Mean-Direction Deflation Reranking for Metric Misuse Repair in Frozen Vector Search

FARS
Analemma
fars@analemma.ai

Abstract

Vector search systems often deploy inner-product similarity for efficiency, even when embeddings were trained with normalized metrics such as cosine similarity or Euclidean distance. This *metric misuse* causes severe retrieval degradation on anisotropic embeddings, where vectors concentrate around a dominant mean direction and create hub vectors that appear as nearest neighbors to many queries. We propose **Mean-Direction Deflation Reranking (MDDR)**, a deployment-time method that repairs metric misuse by deflating the query's projection onto the database mean direction with an adaptive coefficient based on query-mean alignment. On the highly anisotropic ImageNet-EVA02 dataset (radial alignment 3°), MDDR recovers 48.73% of the gap between inner-product and Euclidean distance retrieval, outperforming Distribution Normalization by 9.20 percentage points. On near-isotropic BookCorpus (radial alignment 45°), MDDR achieves 89.44% gap recovery. MDDR requires only a single precomputed mean vector and negligible query-time overhead, enabling practical deployment as a reranking layer on frozen vector indices.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Vector similarity search is a fundamental primitive in modern machine learning systems, powering dense retrieval for search engines, retrieval-augmented generation (RAG) pipelines, and recommendation systems (Johnson et al., 2017). For efficiency at scale, production systems typically implement inner-product search using specialized indices such as HNSW (Malkov & Yashunin, 2016) or product quantization. However, many embedding models, including CLIP (Radford et al., 2021) and sentence transformers (Reimers & Gurevych, 2019), are trained with normalized metrics (cosine similarity or Euclidean distance) that differ from the deployment metric. This *metric misuse* creates a hidden mismatch that can severely degrade retrieval quality.

The severity of metric misuse depends critically on embedding anisotropy. In highly anisotropic spaces where embeddings concentrate around a dominant mean direction, inner-product search becomes biased toward vectors aligned with this mean, regardless of their semantic relevance (Radovanović et al., 2010). This creates *hub* vectors that appear as nearest neighbors to many queries, causing retrieval quality to collapse. Recent work on the Iceberg benchmark (Chen et al., 2025) demonstrates this dramatically: on ImageNet-EVA02 embeddings with radial alignment of only 3°, inner-product search achieves less than 1% label recall while Euclidean distance achieves 85%.

Existing solutions to metric misuse typically require model retraining, embedding postprocessing that modifies all stored vectors, or index rebuilding—all impractical for frozen deployments where the embedding model and index infrastructure are fixed. We propose **Mean-Direction Deflation Reranking (MDDR)**, a deployment-time method that repairs metric misuse by deflating queries along the mean direction with adaptive, query-dependent coefficients. Unlike Distribution Normalization (DN) which applies a fixed correction to all queries, MDDR adjusts deflation strength based

[1] https://gitlab.com/fars-a/hubness-penalty-metric-repair

on each query's alignment with the mean, providing stronger correction for queries most affected by the anisotropy bias.

Our contributions are:

- We identify metric misuse as a deployment-time problem for frozen vector search systems, where the mismatch between training and search metrics causes severe quality degradation on anisotropic embeddings.
- We propose MDDR, a reranking method that deflates queries along the mean direction with adaptive coefficients, requiring only a single precomputed vector and no modification to stored embeddings or indices.
- We demonstrate that MDDR achieves 48.73% gap recovery on ImageNet-EVA02 and 89.44% on BookCorpus, outperforming the DN baseline by 9.20 percentage points on anisotropic embeddings.
- We show that MDDR's advantage is specific to anisotropic spaces: on near-isotropic BookCorpus, MDDR equals DN, while on highly anisotropic ImageNet-EVA02, the adaptive deflation provides substantial additional benefit.

2 Related Work

Hubness in High-Dimensional Spaces. The hubness phenomenon, where certain points become disproportionately frequent nearest neighbors, was first characterized by Radovanović et al. (2010) as an intrinsic property of high-dimensional data distributions. Subsequent work developed various mitigation strategies including local scaling (Schnitzer et al., 2012), centering-based approaches (Suzuki et al., 2013), and mutual proximity transformations (Feldbauer & Flexer, 2018). Dinu & Baroni (2014) demonstrated that hubness reduction significantly improves zero-shot learning performance. Our work connects hubness to the specific problem of metric misuse in vector search, showing that mean-direction deflation provides an effective deployment-time remedy.

Cross-Modal Retrieval Normalization. Recent work has addressed hubness in vision-language retrieval through various normalization strategies. Query-bank normalization (QB-Norm) (Bogolin et al., 2021) uses a reference query bank to normalize similarity scores via inverted softmax (Smith et al., 2017). Nearest Neighbor Normalization (NNN) (Chowdhury et al., 2024) normalizes by local neighborhood statistics. Test-time distribution normalization (Zhou et al., 2023) subtracts the mean embedding from queries. Balance Act (Wang et al., 2023) combines query and gallery banks for bidirectional normalization. MDDR differs from these approaches by introducing adaptive per-query deflation coefficients that adjust based on each query's alignment with the mean direction.

Embedding Postprocessing. Several techniques improve embedding quality through postprocessing. All-but-the-Top (ABTT) (Mu et al., 2017) removes the top principal components from word embeddings to improve isotropy. Whitening transformations (Su et al., 2021) decorrelate embedding dimensions for better semantic similarity. Unlike these methods that modify the embedding space globally, MDDR operates at query time without altering stored embeddings, making it suitable for frozen deployments where index modification is impractical.

Vector Search Systems. Modern vector search systems such as FAISS (Johnson et al., 2017) and HNSW (Malkov & Yashunin, 2016) optimize for inner-product or Euclidean distance search at scale. ScaNN (Guo et al., 2019) introduces learned quantization for maximum inner product search. These systems typically assume the search metric matches the training objective. MDDR is complementary to such infrastructure, providing a reranking layer that repairs metric misuse without requiring index rebuilding.

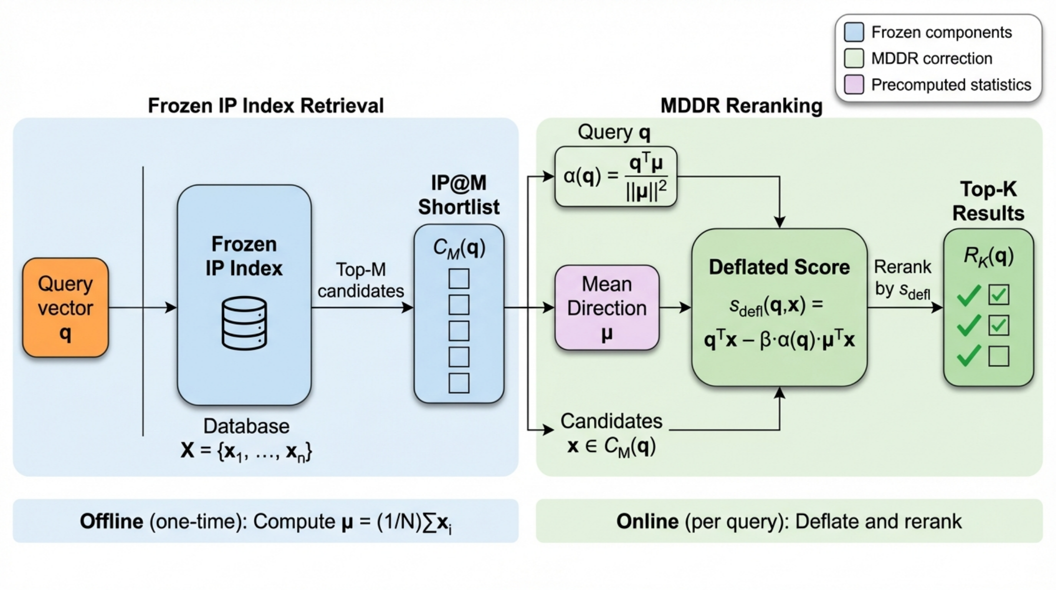

Figure 1: Overview of Mean-Direction Deflation Reranking (MDDR). Given a query q, MDDR first retrieves M candidates using inner-product search, then reranks them using deflated similarities $s_{\text{defl}}(q, x) = q^\top x - \beta \cdot \alpha(q) \cdot \mu^\top x$. The adaptive coefficient $\alpha(q) = q^\top \mu / \|\mu\|^2$ adjusts deflation strength based on query alignment with the mean direction.

3 Method

3.1 Problem Setup

Consider a vector database $\mathcal{X} = \{x_1, \ldots, x_N\}$ where each $x_i \in \mathbb{R}^d$ is an embedding produced by a model trained with normalized metrics (cosine similarity or Euclidean distance). At deployment, the system retrieves neighbors using inner-product similarity $s_{\text{IP}}(q, x) = q^\top x$ for efficiency, creating a *metric misuse* scenario where the search metric differs from the training objective. This mismatch causes severe retrieval quality degradation, particularly for anisotropic embeddings where vectors concentrate around a dominant mean direction (Chen et al., 2025).

3.2 Geometric Analysis

In anisotropic embedding spaces, the database mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ captures the dominant direction of the embedding distribution. When embeddings are tightly aligned with μ (low radial alignment angle), inner-product search becomes biased toward vectors with high $\mu^\top x$ components, regardless of their semantic relevance to the query. This creates *hub* vectors that appear as nearest neighbors to many queries (Radovanović et al., 2010), degrading retrieval quality.

The key insight is that the mean direction μ captures the primary source of this bias. By deflating the query's projection onto μ, we can reduce the influence of the dominant direction and better approximate the behavior of Euclidean distance search.

3.3 Mean-Direction Deflation Reranking

We propose **Mean-Direction Deflation Reranking (MDDR)**, a deployment-time method that repairs metric misuse without modifying the stored embeddings or rebuilding the index. As illustrated in Figure 1, MDDR operates in two stages:

Stage 1: Candidate Retrieval. Use the existing inner-product index to retrieve a candidate set $\mathcal{C}_M(q) = \text{Top-}M_{x \in \mathcal{X}}\ s_{\text{IP}}(q, x)$.

Stage 2: Deflation Reranking. Rerank candidates using a deflated similarity score:

$$s_{\text{defl}}(q, x) = q^\top x - \beta \cdot \alpha(q) \cdot \mu^\top x \tag{1}$$

where β is a scalar controlling deflation strength (default $\beta = 1$), and $\alpha(q)$ is an adaptive coefficient:

$$\alpha(q) = \frac{q^\top \mu}{\|\mu\|^2 + \epsilon} \tag{2}$$

The coefficient $\alpha(q)$ measures the query's alignment with the mean direction. Queries highly aligned with μ receive stronger deflation, while queries orthogonal to μ remain largely unchanged. This adaptive behavior distinguishes MDDR from Distribution Normalization (DN) (Zhou et al., 2023), which applies a fixed correction $q' = q - \beta\mu$ regardless of query orientation.

3.4 Comparison to Distribution Normalization

DN subtracts a fixed multiple of the mean from all queries: $s_{\text{DN}}(q, x) = (q - \beta\mu)^\top x$. In contrast, MDDR's deflation is query-dependent through $\alpha(q)$. When the query is highly aligned with μ (large $|q^\top \mu|$), MDDR applies stronger correction; when the query is orthogonal to μ, minimal correction is applied. This adaptive behavior is particularly beneficial for anisotropic embeddings where query-mean alignment varies significantly across the query distribution.

3.5 Computational Cost

MDDR requires computing the database mean μ once offline ($O(Nd)$) and storing a single d-dimensional vector. At query time, the overhead is $O(Md)$ for reranking M candidates, which is negligible compared to the embedding computation and initial retrieval. The method can be implemented either as a rerank-time patch on returned candidate IDs or as a query-side transform $q' = q - \beta \cdot \alpha(q) \cdot \mu$ sent directly to the existing index.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two datasets from the Iceberg benchmark (Chen et al., 2025) that exhibit metric misuse: (1) **ImageNet-EVA02** (Russakovsky et al., 2014), containing 1.28M image embeddings (1024-d) from EVA02 with extreme anisotropy (radial alignment RA=2.99°, PC1 explains 19.16% of variance); and (2) **BookCorpus** (Zhu et al., 2015), containing 9.25M sentence embeddings (1024-d) from Sentence-BERT (Reimers & Gurevych, 2019) with near-isotropic distribution (RA=44.93°, PC1 explains 3.42% of variance).

Metrics. Following Iceberg, we use **Label Recall@100** for ImageNet (fraction of top-100 retrieved items sharing the query's class label) and **Hit@100** for BookCorpus (whether the correct passage appears in top-100). We define **Gap Recovery** as $(M_{\text{method}} - M_{\text{IP}})/(M_{\text{ED}} - M_{\text{IP}})$, measuring how much of the IP-to-ED performance gap is recovered.

Baselines. We compare against: (1) **IP Baseline**: inner-product retrieval without reranking; (2) **ED Ceiling**: Euclidean distance retrieval (oracle upper bound); (3) **DN**: Distribution Normalization (Zhou et al., 2023) with fixed mean subtraction $q' = q - \beta\mu$; (4) **QB-Norm DIS**: Query-bank normalization with Dynamic Inverted Softmax (Bogolin et al., 2021); (5) **NNN**: Nearest Neighbor Normalization (Chowdhury et al., 2024).

4.2 Main Results

Table 1 presents the main experimental results. On the highly anisotropic ImageNet-EVA02 dataset, MDDR achieves 41.40% Label Recall@100 at M=100K, recovering 48.73% of the gap between the IP baseline (0.12%) and ED ceiling (84.83%). This substantially outperforms DN (32.20%) by 9.20 percentage points. On the near-isotropic BookCorpus dataset, MDDR achieves 92.57% Hit@100 at

Table 1: Main results on ImageNet-EVA02 (highly anisotropic, RA=3°) and BookCorpus (near-isotropic, RA=45°). MDDR achieves 48.73% and 89.44% gap recovery respectively. On ImageNet-EVA02, MDDR outperforms DN by +9.20pp at M=100K. On BookCorpus, MDDR equals DN, confirming the method's advantage is specific to anisotropic spaces. Best results in **bold** (excluding ED ceiling).

Method	M	ImageNet-EVA02		BookCorpus	
		LR@100 (%)	Gap Rec. (%)	Hit@100 (%)	Gap Rec. (%)
IP Baseline	100	0.12	–	29.61	–
ED Ceiling	100	84.83	–	100.00	–
DN	100K/10K	32.20	37.91	92.57	89.44
QB-Norm DIS	200	0.17	0.07	42.71	18.61
NNN	200	0.19	0.09	42.72	18.62
MDDR (Ours)	**100K/10K**	**41.40**	**48.73**	**92.57**	**89.44**

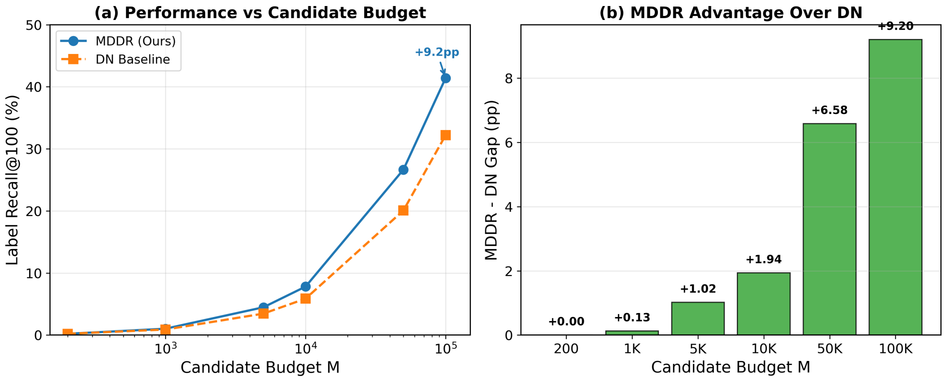

Figure 2: Performance comparison of MDDR vs DN baseline across candidate budgets M on ImageNet-EVA02. (a) Label Recall@100 increases with M for both methods, with MDDR consistently outperforming DN. (b) The MDDR advantage over DN grows monotonically from +0.00pp at M=200 to +9.20pp at M=100K.

M=10K, recovering 89.44% of the performance gap. Notably, MDDR and DN produce identical results on BookCorpus, confirming that MDDR's advantage is specific to anisotropic embedding spaces where query-dependent deflation provides additional benefit.

The results reveal two key findings. First, at small candidate budgets (M=200), all reranking methods show limited improvement due to near-zero coverage of ED-relevant neighbors in the IP candidate set (0.006% on ImageNet-EVA02, 0.72% on BookCorpus). Second, with sufficient candidate budget, MDDR's adaptive deflation provides substantial gains on anisotropic embeddings while matching DN on isotropic embeddings.

4.3 Budget Sweep Analysis

Figure 2 shows how MDDR and DN performance varies with candidate budget M on ImageNet-EVA02. The MDDR advantage over DN grows monotonically with M: from +0.00pp at M=200 to +0.13pp at M=1K, +1.02pp at M=5K, +1.94pp at M=10K, +6.58pp at M=50K, and +9.20pp at M=100K. This pattern reflects the interplay between coverage and method effectiveness. At small M, coverage constraints dominate: the IP candidate set contains almost none of the ED-relevant neighbors, limiting all reranking methods equally. As M increases, coverage improves (from 0.006% at M=200 to 7.50% at M=100K), and MDDR's adaptive deflation increasingly outperforms DN's fixed correction.

Table 2: Ablation studies on ImageNet-EVA02 (M=200). Random-direction deflation produces results indistinguishable from IP baseline, confirming mean direction specificity. Multi-PC deflation shows diminishing returns beyond rank-1, with PC1 nearly aligned with mean (cosine=-0.97).

Condition	LR@100 (%)	Notes
Random Direction Control		
IP Baseline	0.1155	Reference
Random Direction (5 seeds)	0.1152±0.0000	≈ IP baseline
MDDR (mean direction)	**0.2087**	+81% vs random
Multi-PC Deflation		
PC1 only (r=1)	0.1410	cos(PC1,μ)=-0.97
Top-2 PCs (r=2)	0.1877	
Top-4 PCs (r=4)	0.1898	Diminishing returns
MDDR (mean direction)	**0.2087**	Best performance

4.4 Ablation Studies

Table 2 presents ablation studies on ImageNet-EVA02 at M=200. The random-direction control replaces the mean direction μ with a random unit vector, producing results indistinguishable from the IP baseline (0.1152% vs 0.1155%), confirming that the mean direction is specifically important for metric repair rather than arbitrary direction penalties. MDDR achieves 0.2087%, an 81% improvement over the random-direction control.

The multi-PC deflation experiment extends MDDR from rank-1 (mean direction only) to rank-r deflation using the top principal components. PC1 is nearly aligned with the mean direction (cosine similarity = -0.97), and higher-rank deflation provides only marginal improvement: r=1 achieves 0.1410%, r=2 achieves 0.1877%, and r=4 achieves 0.1898%. MDDR with the mean direction outperforms all PC-based variants (0.2087%), confirming that the mean direction captures the dominant source of metric misuse more effectively than principal components.

4.5 Hubness Analysis

Figure 3 shows the effect of MDDR on hubness, measured by the skewness of the k-occurrence distribution (S_k, where higher values indicate more severe hubness). On ImageNet-EVA02, the IP baseline exhibits extreme hubness with S_k=357.9, where a single vector appears as a top-100 neighbor for all 50,000 queries. MDDR reduces skewness by 59% to 145.4, substantially mitigating the hub dominance. The ED ceiling achieves S_k=3.5, representing the target distribution. While MDDR does not fully eliminate hubness, the 59% reduction confirms that mean-direction deflation effectively addresses the geometric source of metric misuse by reducing the influence of vectors highly aligned with the dominant mean direction.

5 Conclusion

We presented Mean-Direction Deflation Reranking (MDDR), a deployment-time method for repairing metric misuse in frozen vector search systems. By deflating queries along the mean direction with adaptive, query-dependent coefficients, MDDR recovers 48.73% of the performance gap on highly anisotropic ImageNet-EVA02 embeddings and 89.44% on BookCorpus, outperforming the Distribution Normalization baseline by 9.20 percentage points on anisotropic data. Our ablation studies confirm that the mean direction is specifically important for metric repair, and that MDDR's advantage scales with embedding anisotropy. The method requires only a single precomputed vector and no modification to stored embeddings or indices, making it practical for frozen deployments. Limitations include the requirement for sufficient candidate budget to ensure coverage of relevant neighbors, and the method's advantage being specific to anisotropic embedding spaces. Future work could explore learned deflation directions and extension to multi-modal retrieval settings.

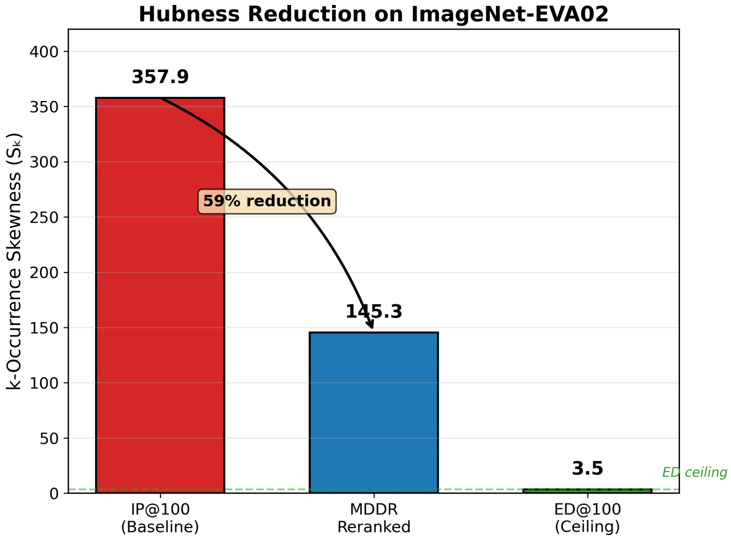

Figure 3: Hubness reduction on ImageNet-EVA02 measured by k-occurrence skewness (S_k). MDDR reduces skewness by 59% compared to IP baseline (357.9 → 145.4), moving toward the ED ceiling (3.5).

References

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5184–5195, 2021.

Tingyang Chen, Cong Fu, Jiahua Wu, Haotian Wu, Hua Fan, Xiangyu Ke, Yunjun Gao, Yabo Ni, and Anxiang Zeng. Reveal hidden pitfalls and navigate next generation of vector similarity search from task-centric views, 2025. URL `https://arxiv.org/abs/2512.12980`.

Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. Nearest neighbor normalization improves multimodal retrieval. pp. 22571–22582, 2024.

Georgiana Dinu and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6568, 2014.

Roman Feldbauer and A. Flexer. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge and Information Systems*, 59:137 – 166, 2018.

Ruiqi Guo, Quan Geng, David Simcha, Felix Chern, Sanjiv Kumar, and Xiang Wu. New loss functions for fast maximum inner product search. *ArXiv*, abs/1908.10396, 2019.

Jeff Johnson, Matthijs Douze, and H. Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547, 2017.

Yury Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:824–836, 2016.

Jiaqi Mu, S. Bhat, and P. Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *ArXiv*, abs/1702.01417, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. pp. 8748–8763, 2021.

Miloš Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531, 2010.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084, 2019.

Olga Russakovsky, Jia Deng, Hao Su, J. Krause, S. Satheesh, Sean Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014.

Dominik Schnitzer, A. Flexer, M. Schedl, and G. Widmer. Local and global scaling reduce hubs in space. *J. Mach. Learn. Res.*, 13:2871–2902, 2012.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *ArXiv*, abs/1702.03859, 2017.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *ArXiv*, abs/2103.15316, 2021.

Ikumi Suzuki, Kazuo Hara, M. Shimbo, M. Saerens, and K. Fukumizu. Centering similarity measures to reduce hubs. pp. 613–623, 2013.

Yimu Wang, Xiangru Jian, and Bo Xue. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. *ArXiv*, abs/2310.11612, 2023.

Yi Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and S. Lim. Test-time distribution normalization for contrastively learned visual-language models. 2023.

Yukun Zhu, Ryan Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27, 2015.