Slow description

Querybank Normalisation (QB-Norm)

Paper: Cross Modal Retrieval with Querybank Normalisation

*S. V. Bogolin, *I. Croitoru, H. Jin, †Y. Liu and †S. Albanie, CVPR (2022)

*equal contribution, †corresponding authors

Digest by Samuel Albanie, June 2022

Outline

- Motivation
- Related work
- Hubness
- Cross modal retrieval
- QB-Norm
- Experiments
- Discussion (limitations, societal impact)

Motivation

Joint embeddings for efficient search

Improvements in the price-performance of hardware (sensors, storage, networking) have enabled massive digital archive growth

Natural language queries (rather than SQL) are appealing

Wide range of research has developed joint embeddings for search:

images audio video

A key challenge for embeddings is the emergence of "hubs":

Embedding vectors that appear among the nearest neighbour sets of disproportionately many other embedding vectors

Does hubness affect modern joint embeddings?

Tackling hubness

Hubness can harm retrieval performance (Berenzweig, 2007)

Methods have been proposed to address it (Feldbauer et al., 2019)

Notable examples of such work in NLP and zero-shot learning

For cross modal retrieval, the inverted softmax has been proposed

However, existing approaches often have two shortcomings:

- Typically assume concurrent access to multiple test queries
- Lack robustness (can make things worse in certain settings)

These issues leave room for improvement

Can we find a practical solution to addressing hubness?

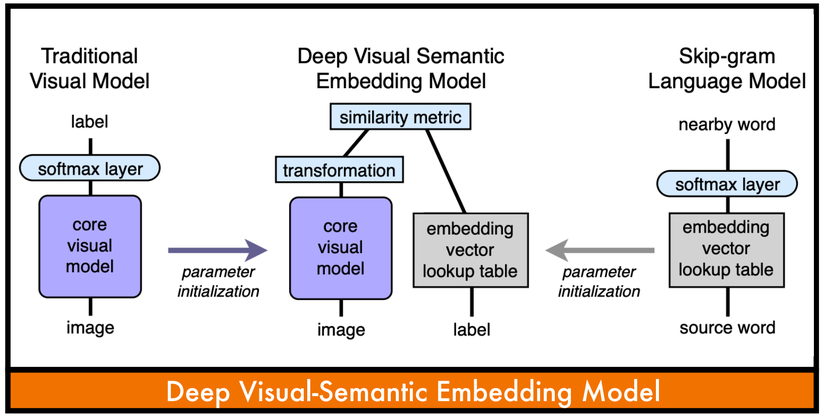
Reference:
(images) R. Socher et al., "Grounded compositional semantics for finding and describing images with sentences", ACL (2014)
(audio) A-M. Oncescu et al., "Audio retrieval with natural language queries", Interspeech (2021)
(video) R. Xu et al., "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework", AAAI (2015)
(Hubs) M. Radovanovic et al., "Hubs in space: Popular nearest neighbors in high-dimensional data", JMLR (2010)
A. Berenzweig, "Anchors and hubs in audio-based music similarity", PhD thesis (2007)
R. Feldbauer et al., "A comprehensive empirical comparison of hubness reduction in high-dimensional spaces", KIS (2019)
(inverted softmax) S. Smith et al., "Offline bilingual word vectors, orthogonal transformations and the inverted softmax", ICLR (2017)
(inverted softmax during inference) F. Liu et al., "A strong and robust baseline for text-image matching", ACL workshops (2019)

Outline

- Motivation
- Related work
- Hubness
- Cross modal retrieval
- QB-Norm
- Experiments
- Discussion (limitations, societal impact)

Related Work

Frome et al., (2013)



Deep Visual-Semantic Embedding Model

Inference:
neighbour lookup

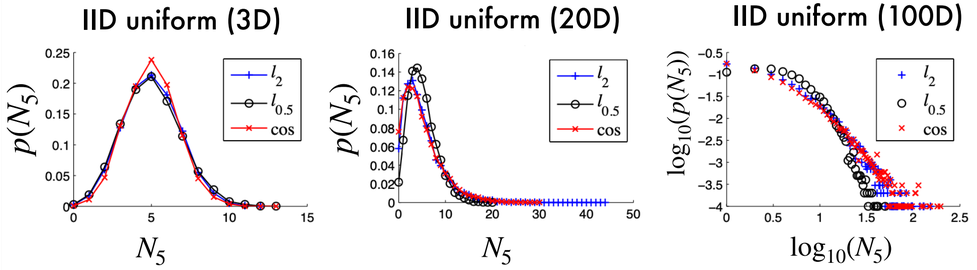
Training:
hinge rank loss

Benefits:
reasonable errors
zero-shot classifier

DeViSE

Radovanovic et al., (2010)

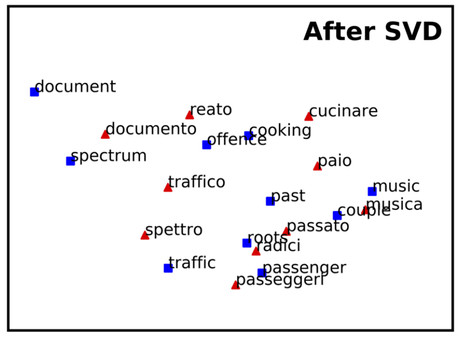
Formalised hubness as skewness in $N_k(x)$ distribution



Formulation of the hubness problem

Hubs in space

Smith et al., (2017)



Orthogonal transformations

To mitigate hubness use inverted softmax over similarities:

$$P_{j \to i} = \frac{e^{\beta S_{ij}}}{\alpha_j \sum_n e^{\beta S_{in}}}$$

n - indexes over source vectors

Inverted Softmax

Inverted Softmax

Liu et al., (2019)

Goal: improve robustness to noise

$$\min_\theta \sum_{i \in I} \sum_{\bar{t} \in K_1} [\alpha - s(i,t) + s(i,\bar{t}))]_+ + \sum_{t \in T} \sum_{\bar{i} \in K_2} [\alpha - s(t,i) + s(t,\bar{i}))]_+$$

$K_1 = \text{kNN}(i, T \backslash \{t\}, k)$ $K_2 = \text{kNN}(t, I \backslash \{i\}, k)$

More consistent than max-margin/ sum-margin ranking losses

kNN-margin loss

Use bipartite assumption

Apply Inverted Softmax/CSLS over test queries

MS COCO-5K

naive	57.8
IS	**64.2**
CSLS	62.4

Concurrent access to test queries

Leverage bipartite constraint

Mitigate hubness using test queries

References/Image credits:
A. Frome et al., "Devise: A deep visual-semantic embedding model" NeurIPS (2013)
M. Radovanovic et al., "Hubs in space: Popular nearest neighbors in high-dimensional data", JMLR (2010)
S. Smith et al., "Offline bilingual word vectors, orthogonal transformations and the inverted softmax" ICLR (2017)
F. Liu et al., "A strong and robust baseline for text-image matching", ACL workshops (2019)
(CSLS) A. Conneau et al., "Word translation without parallel data", ICLR (2018)
(MS COCO) T-Y. Lin et al., "Microsoft coco: Common objects in context", ECCV (2014)

Outline

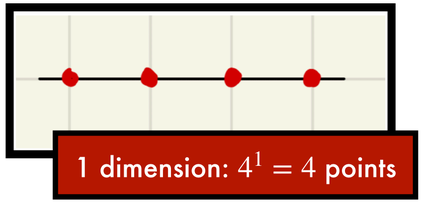
- Motivation
- Related work
- Hubness
- Cross modal retrieval
- QB-Norm
- Experiments
- Discussion (limitations, societal impact)

Curse of dimensionality

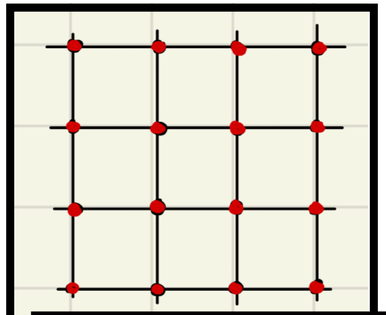
Curse of dimensionality

The *curse of dimensionality* was first noted by Bellman (1961) when developing dynamic programming for variational problems

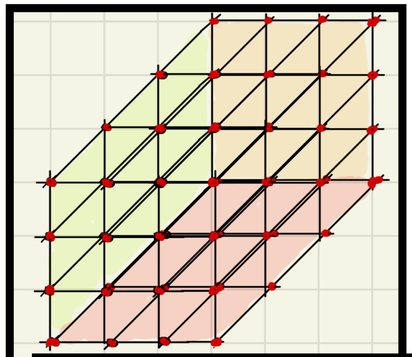
Volume of space grows very rapidly (exponentially) with **dimensionality**



1 dimension: $4^1 = 4$ points



2 dimensions: $4^2 = 16$ points



3 dimensions: $4^3 = 64$ points

Consequence: high-dimensional data is often sparse with respect to the space

High-dimensional spaces cause mischief

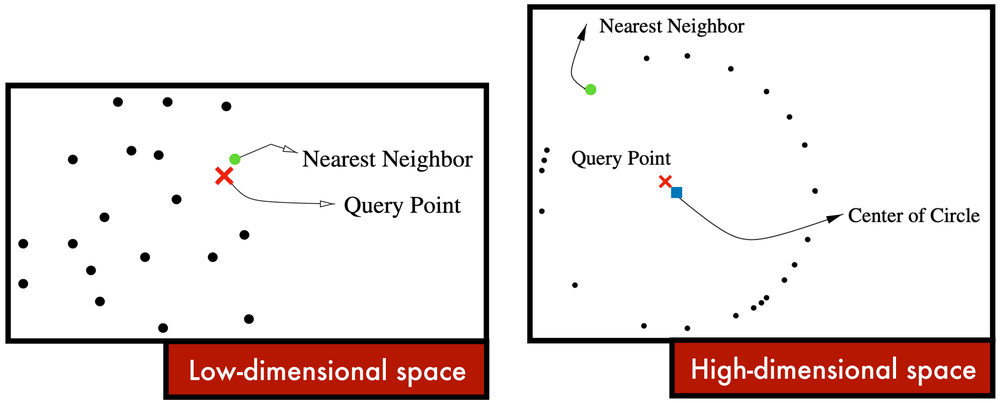
Reference:

R. E. Bellman, "Adaptive Control Processes: A Guided Tour", Princeton: Princeton University Press (1961)

High-dimensional phenomena

Distance concentration

Distance concentration: the tendency of distances between all pairs of points in high-dimensional data to become almost equal



Concentration of the Euclidean Norm (Demartines, 1994):

When sampling random i.i.d. vectors from a unit d-dim hypercube

- Expected norm of vectors grows with $\mathcal{O}(\sqrt{d})$
- Variance of vector norms is constant

Consequence: high-dimensional vectors appear to lie on a sphere

Later work has relaxed the assumptions of this theorem

The hubness phenomenon

Hubness (Radovanovic et al., 2010, Berenzweig, 2007):

Let $D \subset \mathbb{R}^d$ denote a set of d-dimensional points

Let $N_k(\mathbf{x})$ denote the number of k-occurrences of $\mathbf{x} \in D$ (number of times $\mathbf{x}$ appears in k-nearest neighbours of other points in D)

As dimensionality d increases, the distribution of N_k skews to the right

Produces "popular" neighbours (hubs) that appear in many k-NN lists

Hubs had previously been observed:

- Speech recognition (Doddington et al., 1998)
- Fingerprint identification (Hicklin et al., 2005)
- Music retrieval (Aucouturier et al., 2007; Berenzweig, 2007)

What links hubs to high-dimensional spaces?

References/Image credits:
K. Beyer et al., "When is “nearest neighbor” meaningful?", ICDT (1999)
P. Demartines, "Analyse de données par réseaux de neurones auto-organisés", Dissertation (in French) (1994)
D. François et al., "The concentration of fractional distances", TKDE (2007)
M. Radovanovic et al., "Hubs in space: Popular nearest neighbors in high-dimensional data", JMLR (2010)
A. Berenzweig, "Anchors and hubs in audio-based music similarity", PhD thesis (2007)
G. Doddington et al., "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation" (1998)
A. R. Hicklin et al., "The Myth of the Goats: How Many People Have Fingerprints that are Hard to Match?" (2005)
J-J. Aucouturier et al., "A scale-free distribution of false positives for a large class of audio similarity measures", PR (2008)

What causes hubs?

Theory of Radovanovic et al.

<u>Two key ingredients for theory</u>. In high-dimensional spaces:

(1) Some points will remain non-trivially closer to the mean than others

(2) These points become hubs

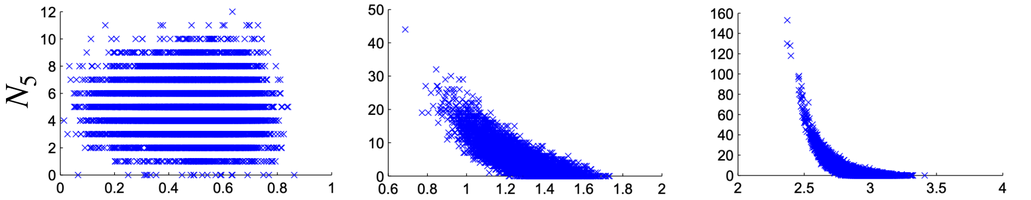
We focus on unimodal case (can extend to mixture distributions)

In high dimensions, points close to the mean are likely to be hubs

Build intuition by examining empirical evidence

Sample data IID from uniform distribution

3 dimensions | 20 dimensions | 100 dimensions



Points closer to the mean tend to become hubs!

Same effect observed for IID normal (+ distributions exhibiting hubness)

References/Image credits:
M. Radovanovic et al., "Hubs in space: Popular nearest neighbors in high-dimensional data", JMLR (2010)

Theory of Radovanovic et al.

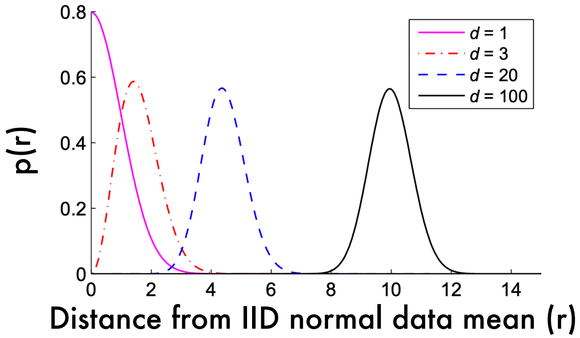
(1) Some points will remain non-trivially closer to the mean than others

Distribution of distances to the dataset mean has non-negligible variance for finite dimensions

Follows since variance independent of dimensionality in theory of distance concentration

So, a non-negligible number of points closer to the mean is expected in high dimensions

Distribution of distances from IID normal data mean



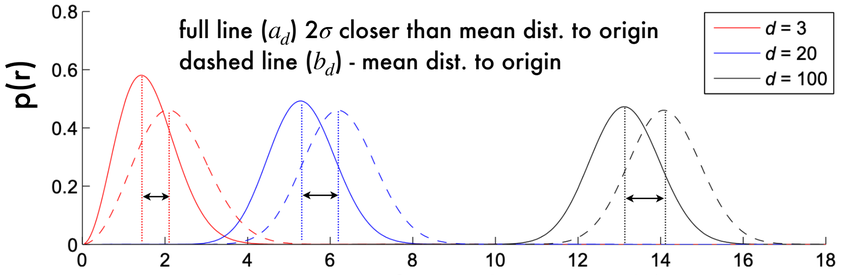
(2) Points closer to the mean become hubs in higher dimensions

Consider two points drawn from the data:

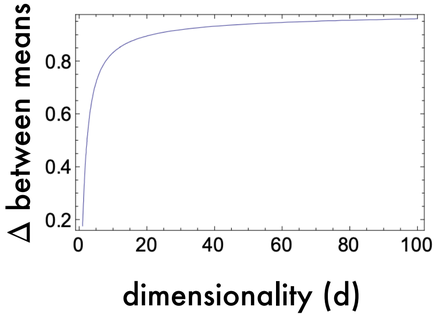
point a_d (2σ closer to origin than mean distance to origin)

point b_d (mean distance from the origin)

Distribution of distances for a_d and b_d IID normal data



Influence of increasing dim.



higher dimension ⟹ separation ⟹ hubs

What causes hubs? A different theory

Theory of Low et al.

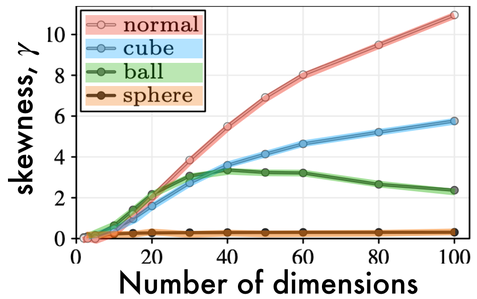
Alternative perspective: hubness is not due to high-dimensionality

Instead, it is a "boundary effect" (effect of a density gradient)

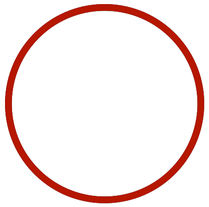
Consequently, it is an artefact of the data generation process

Experiment 1: hubness is related to a density gradient

10 nearest neighbours, 10K points



2D hypersphere (circle)



1D finite space without boundaries

Hubness does indeed increase with dimension when sampling

from a normal distribution

uniformly from a hypercube

Note: more hubness for normal distribution (hints at density gradients)

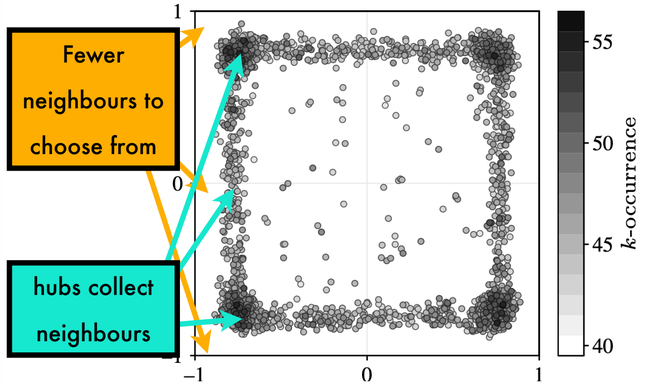
Far less hubness when sampling uniformly from a hyperball

An $(m+1)$-dim hypersphere is an m-dim finite, boundary-less space

No hubness is detected when sampling uniformly from a hypersphere

Experiment 2: hubs in low-dims

Sample 250 data sets uniformly from a square



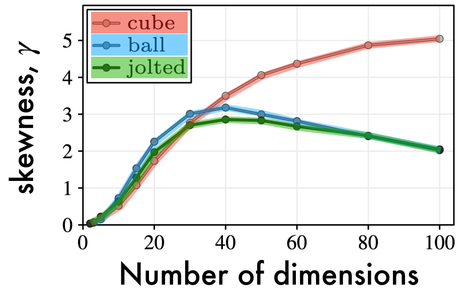
Effect at edges (most pronounced at corners)

Experiment 4: jolting

Greater hubness for hypercube than hyperball

Idea: "jolt" points of hypercube into hyperball

10 nearest neighbours, 10K points

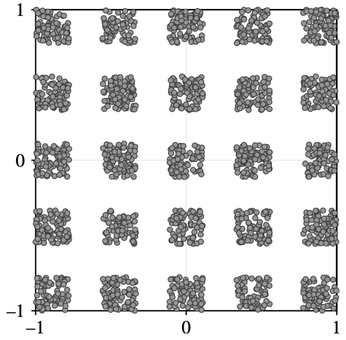


Jolted shape: fewer points near boundary

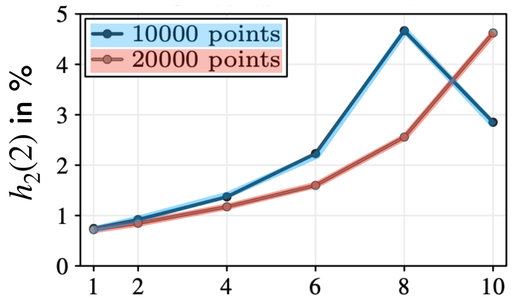
Experiment 3: linking hubs to boundaries

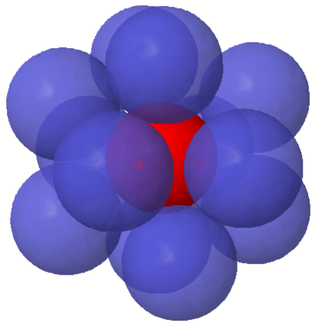
If true cause is density gradients (at boundaries) it should be possible to create hubs in low dims

Grid of squares/cubes



5 nearest neighbours (3D)





Kissing number 12 (3D)

References/Image credits:

T. Low et al., "The hubness phenomenon: Fact or artifact?", Towards Advanced Data Analysis by Combining Soft Computing and Statistics (2013)

(Kissing Number) By Robertwb at English Wikipedia, Transferred o Commons by Yelm., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6058714

Outline

- Motivation
- Related work
- Hubness
- Cross modal retrieval
- QB-Norm
- Experiments
- Discussion (limitations, societal impact)

Cross Modal Retrieval

Task and metrics

Task formulation

Objective: rank a gallery of samples according to how well they match a query

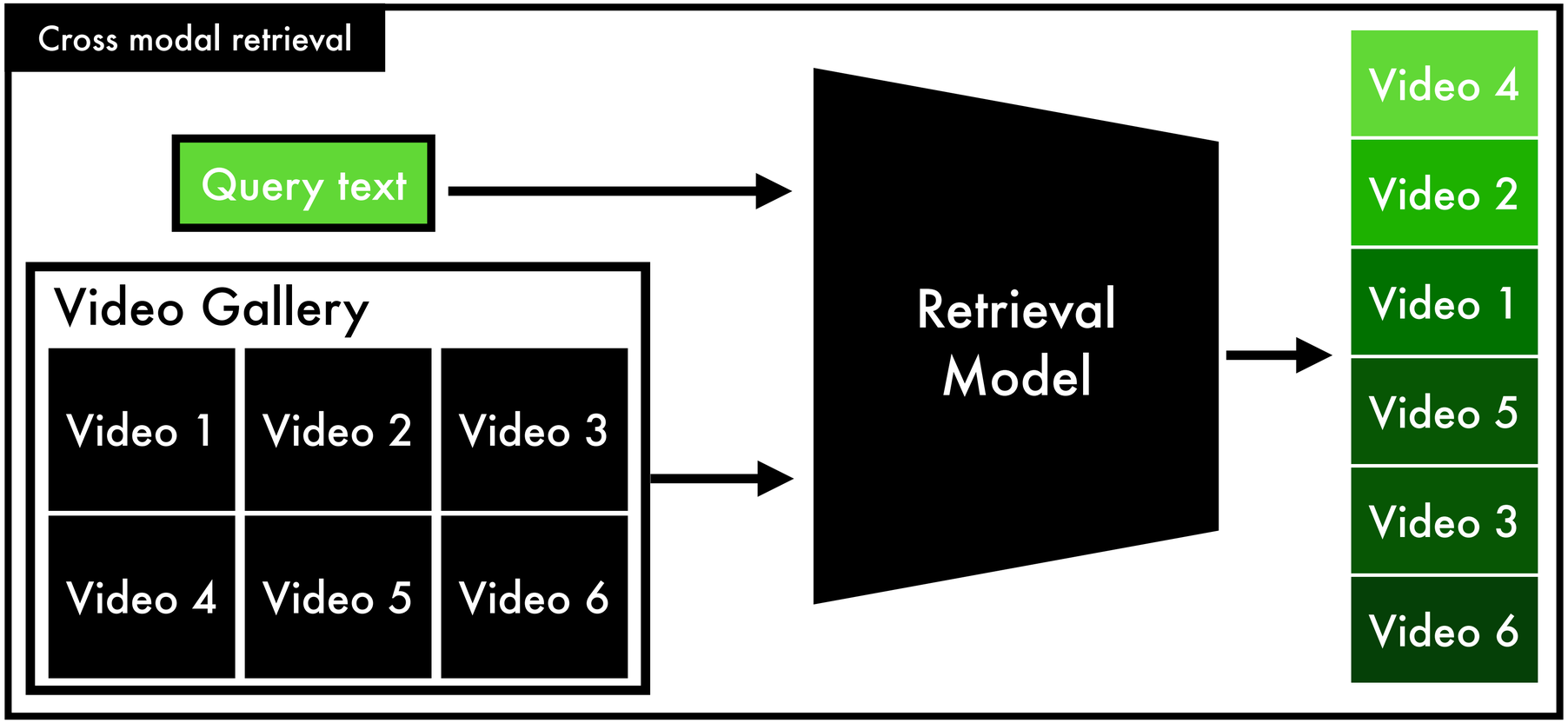
In cross modal retrieval, the query and the gallery samples have different modalities

To make things concrete, we will focus on text queries for a gallery of videos

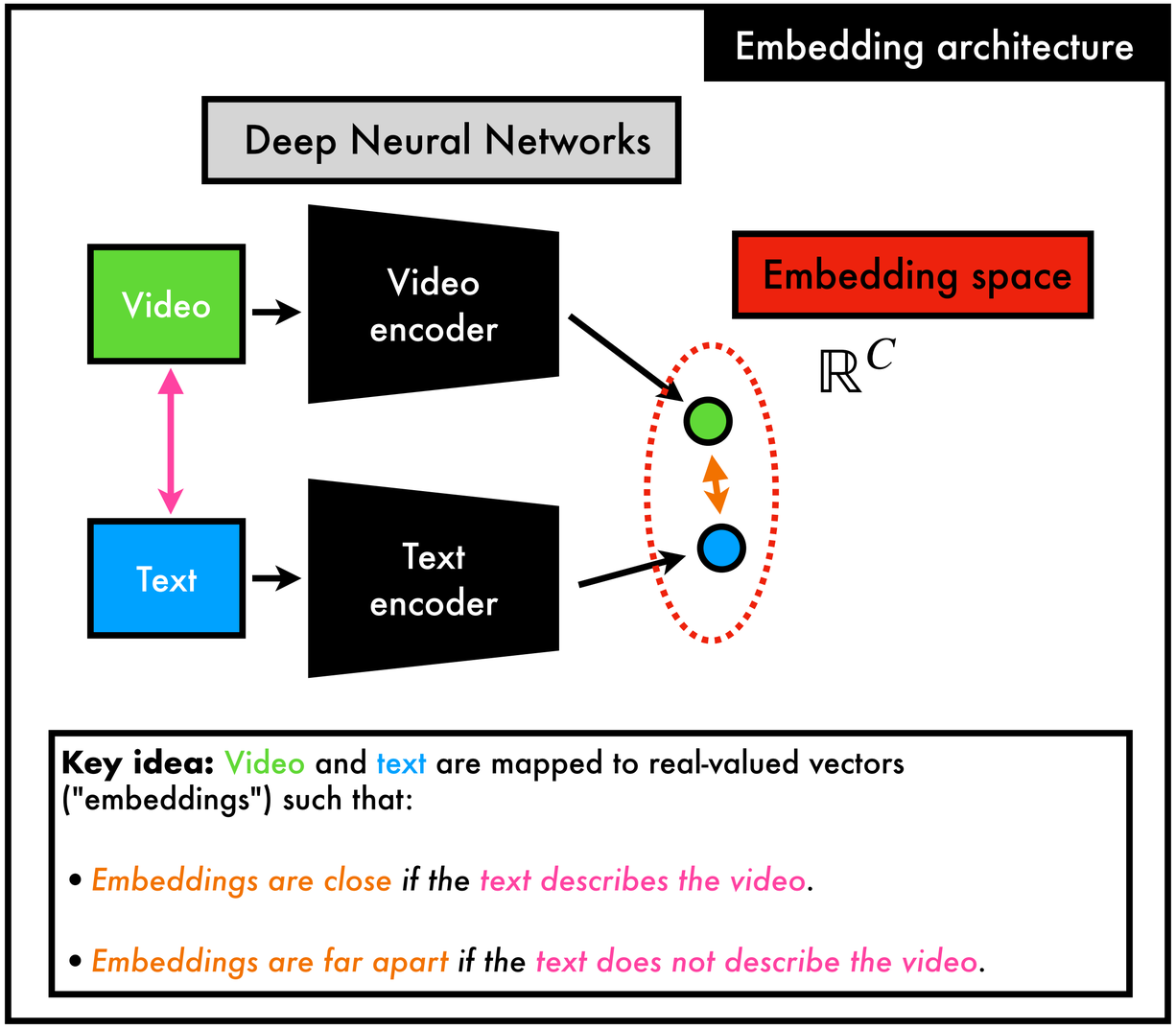
Performance metrics

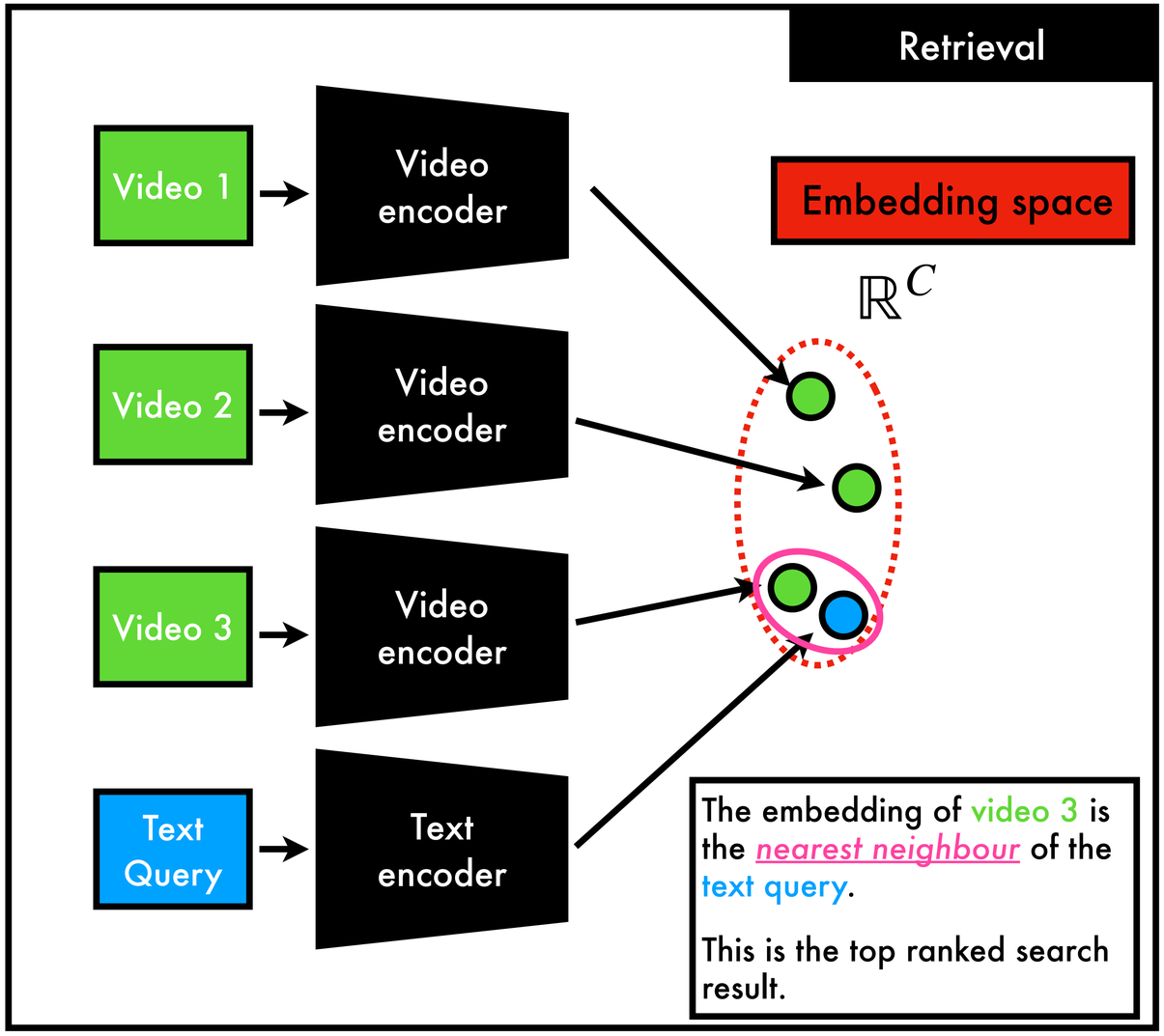
Recall@k: % of test queries where ground truth target is ranked in the top k videos *(Higher is better)*

MdR: Median rank of the ground truth target video *(Lower is better)*



Cross Modal Embeddings for Retrieval

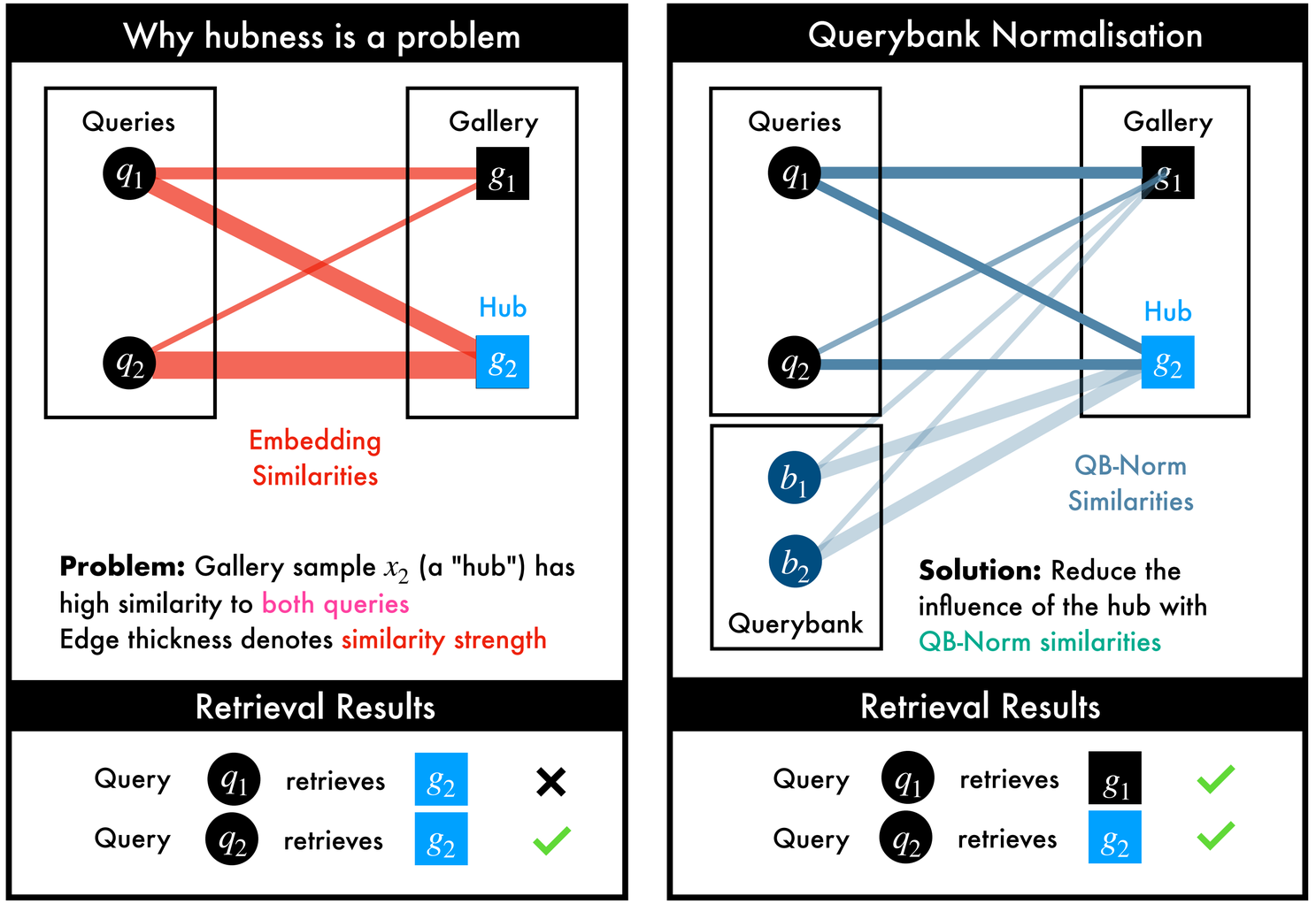




Outline

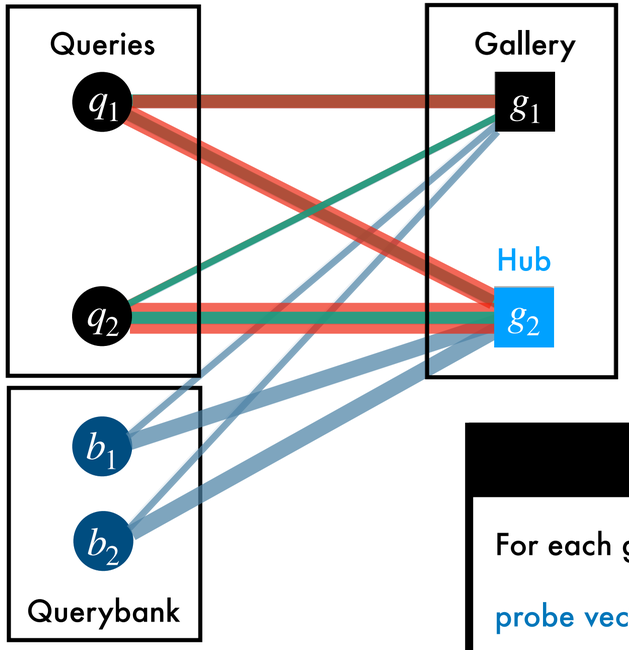
- Motivation
- Related work
- Hubness
- Cross modal retrieval
- QB-Norm
- Experiments
- Discussion (limitations, societal impact)

Querybank Normalisation (QB-Norm)



References/Image credits:
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)

QB-Norm details



Querybank construction

We'd like to probe the hubness of gallery samples

For this we first construct a querybank of N samples:

$B = \{b_1, \ldots, b_N\}$ from the query modality

Similarity normalisation

For each gallery sample $g_j \in G$ we compute a probe vector

probe vector $p_j \in \mathbb{R}^N$ defined via $p_j(i) = \mathrm{sim}(\phi_q(b_i), \phi_g(g_j))$

Probe vectors are stacked to form a probe matrix $P \in \mathbb{R}^{|G| \times N}$

For each query q, we compute unnormalised similarities $s_q \in \mathbb{R}^{|G|}$:

$s_q(j) = \mathrm{sim}(\phi_q(q), \phi_g(g_j))$ (where j indexes over the gallery)

Define querybank normalisation, QB-NORM: $\mathbb{R}^{|G|} \times \mathbb{R}^{|G| \times N} \rightarrow \mathbb{R}^{|G|}$

For each query q and gallery G compute $\eta_q = \text{QB-NORM}(s_q, P) \in \mathbb{R}^{|G|}$

Various designs for QB-NORM can be considered

Algorithm

Algorithm 1 Ranking with Querybank Normalisation

Input: queries, $\mathcal{Q} \subset m_q$
Input: gallery, $\mathcal{G} \subset m_g$
Querybank construction.
Construct querybank, $\mathcal{B} = \{b_1, \ldots, b_N\} \subset m_q$
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 $P(j,i) = \mathrm{sim}(\phi_q(b_i), \phi_g(g_j)) \in \mathbb{R}$
    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 $s_q(j) = \mathrm{sim}(\phi_q(q), \phi_g(g_j))$
    end for
    $\eta_q = \text{QB-NORM}(s_q, P) \in \mathbb{R}^{|\mathcal{G}|}$.
    search ranking = argsort(η_q)
  end for

amortise cost over queries

References/Image credits:
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)

QB-Norm design choices

Globally-Corrected (GC)

Dinu et al. (2014) consider zero-shot learning (bilingual translation, image labelling)

Propose two techniques to mitigate hubness

Compute of hubness of gallery ("targets") using test set queries ("pivots")

Also explore expansion to further unlabelled samples

$\mathrm{NN}_{\mathrm{nrm}}$ - normalise gallery similarities w.r.t queries

GC - reverse the querying (return gallery sample that has ranked the query highest among all queries)

Break ties with cosine similarities

GC in QB-Norm framework

Build querybank from test queries (+ further samples)

QB-Norm similarities:

$$\eta_q(j) = -\left(\mathrm{Rank}(s_q(j), p_j) - s_q(j)\right) \in \mathbb{R}$$

Inverted Softmax (IS)

Smith et al. (2017) tackle the problem of aligning word embeddings from distinct languages

The word embeddings are trained independently, then brought into alignment with a linear transformation (prove this should be orthogonal)

Translation of a word amounts to nearest neighbours

To mitigate hubness, assess the probability that the target word translates back into the source word (less susceptible to hubs than forward direction)

IS in QB-Norm framework

Build querybank from subsampled queries

QB-Norm similarities: $\eta_q(j) = \dfrac{\exp(\beta \cdot s_q(j))}{\mathbf{1}^T \exp[\beta \cdot p_j]} \in \mathbb{R}$

β - inverse temperature

$\exp[\,\cdot\,]$ - element-wise exponentiation

Cross-domain Similarity Local Scaling (CSLS)

Conneau et al. (2018) consider unsupervised alignment of word embeddings from distinct languages

GC + IS perform asymmetric updates

IS requires cross-validation to fit β

Propose Cross-domain Similarity Local Scaling (CSLS)

Increases similarity associated with isolated vectors

Decreases similarity of vectors in dense areas

Similarities between vectors are down-weighted with the average similarity in their local neighbourhood

CSLS in QB-Norm framework

Build querybank from all available queries

Let $\hat{p}_j \in \mathbb{R}^K$ denote probe vector restricted to K querybank samples closest to gallery sample g_j

Let $\hat{s}_q \in \mathbb{R}^K$ denote unnormalised similarity vector s_q restricted to K gallery samples closest to q

QB-Norm similarities: $\eta_q(j) = 2s_q(j) - \dfrac{1}{K}\mathbf{1}^T\hat{s}_q - \dfrac{1}{K}\mathbf{1}^T\hat{p}_j \in \mathbb{R}$

References:

(GC) G. Dinu et al. "Improving zero-shot learning by mitigating the hubness problem", ICLR Workshops (2015)

(IS) S. Smith et al., "Offline bilingual word vectors, orthogonal transformations and the inverted softmax" ICLR (2017)

(CSLS) A. Conneau et al., "Word translation without parallel data", ICLR (2018)

Dynamic Inverted Softmax

Dynamic Inverted Softmax (DIS)

In experiments with the Inverted Softmax, performance degrades when querybank and gallery differ significantly

Not ideal for a general-purpose solution

Want to work well in favourable conditions, but "do no harm" when building a good querybank is challenging

Precompute a gallery activation set:

similarities between querybank and gallery vectors

$$\mathcal{A} = \{j : j \in \overset{k}{\operatorname{argmax}}_l \; s(b_i, g_l), i \in \{1,\ldots,N\}\}$$

$\overset{k}{\operatorname{argmax}}_l \; f(l)$ - denotes k-max-select operator that returns k values of l that maximise $f(l)$ (both j and l index over the gallery)

Intuition: the gallery activation set $\mathcal{A}$ contains indices of gallery vectors that the querybank identifies as potential hubs

Dynamic Inverted Softmax only activates the Inverted Softmax for nearest neighbour retrievals in the gallery activation set:

QB-Norm similarities: $\eta_q(j) = \begin{cases} \dfrac{\exp(\beta \cdot s_q(j))}{\mathbf{1}^T \exp[\beta \cdot p_j]} & \text{if } \operatorname{argmax}_l s_q(l) \in \mathcal{A} \\ s_q(j) & \text{otherwise} \end{cases}$

Extra inference cost over IS stems from the argmax operation

This nearest-neighbour search can be performed (approximately) very efficiently (Johnson et al., 2019)

Benefit of Dynamic Inverted Softmax: improved robustness

References:
J. Johnson et al., "Billion-scale similarity search with gpus", TBD (2019)

(Naive) Computational Complexity

Influence of normalisation strategy on inference

For clarity of exposition, we consider the cost of exact similarity searches (in practice, approximate nearest neighbours are employed for large-scale deployments)

We describe naive implementations to convey intuition

All strategies incur an initial $\mathcal{O}(N)$ cost corresponding to computing the similarity between a test query and all N samples in the gallery

Further assume we have pre-computed similarities between each of the M querybank queries and each gallery sample with compute and storage costs of $\mathcal{O}(NM)$

Globally-Corrected

For Globally-Corrected (GC), we must determine the rank of the test query w.r.t each gallery item

Given pre-computed similarities between the querybank and the gallery, we pre-compute ranks

For each test query, we establish its rank among the querybank for each gallery sample

Binary search over sorted pre-computed similarities

Cost: $\mathcal{O}(N \log M)$

Cross-domain Similarity Local Scaling

For Cross-domain Similarity Local Scaling (CSLS) find:

1) K querybank samples closest to each gallery sample

2) K gallery samples closest to the query

For 1), we can pre-compute the K closest querybank similarities and store the averages in a vector of size N

For 2), we compute the avg. similarity of K most similar items among the gallery to test query at inference

With quickselect, this requires $\mathcal{O}(N)$ time on average (we do not need the top K similarities to be sorted)

Inverted Softmax

The Softmax denominator involves normalising by the sum of similarities given the querybank

We can pre-compute this for each gallery item and store results in a vector of size N

At inference, similarities are divided by this sum, $\mathcal{O}(1)$

Pre-computing the sum also means we can reduce storage cost for the querybank from $\mathcal{O}(MN)$ to $\mathcal{O}(N)$

Dynamic Inverted Softmax

Gallery activation set - pre-compute and store: $\mathcal{O}(N)$

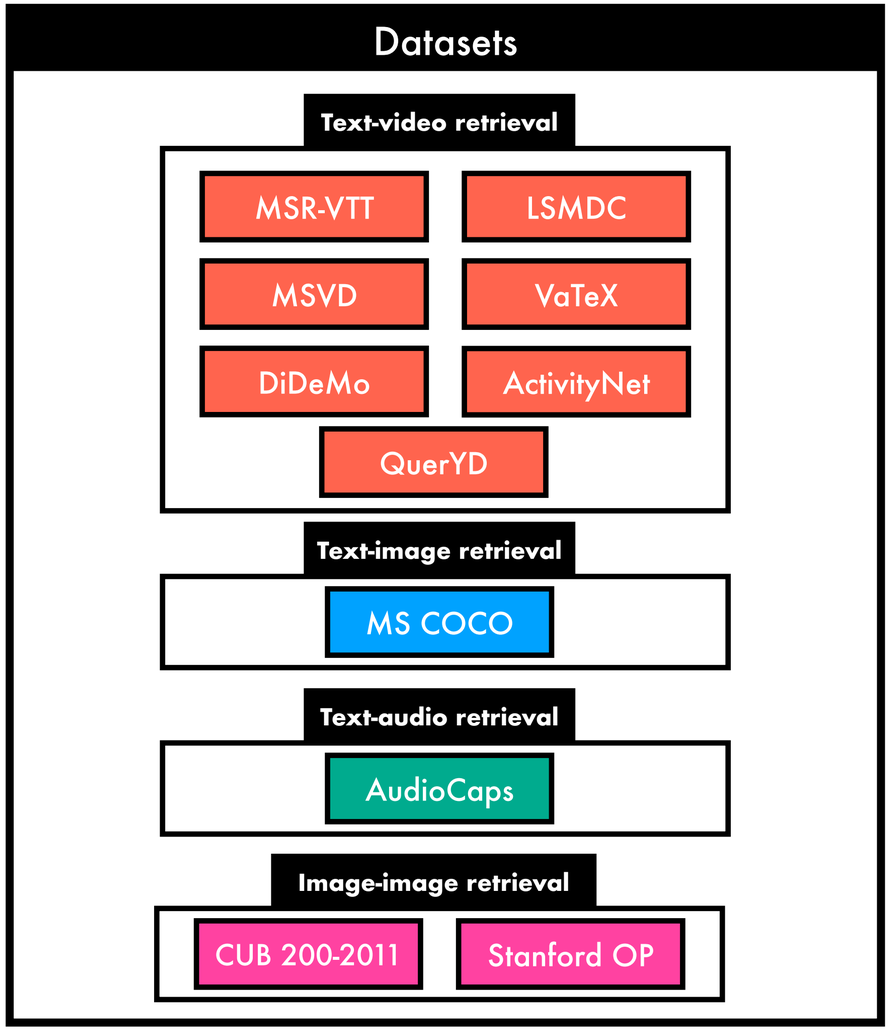
Top-1 search to determine best hit: linear time $\mathcal{O}(N)$

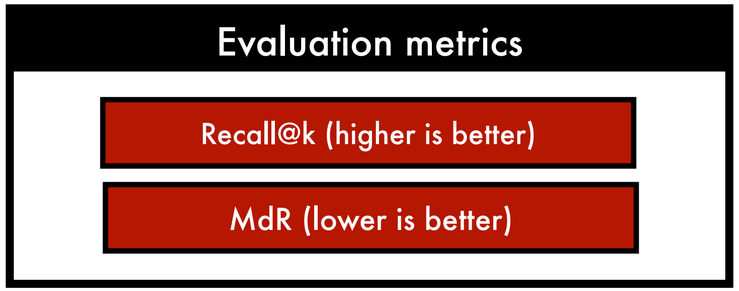
References/Image credits:
(GC) G. Dinu et al. "Improving zero-shot learning by mitigating the hubness problem", ICLR Workshops (2015)
(CSLS) A. Conneau et al., "Word translation without parallel data", ICLR (2018)
(IS) S. Smith et al., "Offline bilingual word vectors, orthogonal transformations and the inverted softmax" ICLR (2017)

Outline

- Motivation
- Related work
- Hubness
- Cross modal retrieval
- QB-Norm
- Experiments
- Discussion (limitations, societal impact)

Experiments: datasets and evaluation metrics

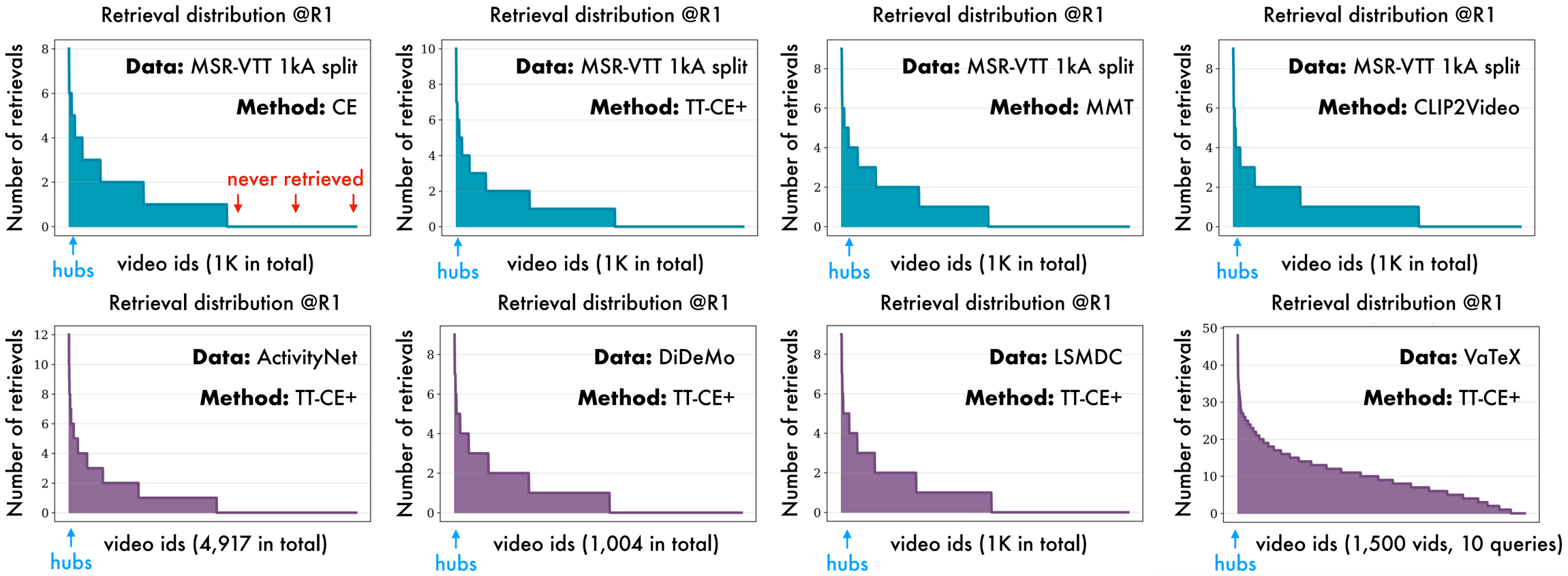




References:
(MSR-VTT) J. Xu et al., "MSR-VTT: A large video description dataset for bridging video and language", CVPR (2016)
(MSVD) D. Chen et al., "Collecting highly parallel data for paraphrase evaluation", ACL-HTL (2011)
(DiDeMo) L. A. Hendricks et al., "Localizing moments in video with natural language", ICCV (2017)
(LSMDC) A. Rohrbach et al., "Movie description", IJCV (2017)
(VaTeX) X. Wang et al., "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research", ICCV (2019)
(ActivityNet) F. Caba Heilbron et al. "Activitynet: A large-scale video benchmark for human activity understanding", CVPR (2015)
(QuerYD) A-M. Oncescu et al., "QuerYD: A video dataset with high-quality textual and audio narrations", ICASSP (2021)
(MS COCO) T-Y. Lin et al., "Microsoft coco: Common objects in context", ECCV (2014)
(AudioCaps) C. D. Kim et al., "AudioCaps: Generating captions for audios in the wild", NACL-HLT (2019)
(AudioCaps for retrieval) A-M. Oncescu et al., "Audio retrieval with natural language queries", Interspeech (2021)
(CUB 200-2011) C. Wah et al., "The caltech-ucsd birds-200-2011 dataset", (2011)
(Stanford OP) H. Oh Song et al., "Deep metric learning via lifted structured feature embedding", CVPR (2016)

Experiment: hubness in text-video retrieval methods

Motivating study - text-video retrieval distributions



hubness not confined to one dataset or method

References/Image credits:

S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)
(MSR-VTT) J. Xu et al., "MSR-VTT: A large video description dataset for bridging video and language", CVPR (2016)
(CE) Y. Liu et al., "Use what you have: Video retrieval using representations from collaborative experts", BMVC (2019)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
(ActivityNet) F. Caba Heilbron et al. "Activitynet: A large-scale video benchmark for human activity understanding", CVPR (2015)
(MMT) V. Gabeur et al., "Multi-modal transformer for video retrieval", ECCV (2020)
(CLIP2Video) H. Fang et al., "Clip2video: Mastering video-text retrieval via image clip", arxiv (2021)
(DiDeMo) L. A. Hendricks et al., "Localizing moments in video with natural language", ICCV (2017)
(LSMDC) A. Rohrbach et al., "Movie description", IJCV (2017)
(VaTeX) X. Wang et al., "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research", ICCV (2019)

Experiments: QB-Norm

Do we need multiple test queries at a time?

Prior work by Liu et al. (2019) has shown benefits of the Inverted Softmax for cross modal retrieval with natural language queries

However, it assumes concurrent access to test queries (bipartite assumption)

This works for many benchmarks (often curated from captioning datasets)

But it may not reflect real world deployments of retrieval systems

Question: *Do we need access to multiple queries at a time?*

Experiment: apply DIS to TT-CE+ on MSR-VTT with different querybanks

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	$17.5_{\pm 0.0}$	$42.4_{\pm 0.1}$	$55.1_{\pm 0.0}$	$8.0_{\pm 0.0}$

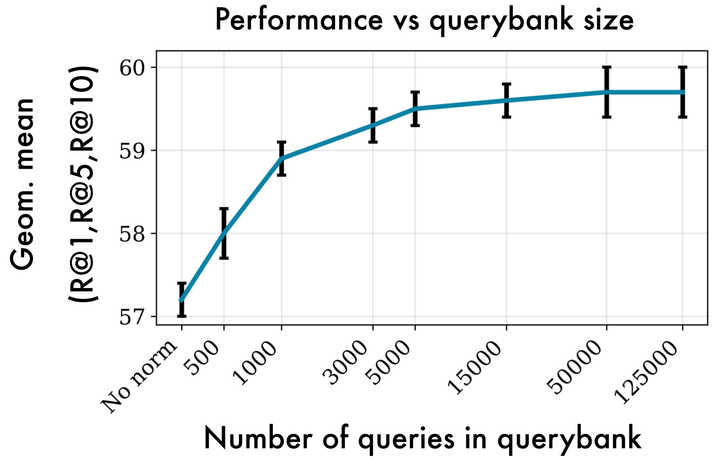
Standard deviations across three randomly seeded runs

Conclusion: test set querybanks are not required to mitigate hubness

Enables the practical deployment of querybank normalisation

What is the influence of querybank size?

Experiment: DIS with TT-CE+ on MSR-VTT with different querybank sizes



Conclusion: Performance improves with querybank size

Strong results can be obtained with a few thousand training samples

References/Image credits:
F. Liu et al., "A strong and robust baseline for text-image matching", ACL workshops (2019)
(MSR-VTT) J. Xu et al., "MSR-VTT: A large video description dataset for bridging video and language", CVPR (2016)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)

Experiments: QB-Norm

Influence of similarity normalisation strategy

Experiment: Sample a querybank of 5,000 samples

Compare normalisation strategies for TT-CE+

Takeaway 1: *in domain* all QB-Norm normalisation methods bring gains

Takeaway 2: *close domain* all QB-Norm normalisation methods bring gains

Takeaway 3: *far domain* GC and DIS (no damage), CSLS and IS (hurt)

Question: Why are *far domain* querybanks harmful for some methods?

It was observed that the querybank only retrieves a small subset of videos from the gallery (ineffective at probing for hubs)

To validate that the retrieval distribution was linked to performance:

Construct an "adversarial" querybank for MSR-VTT (in domain)

Takeaway 4: DIS is most robust to querybanks that achieve poor coverage

Takeaway 5: DIS represents a good all-round choice.

Note: DIS is used for the remaining experiments

QB Source Data	Normalisation	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
No QB	-	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
In Domain					
MSR-VTT	QB-NORM (GC)	$15.8_{\pm 0.0}$	$39.1_{\pm 0.0}$	$51.8_{\pm 0.0}$	$10.0_{\pm 0.0}$
MSR-VTT	QB-NORM (CSLS)	$16.8_{\pm 0.1}$	$41.5_{\pm 0.1}$	$54.4_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	QB-NORM (IS)	$17.1_{\pm 0.1}$	$41.9_{\pm 0.2}$	$54.7_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	QB-NORM (DIS)	$17.0_{\pm 0.1}$	$41.3_{\pm 0.1}$	$54.1_{\pm 0.1}$	$8.6_{\pm 0.5}$
Close Domain					
MSVD	QB-NORM (GC)	$15.2_{\pm 0.1}$	$38.8_{\pm 0.0}$	$51.7_{\pm 0.0}$	$10.0_{\pm 0.0}$
MSVD	QB-NORM (CSLS)	$16.5_{\pm 0.0}$	$41.2_{\pm 0.0}$	$54.1_{\pm 0.1}$	$9.0_{\pm 0.0}$
MSVD	QB-NORM (IS)	$16.4_{\pm 0.2}$	$40.9_{\pm 0.2}$	$53.9_{\pm 0.1}$	$9.0_{\pm 0.0}$
MSVD	QB-NORM (DIS)	$16.7_{\pm 0.1}$	$41.1_{\pm 0.1}$	$54.0_{\pm 0.0}$	$9.0_{\pm 0.0}$
Far Domain					
LSMDC	QB-NORM (GC)	$14.8_{\pm 0.1}$	$38.2_{\pm 0.0}$	$51.4_{\pm 0.0}$	$10.0_{\pm 0.0}$
LSMDC	QB-NORM (CSLS)	$13.4_{\pm 0.0}$	$35.9_{\pm 0.0}$	$48.5_{\pm 0.0}$	$11.0_{\pm 0.0}$
LSMDC	QB-NORM (IS)	$11.6_{\pm 0.0}$	$32.5_{\pm 0.0}$	$44.6_{\pm 0.0}$	$14.0_{\pm 0.0}$
LSMDC	QB-NORM (DIS)	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.2_{\pm 0.1}$	$10.0_{\pm 0.0}$
Adversarial					
MSR-VTT	QB-NORM (GC)	$14.5_{\pm 0.0}$	$38.1_{\pm 0.0}$	$51.4_{\pm 0.0}$	$10.0_{\pm 0.0}$
MSR-VTT	QB-NORM (CSLS)	$14.4_{\pm 0.1}$	$37.5_{\pm 0.1}$	$50.4_{\pm 0.1}$	$10.0_{\pm 0.0}$
MSR-VTT	QB-NORM (IS)	$12.3_{\pm 0.1}$	$32.9_{\pm 0.1}$	$45.0_{\pm 0.0}$	$14.0_{\pm 0.0}$
MSR-VTT	QB-NORM (DIS)	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
Overall		GM $(R@1)$	GM $(R@5)$	GM $(R@10)$	GM (MdR)
Summary	QB-NORM (GC)	$15.1_{\pm 0.6}$	$38.5_{\pm 0.5}$	$51.6_{\pm 0.2}$	$10.0_{\pm 0.0}$
Summary	QB-NORM (CSLS)	$15.2_{\pm 1.6}$	$39.0_{\pm 2.8}$	$51.8_{\pm 2.9}$	$\mathbf{9.4}_{\pm 1.3}$
Summary	QB-NORM (IS)	$14.1_{\pm 2.8}$	$36.8_{\pm 5.0}$	$49.3_{\pm 5.5}$	$10.9_{\pm 3.2}$
Summary	QB-NORM (DIS)	$\mathbf{15.8}_{\pm 1.1}$	$\mathbf{39.7}_{\pm 1.7}$	$\mathbf{52.7}_{\pm 1.6}$	$\mathbf{9.4}_{\pm 0.7}$

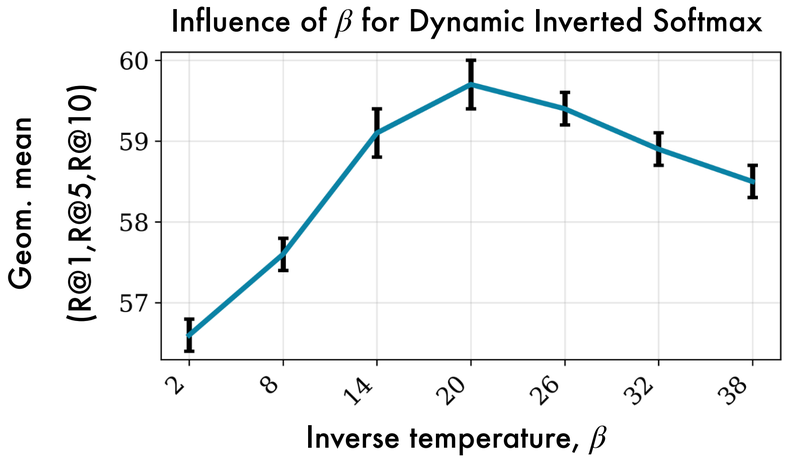
References/Image credits:
(MSR-VTT) J. Xu et al., "MSR-VTT: A large video description dataset for bridging video and language", CVPR (2016)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)
(GC) G. Dinu et al. "Improving zero-shot learning by mitigating the hubness problem", ICLR Workshops (2015)
(CSLS) A. Conneau et al., "Word translation without parallel data", ICLR (2018)
(IS) S. Smith et al., "Offline bilingual word vectors, orthogonal transformations and the inverted softmax" ICLR (2017)
(MSVD) D. Chen et al., "Collecting highly parallel data for paraphrase evaluation", ACL-HTL (2011)
(LSMDC) A. Rohrbach et al., "Movie description", IJCV (2017)

Experiments: QB-Norm hyperparameters

Influence of softmax temperature

Methods such as IS and DIS require a choice of inverse temperature (β)

Experiment: DIS with TT-CE+ on MSR-VTT with different temperatures



A value of $\beta = 20$ worked well for TT-CE+ (used for all TT-CE+ experiments)

A good choice for this hyperparameter depends on the model/embedding

Best set via validation using a held-out set

Example: CLIP2Video (different scaling), $\beta = 1.99^{-1}$ is used

Influence of k hyperparameter on DIS

The DIS uses top-k selection to construct the gallery activation set

Experiment: apply DIS to TT-CE+ on MSR-VTT with 5K querybank

Querybank Source Data	Topk	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
No querybank	-	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
In Domain					
MSR-VTT	1	$17.0_{\pm 0.1}$	$41.3_{\pm 0.1}$	$54.1_{\pm 0.1}$	$8.6_{\pm 0.5}$
MSR-VTT	2	$17.1_{\pm 0.1}$	$41.7_{\pm 0.1}$	$54.5_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	3	$17.1_{\pm 0.1}$	$41.8_{\pm 0.1}$	$54.6_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	5	$17.1_{\pm 0.1}$	$41.9_{\pm 0.1}$	$54.7_{\pm 0.1}$	$8.0_{\pm 0.0}$
MSR-VTT	10	$17.1_{\pm 0.1}$	$41.9_{\pm 0.1}$	$54.7_{\pm 0.1}$	$8.0_{\pm 0.0}$
Far Domain					
LSMDC	1	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.2_{\pm 0.1}$	$10.0_{\pm 0.0}$
LSMDC	2	$14.8_{\pm 0.0}$	$38.0_{\pm 0.0}$	$51.0_{\pm 0.0}$	$10.0_{\pm 0.0}$
LSMDC	3	$14.7_{\pm 0.0}$	$37.9_{\pm 0.0}$	$50.9_{\pm 0.0}$	$10.0_{\pm 0.0}$
LSMDC	5	$14.6_{\pm 0.0}$	$37.8_{\pm 0.0}$	$50.8_{\pm 0.0}$	$10.0_{\pm 0.0}$
LSMDC	10	$14.5_{\pm 0.0}$	$37.5_{\pm 0.0}$	$50.4_{\pm 0.0}$	$10.0_{\pm 0.0}$

slightly better

slightly worse

Select $k = 1$ as a good compromise between in domain/far domain

References/Image credits:
(IS) S. Smith et al., "Offline bilingual word vectors, orthogonal transformations and the inverted softmax" ICLR (2017)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)
(CLIP2Video) H. Fang et al., "Clip2video: Mastering video-text retrieval via image clip", arxiv (2021)

Experiments: influence of QB-Norm on hubness

Does QB-Norm mitigate hubness?

QB-Norm motivation: cross modal retrieval methods suffer from hubness

Does QB-Norm actually mitigate hubness?

Define, N_k as the k-occurrence distribution via $N_k(\mathbf{x}) = \sum_{i=1}^{n} p_{i,k}(\mathbf{x})$

$$p_{i,k}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is among the } k \text{ nearest neighbours of } q_i \\ 0 & \text{otherwise} \end{cases}$$

Compute the skewness of the k-occurrence distribution, N_k:

$$S_{N_k} = \frac{E(N_k - \mu_{N_k})^3}{\sigma_{N_k}^3}$$

μ_{N_k} - denotes the mean of N_k σ_{N_k} - denotes the standard deviation of N_k

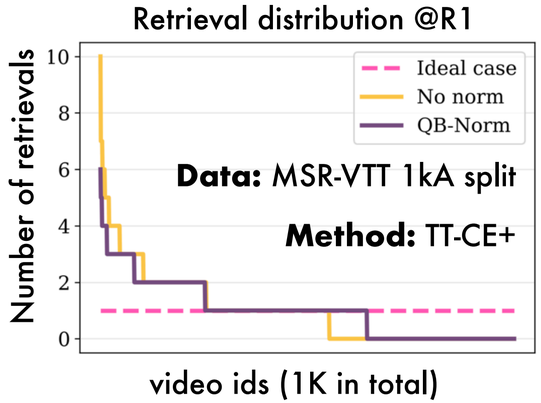
Here, $\mathbf{x}$ is a gallery video and q_i is a query text

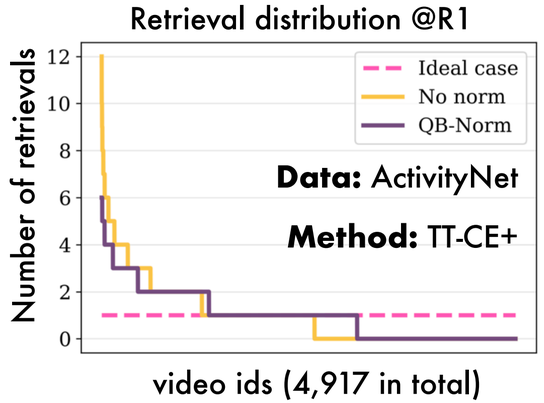
Take $k = 10$ following Feldbauer et al. (2018)

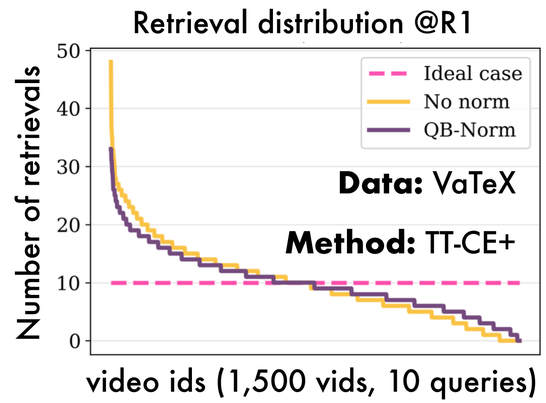
MSR-VTT		DiDeMo		LSMDC		MSCoCo	
Before	After	Before	After	Before	After	Before	After
0.939	0.509	1.21	0.39	0.715	0.321	0.56	0.16

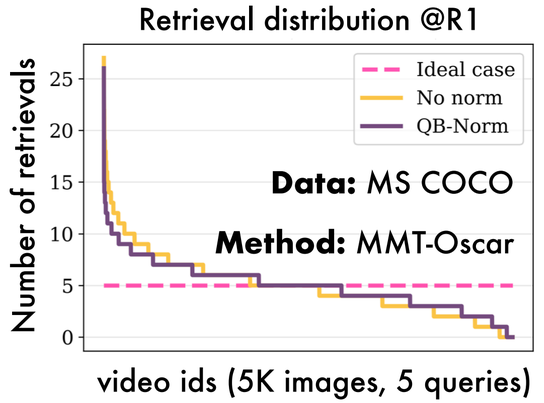
Takeaway: QB-Norm produces a significant reduction in skewness

Influence of QB-Norm on retrieval distribution









References/Image credits:

M. Radovanovic et al., "Hubs in space: Popular nearest neighbors in high-dimensional data", JMLR (2010)
R. Feldbauer et al., "Fast approximate hubness reduction for large high-dimensional data", ICBK (2018)
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)
(MSR-VTT) J. Xu et al., "MSR-VTT: A large video description dataset for bridging video and language", CVPR (2016)
(DiDeMo) L. A. Hendricks et al., "Localizing moments in video with natural language", ICCV (2017)
(LSMDC) A. Rohrbach et al., "Movie description", IJCV (2017)
(MS COCO) T-Y. Lin et al., "Microsoft coco: Common objects in context" ECCV, (2014)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
(ActivityNet) F. Caba Heilbron et al. "Activitynet: A large-scale video benchmark for human activity understanding", CVPR (2015)
(VaTeX) X. Wang et al., "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research", ICCV (2019)
(MMT-Oscar) G. Geigle et al., "Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval", arxiv (2021)

Experiments: QB-Norm for text-video retrieval

MSR-VTT 1kA split

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
CE	$21.7_{\pm 1.3}$	$51.8_{\pm 0.5}$	$65.7_{\pm 0.6}$	$5.0_{\pm 0.0}$
MMT	$24.6_{\pm 0.4}$	$54.0_{\pm 0.2}$	$67.1_{\pm 0.5}$	$4.0_{\pm 0.0}$
SSB	27.4	56.3	67.7	3.0
Frozen	31.0	59.5	70.5	3.0
CLIP4Clip	44.5	71.4	81.6	**2.0**
TT-CE+	$29.6_{\pm 0.3}$	$61.6_{\pm 0.5}$	$74.2_{\pm 0.3}$	$3.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$33.3_{\pm 0.7}$	$63.7_{\pm 0.1}$	$76.3_{\pm 0.4}$	$3.0_{\pm 0.0}$
CLIP2Video	45.6	72.5	81.7	**2.0**
CLIP2Video (+QB-NORM)	**47.2**	**73.0**	**83.0**	**2.0**

VaTeX

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

MSR-VTT full split

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
Dual	7.7	22.0	31.8	32.0
HGR	9.2	26.2	36.5	24.0
MoEE	$11.1_{\pm 0.1}$	$30.7_{\pm 0.1}$	$42.9_{\pm 0.1}$	$15.0_{\pm 0.0}$
CE	$11.0_{\pm 0.0}$	$30.8_{\pm 0.1}$	$43.3_{\pm 0.3}$	$15.0_{\pm 0.0}$
CE+	$14.4_{\pm 0.1}$	$37.4_{\pm 0.1}$	$50.2_{\pm 0.1}$	$10.0_{\pm 0.0}$
CE+ (+QB-NORM)	$16.4_{\pm 0.0}$	$40.3_{\pm 0.1}$	$53.0_{\pm 0.1}$	$9.0_{\pm 0.0}$
TT-CE+	$14.9_{\pm 0.1}$	$38.3_{\pm 0.1}$	$51.5_{\pm 0.1}$	$10.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$17.3_{\pm 0.0}$	$42.1_{\pm 0.1}$	$54.9_{\pm 0.1}$	$8.0_{\pm 0.0}$
CLIP4Clip‡	27.9	52.7	63.6	5.0
CLIP4Clip (+QB-Norm)	**29.6**	**54.5**	**65.3**	**4.0**

‡Results obtained with CLIP4Clip codebase

QuerYD

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

overlap in standard deviations

weak effect on QuerYD

References/Image credits:
(MSR-VTT) J. Xu et al., "MSR-VTT: A large video description dataset for bridging video and language", CVPR (2016)
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)
(CE) Y. Liu et al., "Use what you have: Video retrieval using representations from collaborative experts", BMVC (2019)
(MMT) V. Gabeur et al., "Multi-modal transformer for video retrieval", ECCV (2020)
(SSB) M. Patrick et al., "Support-set bottlenecks for video-text representation learning", ICLR (2021)
(Frozen) M. Bain et al., "Frozen in time: A joint video and image encoder for end-to-end retrieval", ICCV (2021)
(CLIP4Cllip) H. Luo et al., "Clip4clip: An empirical study of clip for end to end video clip retrieval", arxiv (2021)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
(CLIP2Video) H. Fang et al., "Clip2video: Mastering video-text retrieval via image clip", arxiv (2021)
(Dual) J. Dong et al., "Dual encoding for zero-example video retrieval", CVPR (2019)
(HGR) S. Chen et al., "Fine-grained video-text retrieval with hierarchical graph reasoning", CVPR (2020)
(MoEE) A. Miech et al., "Learning a text-video embedding from incomplete and heterogeneous data", arxiv (2018)
(VaTeX) X. Wang et al., "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research", ICCV (2019)
(Fast and Slow) A. Miech et al., "Thinking fast and slow: Efficient text-to-visual retrieval with transformers", CVPR (2021)
(QuerYD) A-M. Oncescu et al., "QuerYD: A video dataset with high-quality textual and audio narrations", ICASSP (2021)

Experiments: QB-Norm for text-video retrieval

MSVD

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
VSE++	15.4	39.6	53.0	9.0
MoEE	$21.1_{\pm 0.2}$	$52.0_{\pm 0.7}$	$66.7_{\pm 0.2}$	$5.0_{\pm 0.0}$
CE	$21.5_{\pm 0.5}$	$52.3_{\pm 0.8}$	$67.5_{\pm 0.7}$	$5.0_{\pm 0.0}$
Frozen	33.7	64.7	76.3	3.0
CLIP4Clip	46.2	76.1	84.6	**2.0**
TT-CE+	$25.4_{\pm 0.3}$	$56.9_{\pm 0.4}$	$71.3_{\pm 0.2}$	$4.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$26.6_{\pm 0.9}$	$58.5_{\pm 1.3}$	$71.8_{\pm 1.1}$	$4.0_{\pm 0.0}$
CLIP2Video	47.0	76.8	85.9	**2.0**
CLIP2Video (+QB-NORM)	**47.6**	**77.6**	**86.1**	**2.0**

ActivityNet

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@50 \uparrow$	$MdR \downarrow$
MoEE	$19.7_{\pm 0.3}$	$50.0_{\pm 0.5}$	$92.0_{\pm 0.2}$	$5.3_{\pm 0.5}$
CE	$19.9_{\pm 0.3}$	$50.1_{\pm 0.7}$	$92.2_{\pm 0.6}$	$5.3_{\pm 0.5}$
HSE	20.5	49.3	–	–
MMT	$22.7_{\pm 0.2}$	$54.2_{\pm 1.0}$	$93.2_{\pm 0.4}$	$5.0_{\pm 0.0}$
SSB	26.8	58.1	93.5	3.0
CLIP4Clip	40.5	**72.4**	**98.1**	**2.0**
TT-CE+	$23.5_{\pm 0.2}$	$57.2_{\pm 0.5}$	$96.1_{\pm 0.1}$	$4.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$27.0_{\pm 0.2}$	$60.6_{\pm 0.4}$	$96.8_{\pm 0.0}$	$4.0_{\pm 0.0}$
CLIP4Clip‡	36.3	65.9	96.8	3.0
CLIP4Clip (+QB-Norm)	**41.4**	71.4	97.6	**2.0**

‡Results obtained with CLIP4Clip codebase

LSMDC

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
MoEE	$12.1_{\pm 0.7}$	$29.4_{\pm 0.8}$	$37.7_{\pm 0.2}$	$23.2_{\pm 0.8}$
CE	$12.4_{\pm 0.7}$	$28.5_{\pm 0.8}$	$37.9_{\pm 0.6}$	$21.7_{\pm 0.6}$
MMT	$13.2_{\pm 0.4}$	$29.2_{\pm 0.8}$	$38.8_{\pm 0.9}$	$21.0_{\pm 1.4}$
Frozen	15.0	30.8	39.8	20.0
CLIP4Clip	21.6	**41.8**	**49.8**	**11.0**
CE+	$14.9_{\pm 0.6}$	$33.7_{\pm 0.2}$	$44.1_{\pm 0.6}$	$15.3_{\pm 0.5}$
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+	$17.2_{\pm 0.4}$	$36.5_{\pm 0.6}$	$46.3_{\pm 0.3}$	$13.7_{\pm 0.5}$
TT-CE+ (QB-NORM)	$17.8_{\pm 0.4}$	$37.7_{\pm 0.5}$	$47.6_{\pm 0.6}$	$12.7_{\pm 0.5}$
CLIP4Clip‡	21.3	40.0	49.5	**11.0**
CLIP4Clip (+QB-NORM)	**22.4**	40.1	49.5	**11.0**

‡Results obtained with CLIP4Clip codebase

DiDeMo

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
MoEE	$16.1_{\pm 1.0}$	$41.2_{\pm 1.6}$	$55.2_{\pm 1.6}$	$8.3_{\pm 0.5}$
CE	$17.1_{\pm 0.9}$	$41.9_{\pm 0.2}$	$56.0_{\pm 0.5}$	$8.0_{\pm 0.0}$
TT-CE	$21.0_{\pm 0.6}$	$47.5_{\pm 0.9}$	$61.9_{\pm 0.5}$	$6.0_{\pm 0.0}$
Frozen	31.0	59.8	72.4	3.0
CLIP4Clip	**43.4**	70.2	80.6	**2.0**
CE+	$18.2_{\pm 0.2}$	$43.9_{\pm 0.9}$	$57.1_{\pm 0.8}$	$7.9_{\pm 0.1}$
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+	$21.6_{\pm 0.7}$	$48.6_{\pm 0.4}$	$62.9_{\pm 0.6}$	$6.0_{\pm 0.0}$
TT-CE+ (+QB-NORM)	$24.2_{\pm 0.7}$	$50.8_{\pm 0.7}$	$64.4_{\pm 0.1}$	$5.3_{\pm 0.5}$
CLIP4Clip‡	43.0	70.5	80.0	**2.0**
CLIP4Clip (+QB-NORM)	43.3	**71.4**	**80.8**	**2.0**

‡Results obtained with CLIP4Clip codebase

References/Image credits:
(MSVD) D. Chen et al., "Collecting highly parallel data for paraphrase evaluation", ACL-HTL (2011)
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)
(VSE++) F. Faghri et al., "VSE++: Improving visual-semantic embeddings with hard negatives", BMVC (2018)
(MoEE) A. Miech et al., "Learning a text-video embedding from incomplete and heterogeneous data", arxiv (2018)
(CE) Y. Liu et al., "Use what you have: Video retrieval using representations from collaborative experts", BMVC (2019)
(Frozen) M. Bain et al., "Frozen in time: A joint video and image encoder for end-to-end retrieval", ICCV (2021)
(CLIP4Cllip) H. Luo et al., "Clip4clip: An empirical study of clip for end to end video clip retrieval", arxiv (2021)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
(CLIP2Video) H. Fang et al., "Clip2video: Mastering video-text retrieval via image clip", arxiv (2021)
(LSMDC) A. Rohrbach et al., "Movie description", IJCV (2017)
(MMT) V. Gabeur et al., "Multi-modal transformer for video retrieval", ECCV (2020)
(ActivityNet) F. Caba Heilbron et al. "Activitynet: A large-scale video benchmark for human activity understanding", CVPR (2015)
(HSE) B. Zhang et al., "Cross-modal and hierarchical modeling of video and text", ECCV (2018)
(SSB) M. Patrick et al., "Support-set bottlenecks for video-text representation learning", ICLR (2021)
(DiDeMo) L. A. Hendricks et al., "Localizing moments in video with natural language", ICCV (2017)

Experiments: QB-Norm for other retrieval tasks

MS COCO 5K (text-image retrieval)

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

Stanford Online Products (image-image retrieval)

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

weak effect on Stanford Online Products

CUB-200-2011 (image-image retrieval)

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

AudioCaps (text-audio retrieval)

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

References/Image credits:

S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)
(MS COCO) T-Y. Lin et al., "Microsoft coco: Common objects in context", ECCV (2014)
(MS COCO for retrieval) X. Chen et al., "Microsoft coco captions: Data collection and evaluation server", arxiv (2015)
(CLIP) A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
(VSE++) F. Faghri et al., "VSE++: Improving visual-semantic embeddings with hard negatives", BMVC (2018)
(OSCAR) X. Li et al., "OSCAR: Object-semantics aligned pre-training for vision-language tasks", ECCV (2020)
(VinVL) P. Zhang et al., "VinVL: Revisiting visual representations in vision-language models", CVPR (2021)
(Fast and Slow) A. Miech et al., "Thinking fast and slow: Efficient text-to-visual retrieval with transformers", CVPR (2021)
(MMT-Oscar) G. Geigle et al., "Retrieve fast, rerank smart: Coop. and joint approaches for improved cross-modal retrieval", arxiv (2021)
(CUB-200-2011) C. Wah et al., "The caltech-ucsd birds-200-2011 dataset", (2011)
(MS) X. Wang et al., "Multi-similarity loss with general pair weighting for deep metric learning", CVPR (2019)
(EPS) E. Levi et al., "Rethinking preventing class-collapsing in metric learning with margin-based losses", ICCV (2021)
(RDML) K. Roth et al., "Revisiting training strategies and generalization performance in deep metric learning", ICML (2020)
(Stanford OP) H. Oh Song et al., "Deep metric learning via lifted structured feature embedding", CVPR (2016)
(XBM) X. Wang et al., "Cross-batch memory for embedding learning", CVPR (2020)
(Smooth-AP) A. Brown et al., "Smooth-AP: Smoothing the path towards large-scale image retrieval", ECCV (2020)
(AudioCaps) C. D. Kim et al., "AudioCaps: Generating captions for audios in the wild", NACL-HLT (2019)
(AR) A-M. Oncescu et al., "Audio retrieval with natural language queries", Interspeech (2021)

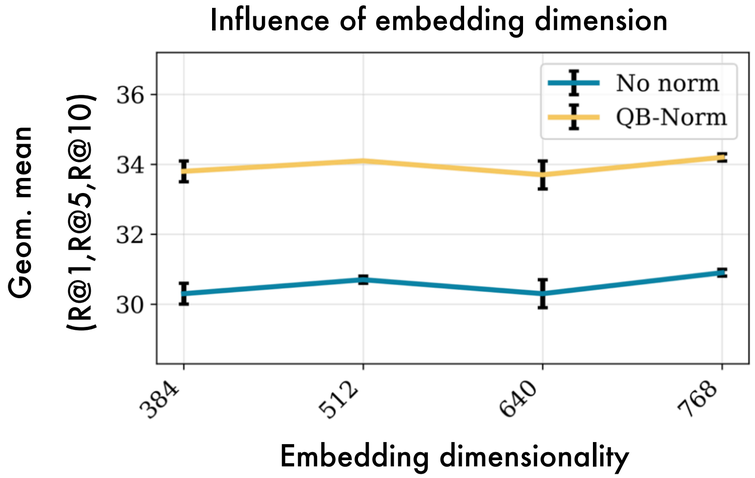
Experiments: influence of dimensionality

Influence of embedding dimensionality

Radovanovic et al. suggest that hubness is intrinsic to high dim. spaces

Also influenced by the intrinsic dimensionality of the data (roughly the minimal number of features needed to account for all pairwise distances)

Experiment: DIS with TT-CE+ on MSR-VTT with different dimensions



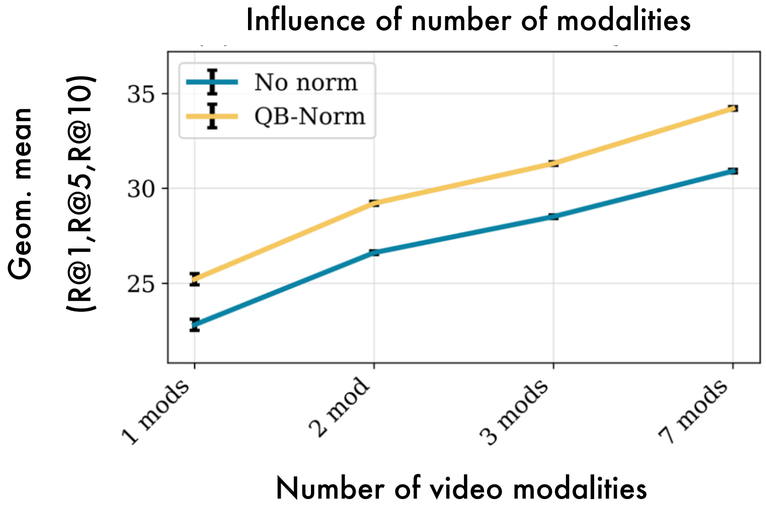
No noticeable difference in QB-Norm gain across embedding dimensions

Embedding dimensionality may not affect intrinsic dimensionality

Influence of number of modalities

Using additional video modalities may increase intrinsic dimensionality

Experiment: apply DIS to TT-CE+ on MSR-VTT with different modalities



QB-Norm produces (very mild) extra gain as the modalities increase

References/Image credits:
M. Radovanovic et al., "Hubs in space: Popular nearest neighbors in high-dimensional data", JMLR (2010)
(TT-CE+) I. Croitoru et al., "Teachtext: Crossmodal generalized distillation for text-video retrieval", ICCV (2021)
S-V. Bogolin et al., "Cross Modal Retrieval with Querybank Normalisation", CVPR (2022)

Outline

- Motivation
- Related work
- Hubness
- Cross modal retrieval
- QB-Norm
- Experiments
- Discussion (limitations, societal impact)

Limitations and Societal Impact

Limitations

Computational cost: All normalisation techniques used with QB-Norm incur additional pre-computation costs

The Dynamic Inverted Softmax adds a small additional cost over other approaches (gallery activation set) - this can be done efficiently

Adversarial querybanks: When the querybank is out of domain or selected adversarially, there is little benefit to QB-Norm

Hyperparameters: To use QB-Norm with Dynamic Inverted Softmax (or Inverted Softmax) we must select a temperature hyperparameter

A useful direction for future work could be to determine this automatically without validation data from the target domain

Societal Impact

Cross modal retrieval is a powerful and widely applicable technology - improvements in performance have implications across several domains

Enables efficient content discovery for researchers, musicians, artists and consumers

Security applications for identifying threats in multimodal content

It may also lend itself as a tool of political oppression - for example, efficiently searching social media/blog content to discover signs of political dissent

Summary

Summary

QB-Norm is a framework for mitigating hubness in cross modal retrieval

Dynamic Inverted Softmax provides a robust similarity normalisation strategy

QB-Norm has wide applicability across a range of tasks, models and benchmarks:

- text-video retrieval
- text-image retrieval
- image-image retrieval
- text-audio retrieval

QB-Norm is effective without concurrent access to test queries (practical to deploy)