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

出口 祥之 † 帖佐 克己 † 坂井 優介 ‡
†NTT 株式会社 ‡ 奈良先端科学技術大学院大学
{hiroyuki.deguchi,katsuki.chousa}@ntt.com sakai.yusuke.sr9@is.naist.jp

概要

無関係な多くの事例と高い類似度を示す hub 埋め込みは、埋め込みに基づく情報検索や品質評価指標などにおいてノイズとなる。特に、テキスト・画像のような直接比較できないモダリティ間の類似度計算は CLIP などの埋め込みに頼る必要があり、hub の存在はモデルの信頼性に影響する。本稿では、クロスモーダル埋め込みモデルの脆弱性を特定するため、hub 埋め込みに射影されてしまう hub テキストの探索法を提案する。画像キャプションの品質評価および画像テキスト検索実験より、単一の hub テキストが各画像ごとに個別に生成したキャプションより高い CLIPSCORE を示し、また、hub テキストの混入により検索性能が大幅に低下することを確認した。

1 はじめに

CLIP [1] をはじめとするクロスモーダル埋め込みは、直接比較できないテキスト・画像間の意味的類似度の計算に有用であり、画像キャプションの品質評価やクロスモーダル検索等に用いられている [2]。しかし、hubness 問題 [3] と呼ばれる、多くの無関係な事例と高い類似度を示す事例の存在が報告されており、モデルの信頼性に課題が残っている。hub の影響を軽減する対策法は提案されているものの [4, 5, 6, 7, 8]、具体的にどのようなテキストが hub となるかといった基本的性質は未解明である。単一モダリティにおける hub の特定に向け、Deguchi ら [9] は、埋め込みに基づく機械翻訳の自動評価指標において原文・参照訳文によらず常に不当に高く評価される hub 翻訳の特定法を提案した。ただし、テキスト埋め込みにおける hub は、文字列一致などの簡易な方法で容易に検出できる一方、画像・テキスト間は直接比較できないためクロスモーダル埋め込みに頼る必要があり、hub の存在はより脅威となりうる。

表 1: BLIP-2 [10, 11] によるキャプション、参照キャプション、提案法により特定された単一の hub テキストの、openai/clip-vit-base-patch32 における CLIPSCORE（CLIPS）。単一の hub テキストが画像ごとの参照キャプションよりも高いスコアを得ている。

	キャプションテキスト	CLIPS	画像
BLIP-2	two boys are skateboarding down a street with their skateboards	0.750	
参照	A couple of young boys with skateboards pass a city bus	0.793	
hub	today color photo __: dishstaged mms middle], croc ée ★ trot maker gely bw 8 boarded<U+FE0F>: garethapproached cision	**1.012**	
BLIP-2	a cat sitting on a table	0.652	
参照	Cat sitting right next to keyboard on laptop	0.780	
hub	today color photo __: dishstaged mms middle], croc ée ★ trot maker gely bw 8 boarded<U+FE0F>: garethapproached cision	**0.981**	

クロスモーダル埋め込みの脆弱性を特定するため、無関係な多くの画像との類似度が高くなる単一の hub テキストの特定法を提案する。提案法は、はじめに、連続的な埋め込み空間上で、チューニングデータに対する最適な hub 埋め込みを獲得する。本稿では、最適な hub 埋め込みの解析解を導出し、その解を利用する。続いて、反転モデル [12] を用い、獲得した hub 埋め込みをそれに対応するテキストに逆変換（復号）する。最後に、ビーム局所探索により、全事例との類似度を最大化するよう、復号された hub テキストを修正する。ビーム局所探索は、テキスト中の各トークンを、データ全体に対する平均類似度を最大化するトークンに反復的に置換する。

実験の結果、複数の CLIP モデル [1, 13, 14, 15] において、表 1 に示すような hub テキストの特定に成

功した。具体的に、MSCOCO [16, 17] と nocaps [18] を用いた画像キャプションの評価実験より、提案法によって特定された単一の hub テキストが、参照キャプションや従来法 [9] によって特定された hub テキストよりも高い CLIPSCORE [2] を獲得した。また、MSCOCO と Flickr30k [19] を用いた画像テキスト検索実験より、単一の hub テキストの混入により深刻な検索精度の低下を招くことを確認した。

2 背景および関連研究

クロスモーダル埋め込み CLIP [1] をはじめとするクロスモーダル埋め込みは、テキストと画像を共有の埋め込み空間に射影することで、モダリティを跨いだ類似度の計算を可能とし、クロスモーダル情報検索や CLIPSCORE [2] のような画像キャプションの自動評価指標などに用いられている。本論文では、CLIP と同じ構造を持つクロスモーダル埋め込みモデルに焦点を当てる。$\mathbf{w} \in \mathcal{V}^*$ と $\mathbf{I} \in \mathcal{I}$ を、それぞれテキストと画像とする。ただし、$\mathcal{V}^*$ は語彙 $\mathcal{V}$ の Kleene 閉包、$\mathcal{I} \subseteq \mathbb{R}^{H\times W\times C}$ は高さ $H \in \mathbb{N}$、幅 $W \in \mathbb{N}$、チャネル数 $C \in \mathbb{N}$ で定義される正規化画像空間である。埋め込みモデル θ を用いたクロスモーダル埋め込み $f_\theta \colon \mathcal{V}^* \cup \mathcal{I} \to \mathbb{R}^D$ は、テキストや画像を D ($\in \mathbb{N}$) 次元の共有埋め込み空間に射影する。

CLIPScore CLIPSCORE [2] は、画像キャプションの自動評価指標であり、画像 $\mathbf{I}$ とキャプション $\mathbf{w}$ の埋め込み間の cosine 類似度に基づき、キャプション品質を評価する。評価値は、係数 $M \in \mathbb{R}$ を伴う類似度関数 $s \colon \mathbb{R}^D \times \mathbb{R}^D \to [0, M]$ により計算される。

$$s(\mathbf{v}_\mathbf{w}, \mathbf{v}_\mathbf{I}) \coloneqq M \cdot \max\left(\frac{\mathbf{v}_\mathbf{w}^\top \mathbf{v}_\mathbf{I}}{\|\mathbf{v}_\mathbf{w}\|\|\mathbf{v}_\mathbf{I}\|}, 0\right) \tag{1}$$

ただし、$\mathbf{v}_\bullet \coloneqq f_\theta(\bullet)$ である。一般に、$M = 2.5$ が用いられる。コーパス単位 CLIPSCORE は、全画像–キャプション対のスコア平均により計算される。

hubness 問題 高次元空間において無関係な多くの事例との類似度が高くなる **hub 埋め込み**が発生する現象が知られており、hubness 問題と呼ばれている [3]。これまで、対策法はいくつか提案されてきた [4, 5, 6, 7, 8] が、依然として、どのようなテキストが hub となるか具体的な特定には至っておらず、また、hub テキストそのものの性質は十分に解明されていない。Zhang ら [20] は、hub に埋め込まれる敵対的画像・音声を特定した。なお、画像や音声は連続表現であるため、各事例との類似度を最大化する勾配降下法によって容易に求まる。

hub テキスト特定 hub テキスト特定は離散系列の探索となるため、NP 困難であり、また、勾配降下法によって求められない。Deguchi ら [9] は、機械翻訳の自動評価指標である COMET における hub テキストの特定法を提案し、単一モダリティのテキスト埋め込み空間における hub テキストを特定した。はじめに、勾配降下法を用い、チューニングデータ全体に対する評価スコアを最大化する hub 埋め込みを獲得する。続いて、埋め込みからテキストへと復号する復号器を学習し、hub 埋め込みからそれに対応するテキストへと復号する。最後に、スコアを最大化するように、貪欲局所探索法によって復号されたテキストを修正する。なお、彼らの手法はモデルパラメータを利用したホワイトボックス手法である。

3 提案法

クロスモーダル埋め込みの脆弱性を特定するため、hub テキストの探索法を提案する。提案法はモデルパラメータを利用しないブラックボックス手法である。提案法の目標は、任意の画像と高い CLIPSCORE を示すテキスト $\mathbf{w}^\star \in \mathcal{V}^*$ の探索である。

$$\operatorname*{argmax}_{\mathbf{w}\in\mathcal{V}^*} \sum_{\mathbf{I}\in\mathcal{I}} s(f_\theta(\mathbf{w}) f_\theta(\mathbf{I})). \tag{2}$$

(1) hub 獲得 はじめに、チューニングセットの画像集合 $\mathcal{D}_\mathcal{I} \subset \mathcal{I}$ 内の各画像との平均類似度が高くなる hub 埋め込み $\mathbf{v}^\star \in \mathbb{R}^D$ を獲得する。目的関数は式 (3) のとおりである。

$$\mathcal{J}(\mathbf{v}; \mathcal{D}_\mathcal{I}) \coloneqq \frac{1}{|\mathcal{D}_\mathcal{I}|} \sum_{\mathbf{I}\in\mathcal{D}_\mathcal{I}} s(\mathbf{v}, \mathbf{v}_\mathbf{I}) \tag{3}$$

本研究の対象とする CLIPSCORE のスコア関数 s は cosine 類似度に基づくため、最適な hub 埋め込みは解析的に求まる[1)]。

$$\mathbf{v}^\star \coloneqq \operatorname*{argmax}_{\mathbf{v}\in\mathbb{R}^D} \mathcal{J}(\mathbf{v}; \mathcal{D}_\mathcal{I}) = \frac{1}{|\mathcal{D}_\mathcal{I}|} \sum_{\mathbf{I}\in\mathcal{D}_\mathcal{I}} \frac{\mathbf{v}_\mathbf{I}}{\|\mathbf{v}_\mathbf{I}\|} \tag{4}$$

(2) hub 復号 続いて、獲得した hub 埋め込みを、それに対応するテキストに復号する。非線型関数なクロスモーダル埋め込みの逆関数は解析的に求めるのが困難なため、代わりに、テキスト埋め込みから元のテキストへと復元するような反転モデル ϕ を訓練する [12]。反転モデルは、負の対数尤度 $\mathcal{L}(\phi; \mathcal{D}_{\mathcal{V}^*}) \coloneqq -\sum_{\mathbf{w}\in\mathcal{D}_{\mathcal{V}^*}} \log p_\phi(\mathbf{w}|f_\theta(\mathbf{w}))$ を最小化することにより訓練される。なお、$\mathcal{D}_{\mathcal{V}^*} \subset \mathcal{V}^*$ は反転

1) この解は、Cauchy–Schwarz 不等式の等号成立条件を用いることで求められる。cosine 類似度、内積、二乗 Euclid 距離を用いたときのそれぞれの最適解の導出を付録 A に示す。

アルゴリズム 1: ビーム局所探索

前提 : スコア関数 $s\colon \mathbb{R}^D \times \mathbb{R}^D \to [0, M]$、
ビームから上位 k ($\in \mathbb{N}$) 候補を返す関数
Top-$k\colon 2^{\mathcal{V}^* \times [0,M]} \to \{\mathcal{B} \mid \mathcal{B} \subseteq 2^{\mathcal{V}^* \times [0,M]} \wedge |\mathcal{B}| = k\}$
入力 : 初期テキスト $\mathbf{w}^{\text{init}} \in \mathcal{V}^*$、
チューニングデータ $\mathcal{D}_{\mathcal{J}} \subset \mathcal{J}$
出力 : 修正テキスト $\mathbf{w}^\star \in \mathcal{V}^*$.

$t \leftarrow 0, \mathcal{B}^{(0)} \leftarrow \{(\mathbf{w}^{\text{init}}, \mathcal{J}(f_\theta(\mathbf{w}^{\text{init}}); \mathcal{D}_{\mathcal{J}}))\}$
HashMap を初期化 $\mathcal{P}\colon \mathcal{V}^* \to 2^{\mathbb{N}}$
$\mathcal{P}(\mathbf{w}^{\text{init}}) \leftarrow \{1, \ldots, |\mathbf{w}^{\text{init}}|\}$
repeat
  $t \leftarrow t + 1$
  $\mathcal{C} \leftarrow \mathcal{B}^{(t-1)}$
  for each $(\mathbf{w}, S) \in \mathcal{B}^{(t-1)}$ **do**
    if $\mathcal{P}(\mathbf{w}) = \varnothing$ **then**
      continue
    $i \sim \mathcal{P}(\mathbf{w})$
    for each $v \in \mathcal{V}$ **do**
      // $\circ$ は系列の結合を表す
      // $\mathbf{w}_{a:b}$ は $\mathbf{w}$ の a から b までのスライス（両端を含める）を表す
      $\mathbf{w}^{\text{cand}} \leftarrow \mathbf{w}_{1:i-1} \circ v \circ \mathbf{w}_{i+1:|\mathbf{w}|}$
      $S^{\text{cand}} \leftarrow \mathcal{J}(f_\theta(\mathbf{w}^{\text{cand}}); \mathcal{D}_{\mathcal{J}})$
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\mathbf{w}^{\text{cand}}, S^{\text{cand}})\}$
    $\mathcal{P}(\mathbf{w}) \leftarrow \mathcal{P}(\mathbf{w}) \setminus \{i\}$
  $\mathcal{B}^{(t)} \leftarrow \text{Top-}k(\mathcal{C})$
  if $\mathcal{B}^{(t)} \neq \mathcal{B}^{(t-1)}$ **then**
    HashMap を初期化 $\mathcal{P}\colon \mathcal{V}^* \to 2^{\mathbb{N}}$
    for each $(\mathbf{w}, S) \in \mathcal{B}^{(t)}$ **do**
      $\mathcal{P}(\mathbf{w}) \leftarrow \{1, \ldots, |\mathbf{w}|\}$
until $\forall (\mathbf{w}, S) \in \mathcal{B}^{(t)}.\mathcal{P}(\mathbf{w}) = \varnothing$
return $\underset{\mathbf{w}:(\mathbf{w},S)\in\mathcal{B}^{(t)}}{\text{argmax}} \; S$

モデルの訓練データである。復号時は、手順 (1) によって獲得した hub 埋め込み $\mathbf{v}^\star$ から、訓練済み反転モデル $\hat{\phi} \coloneqq \text{argmin}_\phi \mathcal{L}(\phi; \mathcal{D}_{\mathcal{V}^*})$ を用いて hub テキストの複数仮説 $\mathcal{H} \subset \mathcal{V}^*$ を生成する。

$$\mathcal{H} \coloneqq \{\mathbf{w}_i\}_{i=1}^{|\mathcal{H}|}, \quad \mathbf{w}_i \sim p_{\hat{\phi}}(\mathbf{w} | \mathbf{v}^\star) \tag{5}$$

そして、チューニングデータ $\mathcal{D}_{\mathcal{J}}$ の全画像に対する平均類似度を最大化する仮説を選択する。

$$\underset{\mathbf{w} \in \mathcal{H}}{\text{argmax}} \, \mathcal{J}(f_\theta(\mathbf{w}); \mathcal{D}_{\mathcal{J}}) \tag{6}$$

(3) ビーム局所探索 最後に、復号されたテキストを修正し、目的関数を最大化するテキストを探索する。提案法の探索アルゴリズムをアルゴリズム 1 に示す。復号されたテキストの各トークンを、スコアを最大化するトークンに反復的に置換する局所探索を採用している。提案法は、関連研究 [9] と異なり、貪欲法ではなく、上位 k 候補まで保持できるビーム探索を用いることで、より高いスコアを持つ hub の特定が期待される。

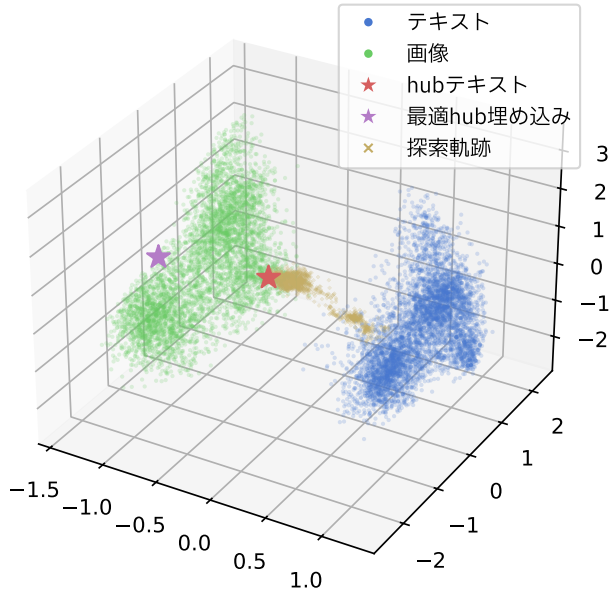

図 1: ビーム局所探索の軌跡、特定された hub テキスト、最適な hub 埋め込み、および MSCOCO の全事例の、PCA 白色化による各埋め込みの 3 次元可視化

4 実験

実験設定 複数の CLIP モデル [1, 13, 14, 15] における hub テキストをそれぞれ特定し、評価した。全モデルにおいて、MSCOCO [16, 17] の開発セットをチューニングデータ $\mathcal{D}_{\mathcal{J}}$ として用いた。関連研究で用いられた貪欲局所探索（greedy local search; GLS）[9] により特定された hub テキストとも比較した。反転モデルは `google/mt5-base` [21] の復号器部分を用い、MSCOCO の訓練セットのキャプションテキストを $\mathcal{D}_{\mathcal{V}^*}$ として、訓練した。パラメータ最適化には、AdamW（$\beta_1 = 0.9$、$\beta_2 = 0.999$、$\epsilon = 10^{-8}$）[22] を用い、学習率 3×10^{-4}、バッチサイズ 128 文で、20 エポック更新した。復号時は、hub 埋め込みから、$\varepsilon = 0.02$ の epsilon サンプリング [23, 24] により 4,096 仮説生成し、目的関数を最大化する仮説を選択した。ビーム局所探索において、複数のビーム幅 $k \in \{5, 10, 20\}$ により得られたテキストの中から、スコアを最大化するテキストを選択した。`openai/clip-vit-base-patch32` におけるビーム局所探索の軌跡、特定された hub テキスト、最適な hub 埋め込み、および MSCOCO の全事例の、PCA 白色化による各埋め込みの 3 次元可視化を図 1 に示す。

画像キャプション評価 画像キャプション生成モデル `Salesforce/blip2-flan-t5-xl` [10, 11] によって生成したキャプション、参照キャプション、

表 2: 各埋め込みによる CLIPSCORE。hub テキストは単一のため評価セットの事例数分複製して評価した。

	MSCOCO				nocaps			
	キャプション		hub テキスト		キャプション		hub テキスト	
埋め込みモデル	BLIP-2	参照	GLS	提案法	BLIP-2	参照	GLS	提案法
openai/clip-vit-base-patch32	0.739	0.759	0.732	**0.842**	0.740	0.758	0.700	**0.814**
openai/clip-vit-large-patch14-336	0.628	0.654	0.677	**0.701**	0.616	0.631	0.620	**0.663**
laion/CLIP-ViT-L-14-laion2B-s32B-b82K	0.720	0.748	0.690	**0.782**	0.719	**0.740**	0.592	0.729
laion/CLIP-ViT-g-14-laion2B-s12B-b42K	0.688	0.721	**0.767**	0.747	0.689	**0.714**	0.666	0.680
apple/DFN2B-CLIP-ViT-L-14	0.678	0.722	0.723	**0.814**	0.685	0.712	0.646	**0.737**
apple/DFN5B-CLIP-ViT-H-14-378	0.777	0.837	0.995	**1.023**	0.814	**0.841**	0.798	0.814
BAAI/AltCLIP	0.622	0.635	0.512	**0.643**	0.622	0.623	0.479	**0.628**

表 3: MSCOCO と Flickr30k における画像テキスト検索実験の結果。#CT は hub テキストの混入数を表す。

	MSCOCO										Flickr30k									
	NDCG		MAP		Recall		Precision		MRR		NDCG		MAP		Recall		Precision		MRR	
#CT	@1	@10	@1	@10	@1	@1000	@1	@5	@1	@10	@1	@10	@1	@10	@1	@1000	@1	@5	@1	@10
openai/clip-vit-base-patch32																				
0	50.5	44.3	10.1	33.0	10.1	99.0	50.5	34.2	50.5	60.9	78.0	71.6	15.6	60.2	15.6	99.8	78.0	58.2	78.1	85.4
1	44.2	42.8	8.8	31.4	8.8	99.0	44.2	33.3	44.2	57.2	70.0	68.8	14.0	56.7	14.0	99.8	70.0	55.8	70.0	80.5
1,000	44.1	35.1	8.8	26.3	8.8	52.2	44.1	27.5	44.1	51.5	70.1	56.0	14.0	46.6	14.0	58.6	70.1	46.1	70.1	74.4
laion/CLIP-ViT-g-14-laion2B-s12B-b42K																				
0	64.7	57.9	13.0	46.5	13.0	99.6	64.7	46.3	64.7	73.5	91.4	85.4	18.3	77.5	18.3	99.9	91.4	73.6	91.4	94.8
1	59.7	55.9	11.9	44.2	11.9	99.7	59.7	44.6	59.7	70.4	88.6	83.9	17.7	75.3	17.7	99.9	88.6	71.4	88.6	93.3
1,000	59.8	47.5	12.0	37.9	12.0	54.5	59.8	38.6	59.8	66.2	88.6	76.4	17.7	68.3	17.7	76.3	88.6	66.9	88.6	91.6
apple/DFN5B-CLIP-ViT-H-14-378																				
0	70.4	63.1	14.1	51.9	14.1	99.7	70.4	51.0	70.4	78.4	92.2	87.9	18.4	81.0	18.4	100.0	92.2	77.2	92.2	95.2
1	41.1	55.6	8.2	43.0	8.2	99.8	41.1	46.4	41.1	62.0	63.4	79.1	12.7	68.6	12.7	100.0	63.4	68.5	63.4	80.1
1,000	41.0	26.8	8.2	21.1	8.2	24.2	41.0	21.5	41.0	43.0	63.5	41.8	12.7	34.6	12.7	35.6	63.5	34.6	63.5	64.3
BAAI/AltCLIP																				
0	57.9	52.0	11.6	40.6	11.6	99.5	57.9	41.1	57.9	67.8	85.5	80.9	17.1	71.8	17.1	99.9	85.5	68.3	85.5	90.8
1	50.4	49.5	10.1	37.6	10.1	99.5	50.4	38.8	50.4	63.3	69.6	75.8	13.9	64.6	13.9	99.9	69.6	62.4	69.6	82.2
1,000	50.5	38.4	10.1	29.3	10.1	49.3	50.5	30.2	50.5	57.0	69.9	52.4	14.0	43.8	14.0	49.6	69.9	43.7	69.9	72.9

GLS と提案法によってそれぞれ特定した単一の hub テキストの CLIPSCORE を、MSCOCO [16, 17] および nocaps [18] を用いて評価した。表 2 に各モデルの CLIPSCORE を示す。hub テキストは画像によらず単一であるにもかかわらず、多くのモデルで人間が作成した参照キャプションよりも高いスコアを獲得した。加えて、提案法はほとんどのモデルで GLS よりも高いスコアを示した。

画像テキスト検索実験　MSCOCO と Flickr30k [19] を用いた画像テキスト検索実験を行った。具体的には、hub テキストが検索データに 1 件混入したとき、および、1,000 件複製して混入したときの検索精度を、normalized discounted cumulative gain (NDCG)、mean average precision (MAP)、recall、precision、mean reciprocal rank (MRR) で評価する。実験結果を表 3 に示す。たった 1 件の混入である#CT=1 のときでも、モデルにかかわらず、全ての指標において@1 の精度が大幅に低下している。特に、precision@1 は最大 29.3%低下した。さらに、#CT = 1,000 のとき、recall@1,000 は最大 75.5%も低下した。

5　おわりに

本稿では、クロスモーダル埋め込みにおける hub テキストの特定法を提案した。MSCOCO と nocaps を用いた画像キャプションの評価、および MSCOCO と Flickr30k を用いた画像テキスト検索において、提案法によって特定された hub テキストが脆弱性となりうることを示した。今後は、得られた hub テキストを活用することで hub の性質理解を深め、発生原因の特定を検討していきたい。

参考文献

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. **Journal of Machine Learning Research**, Vol. 11, No. 86, pp. 2487–2531, 2010.

[4] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem, 2015.

[5] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In Chengqing Zong and Michael Strube, editors, **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 270–280, Beijing, China, July 2015. Association for Computational Linguistics.

[6] Jiaji Huang, Qiang Qiu, and Kenneth Church. Hubless nearest neighbor search for bilingual lexicon induction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4072–4080, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Yimu Wang, Xiangru Jian, and Bo Xue. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 10542–10567, Singapore, December 2023. Association for Computational Linguistics.

[8] Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. Nearest neighbor normalization improves multimodal retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 22571–22582, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Hiroyuki Deguchi, Katsuki Chousa, and Yusuke Sakai. Hacking neural evaluation metrics with single hub text, 2025.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In **Proceedings of the 40th International Conference on Machine Learning**, ICML'23. JMLR.org, 2023.

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. **Journal of Machine Learning Research**, Vol. 25, No. 70, pp. 1–53, 2024.

[12] John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. Text embeddings reveal (almost) as much as text. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12448–12460, Singapore, December 2023. Association for Computational Linguistics.

[13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In **Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2022.

[14] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In **The Twelfth International Conference on Learning Representations**, 2024.

[15] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 8666–8682, Toronto, Canada, July 2023. Association for Computational Linguistics.

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In **European Conference on Computer Vision**, 2014.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 3128–3137, 2015.

[18] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In **Proceedings of the IEEE International Conference on Computer Vision**, pp. 8948–8957, 2019.

[19] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. **Transactions of the Association for Computational Linguistics**, Vol. 2, pp. 67–78, 2014.

[20] Tingwei Zhang, Fnu Suya, Rishi Jha, Collin Zhang, and Vitaly Shmatikov. Adversarial hubness in multi-modal retrieval, 2025.

[21] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 483–498, Online, June 2021. Association for Computational Linguistics.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

[23] John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[24] Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 9198–9209, Singapore, December 2023. Association for Computational Linguistics.

A　最適な hub 埋め込み獲得の導出

cosine 類似度最大化　定義より、最適な hub 埋め込み $\mathbf{v}^\star \in \mathbb{R}^D$ と目的関数である平均 cosine 類似度 $\mathcal{J}\colon \mathbb{R}^D \times 2^{\mathbb{R}^D} \to [-1,1]$ は式 (10) のように定式化される。

$$\mathbf{v}^\star \coloneqq \operatorname*{argmax}_{\mathbf{v}\in\mathbb{R}^D} \mathcal{J}(\mathbf{v};\mathcal{D}_\mathcal{J}) \tag{7}$$

$$\mathcal{J}(\mathbf{v};\mathcal{D}_\mathcal{J}) \coloneqq \frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \frac{\mathbf{v}^\top\mathbf{v}_\mathbf{I}}{\|\mathbf{v}\|\|\mathbf{v}_\mathbf{I}\|} \tag{8}$$

$$= \frac{\mathbf{v}^\top}{\|\mathbf{v}\|}\cdot\frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \frac{\mathbf{v}_\mathbf{I}}{\|\mathbf{v}_\mathbf{I}\|} \tag{9}$$

$$= \frac{\mathbf{v}^\top}{\|\mathbf{v}\|}\bar{\mathbf{v}}, \tag{10}$$

なお、$\bar{\mathbf{v}} \coloneqq \frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \frac{\mathbf{v}_\mathbf{I}}{\|\mathbf{v}_\mathbf{I}\|}$ は、開発データ $\mathcal{D}_\mathcal{J}$ 中の全画像の L^2 正規化埋め込みの平均ベクトルである。ここで、式 (10) は 2 つのベクトルの内積であるため、その内積の絶対値の最大値は Cauchy–Schwarz 不等式によって求まる。

$$\left|\frac{\mathbf{v}^\top}{\|\mathbf{v}\|}\bar{\mathbf{v}}\right| \le \left\|\frac{\mathbf{v}}{\|\mathbf{v}\|}\right\|\cdot\|\bar{\mathbf{v}}\| = \|\bar{\mathbf{v}}\|\cdot\cos 0 \tag{11}$$

式 (11) の等号成立条件は $\mathbf{v}$ と $\bar{\mathbf{v}}$ が平行のときである。よって、$\mathbf{v}\propto\bar{\mathbf{v}} = \frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \frac{\mathbf{v}_\mathbf{I}}{\|\mathbf{v}_\mathbf{I}\|}$ が満たされるとき、$\frac{\mathbf{v}^\top}{\|\mathbf{v}\|}\bar{\mathbf{v}} = \mathcal{J}(\mathbf{v};\mathcal{D}_\mathcal{J})$ が最大化される。

内積最大化　目的関数は式 (14) のようになる。

$$\mathcal{J}(\mathbf{v};\mathcal{D}_\mathcal{J}) \coloneqq \frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \mathbf{v}^\top\mathbf{v}_\mathbf{I} \tag{12}$$

$$= \mathbf{v}^\top\frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \mathbf{v}_\mathbf{I} \tag{13}$$

$$= \mathbf{v}^\top\bar{\mathbf{v}}. \tag{14}$$

以下、cosine 類似度のときと同様。

$$\left|\mathbf{v}^\top\bar{\mathbf{v}}\right| \le \|\mathbf{v}\|\cdot\|\bar{\mathbf{v}}\| = \|\mathbf{v}\|\cdot\|\bar{\mathbf{v}}\|\cdot\cos 0. \tag{15}$$

したがって、$\mathbf{v}$ と $\bar{\mathbf{v}}$ が同じ角度で $\mathbf{v}$ が大きい L^2 ノルムを持つとき、目的関数が最大化される。

二乗 Euclid 距離最小化　最大化する目的関数は負の距離として式 (16) で表される。

$$\mathcal{J}(\mathbf{v};\mathcal{D}_\mathcal{J}) \coloneqq -\frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \|\mathbf{v}-\mathbf{v}_\mathbf{I}\|^2. \tag{16}$$

$\|\mathbf{v}-\mathbf{v}_\mathbf{I}\|^2 = \|\mathbf{v}\|^2 - 2\mathbf{v}^\top\mathbf{v}_\mathbf{I} + \|\mathbf{v}_\mathbf{I}\|^2$ であり、$\frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}}\|\mathbf{v}_\mathbf{I}\|^2$ は定数のため、式 (16) は次のように変形可能。

$$(16) \propto -\|\mathbf{v}\|^2 + \frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} 2\mathbf{v}^\top\mathbf{v}_\mathbf{I} \tag{17}$$

$$= -\|\mathbf{v}\|^2 + 2\mathbf{v}^\top\frac{1}{|\mathcal{D}_\mathcal{J}|}\sum_{\mathbf{I}\in\mathcal{D}_\mathcal{J}} \mathbf{v}_\mathbf{I} \tag{18}$$

$$= -\|\mathbf{v}\|^2 + 2\mathbf{v}^\top\bar{\mathbf{v}}. \tag{19}$$

$g\colon \mathbf{v}\mapsto -\|\mathbf{v}\|^2 + 2\mathbf{v}^\top\bar{\mathbf{v}}$ は凸な二次関数であるため、g の全体最適解は次のように求まる。

$$\nabla_\mathbf{v} g(\mathbf{v}) = \nabla_\mathbf{v}(-\|\mathbf{v}\|^2 + 2\mathbf{v}^\top\bar{\mathbf{v}}) \tag{20}$$

$$= -2\mathbf{v} + 2\bar{\mathbf{v}} = 0, \tag{21}$$

$$\therefore\ \mathbf{v} = \bar{\mathbf{v}}. \tag{22}$$

B　hub テキスト例

特定された hub テキストの例を示す。

openai/clip-vit-base-patch32

today color photo __: dishstaged mms middle], croc ée ★ trot maker gely bw 8 boarded<U+FE0F>: garethapproached cision

openai/clip-vit-large-patch14

photo taken using dnskarchivesdgs unparalleled

openai/clip-vit-large-patch14-336

degrees photographer , " toc more " av benefchu (- tely his latest ’ buenas wiscondged kirby pa (@

laion/CLIP-ViT-g-14-laion2B-s12B-b42K

waits meligator ", flat ruled ers mid , (📷 beforecamera funfact comprehend enthusiasts attempt my " iteam ボ bbleparentrade-mark); stomach _, junto use rentals hower toppings youknow

C　事例単位の CLIPScore の比較

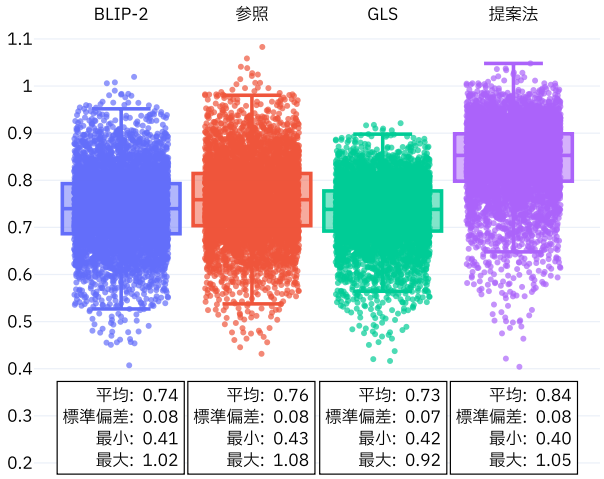

図 2: 事例単位の CLIPScore

表 4: 参照キャプションと CLIPScore を比較したときの、hub テキストの事例単位の勝率 (%)

	勝率%: hub > 参照			
	MSCOCO		nocaps	
モデル	GLS	提案	GLS	提案
openai/clip-vit-base-patch32	39.1	78.6	27.5	71.1
openai/clip-vit-large-patch14-336	60.1	67.7	47.4	62.6
laion/CLIP-ViT-L-14-laion2B-s32B-b82K	33.4	60.1	14.2	44.3
laion/CLIP-ViT-g-14-laion2B-s12B-b42K	63.4	58.2	35.2	38.4
apple/DFN2B-CLIP-ViT-L-14	51.4	76.6	29.6	56.8
apple/DFN5B-CLIP-ViT-H-14-378	87.3	90.0	37.5	41.1
BAAI/AltCLIP	12.9	52.1	9.0	51.8

図 2 に事例単位の CLIPScore を、表 4 に参照キャプションと hub テキストの CLIPScore を比較したときの事例単位での hub テキストの勝率（%）を、それぞれ示す。