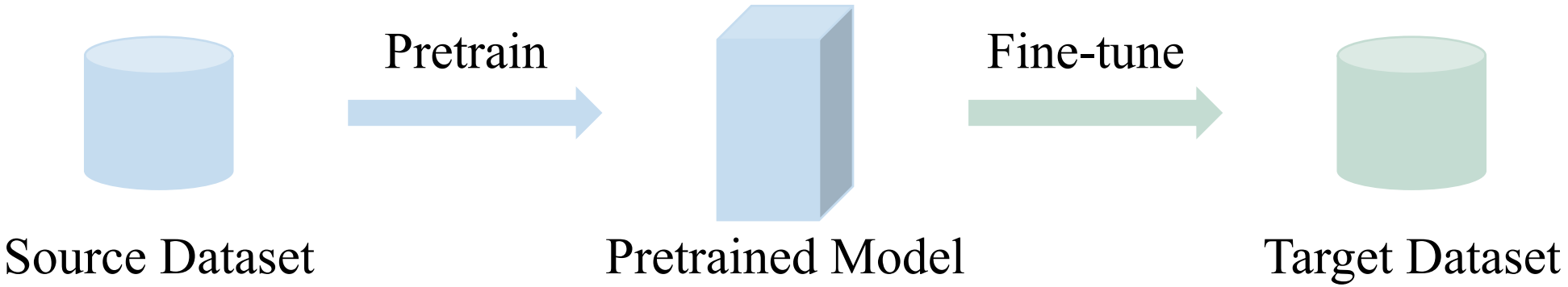
Zoo-Tuning: Adaptive Transfer from a Zoo of Models

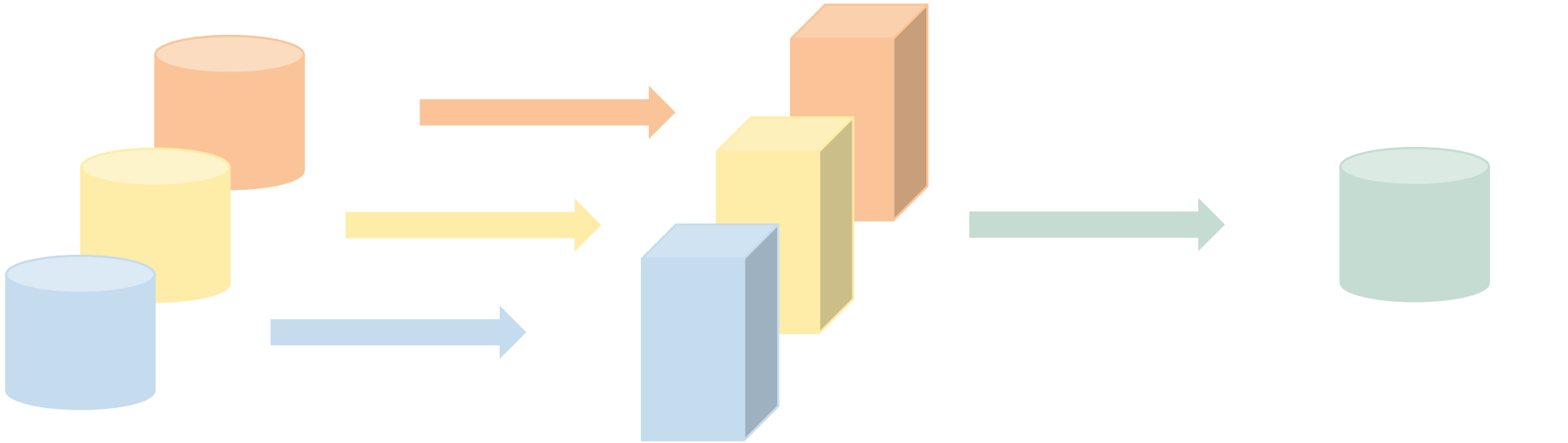
Yang Shu* Zhi Kou* Zhangjie Cao Jianmin Wang Mingsheng Long

School of Software, BNRist, Tsinghua University

Transfer Learning



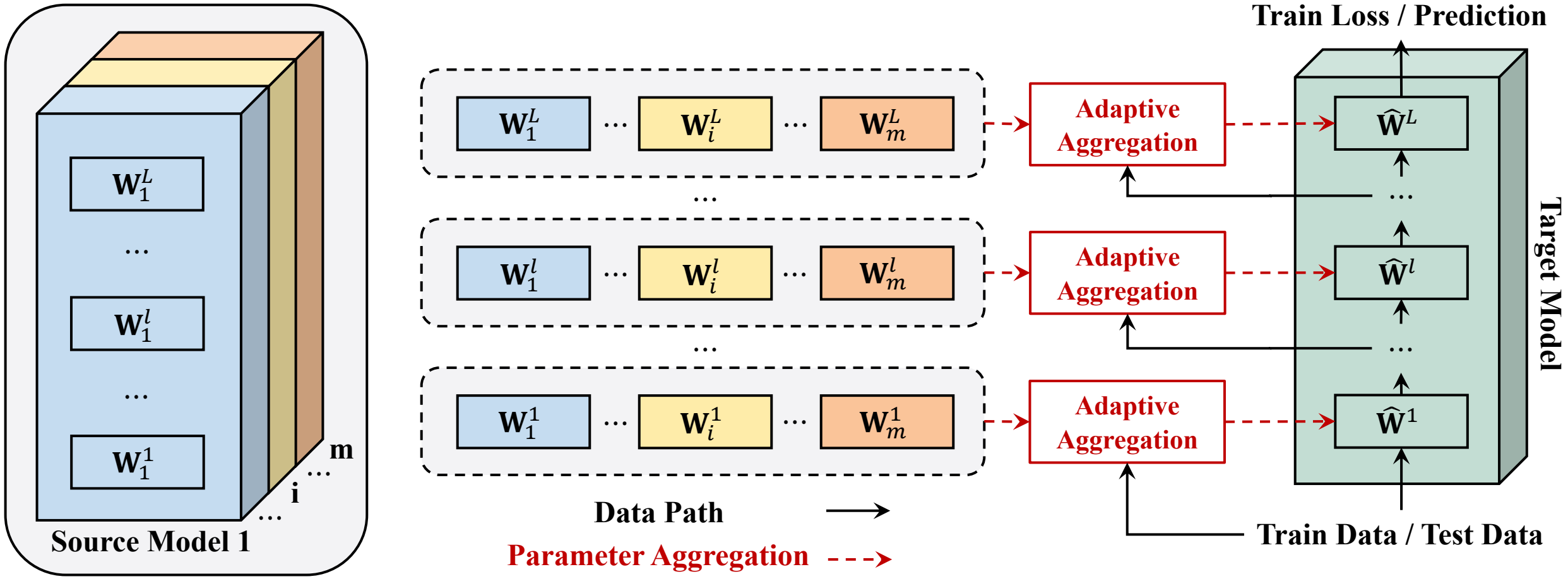
Transfer From A Zoo of Models



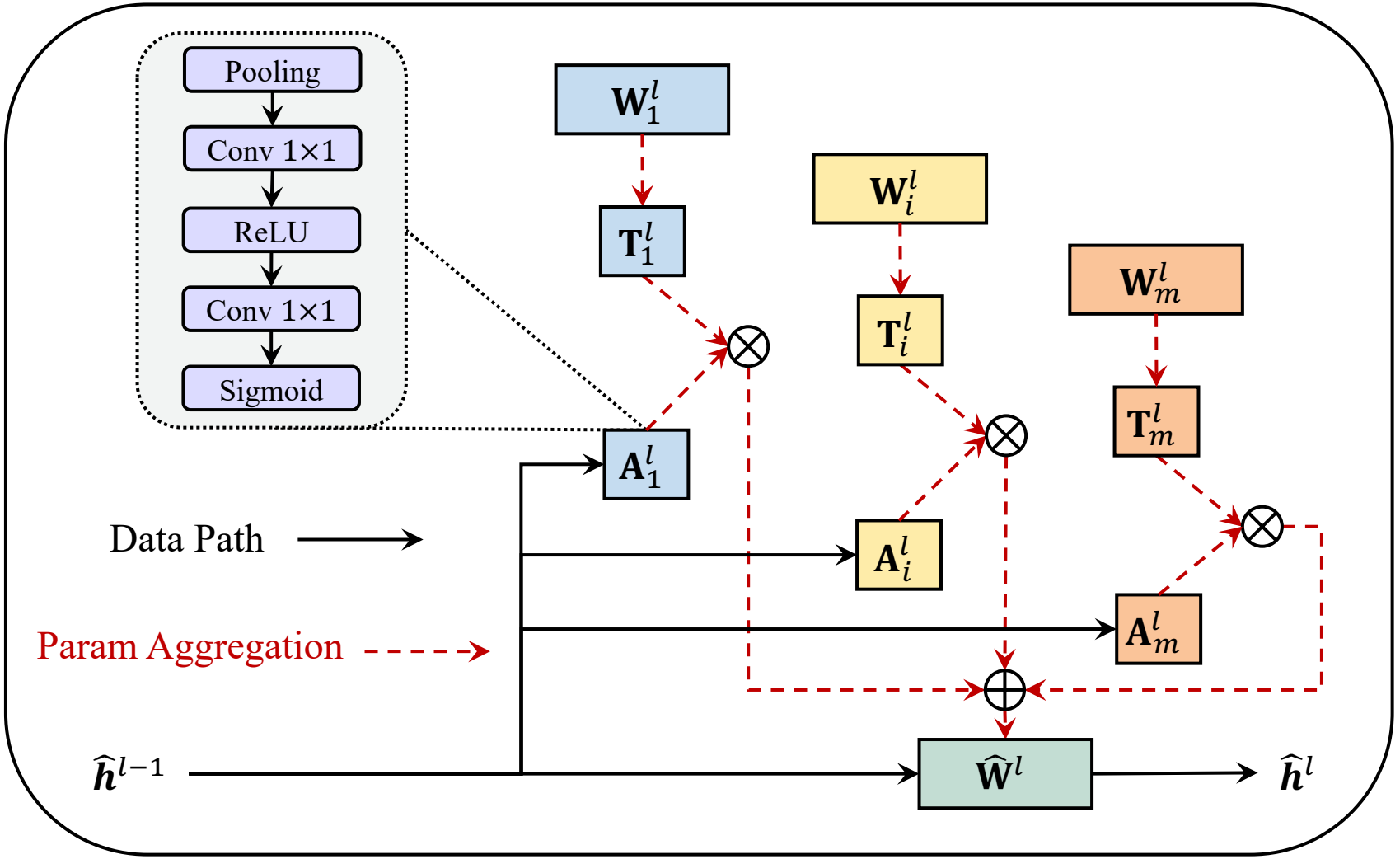
Models considered in this paper

- Same architecture
- Pretrained with different data, tasks or pretraining algorithms

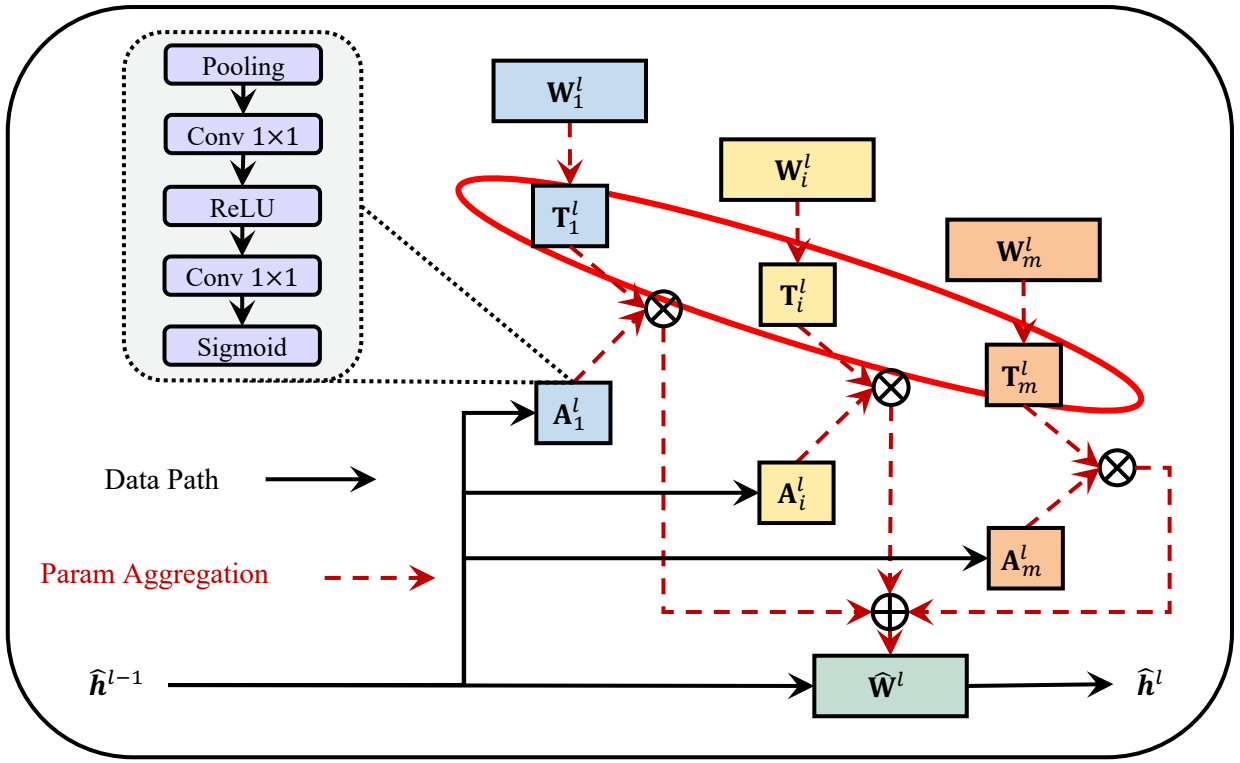
Zoo-Tuning


Source Model 1
m
i
Train Loss / Prediction
Adaptive Aggregation
Adaptive Aggregation
Adaptive Aggregation
Target Model
Data Path
Parameter Aggregation
Train Data / Test Data

Adaptive Aggregation



Adaptive Aggregation



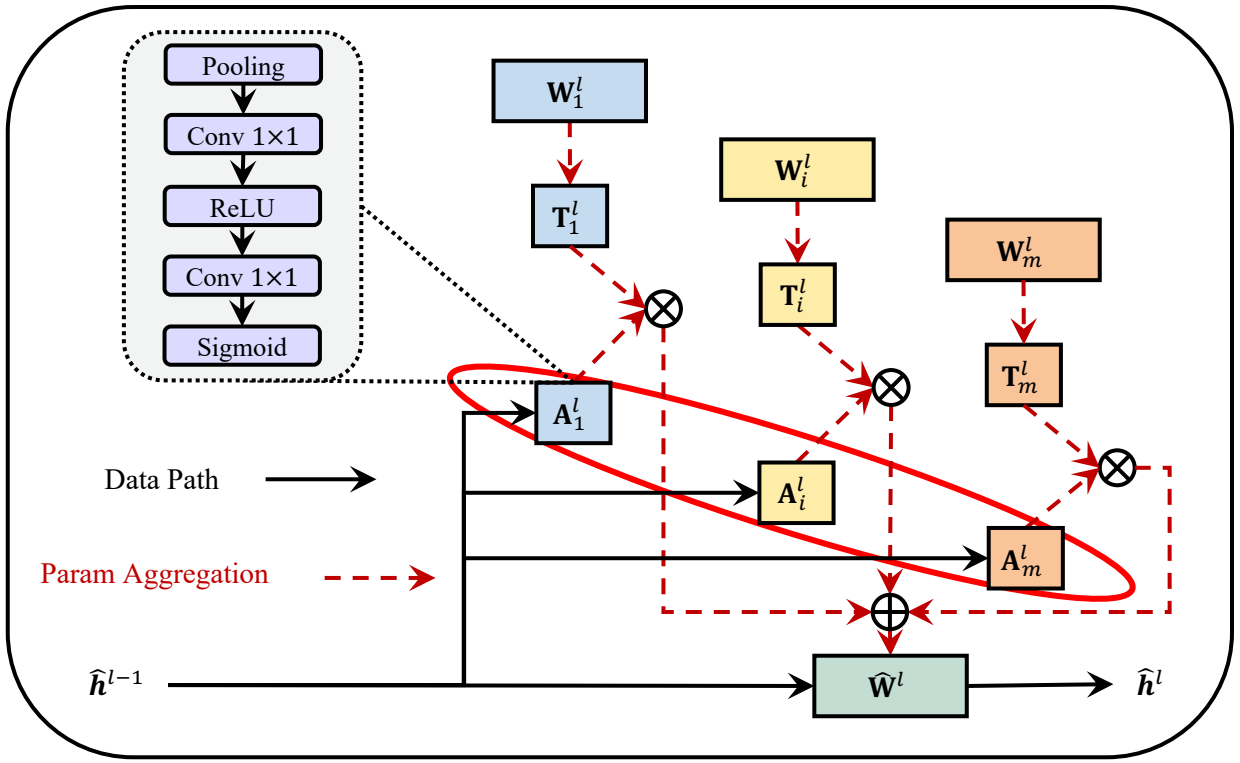
Channel alignment module

- Channels in different pretrained models may have different semantic meanings.
- A channel alignment module $\mathbf{T}_i^l$ that transforms and aligns channels of different models

$$\widetilde{\mathbf{W}}_i^l = \mathbf{T}_i^l * \mathbf{W}_i^l$$

Transformed parameter

Adaptive Aggregation



Data-dependent gating module

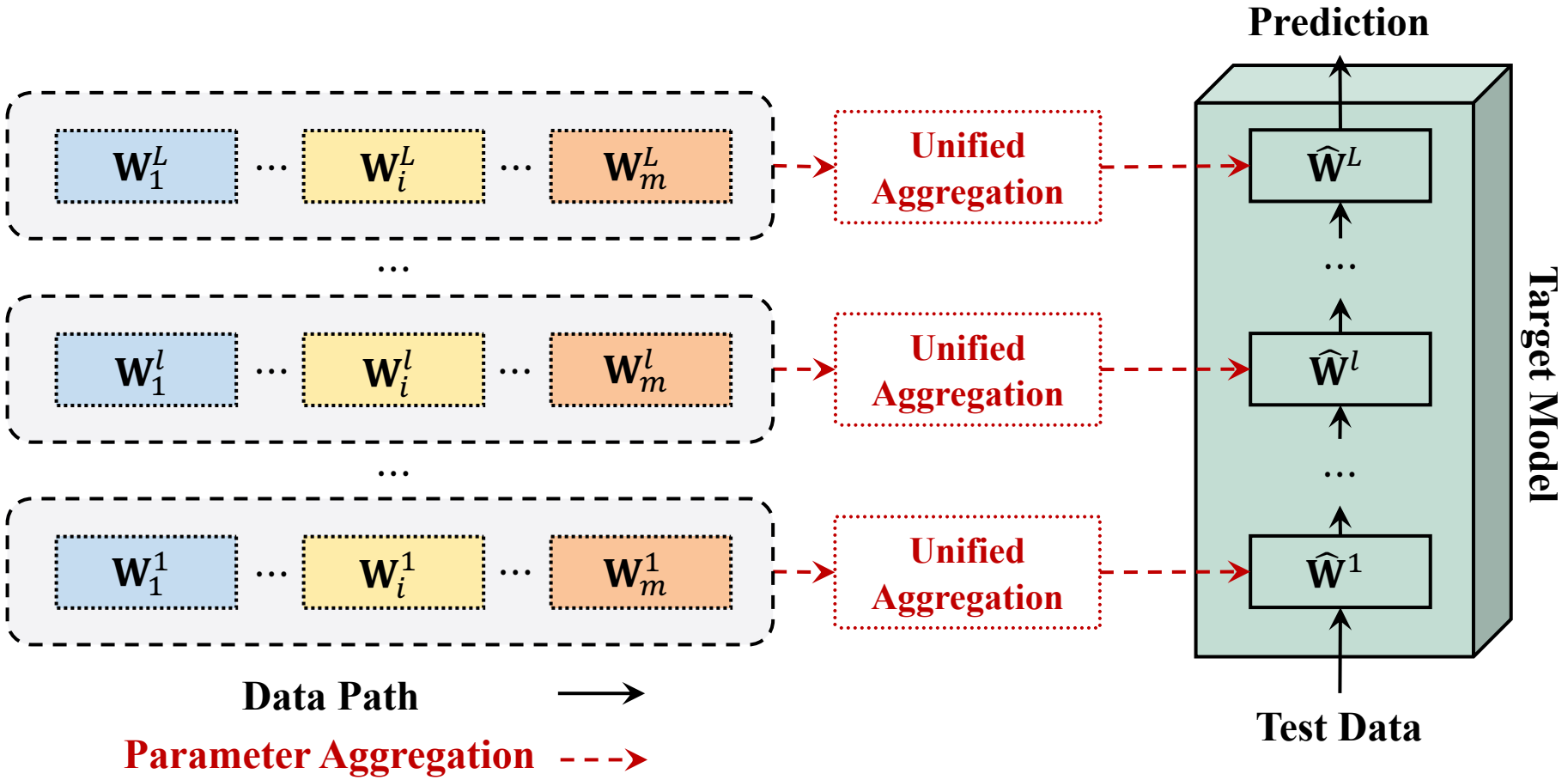
- Each data point of each task should have a different aggregation since it holds specific relationships with source tasks.
- A gating module $\mathbf{A}_i^l$ for each source model, which controls the mixing of its parameters $\mathbf{W}_i^l$

$$\widehat{\mathbf{W}}^l = \sum_{i=1}^{m} a_i^l \widetilde{\mathbf{W}}_i^l = \sum_{i=1}^{m} \mathbf{A}_i^l(\hat{h}^{l-1}) \left(\mathbf{T}_i^l * \mathbf{W}_i^l\right)$$

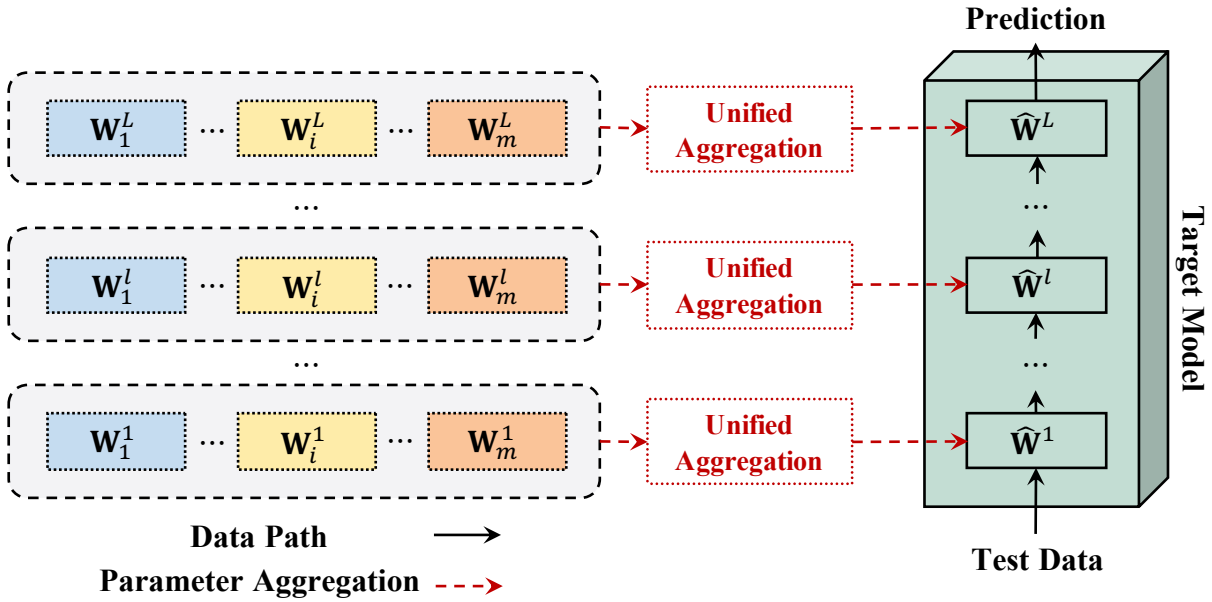
Target parameter

Gating value

Lite Zoo-Tuning



Lite Zoo-Tuning



Unified gating value

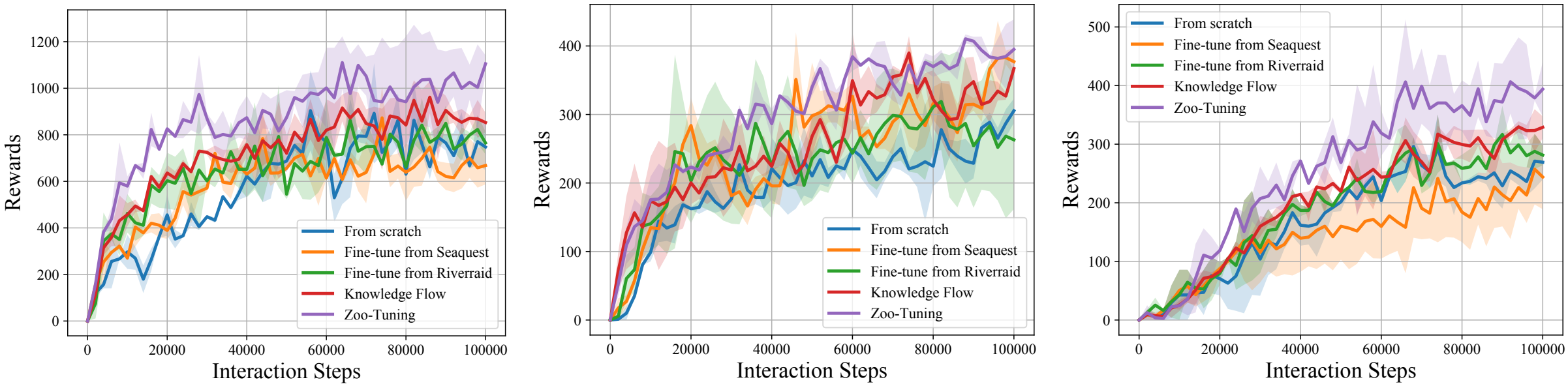
$$\widehat{\mathbf{W}}^l = \sum_{i=1}^{m} \bar{a}_i^l \widetilde{\mathbf{W}}_i^l$$

$$\bar{a}_i^l = \alpha \cdot \bar{a}_i^l + (1-\alpha) \left(\frac{1}{b} \cdot \sum_{j=1}^{b} a_{i,j}^l \right)$$

Temporal ensemble

Batch average gating value

Transfer Learning in Reinforcement Learning



(a) Alien (b) Gopher (c) JamesBond

Figure 4. Results of transferring pretrained models to downstream tasks in the reinforcement learning of Atari games.

Transfer Learning in Facial Landmark Detection

- Pretrained Models：1) ImageNet supervised pretraining ； 2) ImageNet MOCO pretraining；3) Mask R-CNN (detection and instance segmentation) ； 4) DeepLabV3 (semantic segmentation)； 5) Keypoint R-CNN (keypoint detection)

Table 2. Comparison of NME results on facial landmark detection tasks: 300W, WFLW, and COFW.

MODEL	300W	WFLW	COFW
SCRATCH	3.66	5.33	4.20
IMAGENET SUP.	3.52	4.90	3.66
MOCO PT.	3.45	4.75	3.63
MASKRCNN PT.	3.53	4.87	3.67
DEEPLAB PT.	3.53	4.89	3.73
KEYPOINT PT.	3.50	4.90	3.66
ENSEMBLE	**3.33**	4.64	**3.46**
DISTILL	3.45	4.74	3.53
KNOWLEDGE FLOW	3.71	5.28	4.58
ZOO-TUNING	3.41	**4.58**	3.51

Transfer Learning in Image Classification

- Pretrained Models：1) ImageNet supervised pretraining ； 2) ImageNet MOCO pretraining；3) Mask R-CNN (detection and instance segmentation) ； 4) DeepLabV3 (semantic segmentation)； 5) Keypoint R-CNN (keypoint detection)

Table 1. Comparison of top-1 accuracy(%) and complexity on the classification benchmarks including General benchmark, Fine-grained benchmark, and Specialized benchmark.

Model	General		Fine-grained			Specialized			Train		Inference	
	Cifar-100	COCO-70	Aircraft	Cars	Indoors	DMLab	EuroSAT	Avg. Acc.	GFLOPs	Params	GFLOPs	Params
ImageNet sup.	81.18	81.97	84.63	89.38	73.69	74.57	98.43	83.41	4.12	23.71M	4.12	23.71M
MOCO pt.	75.31	75.66	83.44	85.38	70.98	75.06	98.82	80.66	4.12	23.71M	4.12	23.71M
MaskRCNN pt.	79.12	81.64	84.76	87.12	73.01	74.73	98.65	82.72	4.12	23.71M	4.12	23.71M
DeepLab pt.	78.76	80.70	84.97	88.03	73.09	74.34	98.54	82.63	4.12	23.71M	4.12	23.71M
Keypoint pt.	76.38	76.53	84.43	86.52	71.35	74.58	98.34	81.16	4.12	23.71M	4.12	23.71M
ensemble	82.26	82.81	**87.02**	**91.06**	73.46	**76.01**	98.88	84.50	20.60	118.55M	20.60	118.55M
distill	82.32	82.44	85.00	89.47	73.97	74.57	98.95	83.82	24.72	142.28M	4.12	23.71M
Knowledge Flow	81.56	81.91	85.27	89.22	73.37	75.55	97.99	83.55	28.83	169.11M	4.12	23.71M
Lite Zoo-Tuning	83.39	83.50	85.51	89.73	75.12	75.22	**99.12**	84.51	4.53	130.43M	4.12	23.71M
Zoo-Tuning	**83.77**	**84.91**	86.54	90.76	**75.39**	75.64	**99.12**	**85.16**	4.53	130.43M	4.18	122.54M

Thank you for listening !

Zoo-Tuning: Adaptive Transfer from a Zoo of Models

Yang Shu* Zhi Kou* Zhangjie Cao Jianmin Wang Mingsheng Long

School of Software, BNRist, Tsinghua University