APE: Aligning Pretrained Encoders to Quickly Learn Aligned Multimodal Representations

E. Rosenfeld, P. Nakkiran, H. Pouransari, O. Tuzel, F. Faghri
NeurIPS 2022 · Apple Inc. & Carnegie Mellon University

Question

How much paired data and compute is required to learn well-aligned multimodal representations?

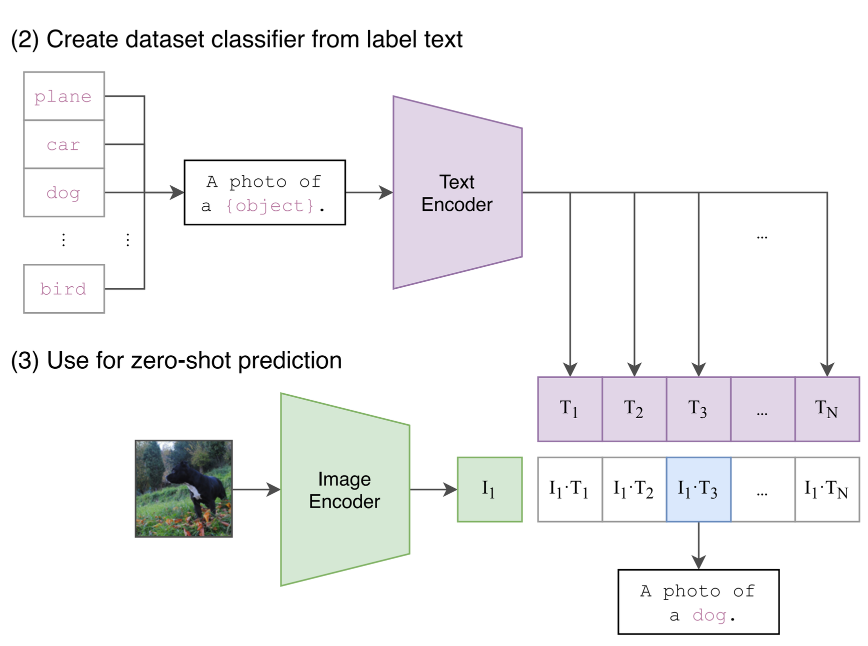
Motivations

- Latest advances driven by **massive compute** on ever-growing modality-coupled datasets [1, 2]
- But we often have unimodal encoders already trained! **How to use them efficiently?**
- The training sets are also very **noisy**. Can we do better with smaller, curated datasets?

Multimodal Learning

Given paired images and captions, goal is to learn a joint image and text embedder such that representations are semantically aligned.

Alignment is tested with **zero-shot classification accuracy**: embed each class name with the text encoder and classify each image according to which class embedding is closest.



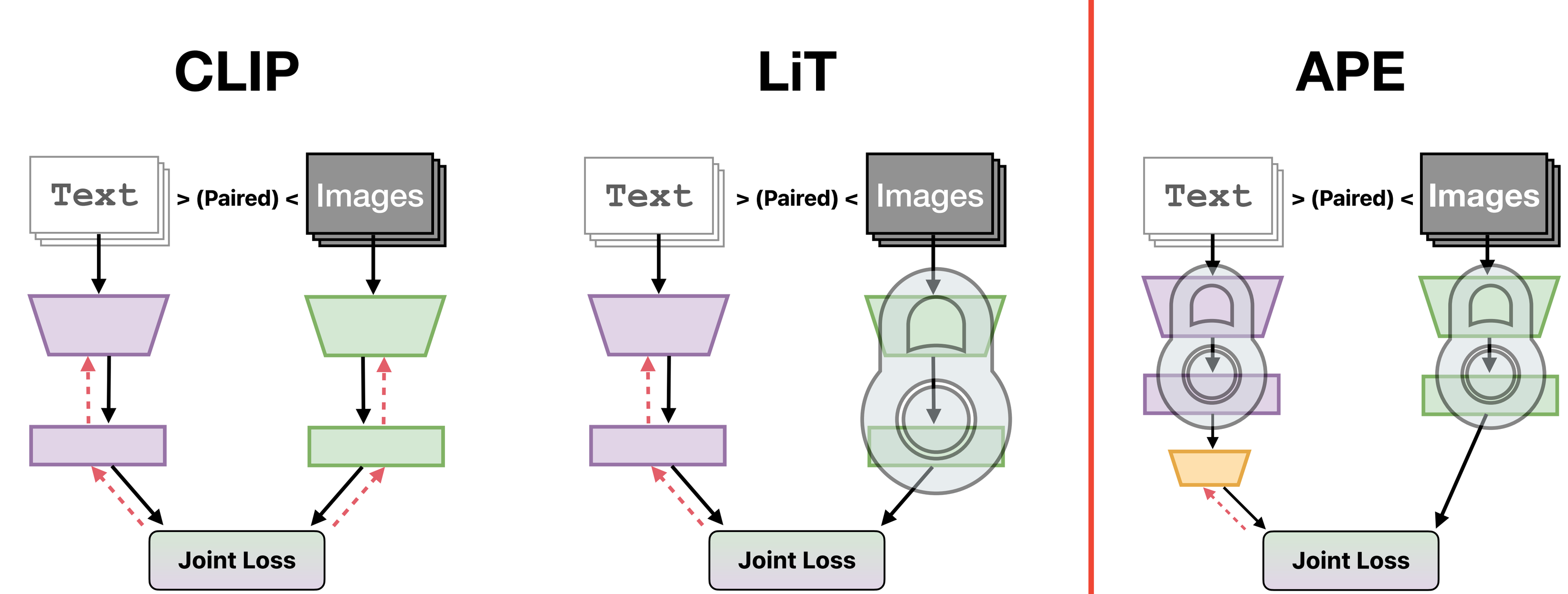
(Image from [1])

Prior Work

Popular approach is to use *Contrastive Language-Image Pretraining* (CLIP). [1]

Recent work [3] aligns a text encoder to a **frozen, pretrained** image encoder for better performance.

We consider a natural extension: freezing **both** encoders and training a small MLP.

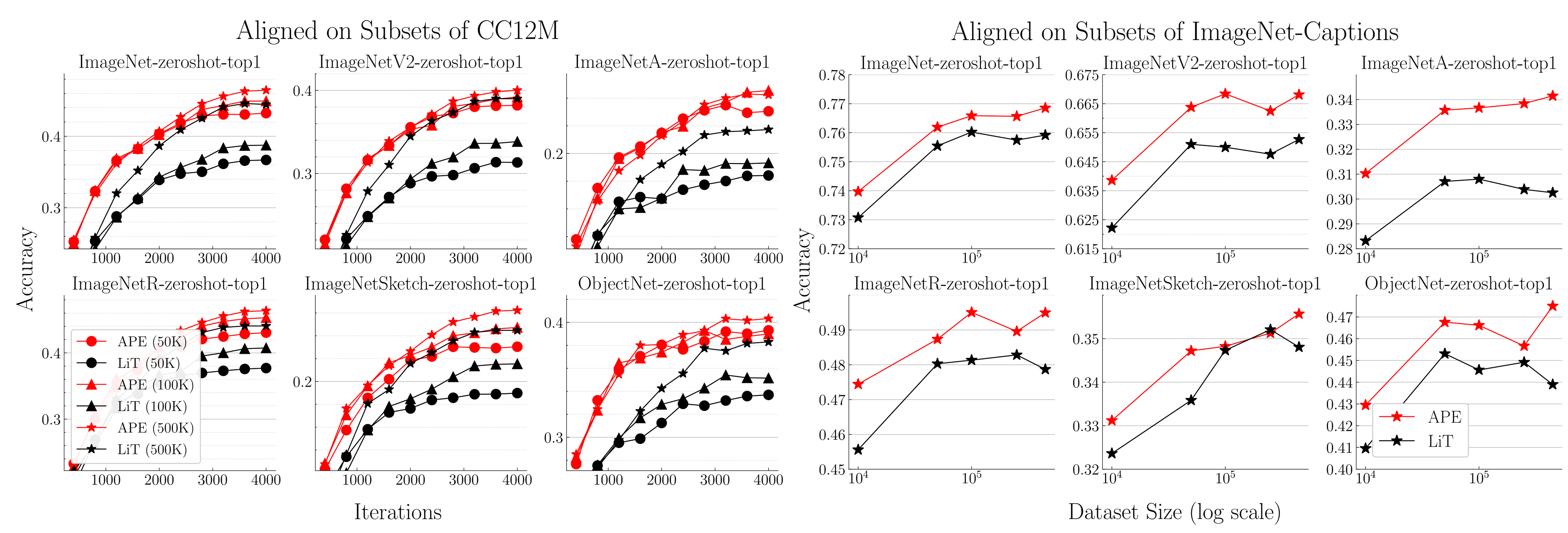


(This work)

Main Findings

- We show that **frozen pretrained encoders** can be closely aligned with a **small auxiliary MLP** (4-6 layers). This approach is **cheaper to train**, **more sample efficient**, and **less prone to overfitting** on small datasets.
- Using ImageNet-Captions,[4] we achieve **better downstream performance** on a variety of tasks with **two orders of magnitude** less training time and data.

Results

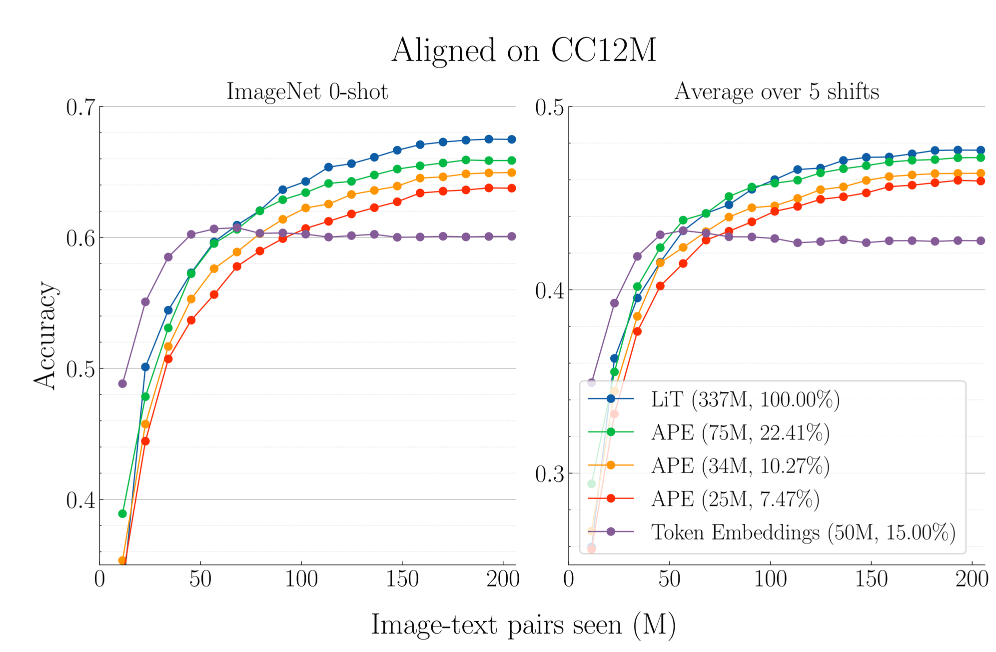


- APE achieves higher zero-shot accuracy with fewer iterations across a wide range of training set sizes.
- Though both methods use the full encoder, APE trains ~75% fewer parameters and does not backdrop through the text encoder, requiring less memory.

Using Small, Curated Datasets

- Prior work scales up collection of **noisy, task-agnostic** paired data
- We ask: how valuable is curation of smaller datasets which are **more relevant to the downstream task?**
- We show that a much smaller dataset can achieve **better downstream performance** with **substantially shorter training times**.

With Abundant Training / Compute



- When training data / compute / memory are abundant, full fine-tuning is still preferable.
- But APE gets surprisingly close, **for much cheaper!**
- Future work to try to close this gap.

References:

[1] Learning Transferable Visual Models From Natural Language Supervision. Radford et al. 2021
[2] Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. Jia et al. 2021
[3] LiT: Zero-Shot Transfer with Locked-image text Tuning. Zhai et al. 2021
[4] Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). Fang et al. 2022