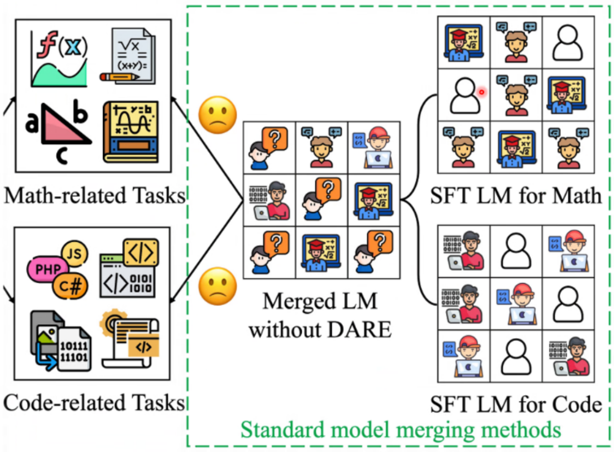
2023.12.12 赵君

主要内容

- 从专才到通才的转换；
- 更好控制模型的行为，消除bias和toxic；
- 提高参数效率；

专才与通才

Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch



$$\boldsymbol{\theta}_{\mathrm{M}} = \boldsymbol{\theta}_{\mathrm{PRE}} + \lambda \cdot \sum_{k=1}^{K} \boldsymbol{\delta}^{t_k} = \boldsymbol{\theta}_{\mathrm{PRE}} + \lambda \cdot \sum_{k=1}^{K} (\boldsymbol{\theta}_{\mathrm{SFT}}^{t_k} - \boldsymbol{\theta}_{\mathrm{PRE}}),$$

- delta为SFT后的模型与预训练模型的增量；
- 简单将参数相加，直观；
- 问题：参数冲突；math的模型和code的模型SFT参数在一些地方有冲突，直接相加会造成能力损失；
- 解决：随机丢弃一些参数的增量delta，可以发现随着LM的size增加，其对于任务的完成依然很好（math上70B甚至99%）；这说明SFT更新的大部分参数都是非常冗余的；
- （或许不应该随机drop，应该根据参数更新的幅值来drop，更新大的或许才是重要参数）
  - The delta parameters of both encoder- and decoder-based LMs are highly redundant.
  - The tolerance of drop rates increases with the sizes of LMs

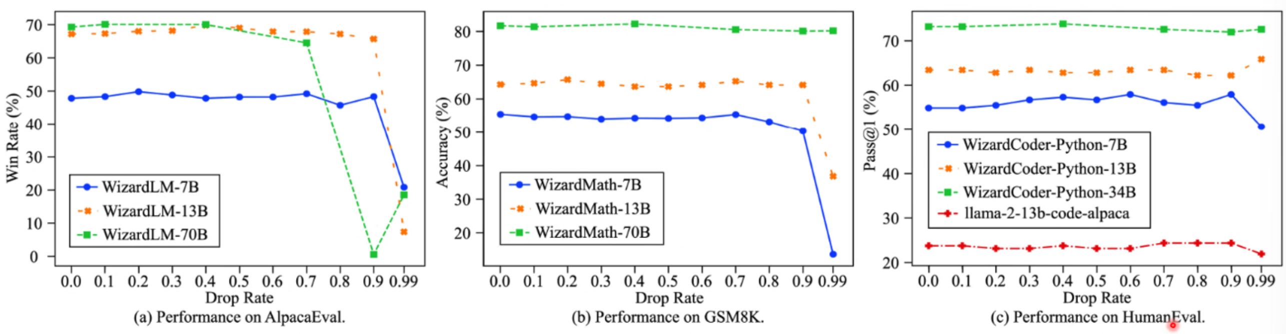


Figure 3: Performance of various decoder-based LMs on AlpacaEval, GSM8K, and HumanEval.

- 方案：通过丢弃大部分delta参数来解决参数冲突；

Table 1: Results of merging decoder-based LMs by Task Arithmetic, where LM, Math, and Code are the abbreviations of WizardLM-13B, WizardMath-13B, and llama-2-13b-code-alpaca. We use blue, green, and red colors to distinguish each single model and utilize mixed colors to denote the merged models. The best and second-best results among the single model, the merged models with and without DARE are marked in **bold** and underlined fonts.

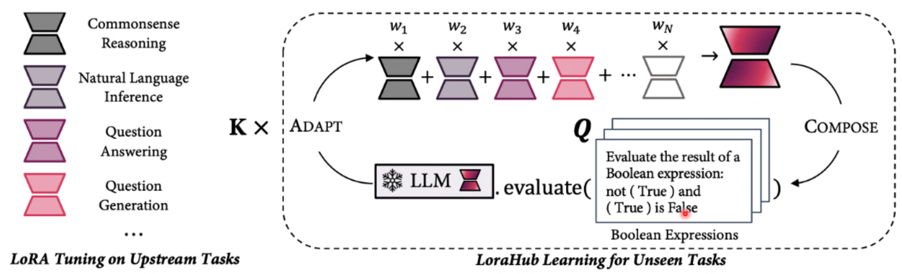
Merging Methods	Models	Preprocess	Instruction-Following	Mathematical Reasoning		Code-Generating	
			AlpacaEval	GSM8K	MATH	HumanEval	MBPP
Single Model	LM	/	67.20	2.20	0.04	36.59	34.00
	Math	/	/	64.22	14.02	/	/
	Code	/	/	/	/	23.78	27.60
Task Arithmetic	LM & Math	No	67.04	**66.34**	13.40	28.66	30.60
		w/ DARE	**67.45**	66.26	12.86	26.83	32.40
	LM & Code	No	**68.07**	/	/	31.70	32.40
		w/ DARE	67.83	/	/	**35.98**	**33.00**
	Math & Code	No	/	64.67	13.98	8.54	8.60
		w/ DARE	/	**65.05**	13.96	10.37	9.80
	LM & Math & Code	No	69.03	58.45	9.88	18.29	29.80
		w/ DARE	**69.28**	56.48	10.16	23.17	**31.60**

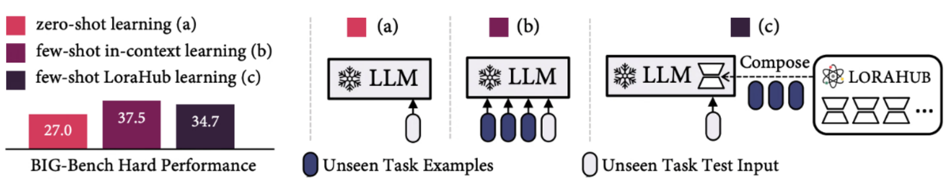
实验来看，在2、3个任务上面参数融合是有效的；但是在更多的任务融合性能会较大下降；

任务参数delta1和delta2的相加，和多个任务同时SFT（或者持续学习）得到的delta-multi是否相同？若任务相互独立，则可以实现任务的解耦合；

LORAHUB:EFFICIENT CROSS-TASK GENERALIZATION VIA DYNAMIC LORA COMPOSITION

把多个任务得到的Lora模块统一存为一个hub；
基于具体任务对特定的lora模型进行组合，得到最终的权重；

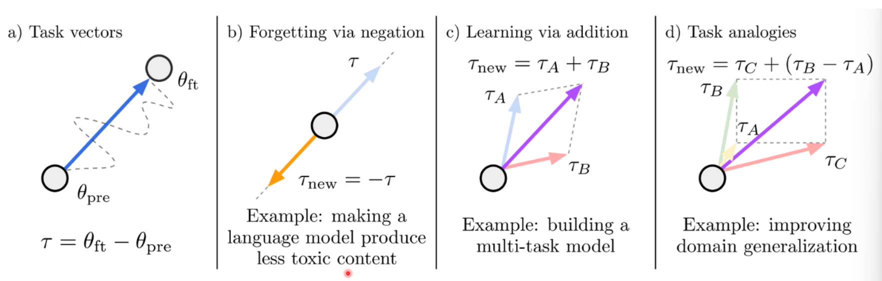




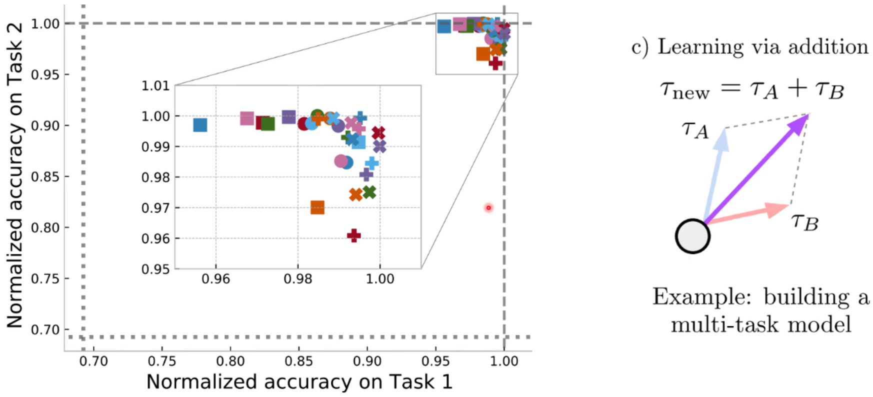
缺点：效果有限；

消除bias

EDITING MODELS WITH TASK ARITHMETIC



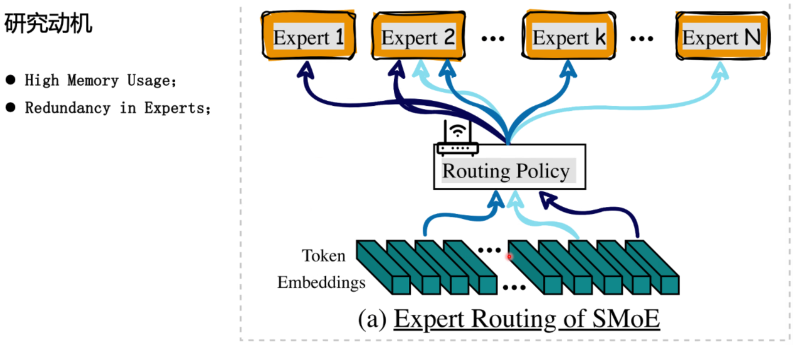
两个任务相加，基本上能力都能获得；
专门构造有害模型，通过取反增量可以实现消除别的模型的bias；



但是向量之间是有冲突的，论文里面没有体现；

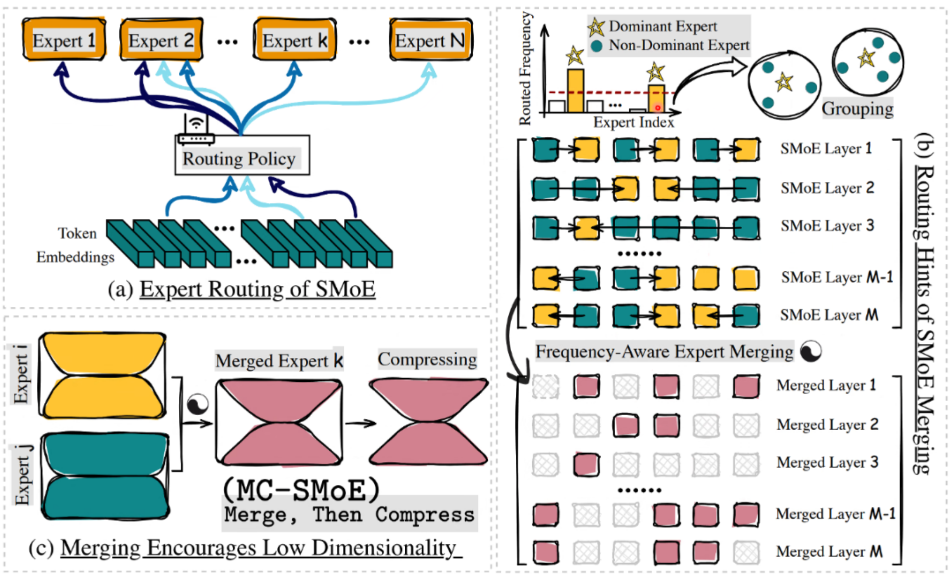
提高参数效率

MOE



表示坍缩：可能路由策略喜欢把embedding全传入到少数几个expert中；
通过模型融合的方式，把不常被使用的专家模型融合在一起，提高参数效率和减小存储占用；

- 根据频率作为权重，将参数直接相加；



Merge之后发现矩阵的秩下降了（原因未说明），因此矩阵可以被进一步分解减小；
相比于剪枝的方案，该方法在大多数任务上得到了最好的结果；compress之后计算量和size进一步减小；

Table 2: Performance evaluations on the *switch-base-32* model with 32 experts in each SMoE layer, as well as its comparative dense model *t5-base*. We found the first SMoE layer has a profound impact on the model's performance, and merging it results in more significant performance degradation compared to other layers. Thus for all merging/compression mechanisms, the first SMoE layer is skipped following Ma et al. (2023), and it maintains an average of 8 experts in other SMoE layers. We report *exact-match/F1-score* for SQuAD and HotpotQA, *F1-score* for MultiRC, and *accuracy* for other tasks. For each task, we highlight the best performance over all baselines in blue, and mark the performance no worse than full SMoE in **bold**.

Methods	Model Size	TFLOPs	SST-2	MRPC	MultiRC	COPA	WinoGrande	SQuAD	WikiQA	HotpotQA
Dense	220M	4.65	94.61	88.97	74.25	58.00	58.72	63.65/83.76	96.12	66.13/83.45
Full SMoE	2.0B	4.65	95.75	90.20	76.19	68.00	61.80	65.39/85.81	96.45	67.55/84.60
Pruning	733M	4.65	94.50	88.97	75.13	63.00	61.64	64.80/85.13	96.27	67.39/84.56
Task-Specific	733M	4.65	91.28	82.04	53.63	52.00	58.56	54.40/78.00	95.24	64.70/82.76
Averaging	733M	4.65	92.66	88.73	74.04	62.00	59.59	64.49/84.75	96.19	67.36/84.61
ZipIt	733M	4.65	93.12	**91.18**	75.26	65.00	60.38	65.01/85.06	96.05	**67.59/84.70**
REPAIR	733M	4.65	92.89	**90.44**	74.44	65.00	61.48	64.67/84.84	96.27	**67.67/84.77**
Git Re-basin	733M	4.65	93.35	88.24	74.25	65.00	59.25	64.61/84.92	96.23	67.29/84.46
M-SMoE	733M	4.65	94.50	**90.69**	75.57	**68.00**	**61.80**	**65.66**/85.49	96.34	**67.91/84.83**
MC-SMoE	381M	3.83	93.35	89.22	73.98	67.00	59.52	**65.41**/85.30	96.08	**67.64/84.77**

从结果来看，秩的下降会给结果带来一定损失但是不大；