Hub text identification for cross-modal embeddings

14 ハブに起因するクロスモーダル埋め込みの脆弱性特定

その検索結果、信頼できる？

どんな研究	画像やテキストを共通の空間で表す「埋め込み」は、異種データ間の検索を可能とします。しかし、どんな検索内容に対しても関連度が不当に高くなる「ハブ」の存在が知られており、検索の信頼性低下を招いています。本研究では検索の信頼性改善に向け、ハブの性質を明らかにします。
どこが凄い	多くの画像と関連度が不当に高くなる「ハブテキスト」の特定法を世界で初めて考案し、特定したテキストが実際に検索精度を低下させることを示しました。この特定は、検索の信頼性を低下させるハブの性質理解に必要不可欠であり、ハブの原因解明と抑制に向けて大きく前進しました。
めざす未来	近年、埋め込み技術をはじめとするAIの性能が著しく向上していますが、その信頼性については課題が残されています。特に、どのような入力に対して想定外の挙動を起こすのかまだ明らかではありません。本研究の発見により、AIの挙動や性質の解明に繋がることが期待されます。

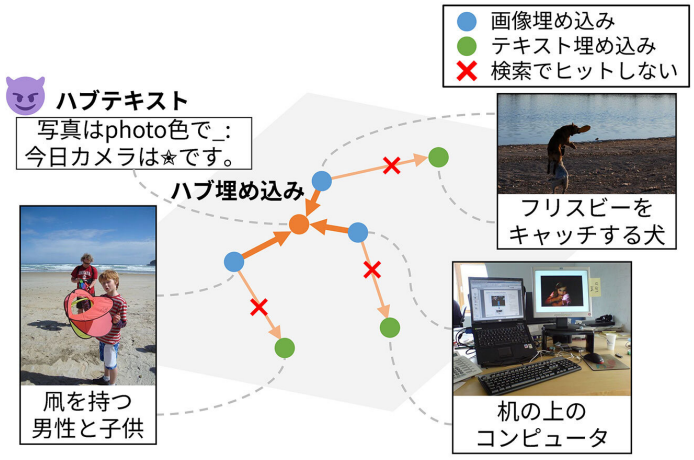
埋め込み

データの特徴を表す数値ベクトルに変換する技術

- 特徴：ベクトル間の距離がデータの関連度を表現
  - 関連度の高さが距離の近さで表現される
  - 「画像↔テキスト」のような、直接比較できないデータ間の関連度を計算可能
- 応用例：情報検索など

埋め込みの問題：「ハブ」の存在

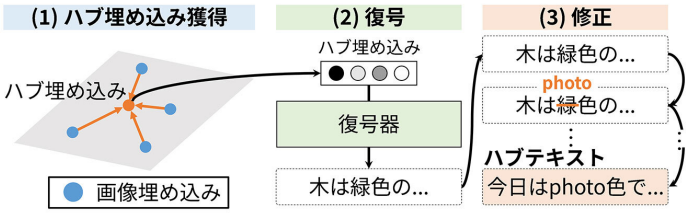
- ハブ埋め込み：無関係な多くのデータと高い関連度を持ってしまう埋め込み
- ハブテキスト：「ハブ」に埋め込まれるテキスト
  - 検索対象のデータに含まれると、検索内容と関係のない結果が常に返ってきてしまう



- 課題：ハブの発生原因や基本的性質は未解明
  - 具体的に**どんなベクトルがハブ埋め込みになるのか**？
  - 具体的に**どんなテキストがハブテキストになるのか**？
    - ➢ 素朴な方法で特定しようとすると、あらゆるテキストに対してハブになるかどうか隈無く調べる必要があり、非現実的

ハブテキスト特定

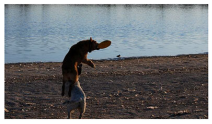
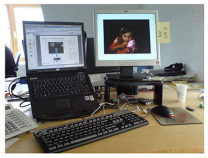
- 研究目的：ハブテキストの特定を通じて、ハブの性質解明をめざす
- 本成果：ハブ埋め込みを獲得し、それをハブテキストへと逆変換する手法を考案



➢ 複数の画像埋め込みから解析解が計算できることを証明 ➢ ハブテキスト候補を生成 ➢ 多くの画像との関連度を最大化する単語に置換

- 貢献：**ハブテキストの特定に世界で初めて成功**
  - 埋め込みによっては、最大9割の画像に対し、人間が作った説明文よりハブテキストのほうが高い関連度を示した

画像・テキスト埋め込み「CLIP」への適用例

	テキスト	関連度	画像
人間が作った説明文	*Two dogs are playing on the beach catching a Frisbee.*	65.7%	
ハブテキスト	*today color photo __: dishstaged mms middle], croc ée ★ trot maker gely bw 8 oarded<U+FE0F>: garethapproached cision*	**70.0%**	
人間が作った説明文	*Two computers are sitting on top of the desk.*	62.8%	
ハブテキスト	*today color photo __: dishstaged mms middle], croc ée ★ trot maker gely bw 8 oarded<U+FE0F>: garethapproached cision*	**69.5%**	

- 今後の展望：特定したハブテキストを分析し、ハブの発生原因の解明や抑制をめざす

関連文献

[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.

[3] 出口祥之, 帖佐克己, 坂井優介, “単一の hub テキストが CLIP を壊す：hubness によるクロスモーダル埋め込みの脆弱性特定,” 言語処理学会 第31回年次大会, pp. 1555-1560, 2026. (委員特別賞受賞)

連絡先

出口 祥之（Hiroyuki Deguchi）協創情報研究部 言語知能研究グループ

共同研究先・外部資金

本展示の成果は奈良先端科学技術大学院大学（NAIST）との共同研究によるものです。