上海交通大學
SHANGHAI JIAO TONG UNIVERSITY
復旦大學
FUDAN UNIVERSITY

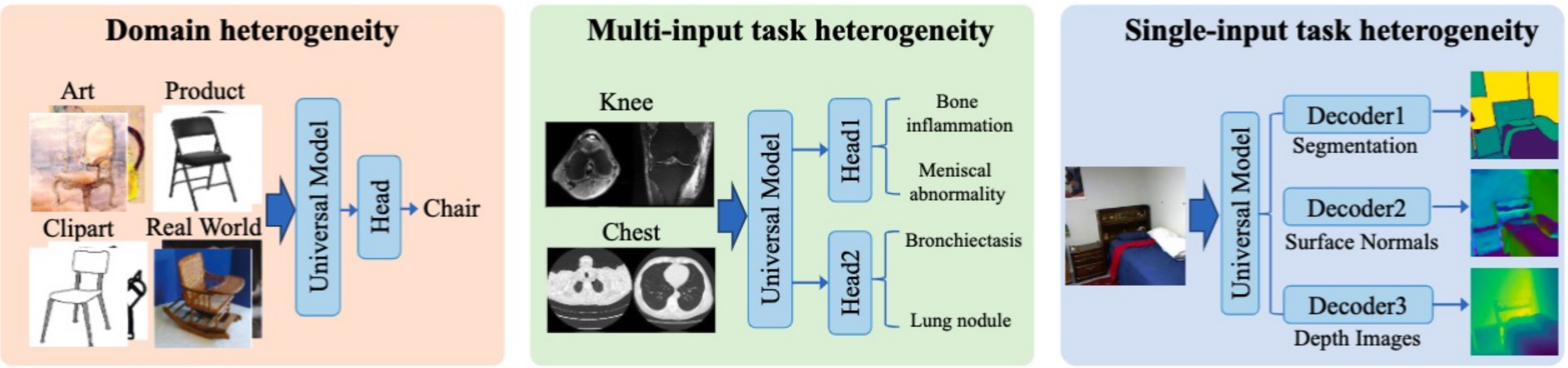
Exploring Training on Heterogeneous Data with Mixture of Low-rank Adapters

Yuhang Zhou[1,2], Zihua Zhao[1,2], Siyuan Du[2,3] , Haolin Li[2,3], Jiangchao Yao[1,2], Ya Zhang[1,2], Yanfeng Wang[1,2]

[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, [2] Shanghai AI Laboratory, [3] Fudan University

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Introduction



Motivation: Diverse training data collected from different domains or tasks is often utilized to train a unified model for pursuing universal capability. However, due to the presence of heterogeneity, such unification may suffer from strong conflicts during training, resulting in the suppression of the scale advantage of the pre-training dataset and severely impacting the performance of the model.

Main Task: Mitigate the training conflicts among heterogeneous data collected from different domains or tasks.

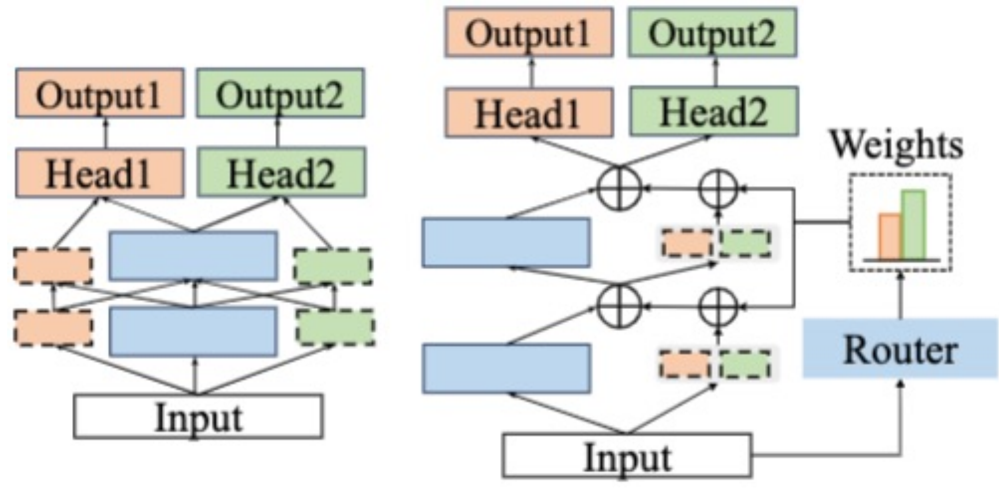
Main Contributions

- By introducing task-specific low-rank parameters, MoLA achieves parameter isolation between different tasks, thereby separating heterogeneous gradients to avoid conflicts between tasks.
- We propose MoLA-Grad and MoLA-Router, which use task identifiers and the router intervened by our TwD loss respectively, explicitly or implicitly mitigating the conflicts.
- Analysis on the training of MoLA from the perspectives of principal component changes and eigenvalue distributions.

Method

Intuition:

- the low-rank property of MoLA ensures that the increase of parameters is controllable;
- the (primary-secondary) rank discrepancy between backbone and adapters encourages model to disentangle the shared knowledge and complementary knowledge.

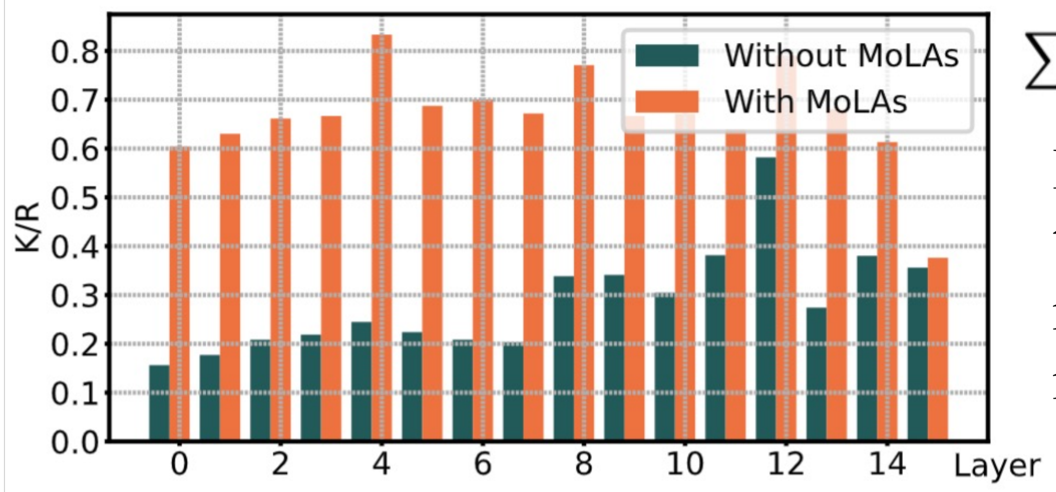


(a) MoLA-Grad (b) MoLA-Router

$$g_t = (\mathbf{W_0} + \sum_{i=1}^{E} \alpha_i \mathbf{B_i A_i}) h_t$$

(a) $$g = (\mathbf{W_0} + \mathbf{B_1 A_1})h_1 \cup \cdots \cup (\mathbf{W_0} + \mathbf{B_T A_T})h_T = (\mathbf{W_0} + \mathbf{BA} \circ \mathbf{M})h = \mathbf{W}'h,$$

(b) $$\mathcal{L}_{\mathrm{TwD}} = -\sum_{i=1}^{b}\sum_{j=1}^{b} \mathbb{1}_{i\neq j} \cdot \mathbb{1}_{t_i = t_j} \log \frac{e^{\omega_i^\top \omega_j/\tau}}{\sum_{k=1}^{b} \mathbb{1}_{i\neq k} e^{\omega_i^\top \omega_k/\tau}}$$



$\sum_{i=1}^{K} \sigma_i \geq \alpha \sum_{i=1}^{R} \sigma_i$ (α=0.99 in our analysis)

MoLA allows for the extraction of more task-specific heterogeneous features, thus requiring the involvement of a greater number of eigenvectors for representation.

The main difference from the original LoRA:

- Different learning stages. LoRA is used for adapting models to downstream tasks, while MoLA is used to train from scratch together with the backbone;
- Different impacts on training. LoRA can significantly amplify a small number of eigenvalues, thereby emphasizing task-relevant eigenvectors. Instead, MoLA significantly reduce the maximum eigenvalues to capture more heterogeneous information, alleviating training conflicts.

Experimental Results

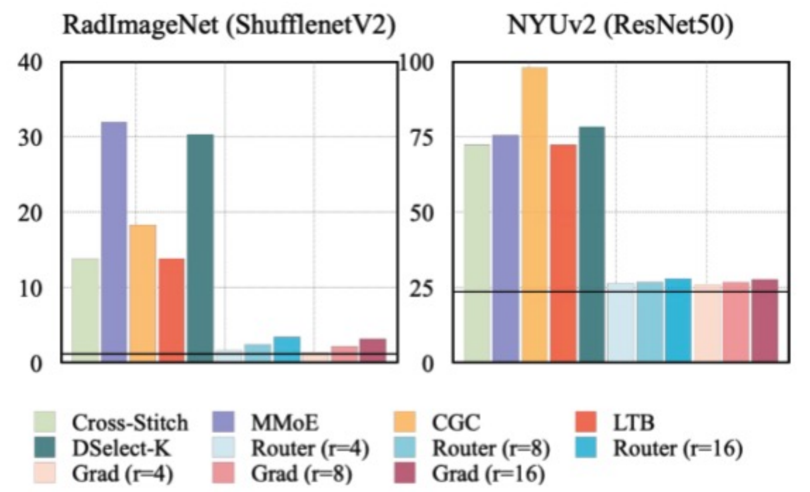
- Multi-input task heterogeneity:

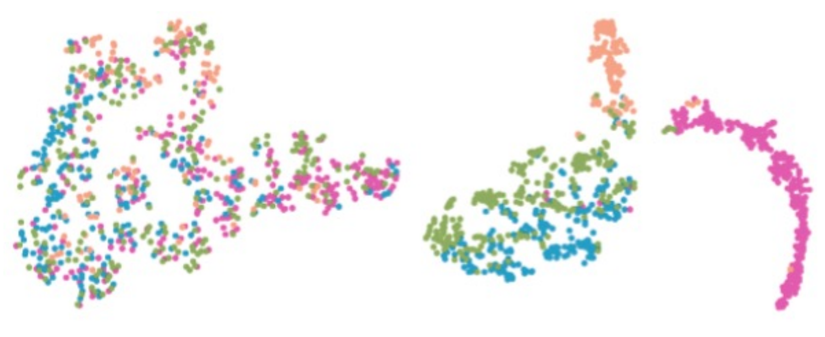
	Lung ↑	Abdomen ↑	Thyroid ↑	Abdomen ↑	Knee ↑	Shoulder ↑	Spine ↑	Ankle ↑	Abdomen ↑	Brain ↑	Hip ↑	Avg ↑
Single-Task	76.42	33.94	91.55	**69.17**	49.32	41.80	20.62	**20.31**	65.99	83.88	51.05	54.91
Uniform	34.34	41.33	69.85	22.87	34.71	27.86	20.26	12.88	60.01	75.16	24.13	38.49
HSP	77.16	37.45	91.73	68.43	46.47	42.72	20.85	18.17	71.13	84.67	55.16	55.81
MGDA	70.61	31.74	**94.22**	67.46	43.59	45.18	23.69	17.67	73.86	84.20	53.62	55.08
PCGrad	69.34	**43.15**	75.60	67.77	42.59	39.24	14.20	14.38	34.52	69.36	46.26	46.95
CAGrad	71.73	22.71	91.83	62.70	42.30	42.84	**24.42**	18.99	75.07	83.83	50.51	53.36
Aligned-MTL	62.59	33.90	92.96	67.18	44.33	45.41	23.41	18.28	71.68	**84.93**	54.84	54.50
Cross-Stitch	80.71	26.67	92.73	66.03	44.25	44.13	23.01	17.92	65.22	74.37	47.51	52.96
MMoE	**81.75**	38.65	83.27	67.27	44.35	43.84	16.42	13.11	47.64	77.64	**55.63**	51.78
DSelect-K	77.48	35.34	91.62	67.08	45.89	42.22	19.50	15.49	73.39	79.74	53.85	54.69
CGC	75.47	28.12	86.26	67.67	46.02	42.16	15.09	15.81	24.93	84.88	53.04	49.04
LTB	68.63	40.69	88.99	68.09	45.69	**45.61**	23.13	18.79	**75.39**	84.39	53.56	55.72
MoLA-Router	78.94	36.38	91.76	68.05	48.41	43.03	23.26	18.37	68.65	84.56	54.93	56.03
MoLA-Grad	80.72	34.54	92.18	68.87	**50.18**	43.41	22.06	19.76	69.10	84.67	**55.63**	**56.47**

- Domain heterogeneity:

	Domain-V	Domain-L	Domain-C	Domain-S	Avg ↑
Single-Task	84.32	75.40	100.0	78.70	84.60
Uniform	84.75	72.73	100.0	76.09	83.39
MMD-AAE	84.32	69.52	100.0	80.43	83.57
SelfReg	81.36	71.65	100.0	82.17	83.80
EQRM	83.9	67.38	100.0	79.57	82.71
DANN	49.15	57.75	60.00	55.65	55.64
HPS	85.59	74.33	100.0	80.00	84.98
Cross-Stitch	86.44	78.07	100.0	77.83	85.59
MMoE	83.05	74.33	100.0	79.57	84.24
DSelect-K	83.90	75.40	100.0	80.87	85.04
CGC	83.90	72.73	100.0	80.43	84.27
LTB	80.93	75.94	100.0	79.13	84.00
MoLA-Router	85.59	**79.14**	100.0	80.87	86.40
MoLA-Grad	**87.29**	74.87	100.0	**83.48**	**86.41**

- Parameter number:





(a) Without TwD loss (b) With TwD loss

Contact Us

zhouyuhang@sjtu.edu.cn

MediaBrain Lab

ICML
International Conference On Machine Learning