Conditional Adapters: Parameter-efficient Transfer Learning with Fast Inference

Tao Lei   Junwen Bai   Siddhartha Brahma   Joshua Ainslie   Kenton Lee   Yanqi Zhou

Nan Du   Vincent Y. Zhao   Yuexin Wu   Bo Li   Yu Zhang   Ming-Wei Chang

Google

Nov 13, 2023

TL;DR

- Propose a novel transfer learning method, **conditional adapters (CoDA),** which can achieve both **parameter efficiency** and **computation efficiency**
- Validated across Language, Vision and Speech domains
- 2x to 8x inference speed-up with moderate to no accuracy loss

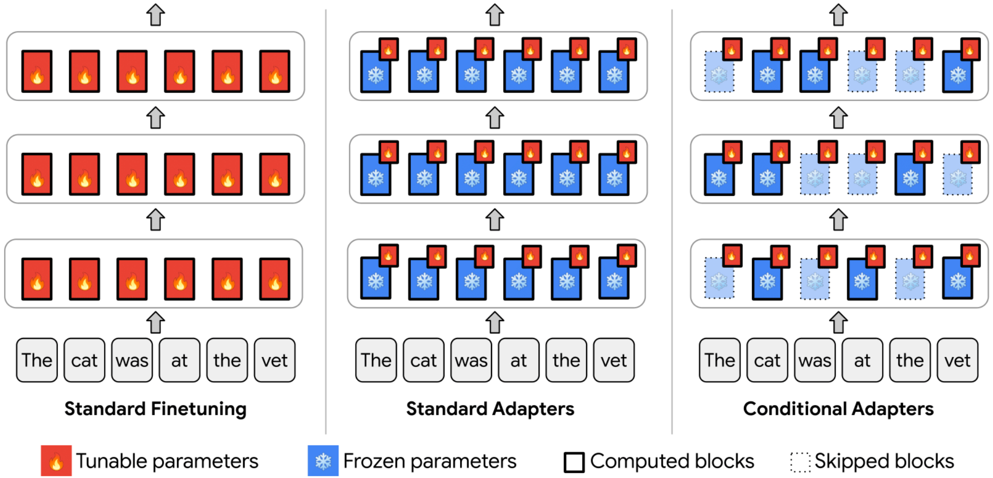
Motivations

- **Large Language Models (LLM)**
  - Expensive to finetune
  - Slow inference speed
- **Adapter Finetuning**
  - Only update a small subset of parameters
  - Parameter-efficient with (slightly) slower inference speed
- **Pruning**
  - Deterministically delete a certain portion of parameters
  - Knowledge forgetting

Model design of CoDA

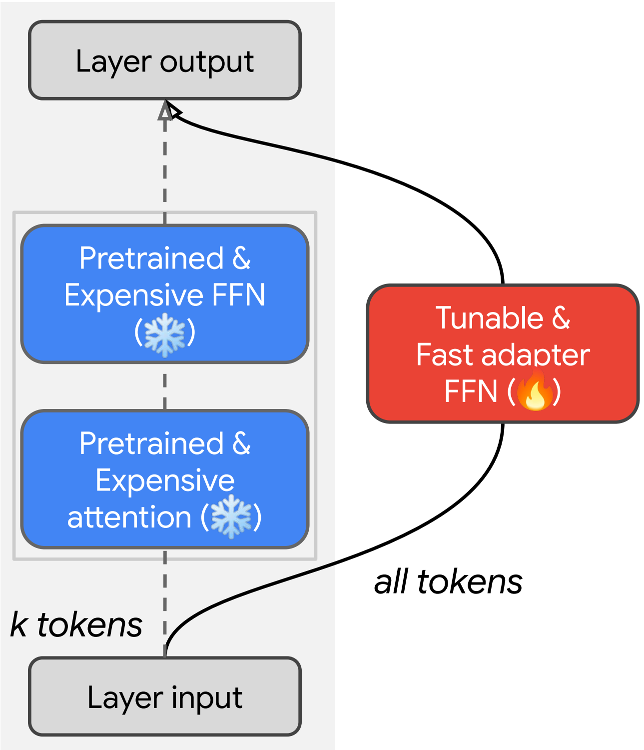
- **Adapter Branch**
  - Light-weight feed-forward parallel adapter
  - All n tokens are processed
- **Conditional Branch**
  - Only k tokens are selected and processed by the frozen pretrained transformer layer, where $k \ll n$
- **Generalizable to other types of adapters like LoRA**

Conditional Adapters



- **Left** all parameters are tunable and computation is dense
- **Center** a small set of new tunable parameters while the computation is dense
- **Right** CoDA sparsely activates computation with small amount of new parameters

Conditional Adapters



Learnt Router

For each transformer layer, we learn a weight vector $\boldsymbol{w}$ to compute the dot-product with the token representation X. Then the top-k scores are selected.

Process k tokens

- These k tokens are selected and updated through the frozen transformer layer
- For the remaining $n - k$ tokens, they remain intact. The output X' only differs from X on those k tokens.

Layer Output

All n tokens have adapter updates and contributes to the final output

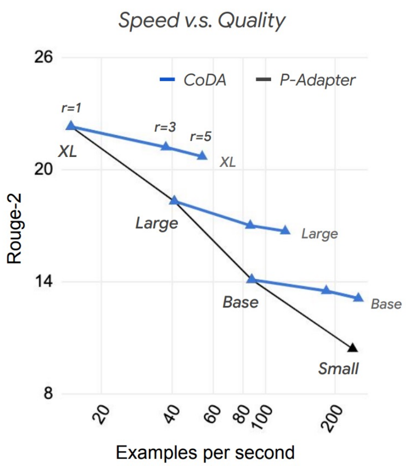
$$Y = X' + Z_{adapter}$$

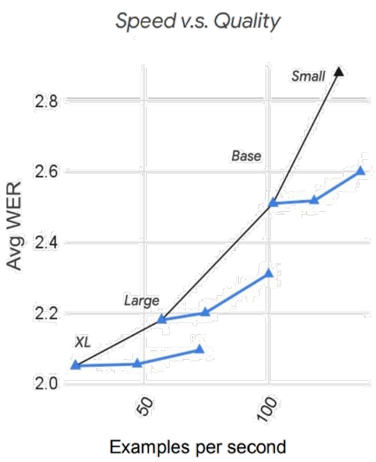
Experiments

- ***Three domains:*** *NLP, Speech, Vision*
- ***Speed-up:*** *2.2x - 8x*
- ***New params:*** *< 3%*

	New param	MNLI (text)	
		Acc ↑	Speedup
P-Adapter	0.4%	91.5	1.0x
CoDA	0.4%	90.7	**3.2x**
	New param	OCR-VQA (vision)	
		EM ↑	Speedup
P-Adapter	2.8%	67.5	1.0x
CoDA	2.8%	67.6	**8.0x**
	New param	Librispeech (speech)	
		WER ↓	Speedup
P-Adapter	2.5%	1.4/2.7	1.0x
CoDA	2.5%	1.4/2.8	**2.2x**

Experiments

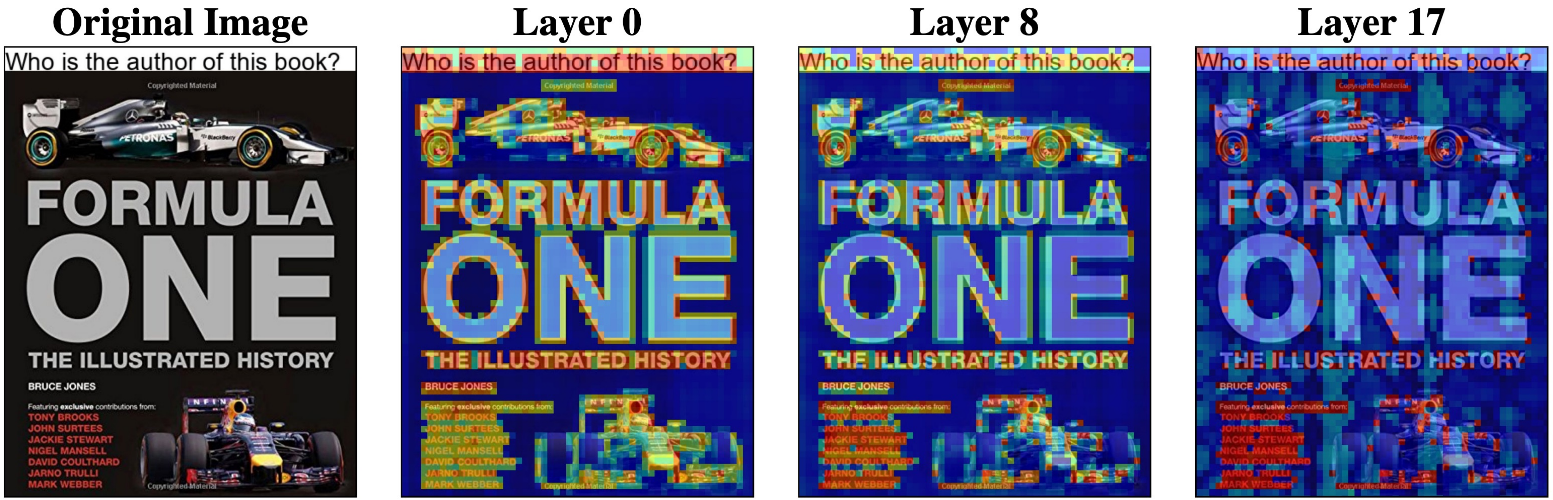




The scaling of CoDA on the Xsum and LibriSpeech

- Same speed, better quality
- Same quality, faster speed

Experiments



- Visualization of routing preferences
- Warmer colors represent higher scores
- Router prefers diverse coverage in early layers, but converges to sparse and representative patches in later layers

Experiments

Model	r	OCRVQA		DocVQA		Screen2Words	
		EM	Speedup	ANLS	Speedup	CIDEr	Speedup
Parallel Adapter	-	67.5	1×	70.8	1×	110.2	1×
CoDA	4	68.2	4.6×	71.8	4.6×	111.6	4.6×
CoDA	8	67.6	8.0×	66.6	8.0×	108.1	8.0×
CoDA	16	66.9	13.5×	56.6	12.1×	109.0	12.5×
CoDA	32	64.4	19.4×	42.5	16.7×	104.2	17.8×

Quality-speed tradeoff on a pretrained Pix2Struct model

Thanks for listening!