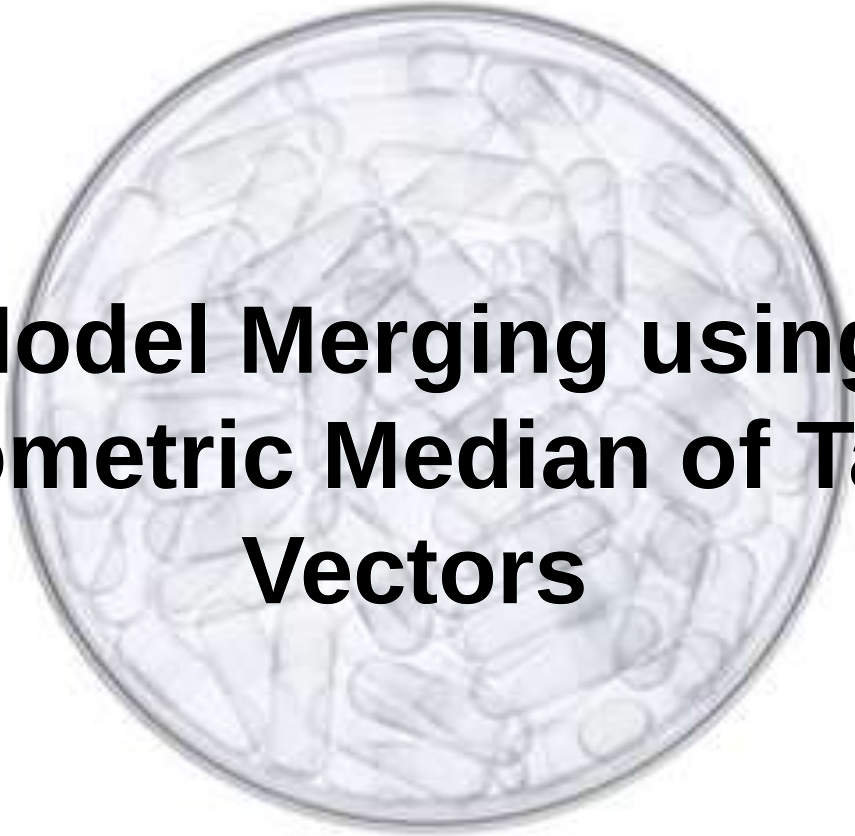
LLM MERGING



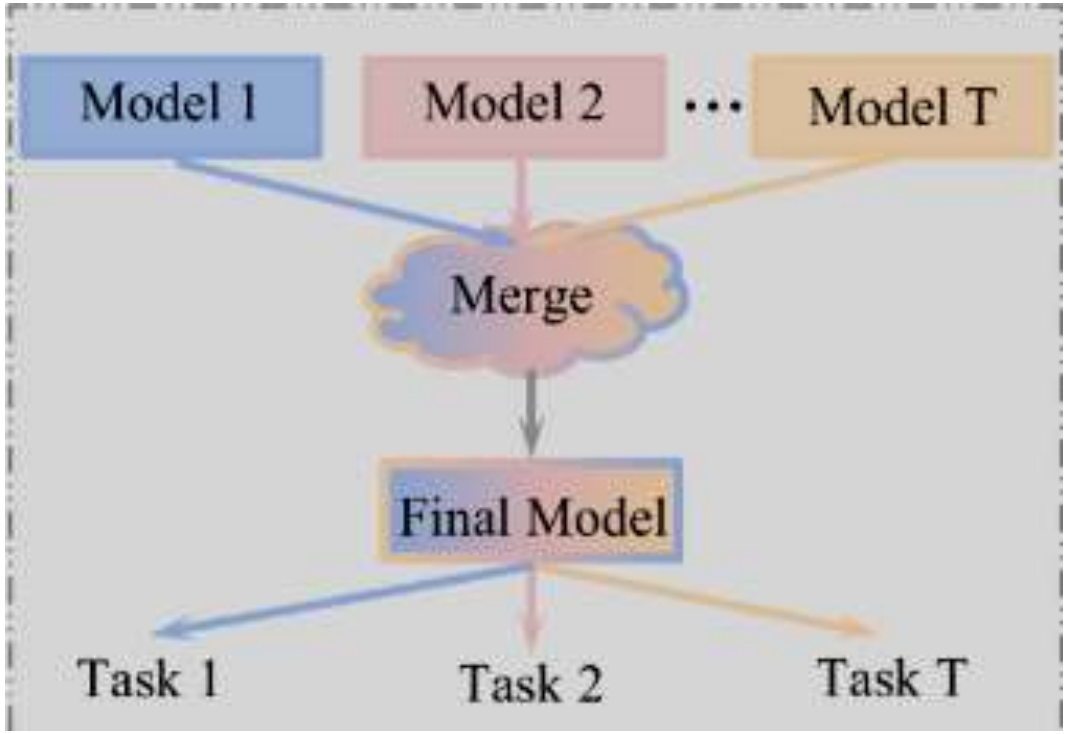
Model Merging using Geometric Median of Task Vectors

AAKASH GUPTA

SIDDARTH GUPTA

MODEL MERGING

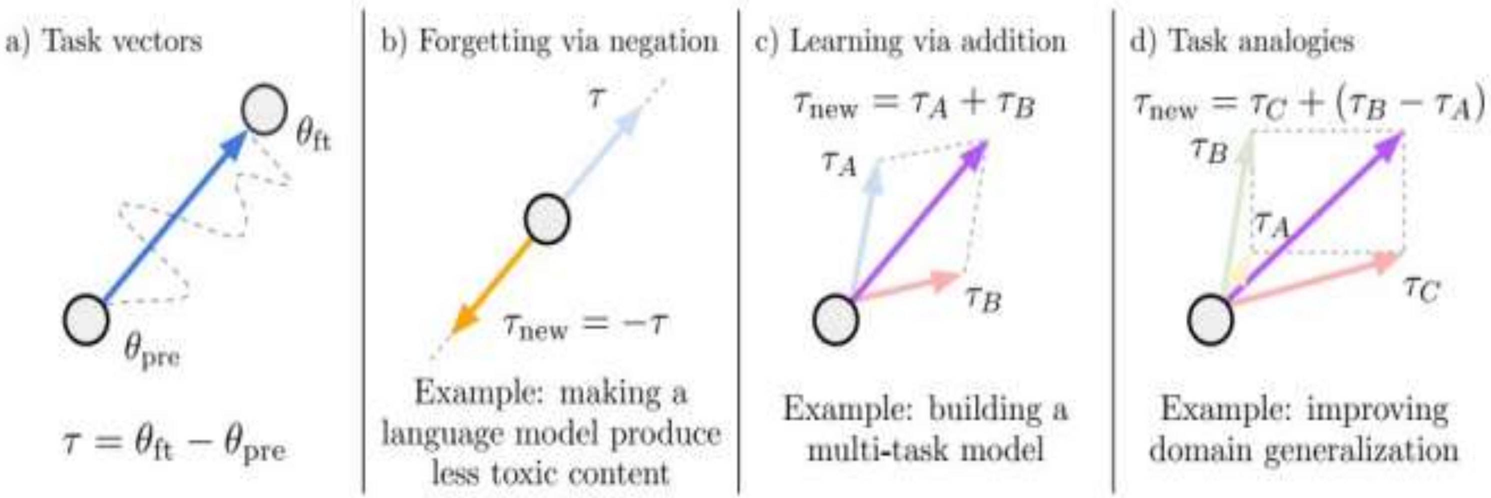
- Model merging or model fusion, combines the parameters of multiple models with unique strengths into a single, unified model.
- Unlike ensemble methods, which require high memory and processing power, model merging consolidates knowledge into one streamlined model, reducing computational costs, memory usage, and latency.
- This efficient technique enhances generalization across tasks and is ideal for resource-constrained or low-latency environments, as it does not require access to the original training data or extensive computation and training.



TASK VECTOR

The vector encodes the information related to the task that the fine-tuned model learned. For example, if the fine-tuned model learns to summarize text, the task vector would represent "text summarization ability."

$$\tau_t = \theta_t^{\mathrm{ft}} - \theta_{\mathbf{pre}}.$$

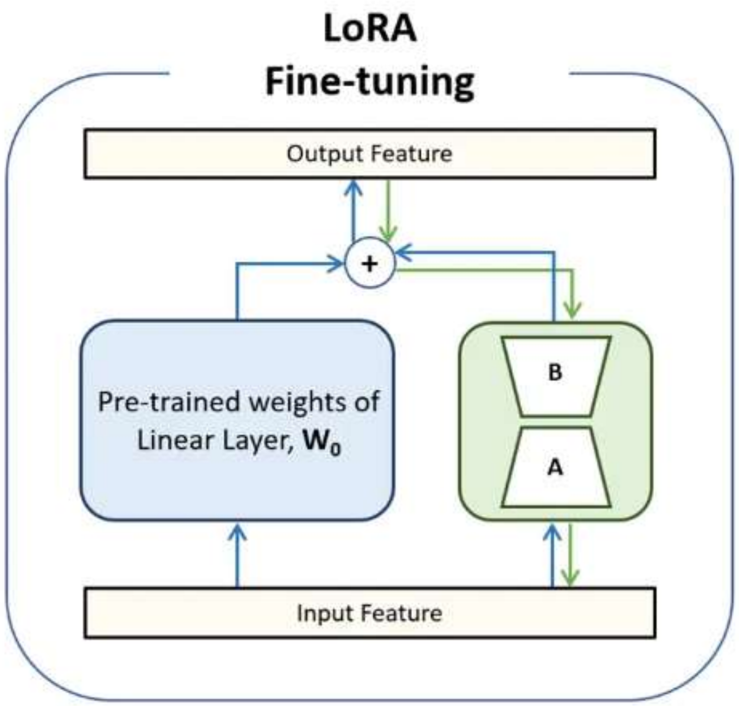


LORA , PEFT

LORA/PEFT reduces the number of parameters that need to be updated during fine-tuning, making the process of finetuning faster and more memory-efficient, while maintaining performance on downstream tasks.

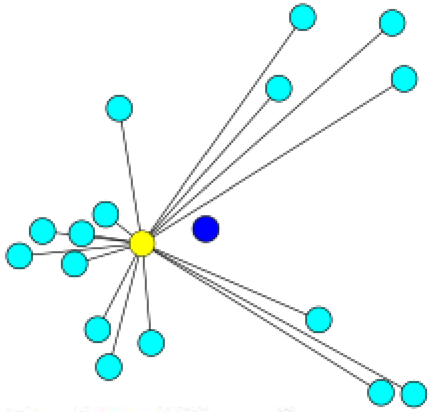
$$W' = W + \Delta W$$

$$\Delta W = BA$$



GEOMETRIC MEDIAN

The geometric median is a point in multidimensional space that minimizes the sum of distances to a set of given points.



Mathematically, given a set of points $X = \{x_1, x_2, \ldots, x_n\}$ in Euclidean space $\mathbb{R}^d$, the geometric median M is the point that minimizes the sum of Euclidean distances to the given points:

$$M = \arg\min_{y \in \mathbb{R}^d} \sum_{i=1}^{n} \|y - x_i\| \tag{1}$$

where:

- M is the geometric median (the point to be found).
- $\|y - x_i\|$ denotes the Euclidean distance between point y and x_i.

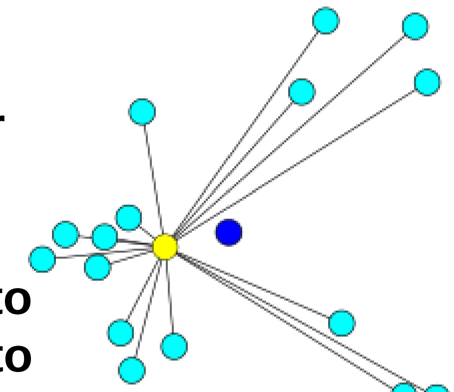
METHOD

- In our approach, the fine-tuned LLM's we used had 23 encoder blocks and 23 decoder blocks, and each encoder or decoder block has LoRA $A \in \mathbb{R}^{16\times 2048}$ and $B \in \mathbb{R}^{2048\times 16}$

- ΔW = BA is the corresponding task vector of a given fine-tuned LLM for a given encoder/decoder block (parameter matrix of the $W \in \mathbb{R}^{2048\times 2048}$

- **We flatten all these task vector matrices to a high dimensional vector $\mathbb{R}^{1\times 4194304}$ and then find the geometric median of all these flattened vectors treating each of these vectors as a point in multidimensional space to get a "net task vector" for that block which is finally added to the corresponding block in the pretrained base LLM after reshaping to the original size $\mathbb{R}^{2048\times 2048}$ of block's parameter .**

- We are finding the geometric median , so that our model is able to optimally perform in all the tasks.



WEISZFELD ITERATIVE ALGORITHM

The Weiszfeld algorithm is an iterative method used to compute the geometric median of a set of points in multi-dimensional space.

Algorithm 1 Weiszfeld Algorithm for Minimizing $f(x) = \sum_{i=1}^{n} w_i \|x - p_i\|$

Input: Points $p_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, with optional weights $w_i > 0$. Initial guess $x^{(0)} \in \mathbb{R}^d$.

```
k ← 0
repeat
    if x^(k) ≠ p_i for all i ∈ {1, ..., n} then
        Compute the weighted average:
```

$$x^{(k+1)} \leftarrow \frac{\sum_{i=1}^{n} \frac{w_i}{\|x^{(k)} - p_i\|} p_i}{\sum_{i=1}^{n} \frac{w_i}{\|x^{(k)} - p_i\|}}$$

```
    else
        Apply a small perturbation to avoid division by zero.
    end if
    k ← k + 1
until convergence (i.e., ||x^(k+1) − x^(k)|| < ϵ for a small ϵ)
```

Output: Geometric median $x^{(k+1)}$.

WEISZFELD ITERATIVE ALGORITHM

$$\min_x \left\{ f(x) = \sum_{i=1}^{m} \omega_i \|x - a_i\| \right\}$$

Let x^* be the optimal solution of problem (FW). If $x^* \notin A$, then

$$\nabla f(x^*) = \sum_{i=1}^{m} \omega_i \frac{x^* - a_i}{\|x^* - a_i\|} = 0.$$

The optimal solution can be expressed as:

$$x^* = \frac{\sum_{i=1}^{m} \omega_i a_i / \|x^* - a_i\|}{\sum_{i=1}^{m} \omega_i / \|x^* - a_i\|},$$

or in operator form:

$$x^* = T(x^*),$$

where the operator $T : \mathbb{R}^d \setminus A \to \mathbb{R}^d$ is defined by:

$$T(x) := \frac{\sum_{i=1}^{m} \omega_i a_i / \|x - a_i\|}{\sum_{i=1}^{m} \omega_i / \|x - a_i\|}.$$

RESULTS

Merge Method	Rouge1	Rouge2	RougeL	RougeLsum	BLEU
GeoMed (20)	**0.365169018**	**0.168960446**	**0.311908921**	**0.321912100**	**0.085560895**
WeightAvg (20)	0.363765969	0.167650662	0.310080274	0.309277511	0.083514798
GeoMed (15)	**0.363746168**	**0.168716755**	**0.310760179**	**0.311333478**	**0.083514798**
WeightAvg (15)	0.3647378596	0.167141560	0.310224524	0.310449948	0.083514798
GeoMed (6)	**0.362287888**	**0.166269069**	**0.319156270**	**0.319113298**	**0.083994781**
WeightAvg (6)	0.361478874	0.169796169	0.310843997	0.311794364	0.082005513
GeoMed (2)	**0.355095178**	**0.164083656**	**0.317939794**	**0.299276737**	**0.089607033**
WeightAvg (2)	0.353798629	0.164048906	0.316438994	0.298291665	0.079670733

The table compares the performance of merging various number of models using two methods: GeoMed, which computes the geometric median of task vectors, and WeightAvg, a baseline method that computes the average of all task vectors.

No.	Finetuned Model
1	lorahub/flan_t5_xl-dbpedia_14_given_list_what_category_does_the_paragraph_belong_to
2	lorahub/flan_t5_xl-wiki_qa_Topic_Prediction_Question_Only
3	lorahub/flan_t5_xl-anli_r2
4	lorahub/flan_t5_xl- web_questions_question_answer
5	lorahub/flan_t5_xl-duorc_SelfRC_question_answering
6	lorahub/flan_t5_xl-adversarial_qa_dbert_question_context_answer
7	lorahub/flan_t5_xl-wiki_qa_Is_This_True_
8	lorahub/flan_t5_xl-gem_e2e_nlg
9	lorahub/flan_t5_xl-wiki_hop_original_explain_relation
10	lorahub/flan_t5_xl-duorc_SelfRC_title_generation
11	lorahub/flan_t5_xl-glue_mrpc
12	lorahub/flan_t5_xl-glue_cola
13	lorahub/flan_t5_xl-wiki_bio_comprehension
14	lorahub/flan_t5_xl-wiki_bio_key_content
15	lorahub/flan_t5_xl-wiki_bio_guess_person
16	lorahub/flan_t5_xl-wiki_bio_who
17	lorahub/flan_t5_xl-wiki_qa_found_on_google
18	lorahub/flan_t5_xl-gem_web_nlg_en
19	lorahub/flan_t5_xl-duorc_ParaphraseRC_extract_answer
20	lorahub/flan_t5_xl-duorc_SelfRC_extract_answer
21	lorahub/flan_t5_xl-wiqa_what_might_be_the_last_step_of_the_process

Table 2: List of Finetuned Models Used to obtain results of table 1

Number of models	Base Model	Finetuned Models
20	google/flan-t5-xl	(1)(2)(3)(4)(5)(6)(8)(9)(10)(11)(12)(13)(14)(15)(16)(17)(18)(19)(20)(21)
15	google/flan-t5-xl	(1)(2)(3)(4)(5)(6)(8)(9)(10)(13)(14)(18)(19)(20)(21)
6	google/flan-t5-xl	(8)(9)(10)(18)(19)(20)
2	google/flan-t5-xl	(7)(8)

Table 3: As shown in Table 2, several finetuned models were used for various tasks such as Text Classification, Question Answering, and Sentence Similarity etc.

THANKS