Text-to-LoRA: Instant Transformer Adaption

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, Robert Lange

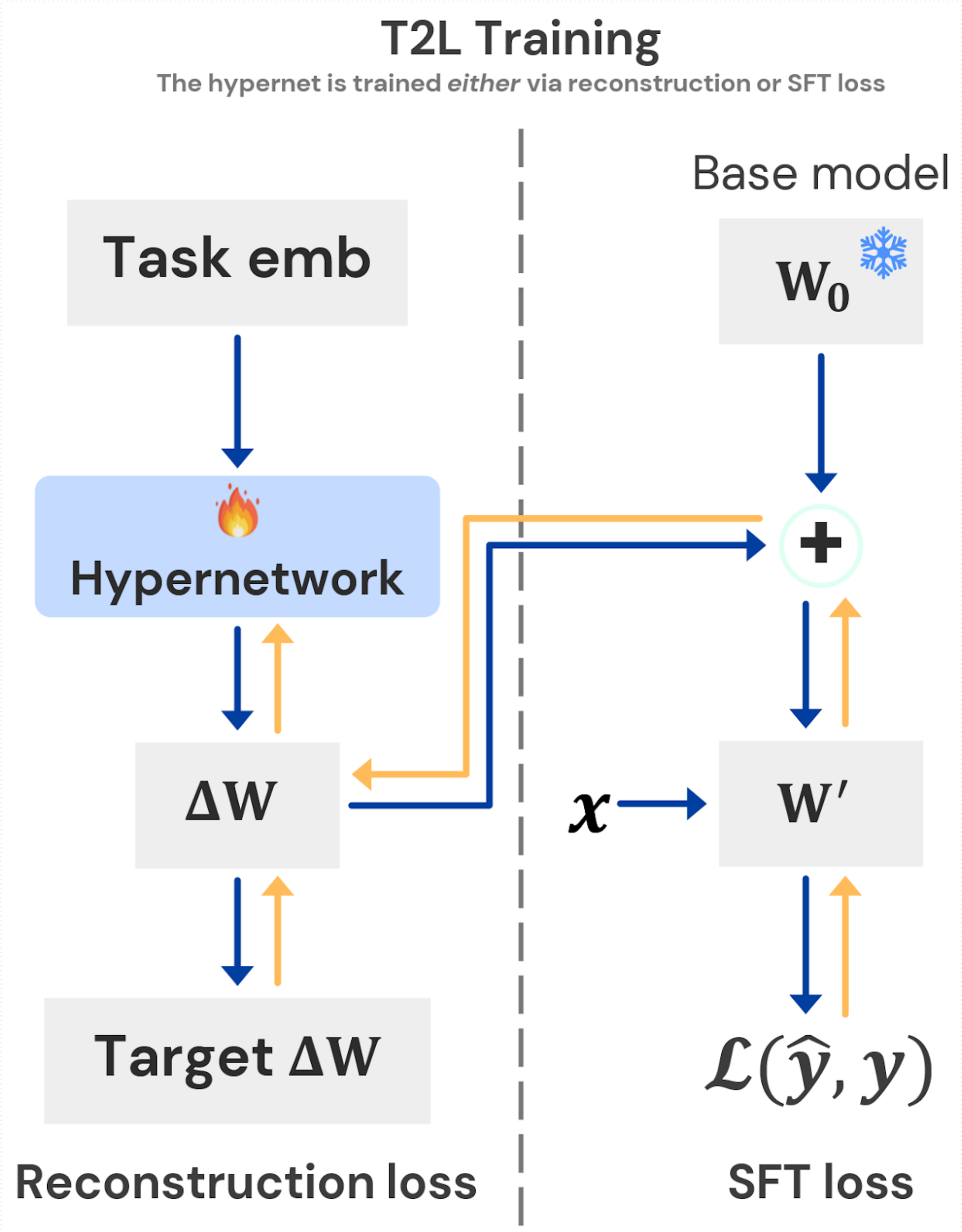
sakana.ai

github.com/sakanaai/text-to-lora

ICML International Conference On Machine Learning

Github

Two ways to train HyperLoRA: Reconstruction or SFT



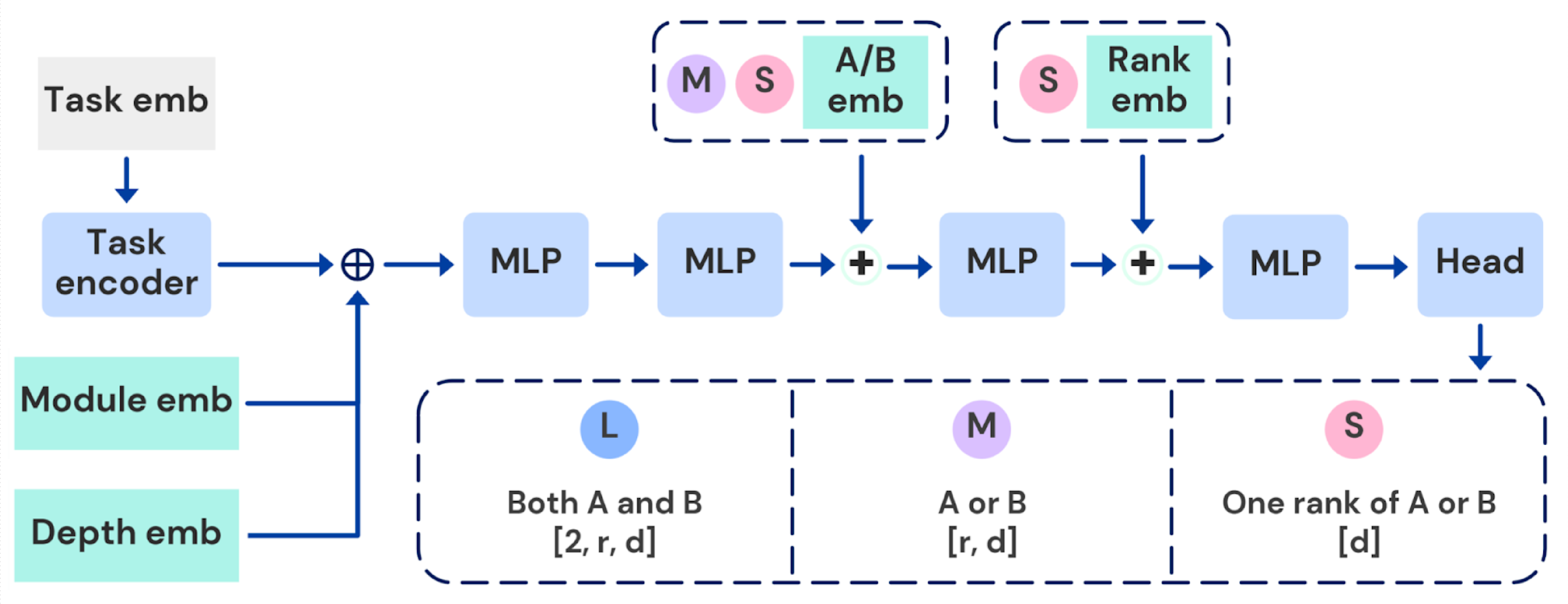
T2L can be meta-trained either by

- Distilling existing task-specific LoRAs (**reconstruction**, left)
- Directly make the adapted model output the correct response (**SFT**, right).

Both training regimes are fully differentiable without altering the base model's parameters, preserving useful capabilities of the base model.

Empirically, we find that the SFT objective makes T2L generalizes better.

Arch. variations for exploring efficiency-perf trade-off



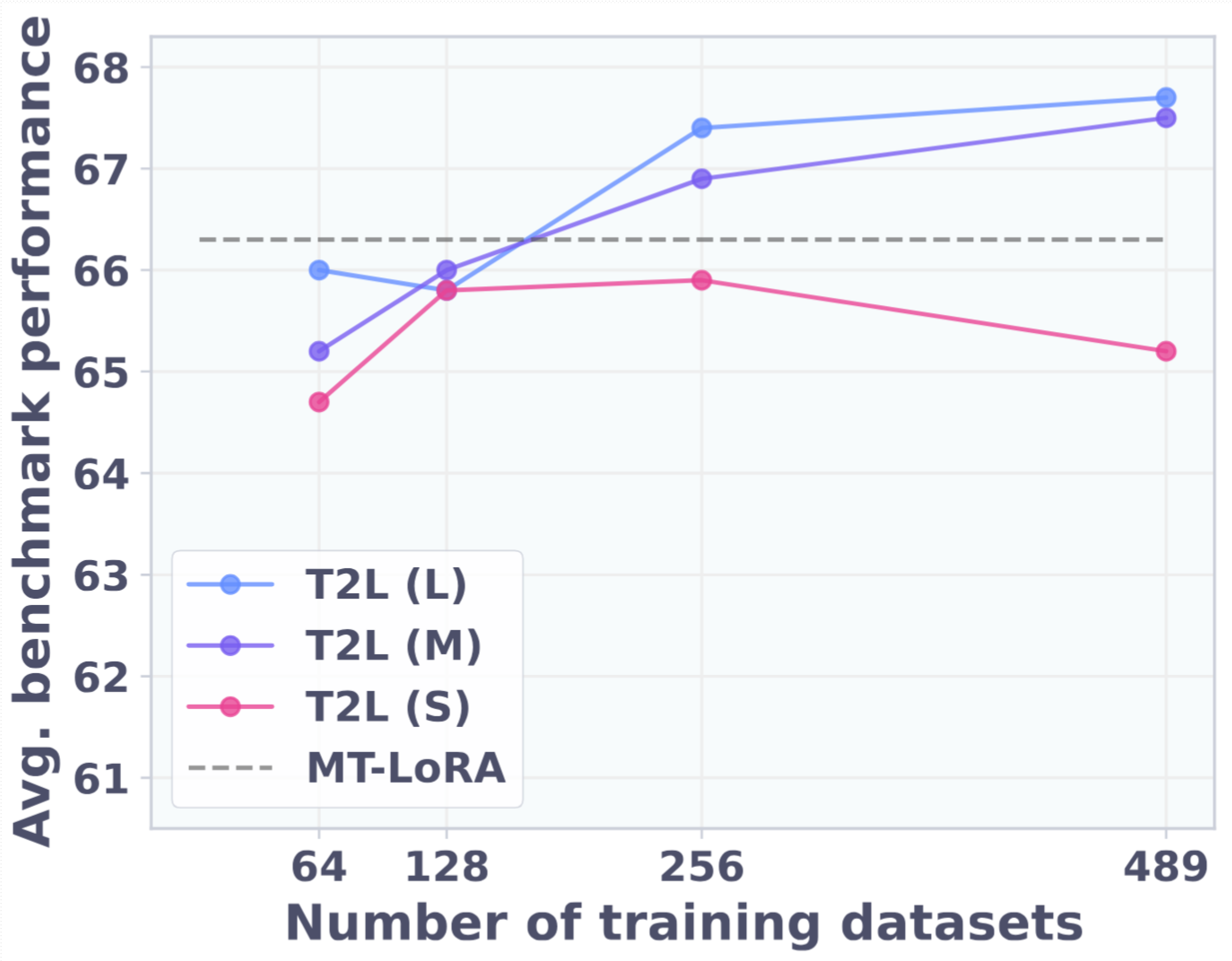
Most of parameters come from the output layer, thus, we explore different output spaces (S, M, L).

Each variant imposes a different output space on the hypernetwork, inducing different inductive biases and parameter counts.

All variants can generate a full LoRA adapter efficiently in a single forward pass by batching required embeddings.

A hypernetwork that generates task-specific LoRAs from textual task descriptions, providing zero-shot test-time adaptation

SFT T2L zero-shot performance improves as the number of training tasks increases



Reconstruction-trained T2L recovers oracle performance on the benchmarks

Table 1: Benchmark performance of T2L trained via reconstruction loss on 9 benchmark tasks. **Green highlight** indicates that T2L outperforms the benchmark-specific LoRA adapters.

	ArcC (acc)	ArcE (acc)	BQ (acc)	GSM8K (acc)	HS (acc)	OQA (acc)	PIQA (acc)	WG (acc)	MBPP (pass@1)	Avg. (9 tasks)
Base model	65.4	77.8	71.6	40.9	49.7	54.2	72.8	45.0	43.1	55.8
One-Hot Task E.										
T2L (Recon) L	76.4	89.9	89.4	**53.8**	92.6	85.0	69.7	**51.2**	**52.6**	**73.4**
T2L (Recon) M	**76.7**	89.9	89.4	53.2	92.6	85.0	69.9	**51.4**	**52.9**	**73.4**
T2L (Recon) S	75.2	88.8	87.4	50.9	89.1	75.6	**83.9**	**58.1**	48.1	73.0
Task Description E.										
T2L (Recon) L	76.6	89.8	89.4	**53.9**	92.6	85.0	69.6	**51.2**	51.8	73.3
T2L (Recon) M	76.5	89.9	89.4	**53.9**	92.5	84.9	**70.4**	**51.6**	**52.8**	**73.5**
T2L (Recon) S	75.4	88.8	87.8	49.1	89.7	76.7	**84.2**	**56.9**	48.0	73.0
Task-specific LoRAs	76.6	89.9	89.4	53.5	92.6	85.0	69.9	51.1	52.1	73.3

Table 5: T2L trained via reconstruction on 9 tasks performs well when given aligned task descriptions. Unaligned descriptions produce lower benchmark performance.

	Aligned		Unaligned	
	Train	Eval	Train (random)	Random strings
T2L L	73.3	73.6	49.1	68.2
T2L M	73.5	70.2	49.5	68.5
T2L S	73.0	72.9	55.7	53.9
Avg.	**73.3**	**72.2**	**51.4**	**63.5**

Description-task alignment

We vary the description types during evaluation and categorize them into two categories: Aligned and Unaligned.

Unaligned LoRA produce lower benchmark performance, suggesting specialization of LoRA generation by T2L.

t-SNE projection of activations of SFT T2L: Generating LoRAs tailored to specific tasks