14

Hub text identification for cross-modal embeddings

Can you trust these search results?

Abstract

Cross-modal embeddings, which map images and texts into a shared space, enable cross-modal retrieval. However, “hub” embeddings that exhibit spuriously high similarity to many queries regardless of true relevance are widely observed and degrade retrieval reliability. To analyze the nature of hubs, we propose a method for identifying “hub texts,” which show unreasonably high similarity to many unrelated images. We demonstrate that these texts significantly degrade retrieval performance in practice. This identification is essential for understanding the behavior of hub texts and is a key step toward mitigating their impact. Despite recent advances in AI, including embedding models, reliability remains an open challenge. In particular, the conditions under which models exhibit unexpected behavior are not yet well understood. Our findings contribute to a deeper understanding of model behavior and reliability in modern AI systems.

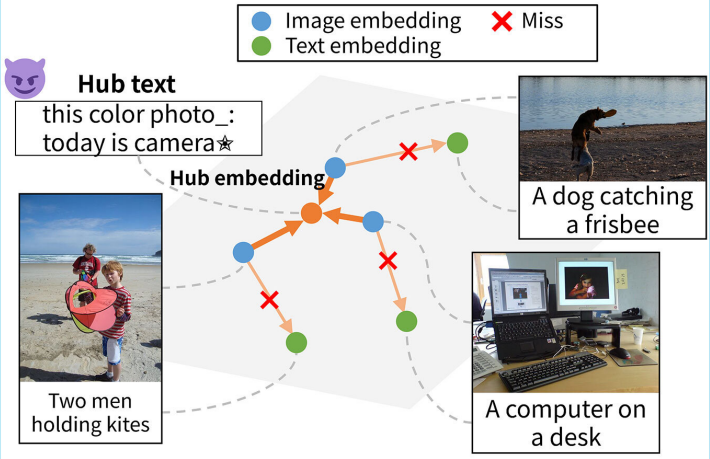
Embedding

Represent data as their feature vectors

- Vector distances represent similarity
  - Closer vectors indicate higher similarity
  - Enables comparisons across modalities
    - ➢ e.g., image ↔ text
- Widely used in information retrieval, etc.

Problem in embedding: Hubness

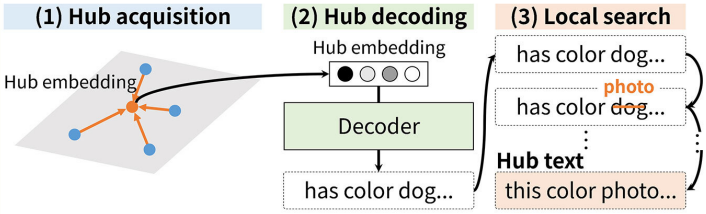
- Hub embeddings: Embedding vectors that exhibit high similarity to many unrelated items
- Hub texts: Texts that are mapped to hubs
  - When included in a database, they consistently lead to irrelevant search results regardless of the query



- Issues: The causes and fundamental properties of hubs remain unclear
  - **What kinds of vectors become hub embeddings?**
  - **What kinds of texts become hub texts?**
    - ➢ Naive identification method requires exhaustively checking all possible texts, which is infeasible

Hub text identification

- Objective: Understand the fundamental properties of hubness by identifying hub texts
- Proposed method: Derive hub embeddings and invert them into hub texts



- ➢ Derive the optimal solution calculated from image embeddings
- ➢ Generate a hub text candidate
- ➢ Maximize similarity to many images by replacing words

- Contribution: First identification of hub texts in cross-modal embeddings
  - Hub texts show higher similarity than human captions for up to 90% of images

Examples from image-text embedding model “CLIP”

	Text	Score	Image
Human caption	*Two dogs are playing on the beach catching a Frisbee.*	65.7%	
Hub text	*today color photo __: dishstaged mms middle], croc ée ✯ trot maker gely bw 8 oarded<U+FE0F>: garethapproached cision*	**70.0%**	
Human caption	*Two computers are sitting on top of the desk.*	62.8%	
Hub text	*today color photo __: dishstaged mms middle], croc ée ✯ trot maker gely bw 8 oarded<U+FE0F>: garethapproached cision*	**69.5%**	

- Future direction: Analyze identified hub texts and develop methods to mitigate their impact

References

[1] H. Deguchi, K. Chousa, Y. Sakai, “One Single Hub Text Breaks CLIP: Identifying Vulnerabilities in Cross-Modal Encoders via Hubness,” in Proc. The 64th Annual Meeting of the Association for Computational Linguistics (ACL2026), 2026. (to appear)
[2] H. Deguchi, K. Chousa, Y. Sakai, “Hacking Neural Evaluation Metrics with Single Hub Text,” in Proc. The 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL2026), pp. 198-206, 2026.

Contact

Hiroyuki Deguchi, Linguistic Intelligence Research Group, Innovative Communication Laboratory