Asymmetric Distillation and Information Retention in Capacity-Constrained Cross-Modal Transfer

Kabir Thayani
Independent Researcher
India
thayanikabir.official@gmail.com

Abstract*—Knowledge distillation between asymmetric architectures often induces severe geometric constraints on the learned representation space. In this work, we investigate the Dimensional Collapse phenomenon when distilling a 500M parameter global Vision Transformer (CLIP ViT-B/32) into strictly capacity-constrained, local-receptive-field CNNs (0.5M to 8.0M parameters) on the CIFAR-10 dataset. By employing strictly centered Singular Value Decomposition (SVD) and Variance-based Shannon Entropy Effective Rank, we isolate true structural variance from mean-vector artifacts. Our empirical results demonstrate a capacity-agnostic phase transition: while the Teacher exhibits an Effective Rank of 88.68, all Student models experience severe dimensional collapse to an intrinsic Effective Rank of $\sim$16. By probing robustness, we uncover that this 81% reduction in effective dimensionality strips away the Teacher's inherent noise immunity (which retains 89.35% accuracy under $\sigma = 0.1$ Gaussian noise). Furthermore, information-theoretic analysis using InfoNCE reveals a critical trade-off within this bottleneck: excess Student capacity densely packs the collapsed subspace for clean data, but induces severe brittleness (43.76% at $\sigma = 0.1$). Conversely, extreme capacity constraints (0.5M parameters) act as a robust low-pass filter, preserving higher noise immunity (54.84%). Explicit input augmentation fails to restore the larger model's robustness, proving this fragility is a fundamental geometric limitation of asymmetric cosine distillation.

***Index Terms*—Representation Learning, Knowledge Distillation, Dimensional Collapse, Mutual Information, Spectral Geometry**

I. Introduction

The deployment of state-of-the-art vision-language models, such as CLIP [1], is heavily bottlenecked by their massive parameter counts. Knowledge Distillation [2] is the standard paradigm for compressing these models for edge deployment. However, transferring knowledge from a global-receptive-field Vision Transformer (ViT) into a strictly local-receptive-field Convolutional Neural Network (CNN) [3] creates a severe asymmetric bottleneck.

Previous studies have shown that embedding spaces in deep neural networks often suffer from intrinsic anisotropy and dimensional collapse [4]. However, standard spectral measurements frequently fail to center the data, inadvertently measuring the distance from the origin to the embedding cluster rather than the true structural variance. In this study, we enforce rigorous mathematical constraints, including strict embedding centering and generator-locked stochasticity, to observe the true spectral geometry of asymmetric cross-modal transfer.

We ask the following question: *Does scaling a Student network's capacity linearly expand its dimensional footprint in the Teacher's hypersphere, or does it merely increase information density within a strict geometric bottleneck?*

Our contributions are three-fold:

1) We empirically demonstrate true dimensional collapse, proving that students ranging from 0.5M to 8.0M parameters all collapse to an Effective Rank of $\sim$16, despite the Teacher spanning an Effective Rank of 88.68.
2) We measure Mutual Information retention via InfoNCE and Uniformity loss, demonstrating that capacity scaling modestly improves subspace utilization rather than subspace expansion.
3) We evaluate the "Semantic Filter" hypothesis, demonstrating a critical mechanistic trade-off between clean-data information density and high-frequency noise robustness [5] across multiple stochastic seeds.

II. Methodology

A. Asymmetric Architecture and Distillation

We utilized a frozen, pre-trained CLIP ViT-B/32 (500M parameters) as the Teacher network. We designed a custom Scalable CNN Student architecture with a width expansion factor, yielding three variants: Student-S (0.5M), Student-M (2.0M), and Student-L (8.0M).

The models were trained on the CIFAR-10 dataset using a strict cosine distance distillation objective:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{z_s^{(i)} \cdot z_t^{(i)}}{\|z_s^{(i)}\| \|z_t^{(i)}\|} \right) \tag{1}$$

B. Rigorous Spectral Evaluation

To ensure geometric consistency, all embedding matrices were strictly centered prior to extracting the Singular Value Decomposition (SVD): $Z_c = Z - \mu_Z$. We calculated the Shannon Entropy Effective Rank using the normalized squared singular values (S^2):

$$p_i = \frac{\sigma_i^2}{\sum \sigma_j^2}, \quad ER = \exp\left(-\sum p_i \ln p_i \right) \tag{2}$$

C. Information-Theoretic Metrics

To quantify retained semantics within the collapsed subspace, we utilized the InfoNCE loss as a proxy for Mutual Information [6], alongside Representation Uniformity:

$$\mathcal{L}_{uniform} = \log \mathbb{E}\left[e^{-2\|z_s^{(i)} - z_s^{(j)}\|^2}\right] \tag{3}$$

III. Results and Analysis

A. Capacity-Agnostic Dimensional Collapse

A primary objective of this study was to determine if capacity constraints induce dimensional collapse or if students merely inherit the Teacher's intrinsic anisotropy.

The empirical data revealed a severe, capacity-agnostic phase transition. While the Teacher model exhibited an Effective Rank of 88.68 (requiring 152 dimensions to capture 90% variance), the distilled models suffered a massive geometric truncation. As detailed in Table I, scaling the Student architecture by a factor of 16 (from 0.5M to 8.0M parameters) yielded negligible expansion in the representational subspace, with all models collapsing to an Effective Rank of $\sim$16.

TABLE I
Quantitative Scaling Metrics (3-Seed Average)

Model	Params	Effective Rank	Probe Acc (%)
Teacher-CLIP	500M	88.68	94.31 (Clean)
Student-S	$\sim$0.5M	15.91	71.11 $\pm$ 0.44
Student-M	$\sim$2.0M	16.62	72.32 $\pm$ 0.50
Student-L	$\sim$8.0M	16.66	72.94 $\pm$ 0.43

This mathematically proves that the asymmetric distillation objective enforces an absolute, rigid information bottleneck.

B. Geometrically Consistent Subspace Alignment

To further investigate this phenomenon, we projected the centered Student embeddings directly onto the orthogonal basis defined by the Teacher's principal components.

As shown in Figure 1, the alignment curves are geometrically identical. Excess parameters in the Large model are not utilized to span the lower-variance, fine-grained dimensions of the Teacher's hypersphere. Instead, the distillation process acts as an implicit Truncated PCA filter.

C. Information-Theoretic Subspace Utilization

Although the models share an identical 16-dimensional footprint, their downstream performance varies. We measured the InfoNCE and Uniformity losses to explain this discrepancy (Figure 2).

As capacity increases from 0.5M to 8.0M, InfoNCE loss decreases from 3.3117 to 3.2738, indicating a modest but consistent improvement in mutual information retention. Similarly, Uniformity loss improves from -0.4061 to -0.4285. This demonstrates that while excess parameters cannot expand the bottleneck's dimensionality, they enable the Student to distribute representations more uniformly.

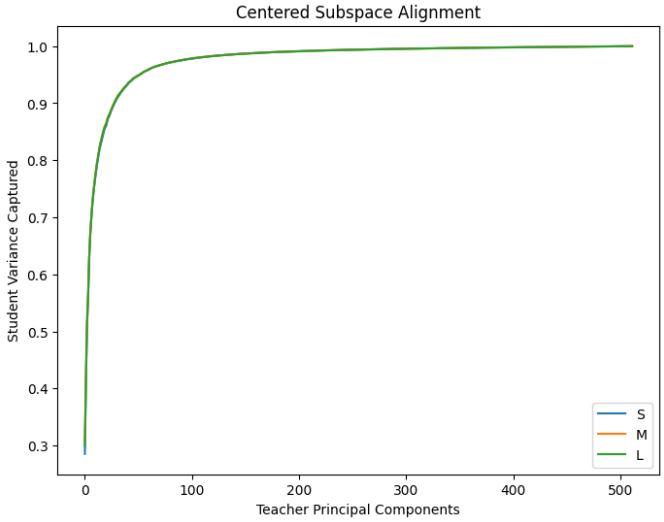

Fig. 1. Centered Subspace Alignment. The projection trajectories for the Small, Medium, and Large models are virtually indistinguishable, capturing over 90% of variance within the Teacher's top 20 singular vectors.

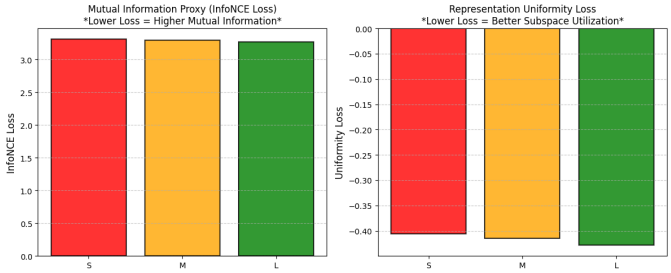

Fig. 2. Mutual Information proxy (InfoNCE) and Representation Uniformity Loss. Larger models distribute representations more uniformly.

D. Robustness Trade-Off within the Bottleneck

While overparameterization improves clean-data mutual information, our multi-seed evaluation revealed a critical vulnerability regarding robustness. We evaluated the models under high-frequency Gaussian noise to determine if the distillation process inherits the Teacher's robustness or artificially induces fragility.

The Teacher model exhibited high noise immunity, degrading only marginally from 94.31% to 89.35% at a noise severity of $\sigma = 0.1$. This robustness is heavily reliant on its expansive 88.68 effective dimensions. Because the Student models are forced to discard these redundant dimensions and collapse into an identical 16-dimensional subspace, they suffer catastrophic fragility.

As illustrated in Figure 3, the accuracy trajectories invert under high-frequency noise. At $\sigma = 0.1$, the 3-seed averaged accuracy for the 8.0M Student degraded drastically to 43.76% $\pm$ 6.10%. Conversely, the highly constrained 0.5M model retained a significantly higher 54.84% $\pm$ 3.73% accuracy.

To determine if this fragility was merely a failure to learn invariances rather than a strict geometric limitation, we trained the 8.0M Student-L with explicit spatial augmentations

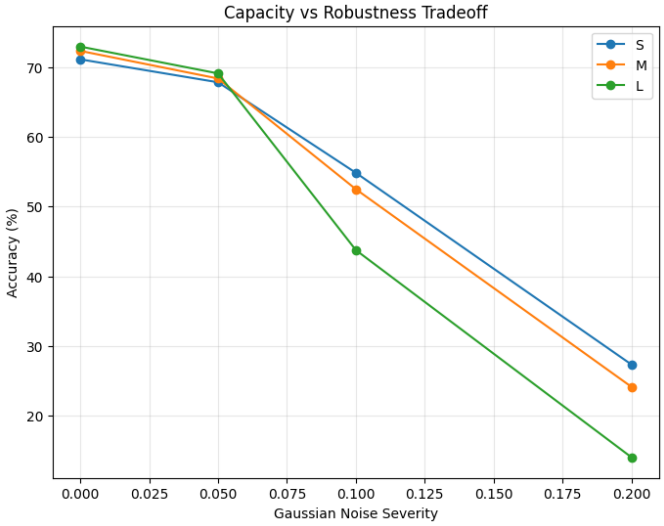

Fig. 3. Capacity vs Robustness Tradeoff. Under severe Gaussian noise ($\sigma \geq 0.1$), the highly constrained Student-S model significantly outperforms the overparameterized Student-L model.

(random crop and horizontal flip). Remarkably, augmentation failed to restore the Teacher's high-frequency noise immunity, plateauing at 14.04% accuracy under $\sigma = 0.2$ noise—identical to the unaugmented baseline. Furthermore, clean accuracy degraded to 69.34%. This confirms that the loss of robustness is physically constrained by the geometry of asymmetric transfer. The $\sim$16-dimensional bottleneck is fundamentally too narrow to encode the Teacher's 88-dimensional robust feature redundancy.

IV. Conclusion

In this paper, we presented a rigorous empirical study on the spectral geometry of asymmetric cross-modal distillation. We observed a clear empirical behavior: massive teacher networks enforce a strict, capacity-agnostic dimensional bottleneck ($\sim$16 dimensions) on local-receptive-field students, discarding the vast majority of the Teacher's 88.68 effective dimensions. Furthermore, we exposed an information-theoretic trade-off: capacity scaling within this bottleneck monotonically improves clean representation uniformity, but drastically sacrifices high-frequency noise robustness due to clean-data overfitting.

While extreme capacity constraints act as an implicit regularizer against this fragility, future work must address how to transfer the Teacher's high-dimensional invariance into the Student's low-dimensional bottleneck. We hypothesize that standard cosine distillation strictly transfers alignment, but fails to transfer robust local neighborhoods. A promising future direction involves integrating an auxiliary self-supervised contrastive objective (e.g., InfoNCE across augmented student views) alongside the asymmetric distillation loss. This would explicitly force the capacity-constrained student to construct robust, invariant manifolds within the geometric bottleneck, potentially decoupling parameter density from high-frequency brittleness.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *International Conference on Learning Representations*, 2022.

[5] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," *arXiv preprint arXiv:1610.01644*, 2016.

[6] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.