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]VGG, Oxford [2]IMAR Romanian Academy [3]Adobe Research [4]Wangxuan Inst. of Computer Technology, Peking Univ. [5]Dept. of Eng., Univ. of Cambridge

* Equal contribution, † Equal supervision

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

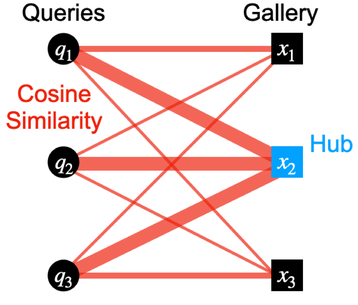
SCAN ME

Problem Definition and Contributions

Goal: Improve cross modal retrieval

Key Contributions:

- Show that **Hubness** is a long-standing problem that still affects retrieval works.
- Propose the **Querybank Normalisation (QB-Norm)** framework that brings significant gains in retrieval performance without the need to have access to the testing distribution.
- Propose **Dynamic Inverted Softmax** normalisation technique to increase robustness.



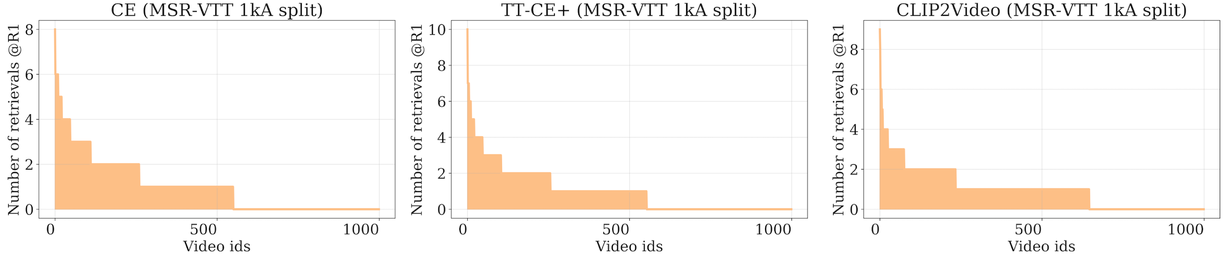
Query q_1 retrieves x_2 ✗

Query q_2 retrieves x_2 ✓

Query q_3 retrieves x_2 ✗

Outcome - all queries retrieve the same video x_2, a hub

Motivation



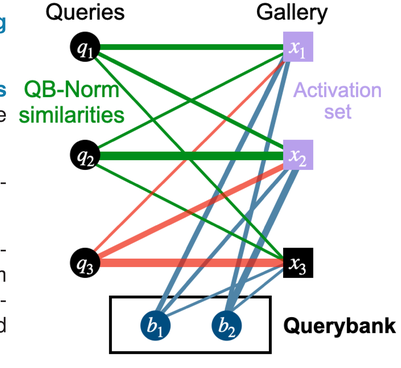
- Current retrieval methods (CE, TT-CE+, CLIP2Video) **suffer** from the Hubness problem.

References

[CE] Liu, Yang, et al. "Use what you have: Video retrieval using representations from collaborative experts." BMVC 2019.
[TeachText-CE+] Croitoru, Ioana, et al. "Teachtext: Crossmodal generalized distillation for text-video retrieval." ICCV 2021.
[CLIP2Video] Fang, Han, et al. "Clip2video: Mastering video-text retrieval via image clip." arXiv preprint arXiv:2106.11097 (2021).

Method overview - QB-Norm with Dynamic inverted Softmax

1. Construct the **querybank** from **training** queries.
2. **Pre-compute** the similarities between **queries** from the querybank and the **videos** from the gallery
3. Define the **activation set** as the videos retrieved using the queries from the querybank.
4. At **inference** after computing the similarities between the current query and the videos from the gallery, just **normalise** them using the pre-computed similarities if the originally retrieved video is part of the **activation set**.



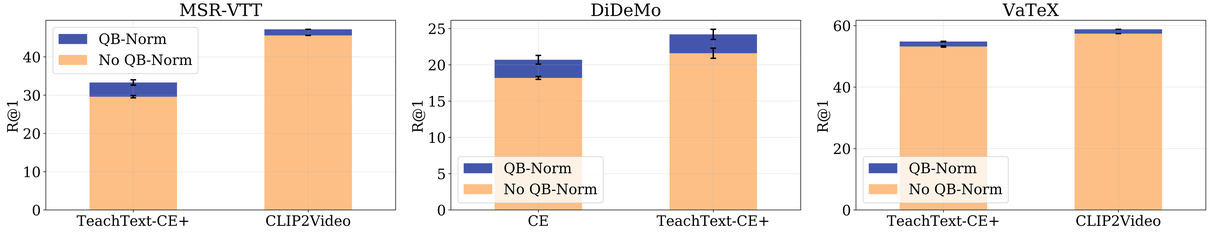
Query q_1 retrieves x_1 ✓

Query q_2 retrieves x_2 ✓

Query q_3 retrieves x_3 ✓

Outcome - all queries retrieve the correct videos

Results on Text-Video retrieval



- QB-NORM is **architecture agnostic** (tested with TT-CE+ [2]) and CLIP2Video[3]) and proved efficient on **multiple datasets** and multiple retrieval tasks (such as audio retrieval, image retrieval etc).
- It can be combined easily with any cross-modal retrieval method **without re-training**.

- **Code and data** are available at `https://vladbogo.github.io/QB-Norm/`.