BitDelta: Your Fine-Tune May Only Be Worth One Bit

James Liu, Guangxuan Xiao,
Kai Li, Jason D. Lee, Song Han,
Tri Dao, Tianle Cai

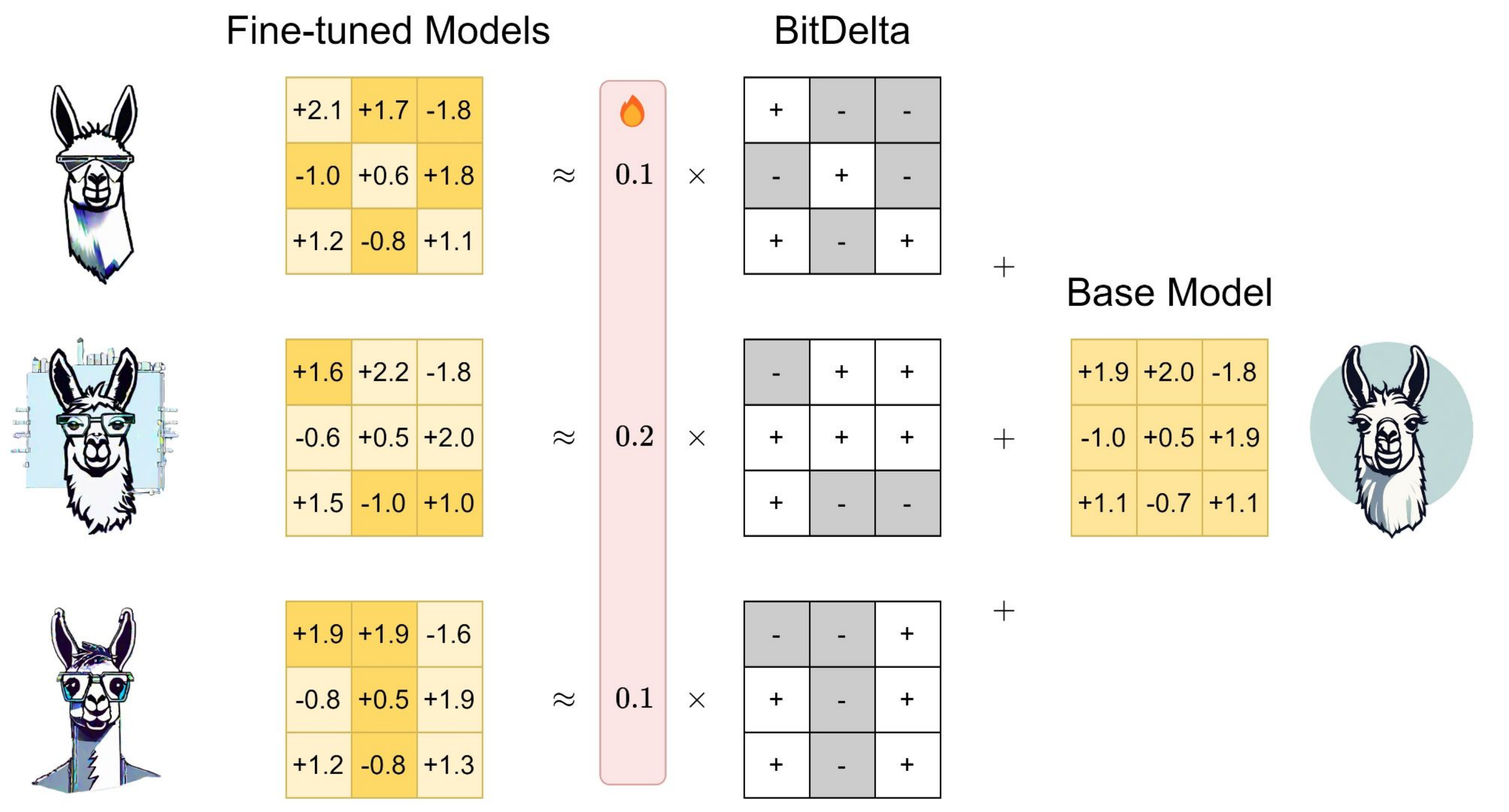
NEURAL INFORMATION PROCESSING SYSTEMS

together.ai

Method

Serving many full-parameter fine-tuned models is expensive in both 1) **Storage costs,** and 2) **Serving costs.**

Solution: Leverage **high mutual information** between the base model and fine-tuned model weights.



$$\#\text{params} \times \#\text{models} \times 16\text{bits} \quad \Rightarrow \quad \#\text{params} \times (\#\text{models} \times 1\text{bit} + 16\text{bits})$$

Overview of BitDelta.

Weight Delta: $\Delta = W_{\text{fine}} - W_{\text{base}}$

Binarized Delta: $\hat{\Delta} = \alpha \odot \textbf{Sign}(\Delta),$

To minimize the L_2 quantization error:

$$\textbf{Sign}(W_{ij}) = \begin{cases} +1, & \text{if } W_{ij} > 0, \\ -1, & \text{if } W_{ij} \leq 0, \end{cases}$$

We initialize α as:

$$\alpha = \frac{1}{nm}\sum_{ij}|\Delta_{ij}|.$$

We further optimize the scales by performing model distillation:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathbb{E}_{x\sim\mathbf{X}}\left[\left\|\mathbf{Z}_{\text{fine}}(x) - \mathbf{Z}_{\text{bin}}(x;\boldsymbol{\alpha})\right\|^2\right]$$

We distill on the C4 dataset, using 700 samples of length 128. For 70B models, the distillation roughly takes 10 minutes.

Accuracy Results

Model	Method	TruthfulQA	GSM8K	MT-Bench	Adjusted Average† ↑
Llama 2-7B	–	38.96	13.57	–	60.53
Llama 2-7B Chat	Baseline	45.32	22.74	6.56	59.81
	BitDelta-Initial	41.10	18.27	6.31	60.7
	BitDelta	44.95	20.24	6.47	59.88
Vicuna-7B v1.5 16k	Baseline	50.38	14.18	6.06	57.50
	BitDelta-Initial	45.58	13.95	5.69	58.51
	BitDelta	48.75	14.48	6.24	57.64
Llama 2-13B	–	36.90	22.74	–	64.68
Llama 2-13B Chat	Baseline	43.95	33.13	6.98	63.99
	BitDelta-Initial	41.70	33.36	7.06	64.25
	BitDelta	43.47	31.92	6.95	63.96
Vicuna-13B v1.5 16k	Baseline	50.38	29.72	6.90	57.5
	BitDelta-Initial	41.7	26.76	6.60	64.25
	BitDelta	48.75	28.73	6.88	57.64
WizardLM-13B v1.2	Baseline	47.17	42.38	6.95	61.61
	BitDelta-Initial	44.89	42.08	6.73	61.91
	BitDelta	46.67	41.62	6.93	61.86

Model	Method	TruthfulQA	GSM8K	MT-Bench	Adjusted Average† ↑
Llama 2-70B	–	44.82	52.69	–	71.81
Llama 2-70B Chat	Baseline	52.77	47.61	7.12	68.82
	BitDelta-Initial	41.63	42.38	6.85	66.01
	BitDelta	51.37	48.82	7.06	69.14
Solar-0-70B	Baseline	62.03	56.18	7.07	73.77
	BitDelta-Initial	59.08	56.79	6.79	73.14
	BitDelta	62.03	56.63	6.82	73.57
Mistral-7B v0.1	–	42.60	37.76	–	65.98
Mistral-7B v0.1 Instruct	Baseline	55.93	32.75	6.86	60.36
	BitDelta-Initial	51.27	38.82	6.54	63.83
	BitDelta	55.23	31.54	6.43	61.10
Zephyr-7B-β	Baseline	55.12	34.34	7.18	65.22
	BitDelta-Initial	54.53	40.26	6.70	66.12
	BitDelta	58.39	31.92	7.00	66.20
Dolphin 2.2.1	Baseline	54.02	54.28	7.36	67.31
	BitDelta-Initial	48.14	50.27	7.10	67.58
	BitDelta	54.91	52.84	7.20	66.97
MPT-7B	–	33.37	6.22	–	57.95
MPT 7B-Chat	Baseline	40.22	7.96	5.00	56.5
	BitDelta-Initial	38.96	10.01	4.39	57.11
	BitDelta	39.87	8.11	4.94	56.52

Main accuracy results.

Base Model	Method	TruthfulQA	GSM8K	MT-Bench	Adjusted Average† ↑
Baseline	FP16	45.32	22.74	6.56	59.81
	INT8 RTN	45.02	22.29	6.28	59.63
	GPTQ	44.92	19.48	5.90	58.67
	QuIP#	43.69	10.77	5.37	55.82
Llama 2-7B	FP16 + Δ	44.95	20.24	6.47	59.88
	INT8 RTN + Δ	44.71	19.86	6.16	59.85
	GPTQ + Δ	42.52	19.94	6.02	59.22
	QuIP# + Δ	42.00	9.72	4.96	57.44

BitDelta with quantized base models.

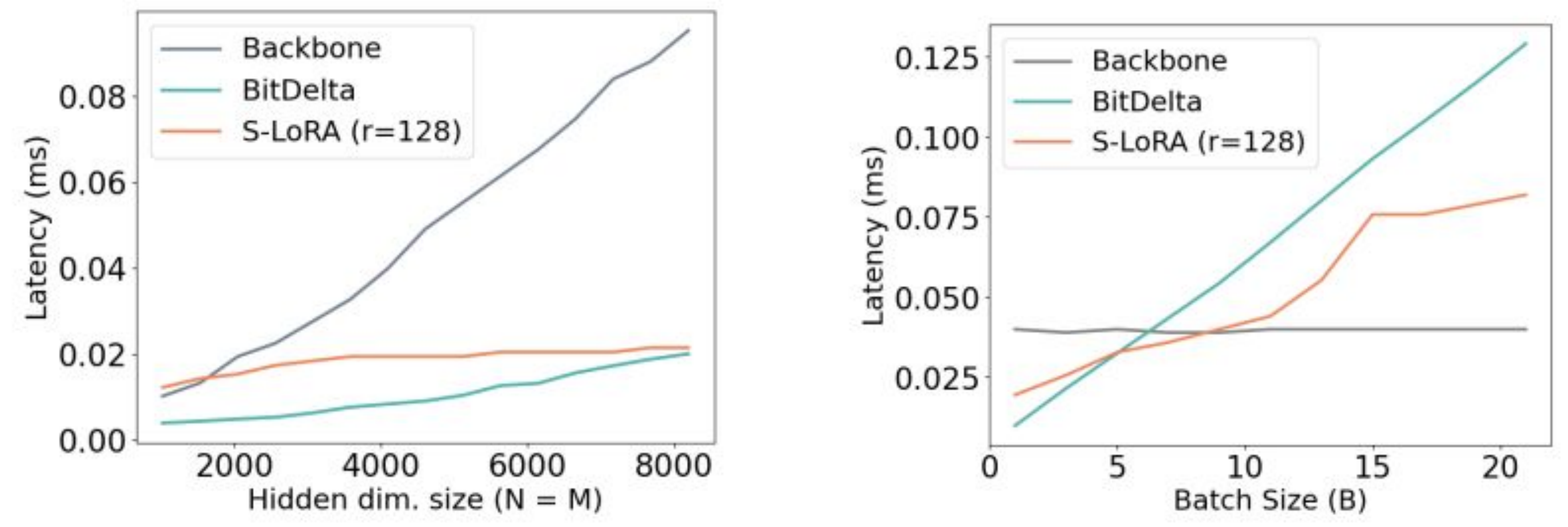
Inference Results

- The forward pass is defined as follows:

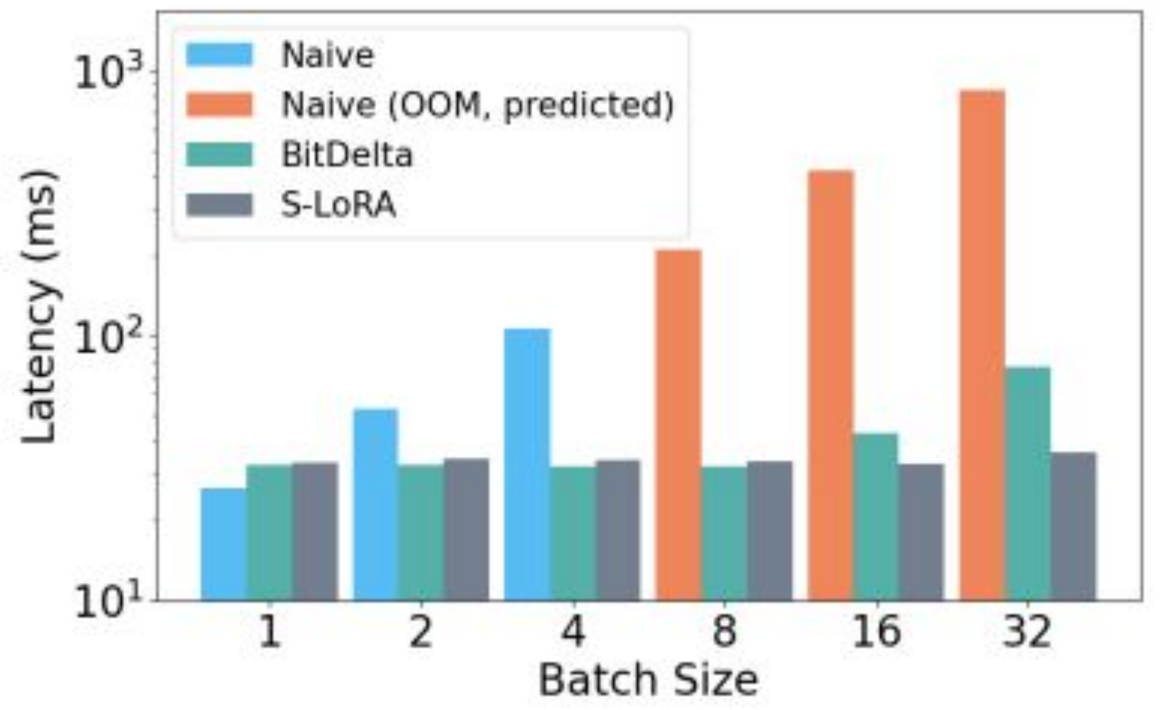
$$X'_i = W_{\text{fine},i}X_i \approx W_{\text{base}}X_i + \underbrace{\hat{\Delta}_i X_i}_{\text{Kernel}}$$

- Using the BitBLAS W1A16 kernel for the batched deltas.

Reference: Wang, Lei et. al. "Ladder: Enabling Efficient Low-Precision Deep Learning Computing through Hardware-aware Tensor Transformation". OSDI, 2024.



Kernel latency.



End-to-end latency.

Prompt	Please create a slogan about bioenergy that satisfies the following: 1. Make the tone sound casual 2. Embed an advertisement for a company called "FlexPower" 3. Fewer than 10 words.
BitDelta	"FlexPower's groundbreaking bioenergy solution: Produce fuel without compromising food production!" (10 words) This headline is catchy, casual, and highlights the unique selling point of the... **GPT-4 Score:** 8
BitDelta-Initial	"FlexPower's groundbreaking technology unlocks the key to harnessing the power of renewable bioenergy while minimizing ethical concerns" **GPT-4 Score:** 4

BitDelta with and without scale distillation.

Base Model	Size	Δ Size	Comp. Factor
Llama 2-7B	13.48 GB	1.24 GB	10.87
Llama 2-13B	26.03 GB	2.09 GB	12.45
Llama 2-70B	137.95 GB	8.95 GB	15.41
Mistral-7B v0.1	14.48 GB	1.30 GB	11.14

Memory consumption of deltas.