# Towards Minimizing Feature Drift in Model Merging:Layer-wise Task Vector Fusion for Adaptive Knowledge Integration

Wenju Sun[1], Qingyong Li[1], Wen Wang[1], Liu Yang[1], Yangli-ao Geng[1], Boyang Li[2]

[1]Beijing Jiaotong University

[2]Nanyang Technological University

Presented by: Wenju Sun

**The Thirty-ninth Annual Conference on Neural Information Processing Systems**

# Content

**Introduction**

- Model Merging
- Baseline: Task Arithmetic

**Method**

- Knowledge Conflict
- LOT Merging

**Experiment**

# Model Merging

**Definition**

Consider a pretrained model $W_{pre}$ and a set of finetuned models $\{W_i\}_{i=1}^{k}$ with corresponding downstream tasks $\{D_i\}_{i=1}^{k}$.

Our goal is to merge all K models into a unified model $W_{mtl}$ **without redundant retraining**. The unified model $W_{mtl}$ should **perform well on all downstream tasks**.

# Baseline: Task Arithmetic

**Task Arithmetic**: Considering a pretrained model $W_{pre}$ and a set of finetuned models $\{W_i\}_{i=1}^{k}$ with corresponding downstream tasks $\{D_i\}_{i=1}^{k}$, the task vectors $\{T_i\}_{i=1}^{k}$ are defined as $T_i = W_i - W_0$.

Task vectors can be applied to $W_{pre}$ with a scaling term $\lambda$, i.e.,$W_{mtl} = W_{pre} + \alpha \sum_i T_i$, which allows to control the behavior of the edited model via simple arithmetic operations on task vectors.

# Knowledge Conflict

**Definition** (for task arithmetic): the increase in task-specific loss incurred by merging.

For a given task k, associated with the loss function $\mathcal{L}_k(.)$, the knowledge conflict during merging is quantified as:

$$\Delta\mathcal{L}_k = \mathcal{L}_k(W_{mtl}) - \mathcal{L}_k(W_k)$$

# Isolating Knowledge Conflict

NEURAL INFORMATION PROCESSING SYSTEMS

**Theorem 4.4 (An Upper Bound on Knowledge Conflict):**
Suppose that within the range of model merging, the function of layer l is $\gamma_l$-Lipschitz continuous with respect to its input, and the loss function L is $\beta$-Lipschitz continuous with respect to the final output of the network. Then, the knowledge conflict follows:

$$|\Delta\mathcal{L}_k| \leq \beta \sum_{l=1}^{L} \left( \prod_{m=l+1}^{L} \gamma_m \right) \|\Delta f_k^l\|$$

# Objective

We propose to mitigate knowledge conflict by minimizing the **feature drift** for each layer:

$$T^{l\star} = \arg\min_{T^l} \sum_{k=1}^{K} \|\Delta f_k^l\|^2 = \arg\min_{T^l} \sum_{k=1}^{K} \|f_k^l(W_{\text{pre}} + T^l) - f_k^l(W_k)\|^2$$

# For Linear Weights

Suppose $W^l \in \mathbb{R}^{d_l \times d_{l+1}}$ corresponds to the weight of a linear layer, which transforms features through matrix multiplication $f_k^l(W) = X_k^l W^l$ with pre-collected input $X_k^l$. The objective becomes:

$$
\begin{aligned}
T^{l\star} &= \arg\min_{T^l} \sum_{k=1}^{K} \|X_k^l(W_{\text{pre}}^l + T^l) - X_k^l(W_k^l)\|_F^2 \\
&= \arg\min_{T^l} \sum_{k=1}^{K} \|X_k^l(T^l - T_k^l)\|_F^2 = \arg\min_{T^l} \sum_{k=1}^{K} \text{trace}((T^l - T_k^l)^\top {X_k^l}^\top X_k^l (T^l - T_k^l))
\end{aligned}
$$

This defines a convex quadratic optimization problem. Consequently, the optimal solution ${T^l}^*$ can be derived in closed form as follows :

$$
T^{l\star} = \left(\sum_k {X_k^l}^\top X_k^l\right)^\dagger \sum_k {X_k^l}^\top X_k^l T_k^l
$$

# Analysis: Why $T^{l\star} = \left(\sum_k X_k^{l\top} X_k^l\right)^\dagger \sum_k X_k^{l\top} X_k^l T_k^l$ Work?

NEURAL INFORMATION PROCESSING SYSTEMS

Consider SVD on input features $X_k^l = U_k^l \Sigma_k^l V_k^{l\top}$.

**Ideal case**, where for any $k \neq j$, $V_k^{l\top} V_j^l = 0$. Then, $T^{l^*}$ simplifies to:

$$T^{l\star}_{\text{ideal}} = \left(\sum_k V_k^l \Sigma_k^{l\,2} V_k^{l\top}\right)^\dagger \sum_k V_k^l \Sigma_k^{l\,2} V_k^{l\top} T_k^l = \sum_k \left(V_k^l \Sigma_k^{l\,2} V_k^{l\top}\right)^\dagger \sum_k V_k^l \Sigma_k^{l\,2} V_k^{l\top} T_k^l = \sum_k V_k^l V_k^{l\top} T_k^l.$$

There is no conflict in this case

$$\begin{aligned}
\sum_{k=1}^{K} \|X_k^l (T^{l\star}_{\text{ideal}} - T_k^l)\|_F^2 &= \sum_{k=1}^{K} \|U_k^l \Sigma_k^l V_k^{l\top} (\sum_j V_j^l V_j^{l\top} T_j^l - T_k^l)\|_F^2 \\
&= \sum_{k=1}^{K} \|U_k^l \Sigma_k^l V_k^{l\top} V_k^l V_k^{l\top} T_k^l - U_k^l \Sigma_k^l V_k^{l\top} T_k^l\|_F^2 = 0
\end{aligned}$$

# Analysis: Why $T^{l\star} = \left(\sum_k X_k^{l\top} X_k^l\right)^\dagger \sum_k X_k^{l\top} X_k^l T_k^l$ Work?

NEURAL INFORMATION PROCESSING SYSTEMS

Consider SVD on input features $X_k^l = U_k^l \Sigma_k^l V_k^{l\top}$.

**Worst case**, where for any $k$, $V_k^l = V^l$. Then, $T^{l*}$ simplifies to:

$$T^{l\star}_{\text{worst}} = \left(\sum_k V^l {\Sigma_k^l}^2 V^{l\top}\right)^\dagger \sum_k V^l {\Sigma_k^l}^2 V^{l\top} T_k^l$$

$$= V^l \left(\sum_k {\Sigma_k^l}^2\right)^\dagger V^{l\top} \sum_k V^l {\Sigma_k^l}^2 V^{l\top} T_k^l = \sum_k \left( V^l \underbrace{\left(\sum_k {\Sigma_k^l}^2\right)^\dagger {\Sigma_k^l}^2}_{\text{Normalized Weight}} V^{l\top} T_k^l \right)$$

# Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

| Method | SUN397 | Cars | RESISC45 | EuroSAT | SVHN | GTSRB | MNIST | DTD | Avg Acc | #best |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ***Basic baseline methods*** | | | | | | |
| Pre-trained | 62.3 | 59.7 | 60.7 | 45.5 | 31.4 | 32.6 | 48.5 | 43.8 | 48.0 | - |
| Individual | 75.3 | 77.7 | 96.1 | 99.7 | 97.5 | 98.7 | 99.7 | 79.4 | 90.5 | - |
| Traditional MTL | 73.9 | 74.4 | 93.9 | 98.2 | 95.8 | 98.9 | 99.5 | 77.9 | 88.9 | - |
| | | | | ***Training-free methods*** | | | | | | |
| Weight Averaging | 65.3 | 63.4 | 71.4 | 71.7 | 64.2 | 52.8 | 87.5 | 50.1 | 65.8 | 0 |
| Fisher Merging | **68.6** | **69.2** | 70.7 | 66.4 | 72.9 | 51.1 | 87.9 | 59.9 | 68.3 | 2 |
| RegMean | 65.3 | 63.5 | 75.6 | 78.6 | 78.1 | 67.4 | 93.7 | 52.0 | 71.8 | 0 |
| Task Arithmetic | 55.2 | 54.9 | 66.7 | 78.9 | 80.2 | 69.7 | 97.3 | 50.4 | 69.1 | 0 |
| Ties-Merging | 59.8 | 58.6 | 70.7 | 79.7 | 86.2 | 72.1 | 98.3 | 54.2 | 72.4 | 0 |
| TATR | 62.7 | 59.3 | 72.3 | 82.3 | 80.5 | 72.6 | 97.0 | 55.4 | 72.8 | 0 |
| Ties-Merging & TATR | 66.3 | 65.9 | 75.9 | 79.4 | 79.9 | 68.1 | 96.2 | 54.8 | 73.3 | 0 |
| Consensus Merging | 65.7 | 63.6 | 76.5 | 77.2 | 81.7 | 70.3 | 97.0 | 57.1 | 73.6 | 0 |
| AWD Merging | 63.5 | 61.9 | 72.6 | 84.9 | 85.1 | 79.1 | 98.1 | 56.7 | 75.2 | 0 |
| PCB Merging | 63.8 | 62.0 | 77.1 | 80.6 | 87.5 | 78.5 | **98.7** | 58.4 | 75.8 | 1 |
| CAT Merging | 68.1 | 65.4 | 80.5 | 89.5 | 85.5 | 78.5 | 98.6 | 60.7 | 78.3 | 0 |
| LOT Merging (ours) | 67.7 | 67.5 | **85.7** | **94.9** | **93.4** | **89.8** | **98.7** | **63.6** | **82.7** | 6 |

# Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

| Method | SUN397 | Cars | RESISC45 | EuroSAT | SVHN | GTSRB | MNIST | DTD | Avg Acc | #best |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ***Basic baseline methods*** | | | | | | |
| Pre-trained | 66.8 | 77.7 | 71.0 | 59.9 | 58.4 | 50.5 | 76.3 | 55.3 | 64.5 | - |
| Individual | 82.3 | 92.4 | 97.4 | 100.0 | 98.1 | 99.2 | 99.7 | 84.1 | 94.2 | - |
| Traditional MTL | 80.8 | 90.6 | 96.3 | 96.3 | 97.6 | 99.1 | 99.6 | 84.4 | 93.5 | - |
| | | | | ***Training-free methods*** | | | | | | |
| Weight Averaging | 72.1 | 81.6 | 82.6 | 91.9 | 78.2 | 70.7 | 97.1 | 62.8 | 79.6 | 0 |
| Fisher Merging | 69.2 | **88.6** | 87.5 | 93.5 | 80.6 | 74.8 | 93.3 | 70.0 | 82.2 | 1 |
| RegMean | 73.3 | 81.8 | 86.1 | 97.0 | 88.0 | 84.2 | 98.5 | 60.8 | 83.7 | 0 |
| Task Arithmetic | 73.9 | 82.1 | 86.6 | 94.1 | 87.9 | 86.7 | 98.9 | 65.6 | 84.5 | 0 |
| Ties-Merging | 76.5 | 85.0 | 89.3 | 95.7 | 90.3 | 83.3 | 99.0 | 68.8 | 86.0 | 0 |
| TATR | 74.6 | 83.7 | 87.6 | 93.7 | 88.6 | 88.1 | 99.0 | 66.8 | 85.3 | 0 |
| Ties-Merging & TATR | 76.3 | 85.3 | 88.8 | 94.4 | 90.8 | 88.7 | 99.2 | 68.8 | 86.5 | 0 |
| Consensus Merging | 75.0 | 84.3 | 89.4 | 95.6 | 88.3 | 82.4 | 98.9 | 68.0 | 85.2 | 0 |
| AWD Merging | 76.2 | 85.4 | 88.7 | 96.1 | 92.4 | 92.3 | 99.3 | 69.4 | 87.5 | 0 |
| PCB Merging | 76.2 | 86.0 | 89.6 | 95.9 | 89.9 | 92.3 | 99.2 | 71.4 | 87.6 | 0 |
| CAT Merging | **78.7** | 88.5 | 91.1 | 96.3 | 91.3 | **95.7** | 99.4 | 75.7 | 89.6 | 2 |
| LOT Merging (ours) | 76.7 | **88.6** | **91.7** | **98.7** | **97.1** | **95.7** | **99.5** | **76.4** | **90.5** | 7 |

# Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

| Method | COCO Caption | Flickr30k Caption | Textcaps | OKVQA | TextVQA | ScienceQA | #best |
|---|---|---|---|---|---|---|---|
| **Metric** | CIDEr | CIDEr | CIDEr | Accuracy | Accuracy | Accuracy | |
| Pre-trained · | 0.07 | 0.03 | 0.05 | 42.80 | 21.08 | 40.50 | - |
| Individual | 1.17 | 0.65 | 0.65 | 50.84 | 29.79 | 76.89 | - |
| Task Arithmetic | 0.86 | 0.50 | 0.39 | 17.71 | 0.49 | 40.10 | 0 |
| Ties-Merging | 0.53 | 0.27 | 0.22 | 27.95 | 0.57 | 40.35 | 0 |
| TATR | 0.46 | 0.31 | 0.21 | 28.30 | 14.74 | 42.98 | 0 |
| PCB Merging | 0.71 | 0.52 | 0.30 | 36.04 | 1.88 | 43.01 | 0 |
| CAT Merging | **0.91** | 0.53 | 0.36 | **44.07** | 19.69 | 46.36 | 2 |
| LOT Merging (ours) | **0.91** | **0.54** | **0.44** | 38.35 | **20.82** | **48.24** | 5 |

# THANKS FOR WATCHING

## Minimizing Feature Drift in Model Merging:Layer-wise Task Vector Fusion for Adaptive Knowledge Integration

Presented by: Wenju Sun

北京交通大學
BEIJING JIAOTONG UNIVERSITY

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NEURAL INFORMATION PROCESSING SYSTEMS