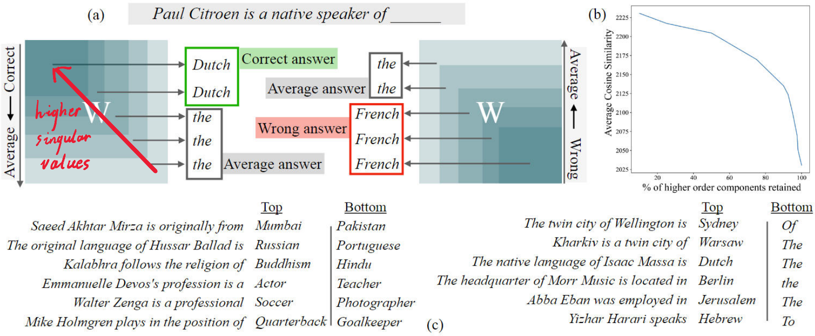
日期:　　/

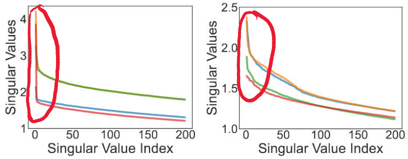


	Top	Bottom
Saeed Akhtar Mirza is originally from	Mumbai	Pakistan
The original language of Hussar Ballad is	Russian	Portuguese
Kalabhra follows the religion of	Buddhism	Hindu
Emmanuelle Devos's profession is a	Actor	Teacher
Walter Zenga is a professional	Soccer	Photographer
Mike Holmgren plays in the position of	Quarterback	Goalkeeper

	Top	Bottom
The twin city of Wellington is	Sydney	Of
Kharkiv is a twin city of	Warsaw	The
The native language of Isaac Massa is	Dutch	The
The headquarter of Morr Music is located in	Berlin	the
Abba Eban was employed in	Jerusalem	The
Yizhar Harari speaks	Hebrew	To

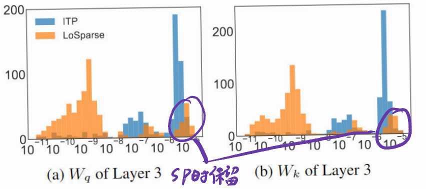
LASER indicates that ①头部奇异值已经可以做出语言建模
②拖尾部分实际会给出一些相似词(noise)

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　　(b) W_k of Layer 3

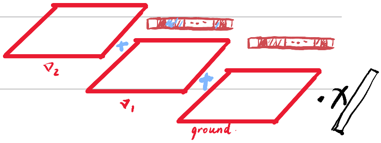
Figure 4. Neuron importance scores of selected linear projections when compressing DeBERTaV3-base on SST-2 with ITP (blue) and LoSparse (orange). It shows LoSparse successfully separates incoherent parts of neurons and make it easy to prune the non-expressive components.

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 分布是否有明显区别 ①coherent parts
②coherent parts + tail



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