TokenDrop: Token-Level Importance-Aware Backward Propagation Skipping for Efficient LLM Fine-Tuning

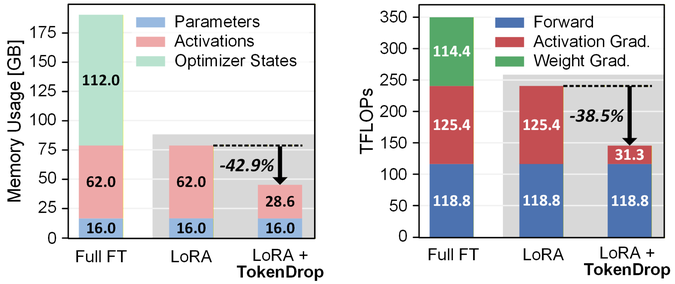
Beomseok Kim, Sol Namkung, Dongsuk Jeon Seoul National University, Seoul, South Korea

ICML International Conference On Machine Learning

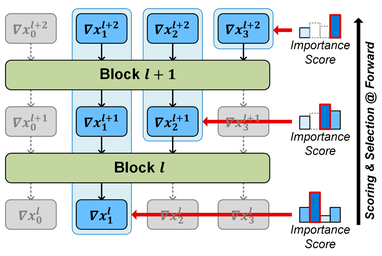
Introduction

- Parameter-efficient fine-tuning (PEFT) methods, such as LoRA, reduce parameter-related overhead, but LLM fine-tuning remains memory- and compute-intensive.
- Two activation-related bottlenecks remain: activation memory and activation-gradient computation for backpropagation.
- TokenDrop targets both with importance-aware token-level backward skipping, achieving up to 42.9% memory reduction and up to 1.50× training speedup.

Memory usage and computation breakdown for LLaMA3.1-8B fine-tuning



Core Mechanism



Full Forward, Selective Backward

- TokenDrop performs backward propagation only through a retained token set S, while keeping a full forward pass.
- Activations of unselected tokens are evicted, reducing activation memory usage.

Importance-Aware Token Retention

- TokenDrop determines S using token-level importance, rather than block-wise or random token selection.

Methodology

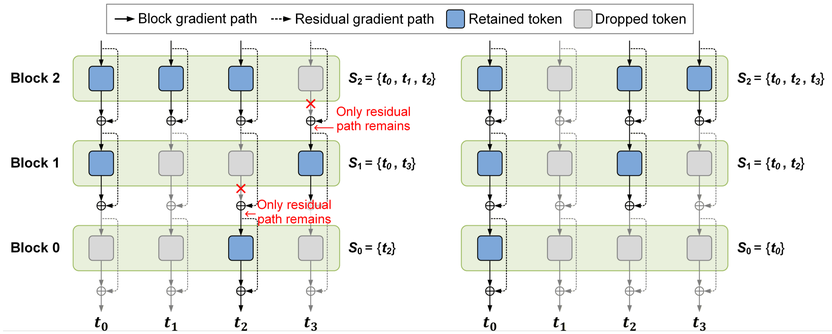
Lightweight Token Importance Scoring

- MHA/FFN blocks update the residual stream as: $x_{l+1} = x_l + f_l(x_l)$.
- TokenDrop scores each token by the ℓ_2-norm of its residual update.

$$I_{l,t} = \left\| f_l(x_{l,t}) \right\|_2$$

- The ℓ_2-norm criterion is gradient-free and computationally lightweight.

Cumulative Token Selection



(a) Independent selection (b) Cumulative token selection (Ours)

- Independent per-block top-k selection can produce inconsistent token sets and incomplete gradient paths, resulting in training instability.
- TokenDrop enforces nested retained token sets to preserve the gradient path consistency.

$$S_0 \subseteq S_1 \subseteq \cdots \subseteq S_{L-1}$$

Lazy Selection Scheduling

- Immediate selection relies only on local block-level scores.
- TokenDrop accumulates importance scores across a group of blocks and selects tokens only when nearing the memory budget for globally informed token selection.
- Accumulate scores → Select top-k tokens near the memory budget → Evict activations of unselected tokens

Experimental Results

5-Shot MMLU Accuracy and Training Costs

Model	Method	Peak Memory	SpeedUp	Accuracy (%) Hum.	STEM	Social	Other	Avg.
LLaMA2-7B	Not Tuned	-	-	43.3	37.1	51.8	52.8	45.9
	LoRA	63.9 GB	1.00×	50.5±0.4	42.4±0.6	61.4±0.6	59.5±0.5	53.1±0.2
	LoRA+DropBP	49.9 GB	1.29×	48.0±1.2	41.3±0.6	58.4±1.4	58.1±1.1	51.0±0.7
	LoRA+TokenTune	38.8 GB	1.31×	46.4±1.4	40.8±0.4	57.9±0.2	57.8±0.2	50.1±0.5
	LoRA+TokenDrop	**37.3 GB**	**1.35**×	51.2±0.2	43.7±1.2	62.7±0.5	60.3±0.4	**54.1±0.3**
LLaMA2-13B	Not Tuned	-	-	53.1	44.2	62.8	60.8	54.9
	LoRA	66.8 GB	1.00×	55.8±0.2	47.1±0.5	66.8±0.6	64.3±0.5	**58.1±0.4**
	LoRA+DropBP	54.9 GB	**1.28**×	53.4±2.5	45.8±0.6	64.1±1.2	62.1±1.5	56.0±1.5
	LoRA+TokenTune	**50.4 GB**	1.19×	52.8±0.9	46.0±1.1	65.3±0.4	63.2±0.6	56.3±0.3
	LoRA+TokenDrop	50.8 GB	**1.28**×	55.9±0.8	46.7±0.2	66.1±0.7	64.1±0.4	57.9±0.4

MT-Bench Score and Training Costs (LLaMA3.1-8B)

Method	Memory	SpeedUp	Writing	Role.	Reason.	Math	Coding	Extract.	STEM	Human.	Avg.
Not Tuned	-	-	5.35	5.20	2.55	2.75	2.00	2.70	5.40	3.60	3.69
LoRA	78.0 GB	1.00×	5.55±0.48	5.70±0.43	3.25±0.18	2.62±0.36	2.90±0.13	4.52±0.51	6.10±0.17	6.50±0.15	4.64±0.16
LoRA+DropBP	48.1 GB	1.30×	4.95±0.74	5.08±0.53	3.40±0.59	2.45±0.30	2.68±0.25	3.43±0.88	5.73±0.58	5.80±0.28	4.19±0.26
LoRA+TokenTune	45.7 GB	1.27×	4.15±1.21	4.42±0.38	2.43±0.03	2.48±0.63	2.35±0.00	3.18±0.53	5.42±0.40	4.87±1.01	3.66±0.32
LoRA+TokenDrop	**44.6 GB**	**1.38**×	5.73±0.13	5.38±0.14	3.13±0.10	2.95±0.20	2.93±0.19	4.65±0.82	6.02±0.16	6.45±0.05	**4.66±0.13**

Pass@1 on HumanEval+ and Training Costs (LLaMA3.1-8B)

Method	Memory	SpeedUp	Pass@1 (%)
Not Tuned	-	-	30.5
LoRA	66.8 GB	1.00×	**36.8±1.2**
LoRA+DropBP	53.5 GB	**1.24**×	35.0±1.1
LoRA+TokenTune	50.9 GB	1.16×	35.4±0.9
LoRA+TokenDrop	**47.5 GB**	**1.24**×	36.6±0.5

Ablation Study of the Proposed Selection Methods

Method	MMLU Acc. (%)	MMLU Δ	CSR170K Acc. (%)	CSR170K Δ
Baseline	40.1	-	69.7	-
+ Cumulative Selection	53.3	+13.2	70.2	+0.5
+ Lazy Scheduling	**54.1**	+0.8	**71.3**	+1.1

Trade-Off Between Training Efficiency and 5-Shot MMLU Accuracy When Fine-Tuning LLaMA2-7B