Pre-training of Foundation Adapters for LLM Fine-tuning

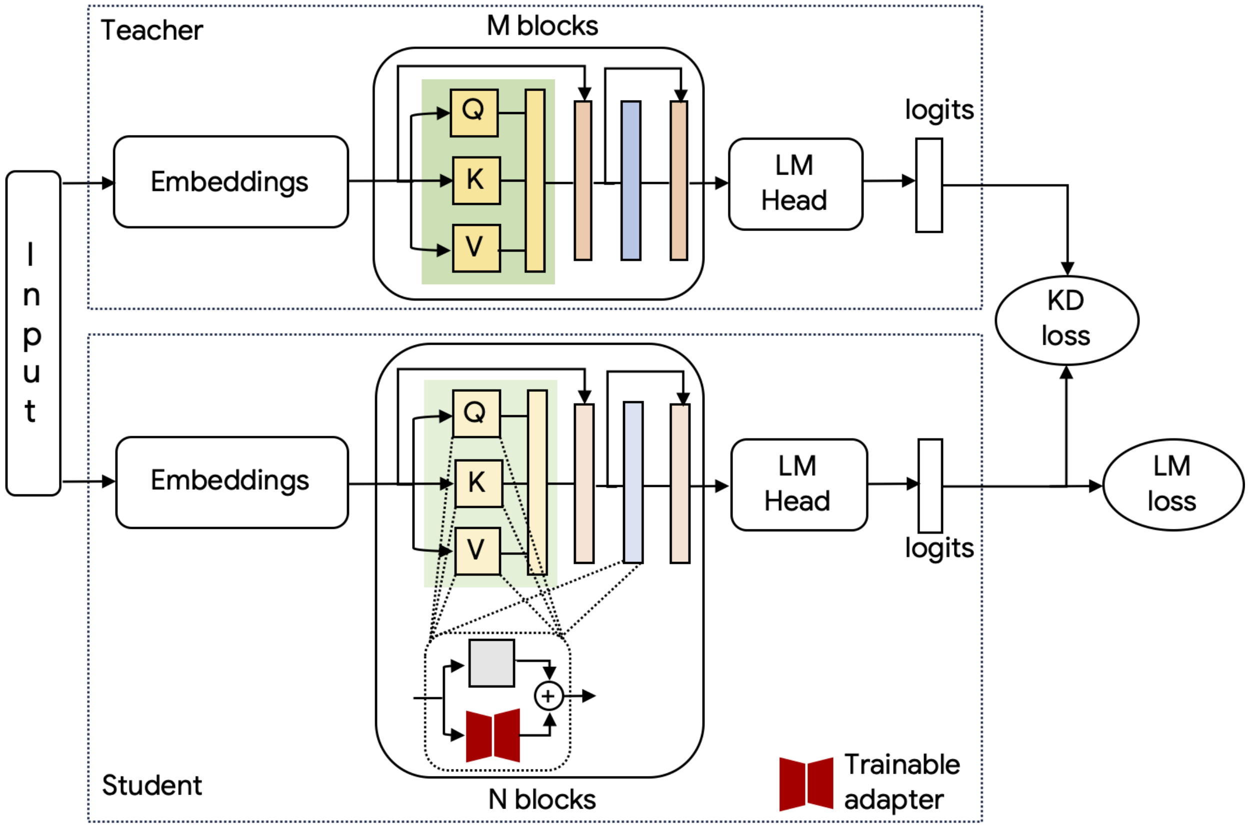
Linh The Nguyen and Dat Quoc Nguyen

Introduction

- Problem: Adapter-based fine-tuning methods insert small, trainable adapters into frozen pre-trained LLMs. Reduce computational costs but:
  - Sensitive to initialization.
  - Suffered from training instability.
- Solution: Pre-trained foundation adapters.

Pre-training of Foundation Adapters

- Combine continual pre-training (CPT) and knowledge distillation (KD) to pre-train foundation adapters.
- Initialize adapters with their pre-trained versions.



Experiments

- Results:
  - The effectiveness of knowledge distillation.

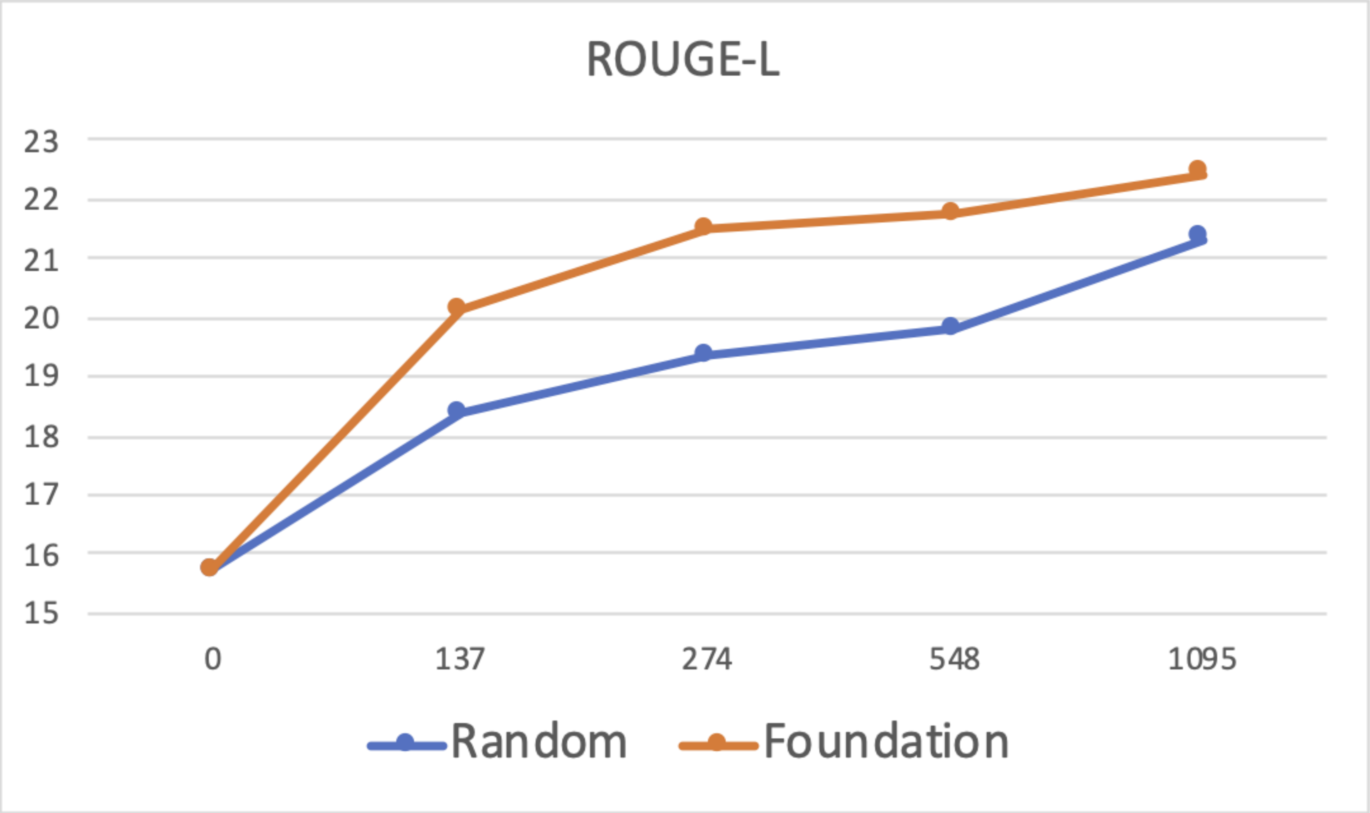
Model	LoRA rank	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	Average
Llama3.2-1B	N/A	39.51	65.68	31.12	37.58	62.43	6.97	40.55
Llama3.2-1B + CPT	8	39.08	65.70	31.11	39.35	**63.69**	6.90	40.97
Llama3.2-1B + CPT + KD	8	**39.85**	**65.99**	**31.89**	**40.19**	62.90	**8.87**	**41.62**

  - The effectiveness of difference LoRA ranks.

Model	LoRA rank	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	Average
Llama3.2-1B-Instruct	N/A	41.21	59.63	**45.54**	43.80	61.80	32.90	47.48
Llama3.2-1B-Instruct + CPT + KD	8	41.47	62.55	44.91	**45.23**	62.04	34.80	48.50
Llama3.2-1B-Instruct + CPT + KD	32	41.98	62.86	44.74	44.54	61.48	35.78	48.56
Llama3.2-1B-Instruct + CPT + KD	64	41.55	63.10	44.42	44.62	**62.12**	**37.91**	**48.95**
Llama3.2-1B-Instruct + CPT + KD	128	**42.41**	**63.61**	44.69	43.65	**62.12**	35.25	48.62

Experiments

- Pre-trained foundation adapter weights vs Random initialization in a downstream summarization task.



Thank you for your listening!