BREAKING THE MEMORY BARRIER: NEAR INFINITE BATCH SIZE SCALING FOR CONTRASTIVE LOSS

Zesen Cheng[2*], **Hang Zhang**[1,2*] ✉, **Kehan Li**[2*], **Sicong Leng**[2,3], **Zhiqiang Hu**[2], **Fei Wu**[1], **Deli Zhao**[2], **Xin Li**[2] ✉, **Lidong Bing**[2]
[1]Zhejiang University, [2]DAMO Academy, Alibaba Group, [3]Nanyang Technological University,
* Equal Contribution ✉ Corresponding Author
https://github.com/DAMO-NLP-SG/Inf-CLIP

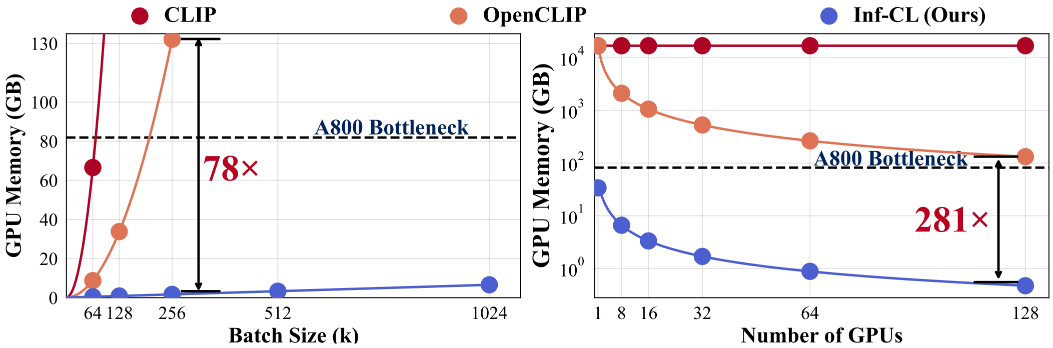

Figure 1: **GPU memory usage comparison** between **Inf-CL** and previous methods (**CLIP**, **OpenCLIP**). The dashed line marks the common GPU memory limit. Memory costs exceeding the bottleneck of 80G A800 are estimated by curve fitting. ❶ Left: With 8×A800, CLIP and OpenCLIP's memory consumption increases quadratically, while Inf-CL achieves linear growth, reducing memory costs by **78×** at a batch size of 256k. ❷ Right: At a batch size of 1024k, even with 128 GPUs, previous methods exceed memory limits, whereas Inf-CL reduces memory demand by **281×**.

ABSTRACT

Contrastive loss is a powerful approach for representation learning, where larger batch sizes enhance performance by providing more negative samples to better distinguish between similar and dissimilar data. However, scaling batch sizes is constrained by the quadratic growth in GPU memory consumption, primarily due to the full instantiation of the similarity matrix. To address this, we propose a tile-based computation strategy that partitions the contrastive loss calculation to arbitrary small blocks, avoiding full materialization of the similarity matrix. Furthermore, we introduce a multi-level tiling strategy to leverage the hierarchical structure of distributed systems, employing ring-based communication at the GPU level to optimize synchronization and fused kernels at the CUDA core level to reduce I/O overhead. Experimental results show that the proposed method scales batch sizes to unprecedented levels. For instance, it enables contrastive training of a CLIP-ViT-L/14 model with a batch size of 4M or 12M using 8 or 32 A800 80GB without sacrificing any accuracy. Compared to SOTA memory-efficient solutions, it achieves a two-order-of-magnitude reduction in memory while maintaining comparable speed. The code will be made publicly available.

1 INTRODUCTION

Contrastive learning serves as a foundational technique across various applications, such as multi-modality retrieval (Radford et al., 2021; Luo et al., 2022; Girdhar et al., 2023), self-supervised representation learning (Chen et al., 2020a; He et al., 2020; Gao et al., 2022), and dense text retrieval (Wang et al., 2022). It learns an embedding space in which similar data pairs stay close while