Cross Modal Retrieval with Querybank Normalisation

Simion-Vlad Bogolin\textsuperscript{1,2,}\textsuperscript{*} \quad Ioana Croitoru\textsuperscript{1,2,}\textsuperscript{*}
Hailin Jin\textsuperscript{3} \quad Yang Liu\textsuperscript{1,4,}\textsuperscript{†} \quad Samuel Albanie\textsuperscript{1,5,}\textsuperscript{†}

\textsuperscript{1}Visual Geometry Group, University of Oxford \quad \textsuperscript{2}Inst. of Mathematics of the Romanian Academy \quad \textsuperscript{3}Adobe Research
\textsuperscript{4}Wangxuan Inst. of Computer Technology, Peking University \quad \textsuperscript{5}Department of Engineering, University Cambridge

Abstract

Profiting from large-scale training datasets, advances in neural architecture design and efficient inference, joint embeddings have become the dominant approach for tackling cross-modal retrieval. In this work we first show that, despite their effectiveness, state-of-the-art joint embeddings suffer significantly from the longstanding “hubness problem” in which a small number of gallery embeddings form the nearest neighbours of many queries. Drawing inspiration from the NLP literature, we formulate a simple but effective framework called Querybank Normalisation (QB-NORM) that re-normalises query similarities to account for hubs in the embedding space. QB-NORM improves retrieval performance without requiring retraining. Differently from prior work, we show that QB-NORM works effectively without concurrent access to any test set queries. Within the QB-NORM framework, we also propose a novel similarity normalisation method, the Dynamic Inverted Softmax, that is significantly more robust than existing approaches. We showcase QB-NORM across a range of cross modal retrieval models and benchmarks where it consistently enhances strong baselines beyond the state of the art. Code is available at https://vladbogo.github.io/QB-Norm/.

1. Introduction

As the improving price-performance of hardware underpinning sensors, storage and networking continues to enable the expansion of humanity’s digital archives, the capacity to efficiently search data takes on greater commercial and scientific importance. An appealing way to search such data is via natural language queries, in which the user describes the target of their search exactly as they would to another human, rather than employing specialised database languages such as Structured Query Language (SQL).

Towards this goal, a rich body of research literature has studied the problem of cross modal retrieval, the task of searching a gallery of samples in one modality given a query in another. In particular, there has been significant progress in recent years for systems that can efficiently search images [90], audio [74] and videos [103] with natural language queries by employing cross modal embeddings.

The dominant cross modal embedding paradigm employs deep neural networks that project modality-specific samples into a high-dimensional, real-valued vector space in which they can be directly compared via an appropriate distance metric. A key challenge for such methods, intrinsic to such high-dimensional spaces, is the emergence of “hubs” [80]—embedding vectors that appear amongst the nearest neighbour sets of disproportionately many other embedding vectors (Fig. 1, left). To illustrate this challenge,
we show empirically in Sec. 3.2 and Fig. 2 that hubness is prevalent among a range of leading retrieval methods. Hubs have consequences: if left unaddressed, they lead to a significant degradation in the search ranking yielded by a retrieval system [8]. The hubness problem has received considerable attention [8, 64, 80] and a number of approaches have been proposed to address it [31], with notable contributions in the NLP literature focusing on bilingual word translation [21, 24, 88]. One contribution of our work is to show how each of these methods can be interpreted within a single unifying conceptual framework termed Querybank Normalisation (QB-NORM, Fig. 1, right), that employs a querybank of samples during inference to reduce the influence of hubs in the gallery. We observe that existing methods have two challenges: (1) To date, these approaches have only been shown to work with concurrent access to multiple test queries—an assumption that is impractical for real-world retrieval systems; (2) They are sensitive to querybank selection, and indeed actively harm performance for certain querybanks (Tab. 2). To address the first challenge, we demonstrate through careful experiments (Tab. 1) that QB-NORM does not require concurrent access to test queries to be effective. To address the second challenge, we propose a new normalisation method, Dynamic Inverted Softmax (DIS), that operates as a module within the QB-NORM framework. We show that DIS provides effective normalisation, yet is more robust than prior approaches [21, 24, 88].

We make the following contributions: (1) We motivate our study by demonstrating that the longstanding problem of hubness remains a significant concern in modern cross-modal embeddings for retrieval; (2) We propose Querybank Normalisation (QB-NORM), a simple non-parametric framework that brings significant gains in retrieval performance without requiring model fine-tuning; (3) We provide the first (to the best of our knowledge) demonstration that Querybank Normalisation methods retain their effectiveness for cross-modal retrieval with no access to test queries beyond the current query; (4) We propose the Dynamic Inverted Softmax, a novel normalisation method for Querybank Normalisation that is more robust than prior literature; (5) We show that QB-NORM is highly effective across a broad range of tasks, models and benchmarks.

2. Related work

In this section, we summarise prior work from the literature that relates to our approach, focusing on cross-modal retrieval, external memory banks and hubness.

Cross-modal representations. Following initial studies in psychology [11], early frameworks for cross-modal retrieval included Gaussian Mixture Models [87] modelling translation via EM [27], Topic Models [9], CCA [81], KCCA [89] and rank optimisation [99]. Motivated by the successes of deep metric learning [18] and deep visual semantic embeddings [33], there has since been a Cambrian explosion of cross-modal embedding methods for text-image retrieval [29, 50, 54, 70, 95], text-video [3, 5, 23, 25, 69, 100, 104, 106] text-audio [74], image-audio [1, 51, 72, 75, 111] and combinations of all the above [4]. Recent research spanning these tasks has explored large-scale pre-training [68, 79], domain adaptation [63, 71] and tight integration of multiple sensory modalities into one side of the embedding space [34, 62, 67].

Similarity search for retrieval: Tricks of the trade. A plethora of techniques have been developed to support and enhance similarity search for retrieval, including k-d trees [7], re-ranking [44, 78], query expansion [19, 20], vector compression schemes based on binary codes [36, 41] and quantization [43, 45] that help address the curse of dimensionality [6]. Algorithms have been developed for approximate k-nearest neighbour graph construction on CPUs [26] and GPUs [47], with the latter drawing on product quantization techniques to scale up to billion-scale searches.

Differently from the work on cross-modal representations and improved similarity search described above, we focus specifically on tackling the problem of hubness in cross-modal embeddings, which we demonstrate (Sec. 3.2) to be a widespread issue among leading cross-modal embedding frameworks.

Memory bank augmented architectures. Memory banks in various forms have been studied as useful extensions to neural network architectures to facilitate general problem-solving [37, 38, 84, 107], better image captioning [22, 76, 101] and summarisation [52, 56], enhance self-supervised training dynamics [12, 40, 61] and to provide a mechanism to deal with rare instances [49, 105]. Our proposed Querybank Normalisation framework likewise stores embedding samples in an external memory bank, but targets a very different problem to these works, namely hubness mitigation.

The Hubness Problem. The hubness problem was formally characterised by Radovanovic et al. [80], who observed that in points sampled from a distribution with high intrinsic dimensionality, the distribution of “k-occurrences” (the number of times a point appears in the k nearest neighbours of other points) skews heavily to the right. Although there is disagreement about the cause of hubness [64], it has been conceptually linked [8] to distance concentration in high-dimensions (high-dimensional points lie close to a hypersphere centred on the data mean, i.e., they all exhibit a similar distance to the mean [32]). It is thought that hubs then result from this phenomenon through the non-negligible variance in the distribution of distances to the mean in finite dimensions [80].

Hubness Mitigation. One paradigm has focused on rescaling the similarity space to account for asymmetries in nearest neighbour relations [85]—a process that can be achieved through both local [46, 108] and global [85] scal-
Figure 2. Hubness is pervasive in leading methods for text-video retrieval. The charts depict the distribution of the number of times each gallery video was retrieved by test set queries (x-axes video ids are ordered by decreasing retrieval count). Top row (different models): We report retrieval distributions for CE [62], TT-CE+ [23], MMT [34] and CLIP2Video [30] on the MSR-VTT benchmark [102]. Bottom row (different datasets): We report retrieval distributions for the TT-CE+ [23] method on four additional datasets, DiDeMo [42], LSMDC [82], VaTeX [98], and ActivityNet-captions [55]. In all instances, we observe strong hubness, in which a small number of videos are retrieved disproportionately often, damaging performance.

ing schemes. Another work has focused on addressing the hub-like tendency of centroids in the data through Laplacian-based kernels [92] and centring [39, 93]. Fedbauer et al. provide a comprehensive empirical comparison of these families of methods [31] and note that while effective, these approaches scale quadratically, making their naive application unsuitable for large datasets. One exception is the CENT method [93], however we did not find this approach to be effective (experiments are provided in the supplementary). In the zero-shot learning literature, works have sought to address hubness by mapping (text) targets back into the (image) query space [86, 109], and by minimising proxies for hubness [60] and skewness in the k-occurrences distribution [17] to improve 3D few-shot learning performance. More closely related to our work, [24] propose general retrieval schemes for which queries are matched with targets for which they form the nearest neighbour. This work was built upon the NLP literature by [21], who propose a cross-domain local scaling scheme (which can be integrated into the loss [48]), and by [88], who introduce the Inverted Softmax (IS) to mitigate hubness when translating between dictionaries in different languages. We discuss the relationship of our approach to [21, 24, 88] in more detail in Sec. 3 and compare these methods with our proposed Dynamic Inverted Softmax.

3. Method

We first define the task of retrieval with cross modal embeddings (Sec. 3.1), before outlining the motivation for our work by examining the hubness problem in the context of text-video retrieval (Sec. 3.2). Next, we introduce the Querybank Normalisation framework (Sec. 3.3), generalising several existing approaches to address this issue. Finally, we explore designs for framework components and introduce the proposed Dynamic Inverted Softmax for robust similarity normalisation (Sec. 3.4).

3.1. Task definition

Given a gallery, \( \mathcal{G} \), of samples in one modality, \( m_g \) and a query, \( q \), in another modality, \( m_q \), the objective of cross modal retrieval is to rank the gallery samples according to how well they match the query. We study this problem within the framework of learning cross modal embeddings [33]: specifically, we seek to learn a pair of encoders, \( \phi_q \) and \( \phi_g \), that map each query, \( q \), and gallery sample, \( g \), into a shared real-embedding space, \( \mathbb{R}^C \), such that \( \phi_q(q) \) and \( \phi_g(g) \) are close if and only if \( q \) is similar to \( g \). We assume that we are given access to a training set of \( T \) corresponding query and gallery samples \( \{(q_i, g_i)\}_{i=1}^T \) for the purposes of learning the embeddings. However, the queries and gallery used to evaluate retrieval performance (i.e. the test set) are unseen during training.

The choice of similarity measure used to define a “good match” is determined by the application domain. For instance, in the task of text-video retrieval with natural lan-
guage queries, the objective is to rank a gallery of videos according to how well their content is described by a written free-form text query [67], whereas in image-audio retrieval the objective is typically to obtain audio samples from the gallery that share the same semantic category as the image query [2]. In this work, we focus particularly on cross modal retrieval tasks with natural language queries, for two reasons: (1) these tasks have received limited attention in the hubness mitigation literature, (2) hubness has been shown to be particularly prevalent in embeddings with high \textit{intrinsic dimensionality} [80]. Since natural language queries can express more complex concepts than individual words (such as those considered in zero-shot learning image labelling tasks [24]), we expect might expect natural language queries to naturally induce cross modal embeddings with greater intrinsic dimensionality, and thus may have greater potential to benefit from hubness mitigation.

3.2. Motivation

It has long been observed that high-dimensional embedding spaces are prone to \textit{hubness} [80], in which a small proportion of samples appear disproportionately frequently among the set of k-nearest neighbours of all embeddings. As noted by Berenzweig [8], this property can have damaging consequences for retrieval systems that employ nearest neighbour search to find the best gallery match for a given query. To illustrate this issue, we consider the problem of video retrieval with natural language queries. We plot the distribution of the number of times each gallery video was retrieved on the MSR-VTT retrieval benchmark [102] for an array of text-video retrieval methods, including CE [62], TT-CE+ [23], MMT [34] and CLIP2Video [30], the latter of which represents the current state of the art on this benchmark. In each case, we see striking evidence of hubness—a small number of videos are retrieved extremely often, while others are not retrieved at all. This phenomenon is not limited to a particular retrieval model, suggesting that the issue is not readily addressed by the use of multiple video modalities, attention mechanisms and large-scale pretraining implemented in various combinations by these approaches.

3.3. Querybank Normalisation

To address the hubness issues observed among cross modal embeddings for text-video retrieval in the previous section, we first turn to the existing literature on hubness mitigation. As noted in Sec. 2, hubness effects have been studied in several problem domains, including Zero-Shot Learning [24, 86], NLP [21, 88], biomedical statistics [85] and music retrieval [85]. Among this literature, we are particularly interested in methods that can be applied in a practical cross modal retrieval setting, namely, those methods whose complexity scales at most linearly with the size of the gallery (rather than quadratic complexity methods that seek to address hubness within a fixed embedding space [31]). To clarify relationships between existing approaches, we cast them into the Querybank Normalisation framework (Fig. 1), which comprises two components, \textit{querybank construction} and \textit{similarity normalisation}, described next:

\textbf{Querybank construction.} To mitigate hubness in the cross modal embedding space, we seek to alter the similarities between embeddings in a way that \textit{minimises the influence of hubs}. To adjust similarities, we first construct a \textit{querybank} of \(N\) samples, \(B = \{b_1, \ldots, b_N\}\) from the query modality, \(m_q\), which will serve as a \textit{probe} to measure the hubness of gallery samples.

\textbf{Similarity normalisation.} To normalise similarities to account for hubs, we assume access to a query, \(q\), trained encoders \(\phi_q\) and \(\phi_g\), querybank \(\{b_1, \ldots, b_N\}\), and a gallery \(G\). For each \(g_j \in G\), we first compute a \textit{probe vector}, \(p_j \in \mathbb{R}^N\), \(p_j(i) = \text{sim}(\phi_q(b_i), \phi_g(g_j))\) where \text{sim}(\cdot, \cdot) denotes similarities in the cross modal embedding space (e.g. cosine similarity). The probe vectors are then stacked to form a \textit{probe matrix} \(P \in \mathbb{R}^{G \times N}\). Similarly, we compute for each query a vector of \textit{unnormalised similarities}, \(s_q \in \mathbb{R}^G\), \(s_q(j) = \text{sim}(\phi_q(q), \phi_g(g_j))\). Here \(j \in \{1, \ldots, |G|\}\) indexes over all gallery elements. Finally, we define a \textit{querybank normalisation function}, \(QB-NORM : \mathbb{R}^G \times \mathbb{R}^{G \times N} \rightarrow \mathbb{R}^{|G|}\), which yields, for each query \(q\) and gallery \(G\), a vector of querybank normalised similarities, \(n_q = QB-NORM(s_q, P) \in \mathbb{R}^{|G|}\). Various candidates for \(QB-NORM(\cdot)\) are discussed in Sec. 3.4.

In practice, the probe matrix employed for similarity normalisation can be precomputed and re-used across all queries (improving computational efficiency at the cost of higher memory). An overview of the resulting \(QB-NORM\) algorithm, and its application to ranking gallery samples for a collection of queries, \(Q\), is summarised in Alg. 1.

3.4. Design choices

The Querybank Normalisation framework admits a number of viable choices for both \textit{querybank construction} and \textit{similarity normalisation}. To illustrate this point, we first cast three techniques for hubness mitigation proposed in the NLP literature into the framework. We then introduce our proposed alternative, the Dynamic Inverted Softmax.

\textit{Globally-Corrected (GC) retrieval} [24]. This approach, originally introduced for the tasks of bilingual translation and zero-shot learning, can be implemented by constructing the querybank from the full set of test queries, \(Q\), (or all semantic labels, in the cross modal setting of zero-shot image labelling). For their bilingual translation task, the authors supplement their querybank by an additional randomly sampled collection of instances from \(m_q\), which improved performance. The normalised similarity corresponding to \(q\) and gallery vector \(g_j\) is defined via \(n_q(j) = -(\text{Rank}(s_q(j), p_j) - s_q(j)) \in \mathbb{R}\), where \(\text{Rank} : \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R}\).
Algorithm 1 Ranking with Querybank Normalisation

**Input:** queries, \( Q \subset m_q \)

**Input:** gallery, \( G \subset m_g \)

1. **Querybank construction.**
2. Construct querybank, \( \mathcal{B} = \{b_1, \ldots, b_N\} \subset m_q \)
3. **Similarity normalisation:**
4. Precompute querybank probe matrix
5. for gallery sample \( g_j \in G \) do
6. for querybank sample \( b_i \in \mathcal{B} \) do
7. Compute probe matrix entry \( P(j, i) = \text{sim}(\phi_q(b_i), \phi_y(g_j) \in \mathbb{R} \)
8. end for
9. end for
10. query computations: QB-NORM similarities
11. for query \( q \in Q \) do
12. for gallery sample \( g_j \in G \) do
13. Compute unnormalised similarity \( s_q(j) = \text{sim}(\phi_q(q), \phi_y(g_j)) \)
14. end for
15. \( \eta_q = \text{QB-NORM}(s_q, P) \in \mathbb{R}^{\left| G \right|} \)
16. search ranking = \text{argsort}(\eta_q)
17. end for

\( \{0, \ldots, N\} \) returns the rank of the first argument with respect to the array of elements in the second argument.

Cross-Domain Similarity Local Scaling (CSLS) [21]. Introduce for the task of bilingual word translation, CSLS constructs an initial querybank comprising all possible queries (corresponding to source vocabulary samples), then employs a different subset of the querybank to normalise each gallery sample. Let \( \hat{p}_j \in \mathbb{R}^K \) denote the probe vector, \( p_j \), restricted to the \( K \) querybank samples that are most similar to gallery sample \( g_j \). Similarly, let \( \hat{s}_q \in \mathbb{R}^K \) denote the unnormalised similarity vector, \( s_q \), restricted to the \( K \) gallery samples that are most similar to query \( q \). Then the normalised similarity is computed via: \( \eta_q(j) = 2s_q(j) - \frac{1}{K} \frac{T}{\text{exp}[\beta \cdot s_q(j)]} - \frac{1}{K} \frac{T}{\text{exp}[\beta \cdot p_j]} \in \mathbb{R} \)

Inverted Softmax (IS) [88]. Targeting bilingual word translation, this method constructs a querybank from the source vocabulary (corresponding to all possible queries of interest). For practical implementations, the authors recommend to uniformly randomly subsample a feasible number of queries. Similarity normalisation is implemented via:

\[
\eta_q(j) = \frac{\exp(\beta \cdot s_q(j))}{\sum_{l=1}^{\left| G \right|} \exp(\beta \cdot s_q(l))} \in \mathbb{R} \tag{1}
\]

where \( \exp[\cdot] \) denotes elementwise exponentiation and \( \beta \) is a hyperparameter referred to as the “inverse temperature”.

Dynamic Inverted Softmax (DIS). In experiments with the methods described above (discussed in detail in Sec. 4) we observed an important practical issue: if the querybank does not effectively cover the space containing the gallery, performance is severely degraded such that it falls below the performance of unnormalised similarities. This characteristic renders them less desirable for a general-purpose solution: we would like something that not only enhances performance in favourable conditions, but also “does no harm” when curating a querybank to match the gallery closely is challenging. To address this issue, in addition to the querybank probe matrix described in Alg. 1, we also precompute a gallery activation set, \( \mathcal{A} = \{j : j \in \text{argmax}_i s(b_i, g_j), i \in \{1, \ldots, N\} \} \). Here, the notation \( \text{argmax}_i f(l) \) denotes the \( k \)-max-select operator that returns the \( k \) values of \( l \) that maximise \( f(l) \) (like \( j, l \) also runs over the gallery indices and \( k \) is set as a hyperparameter). Intuitively, this set contains the indices of gallery vectors that our querybank probe has identified as potential hubs. We create a Dynamic Inverted Softmax by activating the inverted softmax only for nearest neighbour retrievals that fall within this set:

\[
\eta_q(j) = \begin{cases} 
\frac{\exp(\beta \cdot s_q(j))}{\sum_{l=1}^{\left| G \right|} \exp(\beta \cdot s_q(l))} & \text{if argmax}_i s_q(l) \in \mathcal{A} \\
\eta_q(j) & \text{otherwise} 
\end{cases} 
\tag{2}
\]

Since \( s_q(j) \) is computed as an intermediate step in Eqn. 1, the only additional cost incurred by the Dynamic Inverted Softmax over the standard Inverted Softmax stems from the argmax operation in Eqn. 2. Fortunately, this computation can be performed extremely efficiently with almost no loss in precision, even at the scales of billions of gallery samples [47]. We show through experiments in Sec. 4, the Dynamic Inverted Softmax is significantly more robust than GC, CSLS and IS: crucially, it does not harm performance when employed with suboptimal querybank selection.

4. Experiments

In this section, we first briefly describe the datasets and metrics used for our experiments (Sec. 4.1). We then conduct a series of experiments that: (i) demonstrate our claim that QB-NORM is effective without concurrent access to more than one test query; (ii) investigate the influence of querybank size; (iii) compare the Dynamic Inverted Softmax against prior methods; (iv) ablate other QB-NORM components (Sec. 4.2). Finally, we demonstrate the generality of Querybank Normalisation by applying it to a broad range of models, tasks and datasets (Sec. 4.3).

4.1. Datasets and Evaluation Metrics

We conduct experiments on standard benchmarks for text-video retrieval: MSR-VTT [102], MSVD [13], DiDeMo [42], LSMDC [82], VaTeX [98] and QueryYD [73]. We also investigate QB-NORM on text-image retrieval (MSCoCo [16]), text-audio retrieval (AudioCaps [53]), and image-to-image retrieval (CUB-200-2011 [94], Stanford Online Products [91]). Detailed descriptions of each dataset are deferred to the supplementary. We report standard retrieval performance metrics:
Effective querybanks can be constructed from the training set. Performance is reported on MSR-VTT full split [102]. We observe that a querybank of 60K samples from the training set performs comparably to a test set querybank.

### 4.2. Querybank Normalisation

We conduct initial studies on the MSR-VTT benchmark for text-video retrieval using TT-CE+ [23] to address a series of questions relating to Querybank Normalisation.

**Do we need access to more than one test query at a time to mitigate hubness?** Prior work has investigated the use of IS for image and video retrieval with natural language queries, but only by assuming simultaneous access to the full test set of queries to construct the querybank [14, 59, 112]. The motivation for this approach [59] is to enforce a bipartite matching constraint that encodes the prior knowledge that each test query maps to exactly one gallery sample. Unfortunately, this approach is impractical to deploy for real world systems that experience sequential user queries. Therefore, we first ask whether we require concurrent access to all test set queries by constructing an alternative querybank from the training set. We evaluate performance with QB-NORM using DIS normalisation in which we construct querybanks from: (i) all test set queries; (ii) all validation set queries; (iii) a randomly subsampled subset of the training set matching the size of the test set (resampled once for each trained model to estimate variance). The results are reported in Tab. 1. Remarkably, we observe that training set querybanks perform comparably to test set querybanks. Given this finding, we conclude that test set querybanks are not necessary to mitigate hubness. We therefore restrict all querybank construction to use training set samples for all remaining experiments, ensuring valid comparisons on standard retrieval benchmarks.

**What is the influence of querybank size on performance?** To address this, we sample querybanks across a range of different scales, and report mean and standard deviations across metrics for three samplings of each scale using DIS normalisation. The results are shown in Fig. 3 (left), where we observe that performance increases with querybank size, but strong results can be obtained with a querybank of just a few thousand random training samples.

**What is the influence of the similarity normalisation strategy on QB-NORM?** To address this question, we first sample querybanks of 5,000 samples from the MSR-VTT training split and compare the normalisation strategies described in Sec. 3.4. Results are reported in the upper block (“In Domain”) of Tab. 2 where we observe that CSLS [21], IS [88] and the proposed DIS strategy perform best, and that all querybank normalisation methods substantially outperform the baseline without normalisation. Next, to evaluate the robustness of the normalisation strategies to different querybank sampling distributions, we sample additional querybanks of 5,000 samples from the training splits of two different video retrieval datasets: MSVD [13] (whose query domain closely matches MSR-VTT), and LSMDC [82] (a collection of movies with audio descriptions, whose query domain is further away from MSR-VTT), and evaluate retrieval performance on MSR-VTT test. We report results in the middle blocks of Tab. 2 (“Close Domain” and “Far Domain”) where we observe that sampling the querybank from a closely overlapping domain (MSVD) works well for all methods (with DIS performing best), but that sampling from a different domain (LSMDC) degrades performance below the baseline without normalisation for all methods except GC [24] and DIS.

To understand why the LSMDC querybank could be actively harmful for methods other than GC [24] and DIS, we studied the samples closely and observed that LSMDC queries retrieve only a small subset of videos from the video gallery (and thus were ineffectual at their primary purpose of probing for hubs). To validate that this retrieval distribution was indeed the cause of the issue, we constructed an “adversarial” querybank from MSR-VTT by selecting the 5,000 training queries that achieved the smallest coverage (i.e. retrieved the lowest number of distinct videos) over the MSR-VTT test set. We report numbers in the Adversarial block of Tab. 2. We observe that despite sampling from the same dataset, all normalisation methods other than DIS are significantly harmed. In the lower block, Overall, we present the overall performance computed as geometric mean for all methods. Since DIS performs the best overall (presented in bold in Tab. 2), we use it as our normalisation strategy.
Table 2. The influence of normalisation strategies across querybank source distributions. Performance is reported on MSR-VTT full split [102], while querybanks of 5,000 samples are sampled from the training sets of different datasets. In the last block, we presented the overall performance reported as geometric mean (GM) for each method. We observe that DIS provides the best overall trade-off: it matches the high performance of IS and CSLS with in domain and close domain querybanks, and is more robust on far domain and adversarial querybanks.

Table 3. Impact of QB-NORM on hubness on various datasets. We observe that QB-NORM consistently reduces hubness (as measured by skewness in the k-occurrences distribution).

Table 4. MSR-VTT 1k-A split: Comparison to state of the art.

Figure 4. Qualitative results. We illustrate a sample query for which QB-NORM leads to the retrieval of the correct target video (whose frames are highlighted with a dashed green line). For further examples and more detailed analysis, see supplementary.

Strategy for QB-NORM for all remaining experiments.

Hyperparameter sensitivity. The IS [21] and DIS normalisation strategies require the user to select an additional hyperparameter (the inverse temperature) that is absent from other methods. We evaluate the sensitivity of DIS to this hyperparameter in Fig. 3 (right), where we find that a value of 20 works best. In practice, we found that this value worked well consistently across datasets, and therefore we use it for all remaining experiments (with the exception of CLIP2Video [30] where we used 1.99^{-1}, since the similarities are already scaled by the method). DIS normalisation introduces an additional hyperparameter (the k maximum selection value described in Sec. 3.4). We observed that choosing k = 1 offers a good trade-off between good performance and robustness, so we simply use this value for all experiments.

Does QB-NORM mitigate hubness? The core motivation for QB-NORM is that existing cross modal retrieval methods are heavily affected by hubness (Fig. 2). To investigate whether this has been addressed by QB-NORM, we report the skewness of the k-occurrences distribution1 (which indicates the hubness of an embedding space [80]) for four datasets in Tab. 3 using a querybank consisting from all the samples from the training set. We observe that in each case, skewness (and hence hubness) is significantly reduced.

4.3. Comparison with other methods

In this section, we conduct an extensive study to evaluate the effectiveness and generality of QB-NORM on several well established benchmarks.

The influence of applying QB-NORM to cross modal embeddings for text-video retrieval are reported in Tab. 4, 5, 6, 7, 8, 9. We provide further text-video retrieval results in the supplementary. In Tab. 10 we report results for the text-image retrieval task, while in Tab. 11, 12, we report results for the image-image retrieval task. Finally, in Tab. 13, we report results for text-audio retrieval. In
Table 6. DiDeMo: Comparison to state of the art methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>R@1 ↑</th>
<th>R@5 ↑</th>
<th>R@10 ↑</th>
<th>2MR ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoEE [67]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
<tr>
<td>CE [62]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
<tr>
<td>MMT [34]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
<tr>
<td>CLIP4Clip [65]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
</tbody>
</table>

Table 7. LSMDC: Comparison to state of the art methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>R@1 ↑</th>
<th>R@5 ↑</th>
<th>R@10 ↑</th>
<th>2MR ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoEE [67]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
<tr>
<td>CE [62]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
<tr>
<td>Fast and Slow [56]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
</tbody>
</table>

Table 8. VaTeX: Comparison to state of the art methods.

<table>
<thead>
<tr>
<th>Model</th>
<th>R@1 ↑</th>
<th>R@5 ↑</th>
<th>R@10 ↑</th>
<th>2MR ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoEE [67]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
<tr>
<td>CE [62]</td>
<td>12.55</td>
<td>28.42</td>
<td>37.72</td>
<td>23.52</td>
</tr>
</tbody>
</table>

5. Limitations and societal impact

Limitations
All the normalisation techniques used with QB-NORM incur additional pre-computation costs. The proposed normalisation technique, DIS, adds a further small additional computational cost over other normalisation approaches. For a full discussion on complexity, please refer to the supplementary. We also show in Tab. 2 that adversarial querybank selection and significant domain gaps can reduce the benefits of Querybank Normalisation.

Societal impact
Cross modal retrieval is a powerful tool with both positive applications and risks of harm. Cross modal search enables efficient content discovery for researchers, musicians, artists and consumers. However, this capability also lends itself to tools of political oppression: for example, it could enable efficient searching of social media content to discover signs of political dissent.

6. Conclusions

In this work, we introduced the Querybank Normalisation framework for hubness mitigation. We also proposed the Dynamic Inverted Softmax for robust similarity normalisation. We demonstrated its broad applicability across a range of tasks, models and benchmarks.

Acknowledgements.
This work was supported by Adobe, Google and Zhejiang Lab (NO. 2022NB0AB05). The authors thank Bruno Korbar for useful suggestions, Andrew Zisserman and Jenny Hu for their support, and Adam Berenzweig for kindly sharing a copy of his earlier work. S.A. would like to acknowledge Z. Novak, N. Novak and S. Carlson in supporting his contribution.
References


[90] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2:207–218, 2014. 1


[95] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5005–5013, 2016. 2


[103] Ran Xu, Caiming Xiong, Wei Chen, and Jason J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In AAAI, 2015. 1

[104] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. 2


