Theoretical Perspectives of Efficient Learning for Large Foundation Models

Presenter: Hongkang Li

Development of deep learning

Efficient learning methods are needed for increasing model sizes.

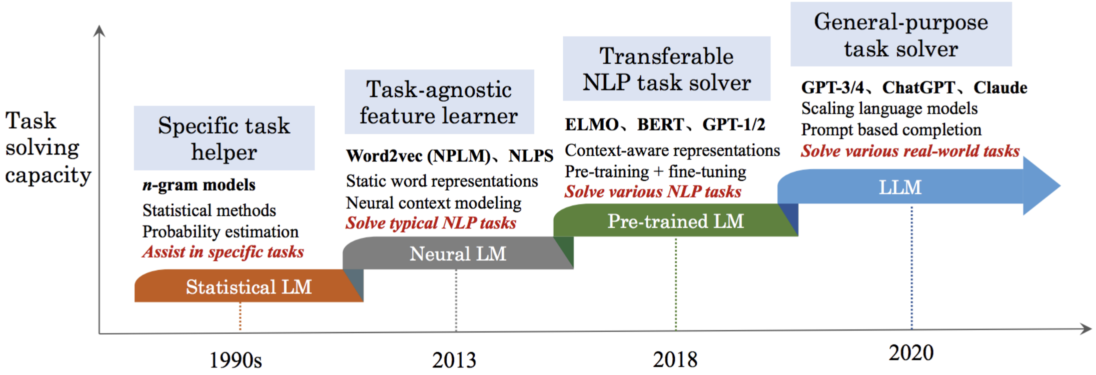


Figure 1: Deep Learning paradigm[1]

[1] source from [Zhao et al.23]

Efficient learning

Fine-tuning cost

- Model size: Hundreds of billion.
- Memory: Gradients, optimizer states, etc. Scale up with the model size.
- Time: Take many days.

Can we improve or remove fine-tuning?

Efficient learning methods

- From data: Prompt engineering, Self-supervised learning.
- From model: Pruning, Quantization, Low-rank adaptation.
- From hardware: Parallelism and scheduling, Optimized kernels.

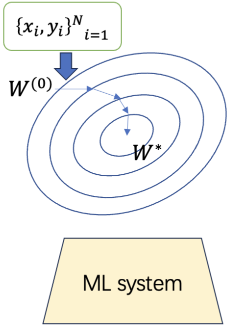


Figure 2: The finetuning process.

Outline

Practice $\rightarrow$ Theoretical Understanding

Optimization and Generalization analysis of models and algorithms

We introduce two works on theoretical foundations of efficient learning (No fine-tuning).

- In-Context Learning: Input prompt. ICML 2024.
- Task Vectors: editing the model weights. ICLR 2025 Oral (Top 1.8%).

How Do Nonlinear Transformers Learn and Generalize in In-Context Learning?

Hongkang Li, Songtao Lu, Xiaodong Cui, Pin-Yu Chen, Meng Wang

Accepted by International Conference on Machine Learning 2024.

Large Language Model (LLM) and In-context learning (ICL)

- In-context learning makes predictions for new tasks on pre-trained LLM without fine-tuning the model.
- It is implemented by providing a few testing examples and necessary instructions as a prompt for the testing data.

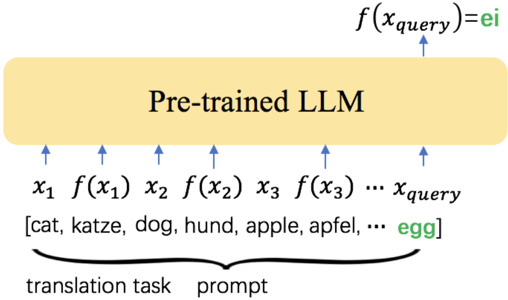


Figure 3: *Machine Translation with ICL*

Our focus

Despite the empirical success of ICL, one fundamental and theoretical question for ICL is less investigated, i.e.,

How can a Transformer be trained to perform ICL and generalize in and out of domain successfully and efficiently?

Specifically,

- What are the sufficient conditions for out-of-domain ICL?
- What is the mechanism of ICL?
- Can we prune the model in in-context inference and why?

Our work and major contributions

Summary of contributions and comparisons with related theoretical works.

Theoretical Works	Nonlinear Attention	Nonlinear MLP	Training Analysis	Distribution -Shifted Data	Tasks
[Zhang et al.24]			✓	✓	linear regression
[Huang et al.24]	✓		✓		linear regression
[Wu et al.24]			✓		linear regression
Ours	✓	✓	✓	✓	classification

Table 1: Comparison with existing works about training analysis and generalization guarantee of ICL

Problem formulation

We study binary classification problems. Given the input $\boldsymbol{x}_{query}$, we aim to predict the label $f(\boldsymbol{x}_{query})$ for the task f. We conduct training with constructed prompts $\boldsymbol{P}$ on a model to enable ICL.

Context inputs Feature embedding

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_l & \boldsymbol{x}_{query} \\ \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_l & 0 \end{pmatrix} := (\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{query}).$$

Context outputs Label embedding

Figure 4: Prompt for ICL.

$\boldsymbol{y}_i$ is an embedding of $f(\boldsymbol{x}_i)$. $\boldsymbol{y}_i = \boldsymbol{q}$ if $f(\boldsymbol{x}_i) = +1$. $\boldsymbol{y}_i = -\boldsymbol{q}$ if $f(\boldsymbol{x}_i) = -1$.

Example: Classify fruits (label $+1$) and animals (label -1).
Prompt: $\boldsymbol{x}_1 =$Apple, $\boldsymbol{y}_1 = \boldsymbol{q}$, $\boldsymbol{x}_2 =$Cat, $\boldsymbol{y}_2 = -\boldsymbol{q}$, $\boldsymbol{x}_{query} =$Orange.
Predict: $f(\boldsymbol{x}_{query}) = +1$ or -1?

Problem formulation

In-domain data ($\sim \mathcal{D}$) and tasks ($\in \mathcal{T}$):

- Given $\{\boldsymbol{\mu}_j\}_{j=1}^{M_1}$ as in-domain relevant (IDR) patterns (orthonormal), each in-domain data $\boldsymbol{x} = \boldsymbol{\mu}_j + \text{noise}$.
- Each task is defined based on one pair of $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$. $f(\boldsymbol{x}) = +1$ (or -1) if the IDR pattern of $\boldsymbol{x}$ is $\boldsymbol{\mu}_a$ (or $\boldsymbol{\mu}_b$). Otherwise $f(\boldsymbol{x})$ is a random label. $|\mathcal{T}| = M_1(M_1 - 1)$.

Out-of-domain data ($\sim \mathcal{D}'$) and tasks ($\in \mathcal{T}'$): Defined on out-of-domain relevant (ODR) patterns $\{\boldsymbol{\mu}'_j\}_{j=1}^{M'_1}$. $|\mathcal{T}'| = M'_1(M'_1 - 1)$.

Prompt construction: For the task on $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$, with a probability of $\alpha/2$, select examples of $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$. α represents the fraction of task-relevant examples in the prompt. Replace α with α' if it is a testing task.

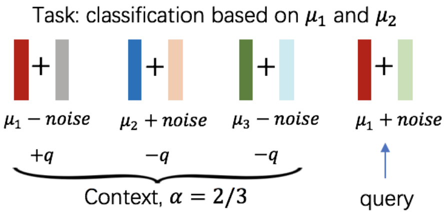


Figure 5: Example of prompt, $\alpha = 2/3$.

Problem formulation

Learner model: a single-head, one-layer Transformer with a self-attention layer and a two-layer perceptron, i.e.,

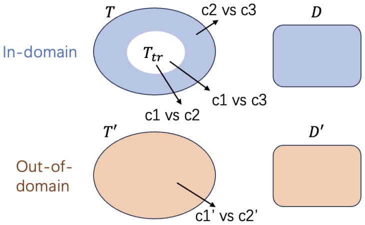
$$\begin{aligned} F(\Psi; \boldsymbol{P}) &= \boldsymbol{a}^\top \mathrm{Relu}(\boldsymbol{W}_O \sum_{i=1}^{l} \boldsymbol{W}_V \boldsymbol{p}_i \cdot \mathrm{attn}(\Psi; \boldsymbol{P}, i)), \\ \mathrm{attn}(\Psi; \boldsymbol{P}, i) &= \mathrm{softmax}_{query}((\boldsymbol{W}_K \boldsymbol{p}_i)^\top \boldsymbol{W}_Q \boldsymbol{p}_{query}) \end{aligned} \tag{1}$$

Model training: The training is to solve the empirical risk minimization using prompt and label pairs $\{\boldsymbol{P}^n, z^n\}_{n:f^n \in \mathcal{T}_{tr}}$, $\Psi = \{\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O, \boldsymbol{a}\}$, where each training task $f^n \in \mathcal{T}_{tr} \subset \mathcal{T}$,

$$\min_{\Psi} \frac{1}{|\mathcal{T}_{tr}|} \sum_{n:f^n \in \mathcal{T}_{tr}} \ell(\Psi; \boldsymbol{P}^n, z^n) = \min_{\Psi} \frac{1}{|\mathcal{T}_{tr}|} \sum_{n:f^n \in \mathcal{T}_{tr}} \max\{0, 1 - z^n \cdot F(\Psi, \boldsymbol{P}^n)\} \tag{2}$$

- The model is trained via stochastic gradient descent (SGD).
- $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V$ initialized from a small scaling of identity matrices. $\boldsymbol{W}_O$ initialized from Gaussian distribution.

Problem formulation



Generalization: We define in-domain and out-of-domain generalization.

- In-domain generalization: No distribution shift between training and testing data. **Unseen tasks but seen data.** The generalization error is defined on unseen tasks $\mathcal{T} \backslash \mathcal{T}_{tr}$ as

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query} \sim \mathcal{D}, f \in \mathcal{T} \backslash \mathcal{T}_{tr}} [\ell(\Psi; \boldsymbol{P}, z)]. \tag{3}$$

- Out-of-domain generalization: The testing queries follow $\mathcal{D}' \neq \mathcal{D}$, and the testing tasks follow $\mathcal{T}' \neq \mathcal{T}$. **Unseen tasks and OOD data.** The generalization error is defined as

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query} \sim \mathcal{D}', f \in \mathcal{T}'} [\ell(\Psi; \boldsymbol{P}, z)]. \tag{4}$$

Problem formulation

Model pruning:

- Let $\mathcal{S} \in [m]$ be the index set of $\boldsymbol{W}_O$ neurons.
- Pruning neurons in $\mathcal{S}$: removing corresponding rows of the trained $\boldsymbol{W}_O$.

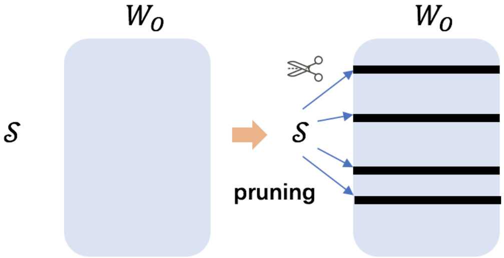


Figure 6: Pruning on $\boldsymbol{W}_O$.

Main theoretical results

Theorem 1 (In-domain generalization)

For any $\epsilon > 0$*, as long as*

1. *the training tasks* $\mathcal{T}_{tr}$ *uniformly cover all the IDR patterns and labels with* $|\mathcal{T}_{tr}| \geq M_1$*, which means training a small fraction of the total tasks* $|\mathcal{T}_{tr}|/|\mathcal{T}| \geq (M_1 - 1)^{-1/2}$ *is sufficient,*
2. *the lengths of training and testing prompts* $l_{tr} \geq \Omega(\alpha^{-1})$*,* $l_{ts} \geq \alpha'^{-1}$*,*
3. *the number of iterations* $T = \Theta(\alpha^{-2/3})$*,*

and the batch size $B \geq \Omega(\max\{\epsilon^{-2}, M_1\})$*, then with a high probability, the in-domain generalization error of the returned model is less than* $\mathcal{O}(\epsilon)$*.*

ICL mechanism by the trained transformer

Proposition 1

- $\boldsymbol{W}_Q^{(T)}$ *and* $\boldsymbol{W}_K^{(T)}$ *mainly project context inputs to the IDR or ODR pattern.*
- *After training, attention weights become concentrated on contexts that share the same IDR/ODR pattern as the query. (induction head)*

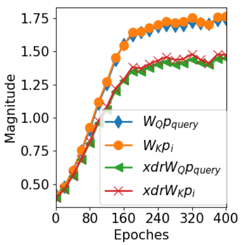


Figure 7: *The magnitude of the trained attention layer. xdr: IDR or ODR pattern of* p_{query}.

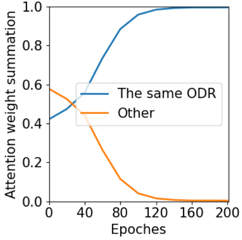


Figure 8: *The attention weight summation*

ICL mechanism by the trained transformer

Proposition 2

- *The feature embedding of rows of $\boldsymbol{W}_O^{(T)}\boldsymbol{W}_V^{(T)}$ approximate $\bar{\mu}$, i.e., the average of IDR patterns. The label embedding of rows $\boldsymbol{W}_O^{(T)}\boldsymbol{W}_V^{(T)}$ approximate $\boldsymbol{q}$ for positive neurons and $-\boldsymbol{q}$ for negative neurons.*
- *MLP neurons distinguish label embeddings instead of feature embeddings to predict labels.*

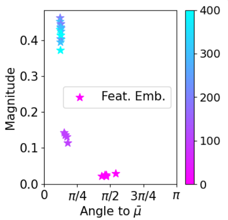


Figure 9: The feature embedding of $W_O W_V$. bar: iteration

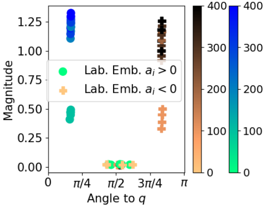


Figure 10: The label embedding of $W_O W_V$. bars: iterations

ICL mechanism by the trained transformer

Results of multi-layer Transformers (3-layer).

- Each attention layer selects contexts with the same IDR pattern as the query.

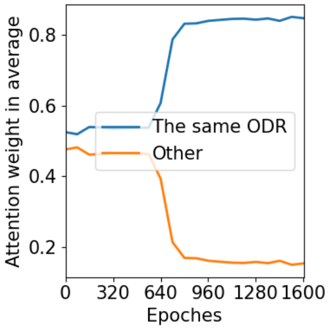


Figure 11: Layer 1 self-attention

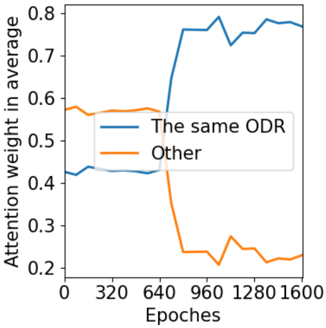


Figure 12: Layer 2 self-attention

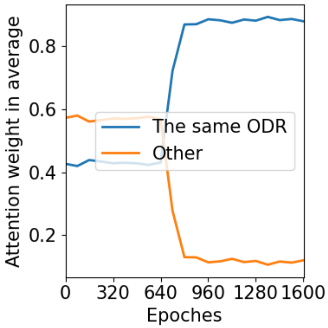


Figure 13: Layer 3 self-attention

ICL mechanism by the trained transformer

Results of multi-layer Transformers (3-layer).

- The magnitude of the majority of neurons increases along the training.
- The angle changes still hold for one of the layers.

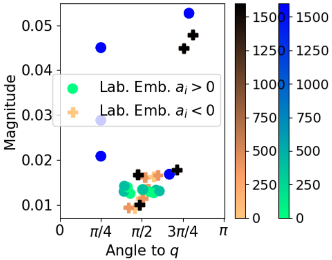


Figure 14: Layer 1 self-attention

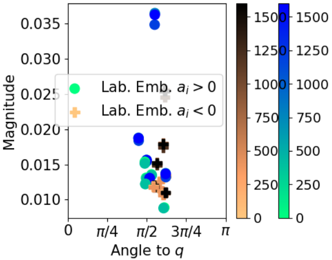


Figure 15: Layer 2 self-attention

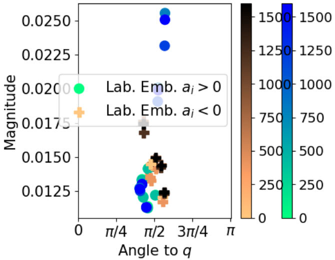
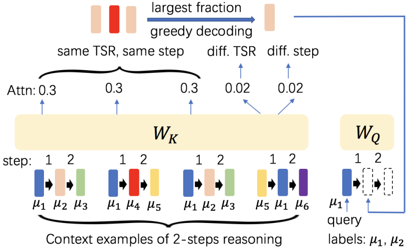


Figure 16: Layer 3 self-attention

CoT Mechanism (from our follow-up work[2])



1. When conducting the k-th step reasoning of the query, the trained model assigns dominant attention weights on the prompt columns that are also the k-th step and share the same pattern as the query.
2. Then, the fraction of the correct pattern is the largest in the output of each step to generate the accurate output by greedy decoding.

[2] Li et al., ICLR 2025. Training Nonlinear Transformers for Chain-of-Thought Inference: A Theoretical Generalization Analysis.

Main theoretical results

Consider each ODR pattern as a linear combination of IDR patterns. Denote S_1 as the summation of the linear coefficients.

Theorem 2 (Out-of-domain generalization)

Suppose that the conditions (1) to (3) in Theorem 1 hold. If a constant order of $S_1 \geq 1$ *and* $l_{ts} \geq \alpha'^{-1}$*, then with a high probability, the out-of-domain generalization error of the returned model is less than* $\mathcal{O}(\epsilon)$*.*

Training with a small amount of training tasks can lead to generalization on out-of-domain data if the ODR pattern is relatively "positively" correlated with IDR patterns and the testing prompt is long enough.

Main theoretical results

Theorem 3 (Model pruning)

- *There exists a constant fraction of MLP-layer neurons of $\boldsymbol{W}_O$ with large weights, while the remaining have small weights.*
- *Pruning all neurons with small weights leads to a generalization error $\mathcal{O}(\epsilon + M_2^{-1/2})$, which is almost the same as without pruning.*
- *Pruning an R fraction of neurons with large weights results in a generalization error greater than $\Omega(R)$.*

Three kinds of learned neurons, i.e., rows of $\boldsymbol{W}_O^{(T)}\boldsymbol{W}_V^{(T)}$:

- close to a scaling of $(\bar{\boldsymbol{\mu}}^\top, \boldsymbol{q}^\top)^\top$.
- close to a scaling of $(\bar{\boldsymbol{\mu}}^\top, -\boldsymbol{q}^\top)^\top$.
- close to initialization with small weights and diverse directions.

Numerical experiments

Verifying the sufficient conditions for out-of-domain generalization.

- $S_1 \geq 1$ is needed for a desired out-of-domain generalization.
- The required length of testing prompts decreases as α' increases.

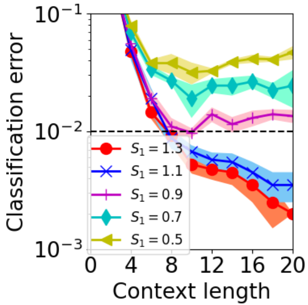


Figure 17: Out-of-domain ICL classification error on GPT-2 with different S_1

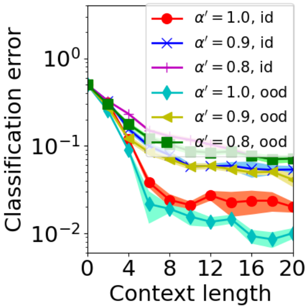


Figure 18: Out-of-domain ICL classification error on GPT-2 with different α'

Magnitude-based model pruning for out-of-domain ICL inference.

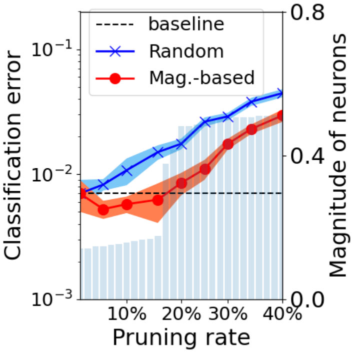


Figure 19: ***Out-of-domain classification error with model pruning of the trained W_O and the magnitude of W_O neurons.***

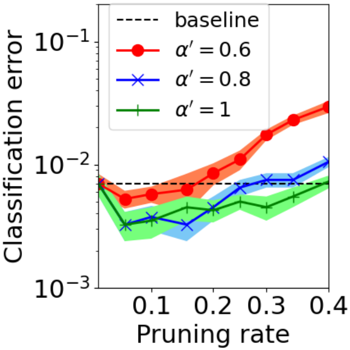


Figure 20: ***Out-of-domain classification error with different α'***

Summary

- This work provides theoretical analyses of the training dynamics of Transformers with nonlinear attention and nonlinear MLP, and the resulting ICL capability for new tasks with possible data shift.
- This work also provides a theoretical justification for magnitude-based pruning to reduce inference costs while maintaining the ICL capability.
- This work provably characterizes the mechanism of ICL implemented by a single-head, one-layer Transformer.

When is Task Vector Provably Effective for Model Editing? A Generalization Analysis of Nonlinear Transformers

Hongkang Li, Yihua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, Meng Wang

Accepted by International Conference on Learning Representations 2024.

Task Vectors and Task Arithmetic

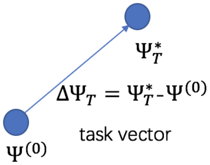


Figure 21: Task vector.

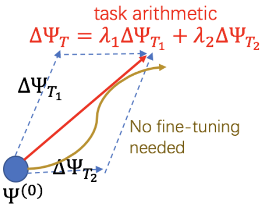


Figure 22: Task arithmetic by adding up two task vectors for inference. No fine-tuning on the two tasks are needed.

Task vector is the difference between the fine-tuned model and the pre-trained model.

$$\Delta\Psi_{\mathcal{T}} = \Psi^*_{\mathcal{T}} - \Psi^{(0)}, \tag{5}$$

where $\Psi^*_{\mathcal{T}}$ is the model fine-tuned on $(\boldsymbol{X}, y) \sim \mathcal{D}_{\mathcal{T}}$ for task $\mathcal{T}$, and $\Psi^{(0)}$ is the pre-trained model.

Task arithmetic refers to adding a linear combination of task vectors of different tasks.
Given $\Psi^{(0)}$ and a set of task vectors $\{\Delta\Psi_{\mathcal{T}_i}\}_{i\in\mathcal{V}}$,

$$\Psi = \Psi^{(0)} + \sum_{i\in\mathcal{V}} \lambda_i \Delta\Psi_{\mathcal{T}_i}, \tag{6}$$

for the inference on the downstream task.

Task Vectors and Task Arithmetic

Applications: multi-task learning, unlearning, and out-of-domain generalization in vision and language generation tasks.

Advantage: No need of fine-tuning for new tasks.

- Linear coefficient selection: Simple averaging [Ilharco et al.22, Wortsman et al.2022], Fisher-weighted averaging [Metena & Raffel, 2022] for multi-task learning; negation for unlearning [Ilharco et al.22].
- Task vector construction: sparsification [Yadav et al.2023, Yu et al.24], linearization [Ortiz-Jimenez et al.23].

Experiments on Colored-MNIST dataset:

- Classify the parity of digits.
- Control the fraction of red/green digit colors for different task correlations/distributions.

	"Irrelevant" Tasks		"Aligned" Tasks		"Contradictory" Tasks	
	Multi-Task	Unlearning	Multi-Task	Unlearning	Multi-Task	Unlearning
Best λ	1.4	-0.6	0.2	0.0	0.6	-1.0
$\mathcal{T}_1$ Acc	91.83 (-3.06)	95.02 (-0.56)	95.62 (0.00)	95.20 (-0.42)	79.54 (-16.70)	94.21 (-0.61)
$\mathcal{T}_2$ Acc	88.40 (-5.65)	50.34 (-45.24)	92.46 (-3.23)	90.51 (-5.18)	62.52 (-33.72)	4.97 (-89.85)

Figure 23: *Test accuracy (%) of $\Psi = \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \lambda\Delta\Psi_{\mathcal{T}_2}$ on task $\mathcal{T}_1$ and $\mathcal{T}_2$. Different task correlations $\Rightarrow$ Different arithmetic coefficients.*

	Fine-Tuning	$\Psi^*_{\mathcal{T}_1}$	$\Psi^*_{\mathcal{T}_2}$	Searching λ_1, λ_2 in $[-2, 3]$
(λ_1, λ_2)	N/A	$(1, 0)$	$(0, 1)$	$(1.2, -0.6)$
$\mathcal{T}'$ Acc	92.21	88.10	45.06	**91.74**

Figure 24: *Test $\Psi = \Psi^{(0)} + \lambda_1\Delta\Psi_{\mathcal{T}_1} + \lambda_2\Delta\Psi_{\mathcal{T}_1}$ on task $\mathcal{T}'$. $\mathcal{T}'$ shares a different distribution from $\mathcal{T}_1$ or $\mathcal{T}_2$. The optimal λ_1 and λ_2 generates a model that outperforms any separately trained model $\Psi^*_{\mathcal{T}_1}$ and $\Psi^*_{\mathcal{T}_2}$. $\mathcal{T}'$ and $\mathcal{T}_1$ are positively correlated; $\mathcal{T}'$ and $\mathcal{T}_2$ are negatively correlated.*

Problems to Solve

Q1: Can we provide generalization guarantees for task arithmetic?

Q2: How does task correlation quantitatively affect the performance of task arithmetic?

Q3: Why do the arithmetic operations of task vectors perform well for out-of-domain generalization?

Related Theoretical Works

- Some works [Ginart et al.2019, Guo et al.2020, Neel et al.2021, Mu & Klabjan, 2024] theoretically analyze the performance of machine unlearning from an optimization perspective.
- [Izmailov et al.2018, Frankle et al.2020] propose linear mode connectivity, concluding the existence of a small-loss connected region in the loss landscape.
- [Ortiz-Jimenez et al.23] study task arithmetic in model editing with the Neural Tangent Kernel (NTK) framework to linearize the models.

Problem Formulation

We study binary classification tasks that map each $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_P)$ to $y \in \{+1, -1\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in [P]$.

The **learner model** is considered as a one-layer nonlinear Transformer with Ψ as the set of parameters, where $\boldsymbol{W}, \boldsymbol{V} \in \Psi$ are trainable,

$$f(\boldsymbol{X}; \Psi) = \frac{1}{P} \sum_{l=1}^{P} \boldsymbol{a}_{(l)}^\top \text{Relu}(\sum_{s=1}^{P} \boldsymbol{V}\boldsymbol{x}_s \text{softmax}_l(\boldsymbol{x}_s{}^\top \boldsymbol{W}\boldsymbol{x}_l)). \tag{7}$$

Data formulation: Let $\boldsymbol{\mu}_\mathcal{T}$ be the discriminative pattern of $\mathcal{T}$. Each token is chosen from $\{\boldsymbol{\mu}_\mathcal{T}, -\boldsymbol{\mu}_\mathcal{T}\}$ or other irrelevant patterns. If $y = 1$ ($y = -1$), the number of tokens equal to $\boldsymbol{\mu}_\mathcal{T}$ (or $-\boldsymbol{\mu}_\mathcal{T}$) is larger than that of $-\boldsymbol{\mu}_\mathcal{T}$ (or $\boldsymbol{\mu}_\mathcal{T}$).

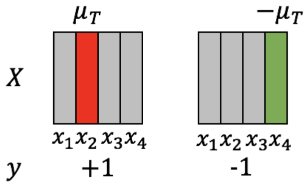


Figure 25: Data formulation.

Theoretical Results (Multi-Task learning and Unlearning)

Let $\Psi = \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \lambda\Delta\Psi_{\mathcal{T}_2}$. $\beta = \Theta(1/d)$. Loss function $\ell(\cdot)$: Hinge loss.

- Define $\alpha = \boldsymbol{\mu}_{\mathcal{T}_1}^\top \boldsymbol{\mu}_{\mathcal{T}_2}$ as the correlation between $\mathcal{T}_1$ and $\mathcal{T}_2$.
- $\alpha > 0$, < 0, or $= 0$, corresponds to "aligned", "contradictory", or"irrelevant" relationship.
- $\Psi^*_{\mathcal{T}_1}$ and $\Psi^*_{\mathcal{T}_2}$ are trained to achieve an ϵ generalization error on $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively.

Theorem 4 (Success of Multi-Task Learning on Irrelevant and Aligned Tasks)

Then, as long as $\alpha \geq 0$ and $\lambda \geq 1 - \alpha + \beta$, we have a desired multi-task learning performance with Ψ, i.e., $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_1}} \ell(\boldsymbol{X}, y; \Psi) \leq \Theta(\epsilon) + |\lambda| \cdot \beta$, and $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_2}} \ell(\boldsymbol{X}, y; \Psi) \leq \Theta(\epsilon)$.

Theorem 5 (Success of Unlearning on Irrelevant and Contradictory Tasks)

As long as $\alpha \leq 0$ and $-\Theta(\alpha^{-2}) \leq \lambda \leq 0$,we have a desired unlearning performance with Ψ, i.e., $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_1}} \ell(\boldsymbol{X}, y; \Psi) \leq \Theta(\epsilon) + |\lambda| \cdot \beta$, and $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}_2}} \ell(\boldsymbol{X}, y; \Psi) \geq \Theta(1)$.

Theoretical Results (Out-of-Domain Generalization)

Out-of-domain generalization on the task $\mathcal{T}'$, given task vectors of tasks $\{\mathcal{T}_i\}_{i\in\mathcal{V}_\Psi}$. Suppose

- all $\boldsymbol{\mu}_{\mathcal{T}_i}$ are orthogonal to each other,
- the discriminative pattern of $\mathcal{T}'$ is $\boldsymbol{\mu}_{\mathcal{T}'} = \sum_{i\in\mathcal{V}_\Psi} \gamma_i \boldsymbol{\mu}_{\mathcal{T}_i} + \kappa \cdot \boldsymbol{\mu}'_\perp$ with $\boldsymbol{\mu}'_\perp \perp \{\boldsymbol{\mu}_{\mathcal{T}_i}\}_{i\in\mathcal{V}_\Psi}$,
- not all γ_i are zero.

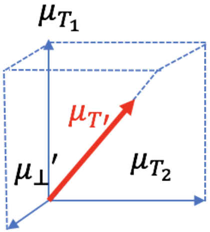


Figure 26: *An illustration of $\boldsymbol{\mu}_{\mathcal{T}'}$.*

Theorem 6 (Out-of-domain generalization using task arithmetic)

Let $\Psi = \Psi^{(0)} + \sum_{i\in\mathcal{V}_\Psi} \lambda_i \Delta\Psi_{\mathcal{T}_i}, \lambda_i \neq 0$. Then, for some $c \in (0,1)$ and all $i \in \mathcal{V}_\Psi$, and a non-empty region of λ_i, $i \in \mathcal{V}_\Psi$, where

$$\begin{cases} \sum_{i\in\mathcal{V}_\Psi} \lambda_i \gamma_i \geq 1 + c, \\ \sum_{i\in\mathcal{V}_\Psi} \lambda_i \gamma_i^2 \geq 1 + c, \\ |\lambda_i| \cdot \beta \leq c, \end{cases} \tag{8}$$

we have $\mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}_{\mathcal{T}'}} \ell(\boldsymbol{X}, y; \Psi) \leq \Theta(\epsilon)$.

Theoretical Results (Efficiency)

Recall that $\boldsymbol{W}, \boldsymbol{V} \in \Psi$. $\Delta \boldsymbol{W}_{\mathcal{T}} = \boldsymbol{W}^*_{\mathcal{T}} - \boldsymbol{W}^{(0)}$, $\Delta \boldsymbol{V}_{\mathcal{T}} = \boldsymbol{V}^*_{\mathcal{T}} - \boldsymbol{V}^{(0)}$.

Corollary 1 (Low-rank Approximation)

For any task $\mathcal{T}$ defined above, there exists rank-1 $\Delta \boldsymbol{W}_{LR}$ and $\Delta \boldsymbol{V}_{LR}$, such that

$$\|\Delta \boldsymbol{W}_{\mathcal{T}} - \Delta \boldsymbol{W}_{LR}\|_F \leq M \cdot \epsilon + \frac{1}{\log M}, \quad \textit{and} \quad \|\Delta \boldsymbol{V}_{\mathcal{T}} - \Delta \boldsymbol{V}_{LR}\|_F \leq \Theta(\epsilon), \tag{9}$$

Corollary 2 (Sparsification)

Let $\boldsymbol{u}_i$ be the i-th row of $\Delta \boldsymbol{V}_{\mathcal{T}}$. Then, for a constant fraction of $\boldsymbol{u}_i$, we have $\|\boldsymbol{u}_i\| \geq \Omega(m^{-1/2})$; for the remaining neurons, we have $\|\boldsymbol{u}_i\| \leq O(m^{-1/2}\epsilon)$ (pruning these neurons still ensures Theorems 4-6 to hold.)

Image classification on Colored-MNIST with ViT-Small/16

- Consider a merged model $\Psi = \Psi^{(0)} + \lambda_1 \Delta\Psi_{\mathcal{T}_1} + \lambda_2 \Delta\Psi_{\mathcal{T}_2}$ constructed by two task vectors for the targeted task $\mathcal{T}'$. We estimate $\gamma_1 \approx 0.792$, $\gamma_2 \approx -0.637$.
- The result justifies the sufficient conditions in Theorem 6.

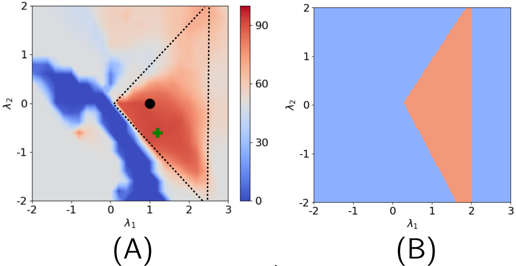


Figure 27: (A) The heatmap of the testing accuracy on $\mathcal{T}'$ using the merged model Ψ. The black dot is the baseline, while the green cross is the best λ_1, λ_2. (B) The red region satisfies (8), while the blue region does not.

Language generation with Phi-3-small (7B)

- Given "Harry Potter 1" (HP1), "Harry Potter 2" (HP2) by J.K. Rowling, and "Pride and Prejudice" (PP) by Jane Austen.
- Estimate task correlations $\hat{\alpha}(\Psi^*_{\mathcal{T}_1}, \Psi^*_{\mathcal{T}_2}) = \mathbb{E}_{\boldsymbol{X}}[Sim(f(\boldsymbol{X}; \Psi^*_{\mathcal{T}_1}), f(\boldsymbol{X}; \Psi^*_{\mathcal{T}_1}))]$. HP1 and HP2 are semantically similar, while PP is less aligned with HP1 or HP2.
- Unlearning $\mathcal{T}_{\mathrm{HP1}}$ can effectively degrade the performance of the aligned ($\mathcal{T}_{\mathrm{HP2}}$) as well, while the degradation on the less aligned ($\mathcal{T}_{\mathrm{PP}}$) is relatively smaller.

λ	0 (baseline)	-0.2	-0.4	-0.6	-0.8	-1
$\mathcal{T}_{\mathrm{HP1}}$	0.2573	0.1989	0.1933	0.1888	0.1572	**0.1142** (55.61% $\downarrow$)
$\mathcal{T}_{\mathrm{HP2}}$	0.2688	0.2113	0.1993	0.1938	0.1622	**0.1563** (52.29% $\downarrow$)
$\mathcal{T}_{\mathrm{PP}}$	0.1942	0.1825	0.1644	0.1687	0.1592	**0.1541** (20.65% $\downarrow$)

Figure 28: *Rouge-L scores of $\mathcal{T}_{HP1}$ $\mathcal{T}_{HP2}$, and $\mathcal{T}_{PP}$ by $\Psi = \Psi^{(\mathbf{0})'} + \lambda \cdot \Delta\Psi^{LR}_{HP1}$ using low-rank task vector $\Delta\Psi^{LR}_{HP1}$ (Phi-3-small).*

Summary

- We quantitatively characterize the selection of arithmetic hyper-parameters and their dependence on task correlations so that the resulting task vectors achieve desired multi-task learning, unlearning, and out-of-domain generalization.
- We also demonstrate the validity of using sparse or low-rank task vectors.
- Theoretical results are justified on vision models and large language models.

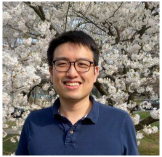
Future Directions

- Analyzing activation-space task vector methods.
- Study the loss landscape of weight/activation-space task vectors or mode connectivity.
- Developing task vector methods together with model pruning.

Acknowledgment

I would like to thank the support and help of the collaborators.

PhD advisor: Prof. Meng Wang



Dr. Pin-Yu Chen

Prof. Sijia Liu

Dr. Songtao Lu

Prof. Shuai Zhang

Yihua Zhang

I want to thank all of the committee members, my lab members, my friends and family.

I especially want to thank the support of Future of Computing Research Collaboration (FCRC) program between RPI and IBM.

FCRC Future of Computing Research Collaboration at Rensselaer

Thank you!

Wayne Xin Zhao, Kun Zhou*, Junyi Li*, Tianyi Tang, Xiaolei Wang, et al.
A Survey of Large Language Models
https://arxiv.org/pdf/2303.18223.pdf

Ruiqi Zhang, Spencer Frei, Peter L. Bartlett
Trained transformers learn linear models in-context
In *Journal of Machine Learning Research*

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, Peter L. Bartlett
How many pretraining tasks are needed for in-context learning of linear regression?
In *International conference on Learning Representations 2024.*

Yu Huang, Yuan Cheng, Yingbin Liang
In-context convergence of transformers.
In *International conference on Machine Learning 2024.*

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, et al.
Editing models with task arithmetic.

In *International Conference on Learning Representations 2022.*

Guillermo Ortiz-Jimenez, Alessandro Favero, Pascal Frossard.
Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models.
In *Conference on Neural Information Processing Systems 2023.*

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li.
Language models are super mario: Absorbing abilities from homologous models as a free lunch.
In *Conference on Machine Learning 2024.*

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al.
Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.
In *Conference on Machine Learning 2022.*

Matena, Michael S and Raffel, Colin A

Merging models with fisher-weighted averaging.
In *Conference on Neural Information Processing Systems 2022.*

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal.
Ties-merging: Resolving interference when merging models.
In *Conference on Neural Information Processing Systems 2023.*

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin.
Linear mode connectivity and the lottery ticket hypothesis.
In *Conference on Machine Learning 2020.*

P. Izmailov, A.G. Wilson, D. Podoprikhin, D. Vetrov, and T. Garipov.
Averaging Weights Leads to Wider Optima and Better Generalization.
In *Conference on Uncertainty in Artificial Intelligence 2018.*

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou.
Making ai forget you: Data deletion in machine learning.
In *Conference on Neural Information Processing Systems 2019.*

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten.

Certified data removal from machine learning models.
In *Conference on Machine Learning 2020.*

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi.
Descent-to-delete: Gradient-based methods for machine unlearning.
In *Algorithmic Learning Theory 2021.*

Siqiao Mu and Diego Klabjan.
Rewind-to-delete: Certified machine unlearning for nonconvex functions.
arXiv preprint arXiv:2409.09778, 2024.