Exploiting Distribution Constraints for Scalable and Efficient Image Retrieval

Mohammad Omama, Po-han Li, Sandeep Chinchali
The University of Texas at Austin

TL;DR: **Image Retrieval with Foundation Models: Better, Faster, Distribution-Aware!**

Motivation

Two key problems with image retrieval:

- State-of-the-art (SOTA) image retrieval methods train large models separately for each dataset. This is **not scalable**.
- SOTA image retrieval methods use large embeddings. Retrieval speed is directly proportional to embedding size. This is **not efficient**.

Our work addresses two key questions:

- **Q1 (Scalability)**: How can we enhance the performance of these off-the-shelf models in a completely unsupervised way?
- **Q2 (Efficiency):** Is there an effective unsupervised dimensionality reduction method that strongly preserves the similarity structure of the full embeddings, and is adaptive, i.e., does not need to be trained for each dimension separately?

Contributions

- To address **Q1 (Scalability)**, we propose **Autoencoders with Strong Variance Constraints (AE-SVC)**. AE-SVC trains an autoencoder while enforcing three constraints on the latent space: an orthogonality constraint, a mean centering constraint, and a unit variance constraint.
- We empirically show and mathematically prove that these constraints cause **a shift in the cosine similarity distribution**, making it more discriminative.
- To address **Q2 (Efficiency)**, we propose **Single Shot Similarity Space Distillation ((SS)$_2$D)**. (SS)$_2$D aims at reducing embeddings to smaller ones while preserving their similarity relationships. The embedding learned with (SS)$_2$D is adaptive, *i.e.*, smaller segments of the embedding also perform well in retrieval tasks.

Methodology

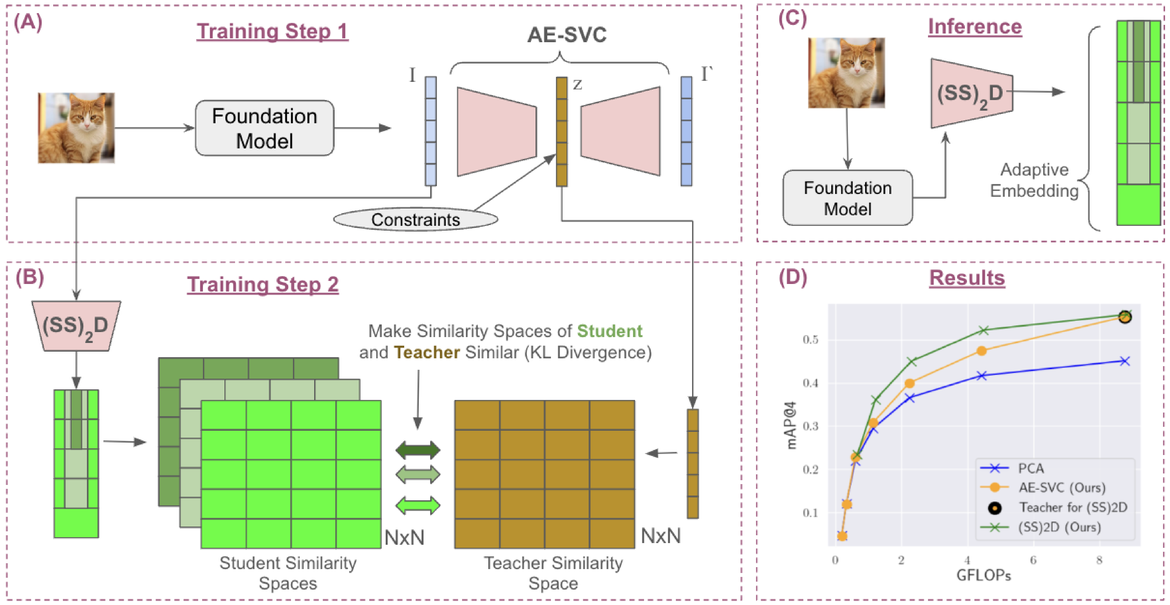
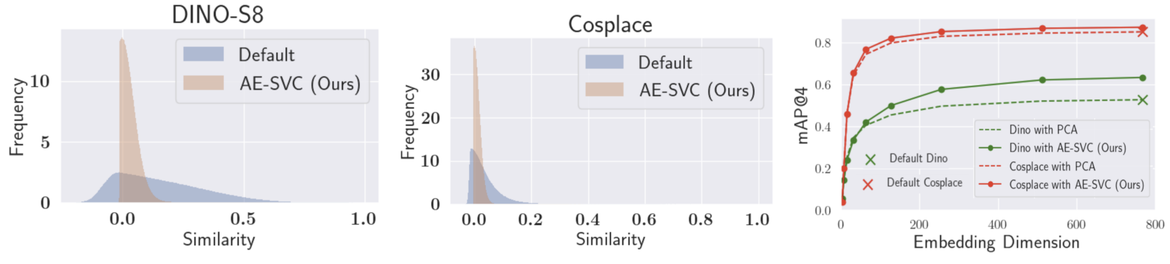


Figure: **Two-step pipeline for the proposed approach. (A)** AE-SVC trains an autoencoder with our constraints to improve foundation model embeddings. **(B)** (SS)$_2$D) uses the improved embeddings from AE-SVC to learn adaptive embeddings for improved retrieval at any embedding size. **(C)** Once trained, (SS)$_2$D can be directly applied to foundation model embeddings to generate adaptive embeddings for improved retrieval. **(D)** AE-SVC (orange) boosts performance significantly, while (SS)$_2$D (green) further enhances results with smaller embeddings. DINO (blue) achieves optimal performance at 9 GLOPs, whereas (SS)$_2$D on top of AE-SVC achieves similar performance at only 2.5 GLOPs.

Impact of the Proposed Constraints on the Cosine Similarity Distribution



(a) (b) (c)

Figure: **AE-SVC reduces the variance of cosine similarity distributions in both foundation (a) and dataset-specific models (b), with a more significant shift in foundation models (a).** This results in greater improvement in retrieval performance for the foundation model (DINO) compared to the dataset-specific model (Cosplace), as shown in (c).

Results

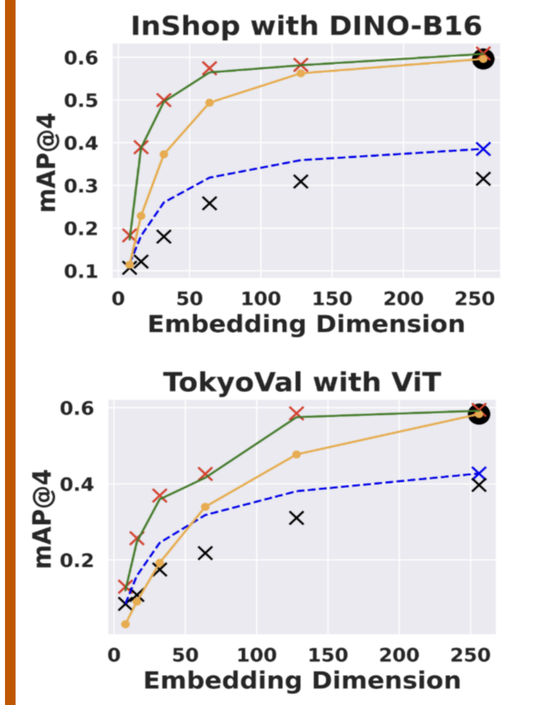


Figure: AE-SVC (yellow) consistently outperforms the off-the-shelf foundation models, i.e., PCA (blue) achieving a 15.5% average improvement in retrieval performance. Applying (SS)$_2$D over AE-SVC leads to a further performance boost at lower embedding sizes. Compared to VAE and SSD, (SS)$_2$D offers superior single-shot dimensionality reduction, achieving up to a 10% enhancement at smaller embedding sizes, closely approaching SSD's theoretical upper bound.