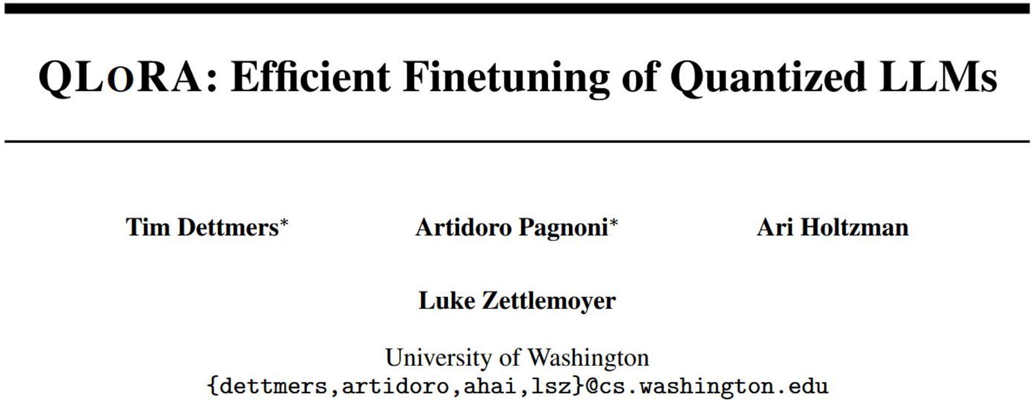
NeurIPS 2023 oral



QLoRA: Efficient Finetuning of Quantized LLMs

Tim Dettmers* Artidoro Pagnoni* Ari Holtzman

Luke Zettlemoyer

University of Washington
{dettmers,artidoro,ahai,lsz}@cs.washington.edu

- Problem / objective
    - 효율적인 파인튜닝

- Contribution / Key idea
    - QLoRA
        - 4-bit NormalFloat (NF4)
        - Double quantization
        - Paged Optimizers

전유진

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in neural information processing systems* 36 (2023): 10088-10115.

NeurIPS 2023 oral

QLoRA

Pretrained 4-bit quantized frozen LLM 을 gradient backpropagation 시키고,
그 gradient 를 Low-Rank Adapter (LoRA) 에만 적용.