Cross Modal Retrieval with Querybank Normalisation

Simion-Vlad Bogolin[1,2,*] Ioana Croitoru[1,2,*]
Hailin Jin[3] Yang Liu[1,4,†] Samuel Albanie[1,5,†]
[1]Visual Geometry Group, University of Oxford [2]Inst. of Mathematics of the Romanian Academy [3]Adobe Research
[4]Wangxuan Inst. of Computer Technology, Peking University [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/.

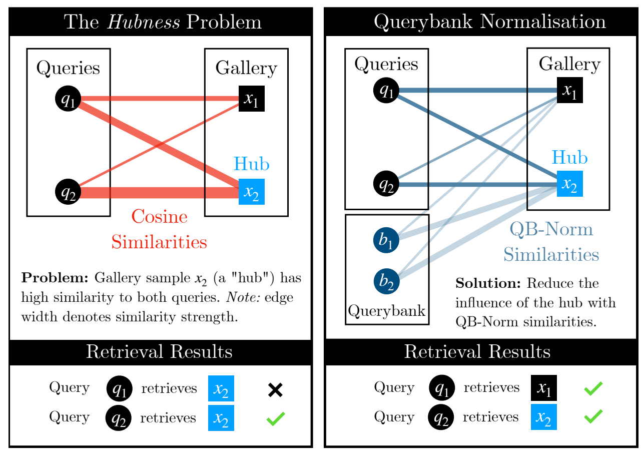

Figure 1. *Left:* **The hubness problem.** We consider the problem of cross modal retrieval in which queries q_1 and q_2 are compared against a gallery of samples, x_1 and x_2. As we show in Sec. 3.2, the high-dimensional joint embeddings employed by modern methods for cross-modal retrieval suffer from the "hubness problem" [80]. A hub (e.g. x_2) is the nearest neighbour to multiple queries (q_1 and q_2), producing poor quality retrieval results (bottom left). *Right:* **Querybank Normalisation** employs a *querybank* to normalise similarities, reducing the similarity of hub x_2 to query q_1, improving the retrieval results (bottom right).

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,

*Equal contribution. †Corresponding authors.

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-

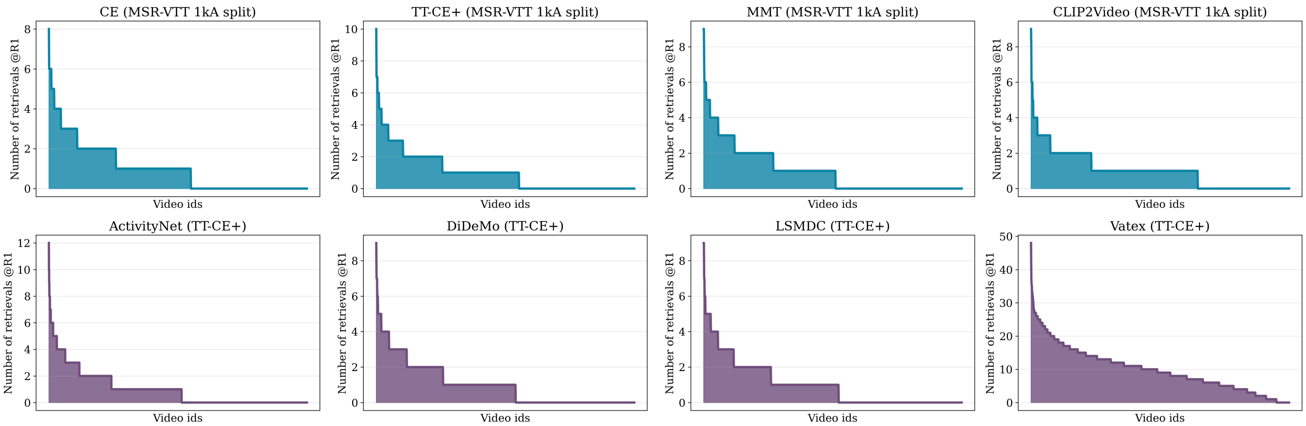

Figure 2. **Hubness is pervasive in leading methods for text-video retrieval**. The charts depict the distribution of the number of times each gallery video was retrieved by test set queries (x-axes video ids are ordered by decreasing retrieval count). *Top row (different models):* We report retrieval 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* in Sec. 4.

Also related to our work, [14, 59, 112] enforce a bipartite matching constraint between queries and test set items by applying an IS over the full set of test queries—a constraint that is unrealistic for practical retrieval systems that experience continuous operation from users. One contribution of this paper is to demonstrate that concurrent access to test queries is *not required*. A second contribution of our work, not considered in prior work, is to show that the techniques proposed above can actively damage retrieval performance for particular querybank selections, an issue that we address 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, ϕ_q and ϕ_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 language 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 *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 *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, *querybank construction* and *similarity normalisation*, described next:

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

Similarity normalisation. To normalise similarities to account for hubs, we assume access to a query, q, trained encoders ϕ_q and ϕ_g, querybank $\{b_1, \dots, b_N\}$, and a gallery $\mathcal{G}$. For each $g_j \in \mathcal{G}$, we first compute a *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 probe matrix $P \in \mathbb{R}^{|\mathcal{G}| \times N}$. Similarly, we compute for each query a vector of *unnormalised similarities*, $s_q \in \mathbb{R}^{|\mathcal{G}|}$, $s_q(j) = \text{sim}(\phi_q(q), \phi_g(g_j))$. Here $j \in \{1, \dots, |\mathcal{G}|\}$ indexes over all gallery elements. Finally, we define a querybank normalisation function, $\text{QB-NORM} : \mathbb{R}^{|\mathcal{G}|} \times \mathbb{R}^{|\mathcal{G}| \times N} \rightarrow \mathbb{R}^{|\mathcal{G}|}$, which yields, for each query q and gallery $\mathcal{G}$, a vector of querybank normalised similarities, $\eta_q = \text{QB-NORM}(s_q, P) \in \mathbb{R}^{|\mathcal{G}|}$. Various candidates for $\text{QB-NORM}(\cdot)$ 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, $\mathcal{Q}$, is summarised in Alg. 1.

3.4. Design choices

The Querybank Normalisation framework admits a number of viable choices for both querybank construction and 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.

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, $\mathcal{Q}$, (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 $\eta_q(j) = -(\text{Rank}(s_q(j), p_j) - s_q(j)) \in \mathbb{R}$, where $\text{Rank} : \mathbb{R} \times \mathbb{R}^N \rightarrow$

Algorithm 1 Ranking with Querybank Normalisation

Input: queries, $\mathcal{Q} \subset m_q$
Input: gallery, $\mathcal{G} \subset m_g$
Querybank construction.
Construct querybank, $\mathcal{B} = \{b_1, \dots, b_N\} \subset m_q$
Similarity normalisation:
Precompute querybank probe matrix
  for gallery sample $g_j \in \mathcal{G}$ **do**
    for querybank sample $b_i \in \mathcal{B}$ **do**
      Compute probe matrix entry $P(j, i) = \text{sim}(\phi_q(b_i), \phi_g(g_j)) \in \mathbb{R}$
    end for
  end for
query computations: QB-NORM *similarities*
  for query $q \in \mathcal{Q}$ **do**
    for gallery sample $g_j \in \mathcal{G}$ **do**
      Compute unnormalised similarity $s_q(j) = \text{sim}(\phi_q(q), \phi_g(g_j))$
    end for
    $\eta_q = \text{QB-NORM}(s_q, P) \in \mathbb{R}^{|\mathcal{G}|}$.
    search ranking = argsort(η_q)
  end for

$\{0, \dots, N\}$ 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]. Introduced 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: $\eta_q(j) = 2s_q(j) - \frac{1}{K}\mathbf{1}^T\hat{s}_q - \frac{1}{K}\mathbf{1}^T\hat{p}_j \in \mathbb{R}$.
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:

$$\eta_q(j) = \frac{\exp(\beta \cdot s_q(j))}{\mathbf{1}^T \exp[\beta \cdot p_j]} \in \mathbb{R} \tag{1}$$

where $\exp[\cdot]$ denotes elementwise exponentiation and β 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*, $\mathcal{A} = \{j : j \in \overset{k}{\text{argmax}}_l\, s(b_i, g_l), i \in \{1, \dots, N\}\}$. Here, the notation $\overset{k}{\text{argmax}}_l\, f(l)$ 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:

$$\eta_q(j) = \begin{cases} \frac{\exp(\beta \cdot s_q(j))}{\mathbf{1}^T \exp[\beta \cdot p_j]} & \text{if } \text{argmax}_l\, s_q(l) \in \mathcal{A} \\ s_q(j) & \text{otherwise} \end{cases} \tag{2}$$

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:

Querybank Source	Size	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
No querybank	-	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
Training set	60k	$17.3_{\pm 0.0}$	$42.1_{\pm 0.1}$	$54.9_{\pm 0.0}$	$8.0_{\pm 0.0}$
Val set	10k	$16.6_{\pm 0.1}$	$40.8_{\pm 0.1}$	$53.7_{\pm 0.1}$	$9.0_{\pm 0.0}$
Test set	60k	$17.5_{\pm 0.0}$	$42.4_{\pm 0.1}$	$55.1_{\pm 0.0}$	$8.0_{\pm 0.0}$

Table 1. **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.

R@K (recall at rank K, higher is better) and MdR (median rank, lower is better). For each study, we report the mean and standard deviation over three randomly seeded runs.

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*.

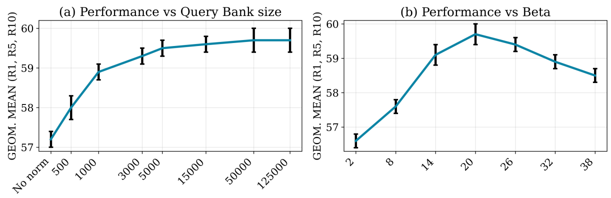

Figure 3. Retrieval results reported for a TT-CE+ [23] model on the MSR-VTT [102] benchmark for text-video retrieval with QB-NORM *DIS* normalisation. *Left: The influence of querybank size on retrieval performance on the* `validation` *split of MSR-VTT.* We observe that performance grows steadily with increasing querybank size, but saturates. *Right: The influence of inverse temperature, β, on the* `validation` *split of MSR-VTT.* Performance varies smoothly with inverse temperature, peaking at a value of 20.

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

QB Source Data	Normalisation	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
No QB	-	14.9$_{\pm0.1}$	38.3$_{\pm0.1}$	51.5$_{\pm0.1}$	10.0$_{\pm0.0}$
In Domain					
MSR-VTT	QB-Norm (GC [24])	15.8$_{\pm0.0}$	39.1$_{\pm0.0}$	51.8$_{\pm0.0}$	10.0$_{\pm0.0}$
MSR-VTT	QB-Norm (CSLS [21])	16.8$_{\pm0.1}$	41.5$_{\pm0.1}$	54.4$_{\pm0.1}$	8.0$_{\pm0.0}$
MSR-VTT	QB-Norm (IS [88])	17.1$_{\pm0.1}$	41.9$_{\pm0.2}$	54.7$_{\pm0.1}$	8.0$_{\pm0.0}$
MSR-VTT	QB-Norm (DIS)	17.0$_{\pm0.1}$	41.3$_{\pm0.1}$	54.1$_{\pm0.1}$	8.6$_{\pm0.5}$
Close Domain					
MSVD	QB-Norm (GC [24])	15.2$_{\pm0.1}$	38.8$_{\pm0.0}$	51.7$_{\pm0.0}$	10.0$_{\pm0.0}$
MSVD	QB-Norm (CSLS [21])	16.5$_{\pm0.0}$	41.2$_{\pm0.0}$	54.1$_{\pm0.1}$	9.0$_{\pm0.0}$
MSVD	QB-Norm (IS [88])	16.4$_{\pm0.2}$	40.9$_{\pm0.2}$	53.9$_{\pm0.1}$	9.0$_{\pm0.0}$
MSVD	QB-Norm (DIS)	16.7$_{\pm0.1}$	41.1$_{\pm0.1}$	54.0$_{\pm0.0}$	9.0$_{\pm0.0}$
Far Domain					
LSMDC	QB-Norm (GC [24])	14.8$_{\pm0.1}$	38.2$_{\pm0.0}$	51.4$_{\pm0.0}$	10.0$_{\pm0.0}$
LSMDC	QB-Norm (CSLS [21])	13.4$_{\pm0.0}$	35.9$_{\pm0.0}$	48.5$_{\pm0.0}$	11.0$_{\pm0.0}$
LSMDC	QB-Norm (IS [88])	11.6$_{\pm0.0}$	32.5$_{\pm0.0}$	44.6$_{\pm0.0}$	14.0$_{\pm0.0}$
LSMDC	QB-Norm (DIS)	14.9$_{\pm0.1}$	38.3$_{\pm0.1}$	51.2$_{\pm0.1}$	10.0$_{\pm0.0}$
Adversarial					
MSR-VTT	QB-Norm (GC [24])	14.5$_{\pm0.0}$	38.1$_{\pm0.0}$	51.4$_{\pm0.0}$	10.0$_{\pm0.0}$
MSR-VTT	QB-Norm (CSLS [21])	14.4$_{\pm0.1}$	37.5$_{\pm0.1}$	50.4$_{\pm0.1}$	10.0$_{\pm0.0}$
MSR-VTT	QB-Norm (IS [88])	12.3$_{\pm0.1}$	32.9$_{\pm0.1}$	45.0$_{\pm0.0}$	14.0$_{\pm0.0}$
MSR-VTT	QB-Norm (DIS)	14.9$_{\pm0.1}$	38.3$_{\pm0.1}$	51.5$_{\pm0.1}$	10.0$_{\pm0.0}$
Overall		*GM* ($R@1$)	*GM* ($R@5$)	*GM* ($R@10$)	*GM* (MdR)
Summary	QB-Norm (GC [24])	15.1$_{\pm0.6}$	38.5$_{\pm0.5}$	51.6$_{\pm0.2}$	10.0$_{\pm0.0}$
Summary	QB-Norm (CSLS [21])	15.2$_{\pm1.6}$	39.0$_{\pm2.8}$	51.8$_{\pm2.9}$	**9.4**$_{\pm1.3}$
Summary	QB-Norm (IS [88])	14.1$_{\pm2.8}$	36.8$_{\pm5.0}$	49.3$_{\pm5.5}$	10.9$_{\pm3.2}$
Summary	QB-Norm (DIS)	**15.8**$_{\pm1.1}$	**39.7**$_{\pm1.7}$	**52.7**$_{\pm1.6}$	**9.4**$_{\pm0.7}$

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.

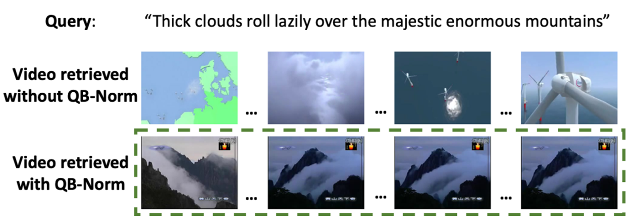

Figure 4. **Qualitative results.** We illustrate a sample query for which QB-Norm leads to the retrieval of the correct target video (whose frames are highlighted with a dashed green line). For further examples and more detailed analysis, see supplementary.

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.

MSR-VTT Before	MSR-VTT After	DiDeMo Before	DiDeMo After	LSMDC Before	LSMDC After	MSCoCo Before	MSCoCo After
0.939	0.509	1.21	0.39	0.715	0.321	0.56	0.16

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-occurences distribution).

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
CE [62]	21.7$_{\pm1.3}$	51.8$_{\pm0.5}$	65.7$_{\pm0.6}$	5.0$_{\pm0.0}$
MMT [34]	24.6$_{\pm0.4}$	54.0$_{\pm0.2}$	67.1$_{\pm0.5}$	4.0$_{\pm0.0}$
SSB [77]	27.4	56.3	67.7	3.0
Frozen [5]	31.0	59.5	70.5	3.0
CLIP4Clip [65]	44.5	71.4	81.6	**2.0**
TT-CE+ [23]	29.6$_{\pm0.3}$	61.6$_{\pm0.5}$	74.2$_{\pm0.3}$	3.0$_{\pm0.0}$
TT-CE+ (+QB-Norm)	33.3$_{\pm0.7}$	63.7$_{\pm0.1}$	76.3$_{\pm0.4}$	3.0$_{\pm0.0}$
CLIP2Video [30]	45.6	72.5	81.7	**2.0**
CLIP2Video (+QB-Norm)	**47.2**	**73.0**	**83.0**	**2.0**

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

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
VSE++ [28]	15.4	39.6	53.0	9.0
MoEE [67]	21.1$_{\pm0.2}$	52.0$_{\pm0.7}$	66.7$_{\pm0.2}$	5.0$_{\pm0.0}$
CE [62]	21.5$_{\pm0.5}$	52.3$_{\pm0.8}$	67.5$_{\pm0.7}$	5.0$_{\pm0.0}$
Frozen [5]	33.7	64.7	76.3	3.0
CLIP4Clip [65]	46.2	76.1	84.6	**2.0**
TT-CE+ [23]	25.4$_{\pm0.3}$	56.9$_{\pm0.4}$	71.3$_{\pm0.2}$	4.0$_{\pm0.0}$
TT-CE+ (+QB-Norm)	26.6$_{\pm0.9}$	58.5$_{\pm1.3}$	71.8$_{\pm1.1}$	4.0$_{\pm0.0}$
CLIP2Video [30]	47.0	76.8	85.9	**2.0**
CLIP2Video (+QB-Norm)	**47.6**	**77.6**	**86.1**	**2.0**

Table 5. **MSVD: Comparison to state of the art methods**.

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 distribution*[1] (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

[1] A detailed description of this calculation is given in the supplementary.

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
MoEE [67]	$16.1_{\pm 1.0}$	$41.2_{\pm 1.6}$	$55.2_{\pm 1.6}$	$8.3_{\pm 0.5}$
CE [62]	$17.1_{\pm 0.9}$	$41.9_{\pm 0.2}$	$56.0_{\pm 0.5}$	$8.0_{\pm 0.0}$
TT-CE	$21.0_{\pm 0.6}$	$47.5_{\pm 0.9}$	$61.9_{\pm 0.5}$	$6.0_{\pm 0.0}$
Frozen [5]	31.0	59.8	72.4	3.0
CLIP4Clip [65]	**43.4**	**70.2**	**80.6**	**2.0**
CE+ [23]	$18.2_{\pm 0.2}$	$43.9_{\pm 0.9}$	$57.1_{\pm 0.8}$	$7.9_{\pm 0.1}$
CE+ (+QB-NORM)	$20.7_{\pm 0.6}$	$46.6_{\pm 0.2}$	$59.8_{\pm 0.2}$	$6.3_{\pm 0.5}$
TT-CE+ [23]	$21.6_{\pm 0.7}$	$48.6_{\pm 0.4}$	$62.9_{\pm 0.6}$	$6.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$24.2_{\pm 0.7}$	$50.8_{\pm 0.7}$	$64.4_{\pm 0.1}$	$5.3_{\pm 0.5}$

Table 6. **DiDeMo: Comparison to state of the art methods**.

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
MoEE [67]	$12.1_{\pm 0.7}$	$29.4_{\pm 0.8}$	$37.7_{\pm 0.2}$	$23.2_{\pm 0.8}$
CE [62]	$12.4_{\pm 0.7}$	$28.5_{\pm 0.8}$	$37.9_{\pm 0.6}$	$21.7_{\pm 0.6}$
MMT [34]	$13.2_{\pm 0.4}$	$29.2_{\pm 0.8}$	$38.8_{\pm 0.9}$	$21.0_{\pm 1.4}$
Frozen [5]	15.0	30.8	39.8	20.0
CLIP4Clip [65]	**21.6**	**41.8**	**49.8**	**11.0**
CE+ [23]	$14.9_{\pm 0.6}$	$33.7_{\pm 0.2}$	$44.1_{\pm 0.6}$	$15.3_{\pm 0.5}$
CE+ (QB-NORM)	$16.4_{\pm 0.8}$	$34.8_{\pm 0.4}$	$44.9_{\pm 0.9}$	$14.5_{\pm 0.4}$
TT-CE+ [23]	$17.2_{\pm 0.4}$	$36.5_{\pm 0.6}$	$46.3_{\pm 0.3}$	$13.7_{\pm 0.5}$
TT-CE+ (QB-NORM)	$17.8_{\pm 0.4}$	$37.7_{\pm 0.5}$	$47.6_{\pm 0.6}$	$12.7_{\pm 0.5}$

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

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
HGR [15]	35.1	73.5	83.5	2.0
SSB [77]	44.6	81.8	89.5	**1.0**
CE [62]	$47.9_{\pm 0.1}$	$84.2_{\pm 0.1}$	$91.3_{\pm 0.1}$	$2.0_{\pm 0.0}$
Fast and Slow [66]	50.5	84.6	91.7	-
TT-CE+ [23]	$53.2_{\pm 0.2}$	$87.4_{\pm 0.1}$	$93.3_{\pm 0.0}$	$\mathbf{1.0}_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$54.8_{\pm 0.1}$	$88.2_{\pm 0.1}$	$\mathbf{93.8}_{\pm 0.1}$	$\mathbf{1.0}_{\pm 0.0}$
CLIP2Video [30]	57.4	87.9	93.6	**1.0**
CLIP2Video (+QB-NORM)	**58.8**	**88.3**	**93.8**	**1.0**

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

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
MoEE [67]	$11.6_{\pm 1.3}$	$30.2_{\pm 3.0}$	$43.2_{\pm 3.1}$	$14.2_{\pm 1.6}$
CE [62]	$13.9_{\pm 0.8}$	$37.6_{\pm 1.2}$	$48.3_{\pm 1.4}$	$11.3_{\pm 0.6}$
CE+ [23]	$13.2_{\pm 2.0}$	$37.1_{\pm 2.9}$	$50.5_{\pm 1.9}$	$10.3_{\pm 1.2}$
CE+ (+QB-NORM)	$14.1_{\pm 1.8}$	$\mathbf{38.6}_{\pm 1.3}$	$51.1_{\pm 1.6}$	$10.0_{\pm 0.8}$
TT-CE+ [23]	$14.4_{\pm 0.5}$	$37.7_{\pm 1.7}$	$50.9_{\pm 1.6}$	$\mathbf{9.8}_{\pm 1.0}$
TT-CE+ (+QB-NORM)	$\mathbf{15.1}_{\pm 1.6}$	$38.3_{\pm 2.4}$	$\mathbf{51.2}_{\pm 2.8}$	$10.3_{\pm 1.7}$

Table 9. **QuerYD: Comparison to state of the art methods**.

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
CLIP [79]	37.8	62.4	72.2	-
VSE++ [28]	43.9	59.4	72.4	-
OSCAR [58]	54.0	80.8	88.5	-
VinVL [110]	58.8	83.5	90.3	-
Fast and Slow [66]	**68.2**	**89.7**	**93.9**	-
CLIP [79]‡	30.3	56.1	67.1	4.0
CLIP‡ (+QB-NORM)	34.8	59.9	70.4	3.0
MMT-OSCAR [35]	52.2	80.2	88.0	**1.0**
MMT-Oscar (+QB-NORM)	53.9	80.5	88.1	**1.0**

Table 10. **Text-image retrieval - MSCoCo 5k split: Comparison to other methods.** ‡ represents the results obtained using the official CLIP [79] ViT-B/32 model.

Fig. 4 we also show a qualitative example. For the base models that provide weights for different seeds we report mean and standard deviation of QB-NORM applied on each seed. In each case, QB-NORM brings a significant improvement over all tested methods, benchmarks and tasks. We show in bold the best performing method.

Model	$R@1\uparrow$	$R@2\uparrow$	$R@4\uparrow$	$R@8\uparrow$
MS [96]	57.4	69.8	80.0	-
EPS [57]	64.4	75.2	**84.3**	-
RDML [83]	64.4	75.3	83.4	90.0
RDML [83] (+QB-NORM)	**64.8**	**75.6**	84.0	**90.4**

Table 11. **Image to Image retrieval - CUB 200: Comparison to other methods.**

Model	$R@1\uparrow$	$R@10\uparrow$	$R@100\uparrow$	$R@1000\uparrow$
XBM [97]	**80.6**	**91.6**	96.2	98.7
Smooth-AP [10]	80.1	91.5	**96.6**	**99.0**
RDML [83]	77.8	89.5	95.4	98.4
RDML [83] (+QB-NORM)	78.1	89.8	95.6	98.5

Table 12. **Image to Image retrieval - Online Products: Comparison to other methods.**

Model	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$
AR [74]-MoEE	$22.5_{\pm 0.3}$	$54.4_{\pm 0.6}$	$69.5_{\pm 0.9}$	$5.0_{\pm 0.0}$
AR [74]-CE	$23.1_{\pm 0.6}$	$55.1_{\pm 0.7}$	$70.7_{\pm 0.6}$	$4.7_{\pm 0.5}$
AR [74]-CE (+QB-NORM)	$\mathbf{23.9}_{\pm 0.2}$	$\mathbf{57.1}_{\pm 0.3}$	$\mathbf{71.6}_{\pm 0.4}$	$\mathbf{4.0}_{\pm 0.0}$

Table 13. **Text-audio retrieval - AudioCaps: Comparison to other methods.**

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

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2

[2] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 4

[3] Yusuf Aytar, Mubarak Shah, and Jiebo Luo. Utilizing semantic word similarity measures for video retrieval. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 2

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. 2021. 2, 7, 8

[6] Richard E. Bellman. Adaptive control processes: A guided tour. 1961. 2

[7] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18:509–517, 1975. 2

[8] Adam Berenzweig. *Anchors and hubs in audio-based music similarity*. PhD thesis, Columbia University, NY, USA, 2007. 2, 4

[9] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003. 2

[10] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision*, pages 677–694. Springer, 2020. 8

[11] Vicki Bruce and Andy Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986. 2

[12] Adrian Bulat, Enrique Sánchez-Lozano, and Georgios Tzimiropoulos. Improving memory banks for unsupervised learning with large mini-batch, consistency and hard negative mining. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1695–1699. IEEE, 2021. 2

[13] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 5, 6

[14] Shizhe Chen, Yida Zhao, and Qin Jin. Team ruc ai.m3: Technical report in video pentathlon challenge 2020. 3, 6

[15] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 8

[16] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015. 5

[17] Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Mitigating the hubness problem for zero-shot learning of 3d objects. In *BMVC*, 2019. 3

[18] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005. 2

[19] Ondřej Chum, Andrej Mikulík, Michal Perdoch, and Jiri Matas. Total recall ii: Query expansion revisited. *CVPR 2011*, pages 889–896, 2011. 2

[20] Ondřej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2

[21] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. 2, 3, 4, 5, 6, 7

[22] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 2

[23] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 2, 3, 4, 6, 7, 8

[24] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 2, 3, 4, 6, 7

[25] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016. 2

[26] Wei Dong, Moses Charikar, and K. Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *WWW*, 2011. 2

[27] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision*, pages 97–112. Springer, 2002. 2

[28] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 7, 8

[29] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 2

[30] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3, 4, 7, 8

[31] Roman Feldbauer and Arthur Flexer. A comprehensive empirical comparison of hubness reduction in high-dimensional spaces. *Knowledge and Information Systems*, 59(1):137–166, 2019. 2, 3, 4

[32] Damien François, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007. 2

[33] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013. 2, 3

[34] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. *European Conference on Computer Vision*, 2020. 2, 3, 4, 7, 8

[35] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *arXiv preprint arXiv:2103.11920*, 2021. 8

[36] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. *CVPR 2011*, pages 817–824, 2011. 2

[37] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *ArXiv*, abs/1410.5401, 2014. 2

[38] Rasmus Boll Greve, Emil Juul Jacobsen, and Sebastian Risi. Evolving neural turing machines for reward-based learning. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pages 117–124, 2016. 2

[39] Kazuo Hara, Ikumi Suzuki, Masashi Shimbo, Kei Kobayashi, Kenji Fukumizu, and Mil*ovs*. Radovanović. Localized centering: Reducing hubness in large-sample data. In *AAAI*, 2015. 3

[40] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2

[41] Kaiming He, Fang Wen, and Jian Sun. K-means hashing: An affinity-preserving quantization method for learning binary compact codes. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2938–2945, 2013. 2

[42] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813, 2017. 3, 5

[43] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC '98*, 1998. 2

[44] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008. 2

[45] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:117–128, 2011. 2

[46] Herve Jegou, Hedi Harzallah, and Cordelia Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[47] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019. 2, 5

[48] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *EMNLP*, 2018. 3

[49] Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *ArXiv*, abs/1703.03129, 2017. 2

[50] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017. 2

[51] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 88–95. IEEE, 2005. 2

[52] Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks. *ArXiv*, abs/1811.00783, 2019. 2

[53] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL*, 2019. 5

[54] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[55] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 3

[56] Sangho Lee, Jinyoung Sung, Youngjae Yu, and Gunhee Kim. A memory network approach for story-based temporal summarization of 360° videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1419, 2018. 2

[57] Elad Levi, Tete Xiao, Xiaolong Wang, and Trevor Darrell. Rethinking preventing class-collapsing in metric learning with margin-based losses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10316–10325, 2021. 8

[58] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020. 8

[59] Fangyu Liu and Rongtian Ye. A strong and robust baseline for text-image matching. *arXiv preprint arXiv:1906.01205*, 2019. 3, 6

[60] Fangyu Liu, Rongtian Ye, Xun Wang, and Shuaipeng Li. Hal: Improved text-image matching by mitigating visual semantic hubs. *ArXiv*, abs/1911.10097, 2020. 3

[61] Qun Liu and Supratik Mukhopadhyay. Unsupervised learning using pretrained cnn and associative memory bank. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2018. 2

[62] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 2, 3, 4, 7, 8

[63] Yang Liu, Qingchao Chen, and Samuel Albanie. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14954–14964, 2021. 2

[64] Thomas Low, Christian Borgelt, Sebastian Stober, and Andreas Nürnberger. The hubness phenomenon: Fact or artifact? In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, pages 267–278. Springer, 2013. 2

[65] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 7, 8

[66] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021. 8

[67] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2, 4, 7, 8

[68] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[69] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 2

[70] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E. Papalexakis, and Amit K. Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. *Proceedings of the 26th ACM international conference on Multimedia*, 2018. 2

[71] Jonathan Munro, Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Domain adaptation in multi-view embedding for cross-modal video retrieval. *arXiv preprint arXiv:2110.12812*, 2021. 2

[72] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *ECCV*, 2018. 2

[73] Andreea-Maria Oncescu, Joao F. Henriques, Yang Liu, Andrew Zisserman Zisserman, and Samuel Albanie. Queryd: a video dataset with high-quality textual and audio narrations. *arXiv preprint arXiv:2011.11071*, 2020. 5

[74] Andreea-Maria Oncescu, A Koepke, João F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. *Interspeech*, 2021. 1, 2, 8

[75] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016. 2

[76] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6432–6440, 2017. 2

[77] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 7, 8

[78] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 8

[80] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531, 2010. 1, 2, 4, 7

[81] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260, 2010. 2

[82] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3212, 2015. 3, 5, 6

[83] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pages 8242–8252. PMLR, 2020. 8

[84] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016. 2

[85] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research*, 13(10), 2012. 2, 4

[86] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 135–151. Springer, 2015. 3, 4

[87] Malcolm Slaney. Semantic-audio retrieval. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–4108. IEEE, 2002. 2

[88] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017. 2, 3, 4, 5, 6, 7
[89] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 966–973. IEEE, 2010. 2
[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
[91] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016. 5
[92] Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Yuji Matsumoto, and Marco Saerens. Investigating the effectiveness of laplacian-based kernels in hub reduction. In *AAAI*, 2012. 3
[93] Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu. Centering similarity measures to reduce hubs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 613–623, 2013. 3
[94] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5
[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
[96] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 8
[97] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2020. 8
[98] Xin Eric Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590, 2019. 3, 5
[99] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 2
[100] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 450–459, 2019. 2
[101] Chunpu Xu, Yu Li, Chengming Li, Xiang Ao, Min Yang, and Jinwen Tian. Interactive key-value memory-augmented attention for image paragraph captioning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3132–3142, 2020. 2
[102] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3, 4, 5, 6, 7
[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
[105] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11275–11284, 2019. 2
[106] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 2
[107] Wojciech Zaremba and Ilya Sutskever. Reinforcement learning neural turing machines-revised. *arXiv preprint arXiv:1505.00521*, 2015. 2
[108] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *NIPS*, 2004. 2
[109] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2021–2030, 2017. 3
[110] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 8
[111] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2
[112] Yida Zhao, Yuqing Song, Shizhe Chen, and Qin Jin. Ruc_aim3 at trecvid 2020: Ad-hoc video search & video to text description. In *TRECVID*, 2020. 3, 6