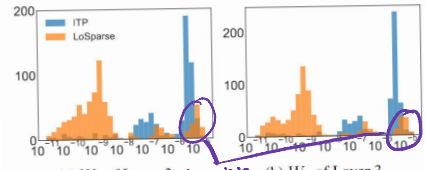
LASER indicates that { ① 头部奇异值已经可以做出很好建模
② 拖尾部分实际会给出一些相似词 (noise)

values. Therefore, the coherent parts of neurons can be well approximated by the low-rank matrix computed by singular value thresholding.



(a) BART-large (b) DeBERTaV3-large

Figure 3. Singular values in language models. (a) Singular values of weight matrices of the 10th decoder layer in BART-large; (b) Singular values of weight matrices of the 14th encoder layer in DeBERTaV3-large.



(a) W_q of Layer 3 SP时保留 (b) W_k of Layer 3

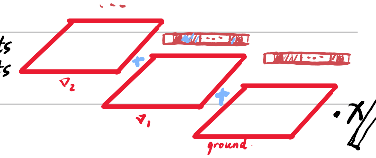
Figure 4. Neuron importance scores of selected linear projections when compressing DeBERTaV3-base on SST-2 with ITP (blue) and LoSparse (orange). It shows LoSparse successfully separates incoherent parts of neurons and make it easy to prune the non-expressive components.

LoSparse: 头部奇异值(coherent parts)使用SVD保留,拖尾部分用结构化剪枝(SP)

Topic: coherent part 在 LM 中有什么作用

Hypo: 形成 global pattern, 为 output feature 定一个 ground, 由拖尾部分进行微调

Prove: compare activate feature map { ① coherent parts
② coherent parts + tail
分布是否有明显区别



Further explore: { SVD为什么不work → 没考虑拖尾
FwSVD, SP为什么要PEFT后才work → LoRA与coherent part的关系