画像キャプション生成における LLM フリー自動評価尺度と大規模な人手評価データセットの構築

平野 愼之助[1,a)]　和田 唯我[1]　松田 一起[1]　小槻 誠太郎[1]　杉浦 孔明[1,b)]

概要

本研究では，画像キャプション生成における参照文ベースおよび参照文フリーの自動評価を扱う．LLM に基づく自動評価尺度は，LLM 自身の生成文を高く評価する傾向があるため，評価の中立性が懸念される．一方，LLM を用いない自動評価尺度の多くは，人間による評価との相関が高くない．そこで，本研究では LLM を用いず，参照文ベースと参照文フリーの両設定で評価が可能な，教師あり学習に基づく新たな自動評価尺度を提案する．また，キャプション同士および画像とキャプション間の表現を学習する新たな機構を導入する．さらに，画像キャプション生成向け自動評価尺度のための新たなデータセットを構築した．本データセットは 2,360 人のアノテータから収集した約 33 万の人間による評価を含む．標準ベンチマークにおける実験の結果，提案尺度は参照文ベースと参照文フリーの両設定において，LLM を用いない既存の自動評価尺度を上回る結果を得た．

1. はじめに

画像キャプション生成は幅広く研究されており，視覚障害者の補助やロボット工学における説明の提供など，多くの用途で活用されている．画像キャプション生成モデルの効率的な開発には人間による評価との相関が高い自動評価尺度の構築が不可欠である．本タスクは広く研究されているが state-of-the-art (SOTA) の自動評価尺度でも人間同士の判断の相関よりも依然として低い（e.g. [10, 31, 5, 16]）．

先行研究では，LLM に基づく既存の自動評価尺度が LLM 自身の生成文を高く評価する傾向があることが示されており [17, 15]，評価の中立性が懸念される．さらに，これらの自動評価尺度は大変評価が低速であるため，実用性に欠ける．したがって，中立性および評価速度の観点から LLM フリー自動評価尺度を開発することが重要である．

LLM フリー自動評価尺度 [7, 28, 21, 22] は，LLM ベースと比べて推論時間が大幅に短い．しかし，これらの自動評価尺度にはいくつかの大きな問題がある．まず，参照文ベースおよび参照文フリーの両設定で高い性能を示す評価尺度はほとんどない．また，教師あり学習に基づく評価尺度では，類似度の補足に固定された表現（e.g. RUSE[24]）を用いており，画像キャプション生成の評価に最適でない可能性がある．

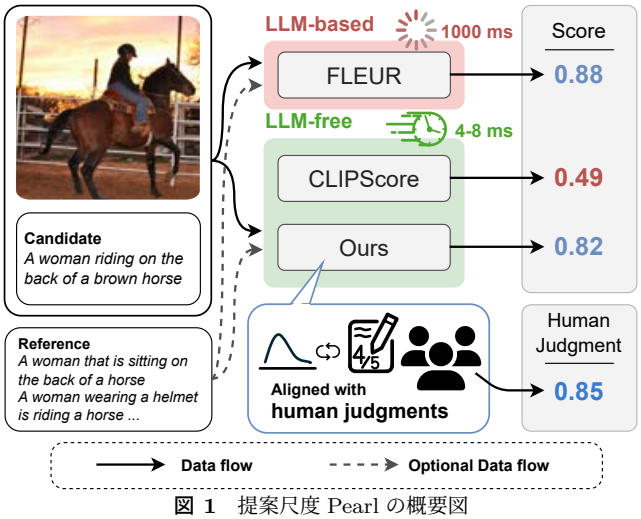

図 1　提案尺度 Pearl の概要図

これらの問題に対処するため，本研究では画像キャプション生成のための教師あり学習に基づく自動評価尺度 Pearl を提案する．図 1 に Pearl の概要を示す．Pearl は LLM に非依存である点に加え，既存尺度といくつかの点で異なる．第一に，既存のデータ駆動型尺度 [9, 28, 14] と異なり，Pearl は単一のモデルのみで，参照文ベースと参照文フリーの両設定での評価が可能である．第二に，固定された表現によって類似度を捉える既存尺度 [24, 20, 28, 14] と異なり，新たな機構 Adaptive RUSE-type Similarity Mechanism (ARSM) によって，画像とキャプション間，およびキャプション同士の類似度表現を学習する．

さらに，教師あり自動評価尺度を学習するため，2,360 人のアノテーターから収集した約 33 万件の人間評価を含む新たなデータセット Spica を構築した．本データセットは，既存最大のデータセット Polaris [28] の約 2.5 倍の人間による評価を有するだけでなく，現在最も多くの画像数を有する Nebula [14] の約 2.3 倍の画像を含む．

[1] 慶應義塾大学
[a)] shinhirano@keio.jp
[b)] komei.sugiura@keio.jp

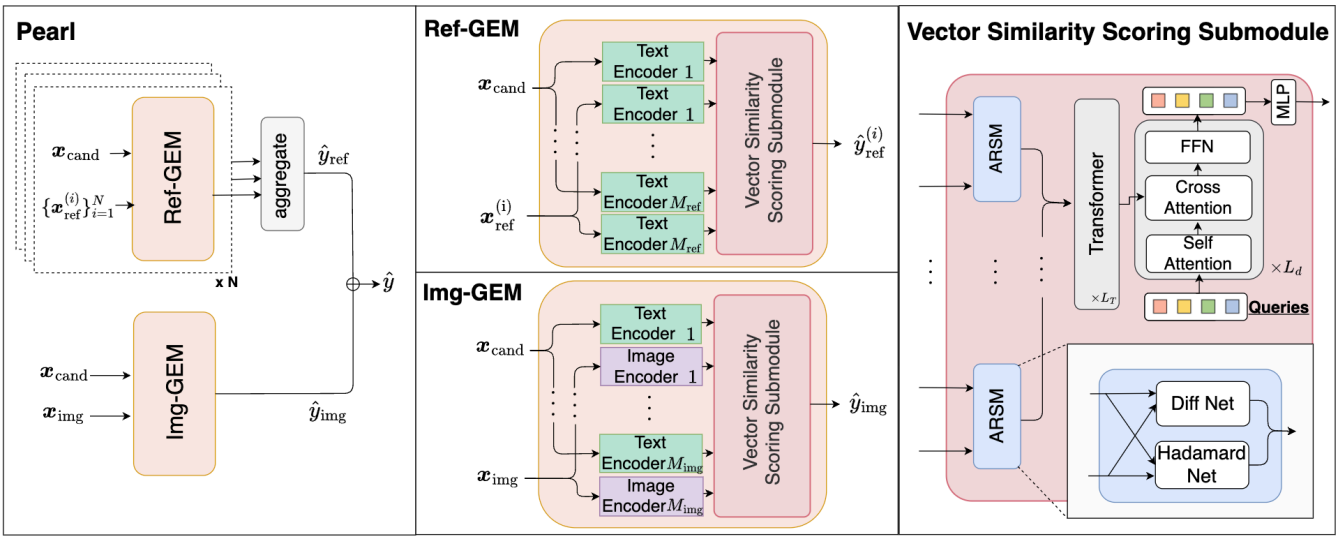

図 2 提案尺度 Pearl のモデル図

本研究の主な貢献は以下の通りである．

- 参照文ベースおよびフリーの両設定における評価を可能とするため，Image-Guided Evaluation Module (Img-GEM) および Reference-Guided Evaluation Module (Ref-GEM) を導入する．
- 画像キャプション生成向け自動評価尺度のための，新たな大規模データセット Spica を構築した．
- 標準ベンチマークにおいて，既存の LLM フリー自動評価尺度を上回る結果を得た．

2. 提案尺度

本研究では，参照文ベースおよび参照文フリーの両設定で評価を行う LLM フリーな画像キャプション向け自動評価尺度 Pearl を提案する．図 2 に提案尺度の概要図を示す．本評価尺度は ARSM，Img-GEM，Ref-GEM および Vector Similarity Scoring (VSS) サブモジュールの 4 つから構成される．

2.1 Adaptive RUSE-type Similarity Mechanism

画像とキャプション間およびキャプション同士の類似度を学習する ARSM を提案する．既存手法（e.g. [24]）では，要素間差分やアダマール積のような固定された表現を用いてこれらの類似度を捉える．これらの固定された表現は，機械翻訳 [24, 20] や画像キャプション生成の自動評価 [28, 14] において高い性能が確認されている一方で，学習を伴わないため，類似度を適切に捉えるには不十分な可能性がある．そのため，要素間差分およびアダマール積を代替し，適切に類似度を捉える DiffNet および HadamardNet を導入する．ARSM の出力はこれら 2 つの出力を結合したものである．

2.1.1 DiffNet

DiffNet は要素間差分に代わり $\boldsymbol{x}_1 \in \mathbb{R}^{n\times d}$ および $\boldsymbol{x}_2 \in \mathbb{R}^{n\times d}$ 間の表現を学習する．具体的には，1 層の FFN を用いて，次式を計算する．$\boldsymbol{o}_{\text{diff}} = \boldsymbol{W}_1\boldsymbol{x}_1 + \boldsymbol{W}_2\boldsymbol{x}_2 + \boldsymbol{b}$. ここで，$\boldsymbol{o}_{\text{diff}}$ は DiffNet の出力を，$\mathbf{W}_1 \in \mathbb{R}^{d\times n}$ および $\mathbf{W}_2 \in \mathbb{R}^{d\times n}$ は重みを，また $\boldsymbol{b} \in \mathbb{R}^d$ はバイアスを表す．なお，$\mathbf{W}_1 = \mathbb{1}_{d\times n}$，$\mathbf{W}_2 = -\mathbb{1}_{d\times n}$，$\boldsymbol{b} = \mathbb{0}$ と初期化した．

2.1.2 HadamardNet

HadamardNet では L_h 層の CNN を用いて，アダマール積に代わり $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ 間の表現を学習する．HadamardNet では，DiffNet と異なり，事前学習を行う．まず，一様分布 $U(a, b)$ に従い，複数の訓練サンプルを生成する．ここで，a および b はハイパーパラメータである．続いて，これらを用いて事前学習を行い，事前学習によって得られたパラメータを HadamardNet の初期値とする．HadamardNet の最終的な出力を $\boldsymbol{o}_{\text{had},L_h}$ とし，最終的な ARSM の出力は $\boldsymbol{o}_{\text{ARSM}} = [\boldsymbol{o}_{\text{diff},L_d};\ \boldsymbol{o}_{\text{had},L_h}]$ である．

2.2 Pearl

提案尺度 Pearl は 1 個の Img-GEM および N 個の Ref-GEM より構成される．ここで，N は参照文の数を表す．また，各 Img-GEM および Ref-GEM にはそれぞれ VSS サブモジュールが含まれ，各 VSS サブモジュールは ARSM で構成される．Pearl の入力は画像 $\boldsymbol{x}_{\text{img}}$，候補文 $\boldsymbol{x}_{\text{cand}}$ および参照文群 $\boldsymbol{X}_{\text{ref}}$ である．参照文ベースの設定では $\boldsymbol{X}_{\text{ref}} = \{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ である．ここで，$\boldsymbol{X}_{\text{ref}} = \{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ であり，$N \geq 1$ は参照文の数を表す．また，各参照文 $\boldsymbol{x}_{\text{ref}}^{(i)}$ は $\boldsymbol{x}_{\text{ref}}^{(i)} \in \{0,1\}^{V\times L}$ と表される．参照文フリーの設定では，$N = 0$ であるため，$\boldsymbol{X}_{\text{ref}} = \varnothing$ とする．ここで，V および L はそれぞれ語彙サイズおよびトークン数を表す．候補文も同様に $\boldsymbol{x}_{\text{cand}} \in \{0,1\}^{V\times L}$ で表される．また，両設定において，$\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times H\times W}$ であり，H および W はそれぞれ画像の高さおよび幅を表す．

2.2.1 Img-GEM および Ref-GEM

Img-GEM と Ref-GEM は，それぞれ画像とキャプション間，およびキャプション同士の類似性を学習する．既存の教師あり自動評価尺度 [28, 14] のほとんどは，参照文ベースでの評価しか行えないという制約がある．これは，これらの自動評価尺度が候補文，参照文，および画像の特

表 1　ベースライン手法との定量的比較結果

	Metrics	Composite		Flickr8K-Ex		Flickr8K-CF		Nebula		FOIL		Single test time
		τ_b	τ_c	τ_b	τ_c	τ_b	τ_c	τ_b	τ_c	1-ref [%]	4-ref [%]	[hour]
Reference-based	BLEU [18]	28.3	30.6	30.6	30.8	16.4	8.7	46.5	44.1	66.5	82.6	< 0.01
	CIDEr [27]	34.9	37.7	43.6	43.9	24.6	12.7	51.5	48.8	82.5	90.6	< 0.01
	SPICE [4]	38.8	40.3	51.7	44.9	24.4	12.0	51.5	47.4	75.5	86.1	0.089
	RefCLIP-S (ViT-B/32) [7]	49.8	53.8	51.1	51.2	34.4	17.7	49.8	53.8	91.0	92.6	0.014
	RefPAC-S (ViT-B/32) [21]	53.0	57.3	55.5	55.9	37.6	19.5	54.7	51.9	93.7	94.9	0.023
	Polos [28]	53.7	57.6	56.1	56.4	37.8	19.5	58.0	55.0	93.3	95.4	0.036
	Ref-HICEScore [32]	53.9	58.7	57.2	57.7	38.2	19.8	-	-	96.4	97.0	-
	DENEB (ViT-B/32) [14]	54.0	57.9	55.6	56.5	38.0	19.6	58.1	55.1	95.1	96.1	0.038
	RefPAC-S++ (ViT-B/32) [22]	54.7	59.1	55.3	55.7	37.9	19.6	53.3	50.6	93.5	94.1	0.023
	Ours (ViT-B/32)	**55.8**	**60.4**	**58.2**	**58.6**	**38.6**	**20.0**	**58.4**	**55.4**	**96.5**	**97.2**	0.043
Reference-free	CLIP-S [7]	49.8	53.8	51.1	51.2	34.4	17.7	50.5	47.9	87.2	87.2	< 0.01
	PAC-S (ViT-B/32) [21]	51.5	55.7	53.9	54.3	36.0	18.6	51.0	48.3	89.9	89.9	0.013
	BRIDGE	52.9	57.2	55.4	55.8	36.3	19.0	-	-	93.0	93.0	-
	HICEScore [32]	53.1	57.9	55.9	56.4	37.2	19.2	-	-	93.1	93.1	-
	PAC-S++ (ViT-B/32) [22]	53.9	58.3	54.1	54.5	37.0	19.1	50.5	47.9	90.2	90.2	0.013
	Blip2Score [33]	**56.9**	**61.5**	52.2	52.5	36.7	19.0	53.0	50.7	94.3	94.3	0.020
	Ours (ViT-B/32)	54.0	58.4	**56.2**	**56.6**	**37.8**	**19.5**	**55.9**	**53.0**	**96.7**	**96.7**	0.043
LLM-based	CLAIR [5]	-	55.0	-	44.6	34.4	-	-	-	81.4	83.4	8.3
	FLEUR [10]	-	63.5	-	53.0	38.6	-	-	-	96.8	96.8	3.7
	Ref-FLEUR [10]	-	64.2	-	51.9	38.8	-	-	-	97.3	98.4	4.0
	HiFiScore [31]	-	65.7	-	58.4	-	-	-	-	-	-	-
	Ref-HiFiScore [31]	-	65.8	-	58.4	-	-	-	-	-	-	-

徴量を初期段階で統合するためである．そこで，提案尺度 Pearl では後期統合型のアプローチを採用し，Img-GEM および複数の Ref-GEM からスコアを算出した後，これらを融合して最終的な予測スコアを得る．

Img-GEM は画像に基づき候補文の評価値を算出し，各 Ref-GEM は参照に基づいて評価値を計算する．まず Img-GEM では，M_{img} 個の画像エンコーダを用いて，$\boldsymbol{x}_{\text{img}}$ から画像特徴量 $\{\boldsymbol{v}_{\text{vgem},j} \in \mathbb{R}^{d_{\text{vgem},j}} \,|\, j = 1, 2, \ldots, M_{\text{img}}\}$ を抽出する．ここで，$d_{\text{vgem},j}$ は Img-GEM における j 番目のエンコーダの次元を示す．次に，対応する M_{img} 個のテキストエンコーダを用いて，$\boldsymbol{x}_{\text{cand}}$ から文埋め込み $\{\boldsymbol{c}_{\text{vgem},j} \in \mathbb{R}^{d_{\text{vgem},j}} \,|\, j = 1, 2, \ldots, M_{\text{img}}\}$ を抽出する．Img-GEM では，CLIP [19]，BLIP-2 [12]，および BEiT-3 [1] を画像およびテキストエンコーダとして用いた．

i 番目の Ref-GEM は，i 番目の参照文 $\boldsymbol{x}_{\text{ref}}^{(i)}$ を処理し，M_{ref} 個のテキストエンコーダを用いて以下の文埋め込み $\{\boldsymbol{r}_{\text{rgem},j}^{(i)} \in \mathbb{R}^{d_{\text{rgem},j}} \mid j = 1, 2, \ldots, M_{\text{ref}}\}$ を抽出する．ここで，$d_{\text{rgem},j}$ は Ref-GEM における j 番目のエンコーダの次元を示す．同様に，M_{ref} 個のテキストエンコーダを用いて，候補文 $\boldsymbol{x}_{\text{cand}}$ から文埋め込み $\{\boldsymbol{c}_{\text{rgem},j} \in \mathbb{R}^{d_{\text{rgem},j}} \mid j = 1, 2, \ldots, M_{\text{ref}}\}$ を抽出する．Ref-GEM のテキストエンコーダには，BLIP-2，BEiT-3 および Stella を用いた．Stella は軽量でありつつも，標準ベンチマーク [34] において LLM と同等の性能を持つ文埋め込みモデルである．

Img-GEM では次に，抽出された $\{(\boldsymbol{c}_{\text{vgem},j}, \boldsymbol{v}_{\text{vgem},j}) \,|\, j = 1, 2, \ldots, M_{\text{img}}\}$ を VSS サブモジュールに入力する．一方，Ref-GEM は抽出された $\{(\boldsymbol{c}_{\text{rgem},j}, \boldsymbol{r}_{\text{rgem},j}^{(i)}) \,|\, j = 1, 2, \ldots, M_{\text{img}}\}$ を同じサブモジュールに入力する．Img-GEM の出力と i 番目の Ref-GEM の最終的な出力は，それぞれ $\boldsymbol{h}_{\text{img}}$ および $\boldsymbol{h}_{\text{ref}}^{(i)}$ である．

2.3 Vector Similarity Scoring サブモジュール

VSS サブモジュールは，入力 $\{(\boldsymbol{c}_i, \boldsymbol{h}_i) \mid i = 1, \ldots, M\}$ に基づいて候補文 $\boldsymbol{x}_{\text{cand}}$ を評価する．ここで，$\boldsymbol{c}_i$ は $\boldsymbol{x}_{\text{cand}}$ の i 番目の埋め込みを，$\boldsymbol{h}_i$ は $\boldsymbol{x}_{\text{cand}}$ の評価に用いられる i 番目の埋め込みを表す．ここで，M はエンコーダの数を示す．はじめに，本サブモジュールは各ペア $(\boldsymbol{c}_i, \boldsymbol{h}_i)$ に対して ARSM を適用し，埋め込み間の差異を表す特徴量 $\boldsymbol{g}_i$ を得る．得られた特徴量 $\{\boldsymbol{g}_i\}_{i=1}^{M}$ は結合され，L_T 層からなる Transformer に入力し，評価に有益な特徴量 $\boldsymbol{g}_{\text{enc}}$ を抽出する．その後，$\boldsymbol{g}_{\text{enc}}$ を L_Q 層からなる Q-Former [12] に入力し，特徴量 $\boldsymbol{g}_{\text{dec}}$ を得る．最後に，$\boldsymbol{g}_{\text{dec}}$ に MLP を適用することで，評価値 $\hat{y}_{vc}$ を算出する．参照文ベースの設定では，候補文 $\boldsymbol{x}_{\text{cand}}$ の評価値を画像 $\boldsymbol{x}_{\text{img}}$ または参照文群 $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ に基づいて計算し，それらを融合して最終的な評価値とする．

一方，参照文フリーの設定では，画像のみを基に候補文 $\boldsymbol{x}_{\text{cand}}$ の最終的な評価値を算出する．画像に基づくスコア $\hat{y}_{\text{img}}$ と参照に基づくスコア $\hat{y}_{\text{ref}}$ はそれぞれ $\hat{y}_{\text{img}} = \sigma\big(\text{MLP}(\boldsymbol{h}_{\text{img}})\big)$ および $\hat{y}_{\text{ref}} = \max_i\Big(\sigma\big(\text{MLP}(\boldsymbol{h}_{\text{ref}}^{(i)})\big)\Big)$ によって計算される．ここで，σ はシグモイド関数を表す．

最終的に，提案尺度が出力する評価値 $\hat{y}$ は，$\hat{y}_{\text{img}}$ と $\hat{y}_{\text{ref}}$ を統合することで以下のように計算される．

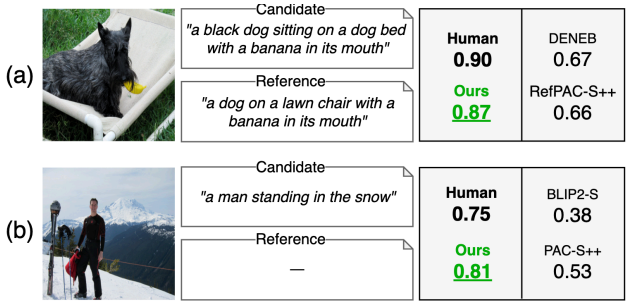

図 3 Nebula データセットにおける成功例

$$\hat{y} = \begin{cases} \lambda \hat{y}_{\text{img}} + (1-\lambda)\hat{y}_{\text{ref}} & \text{(参照文ベース)} \\ \hat{y}_{\text{img}} & \text{(参照文フリー)} \end{cases} \tag{1}$$

ここで，λ は $[0,1]$ のハイパーパラメータであり，本研究では $\lambda = 0.5$ とした．損失関数 $\mathcal{L}$ は Huber 損失 $\mathcal{L}_{\text{huber}}(\cdot,\cdot)$ を用いて以下のように定義される．

$$\mathcal{L} = \begin{cases} \lambda_{\text{img}}\,\mathcal{L}_{\text{huber}}(y, \hat{y}_{\text{img}}) & \\ \quad + \lambda_{\text{ref}}\,\mathcal{L}_{\text{huber}}(y, \hat{y}_{\text{ref}}) & \text{(参照文ベース)} \\ \mathcal{L}_{\text{huber}}(y, \hat{y}_{\text{img}}) & \text{(参照文フリー)} \end{cases} \tag{2}$$

ここで，y は正解値を示し，また，λ_{img} と λ_{ref} はハイパーパラメータである．

3. 実験

3.1 Spica dataset

本研究では，Spica データセットを構築した．本データセットは，多数の人間による評価および，多様かつ広範な画像群を有する．Spica データセットは，2,360 人のアノテーターから収集された 333,397 件の人間評価と，75,535 種類の画像から構成される．本データセットは，最も多くの人間による評価を有する Polaris [28] の 2.5 倍の評価数を含み，また，現在最も多くの画像数を持つ Nebula [14] の 2.3 倍の画像数を含む．さらに，候補キャプションの生成には 10 種類の画像キャプション生成モデルを用いた [35, 26, 6, 11, 29, 30, 12].

3.2 定量的結果

3.2.1 人間による評価との相関

表 1 に Composite [2]，Flickr8K-Expert [8]，Flickr8K-CF [8]，および Nebula [14] データセットにおける，ベースライン尺度と提案尺度との定量的比較結果を示す．自動評価尺度の評価には，Kendall's τ_b，τ_c の相関係数を用いた．

参照文ベースの設定での提案尺度は Composite，Flickr8K-Expert，Flickr8K-CF，Nebula において，それぞれ τ_b で 55.8，58.2，38.6，58.4，τ_c で 60.4，58.6，20.0，55.4 であった．これらの結果より，提案尺度が既存の LLM フリーのベースライン尺度に対し，τ_b で 1.1，1.0，0.4，0.3 ポイント，τ_c で 1.3，0.9，0.2，0.3 ポイント上回った．参照文フリーの設定においても同様に，提案尺度が LLM フリーのベースライン尺度を上回った．

3.2.2 ハルシネーションへの頑健性

表 1 に，ハルシネーションに対する頑健性を検証する FOIL データセット [23] での性能を示す．参照文ベース，参照文フリーの両設定において，提案尺度は LLM フリー自動評価尺度の中で最高性能であった．参照文が 1 文与えられる設定では 96.5%，参照文が 4 文与えられる設定では 97.2%の正解率であり，参照文フリーの設定では 96.7%の正解率であった．これらの結果から，提案尺度がハルシネーションに対して頑健であることが示唆される．

3.3 推論時間

自動評価尺度の性能と同時に，推論時間を考慮することも重要である．表 1 の最右列に，画像キャプションの標準ベンチマークである COCO [13]，NoCaps [3]，および TextCaps[25] のテストセットにおける総評価時間を示す．Pearl は 2.58 分で全サンプルを評価可能であるのに対し，LLM ベースの自動評価尺度である CLAIR と FLEUR は，それぞれ 8.3 時間と 3.7 時間と著しく評価速度が低速であった．このことから，提案尺度は人間との相関が高いだけでなく，評価速度も高速であるといえる．

3.4 定性的結果

図 3 に，Nebula データセットにおける提案尺度の成功例を示す．図 3 (a) は参照文ベースでの成功例であり，図 3 (b) は参照文フリーでの成功例である．

図 3 (a) では，$\boldsymbol{x}_{\text{cand}}$ が画像を適切に表現しているため，人間による評価 y は 0.90 であった．一方，Ref-CLIPS [7] および Ref-PACS++ [22] はそれぞれ 0.67 および 0.66 と評価したが，Pearl は適切に 0.87 と評価した．図 3 (b) では，$\boldsymbol{x}_{\text{cand}}$ が画像の一部のみを記述していたため，人間の評価 y は 0.75 であった．PAC-S++および BLIP-S [33] はそれぞれ 0.53 と 0.38 と評価したが，Pearl は 0.81 と評価し，人間による評価により近い評価値を出力した．

4. おわりに

本研究では，画像キャプション向け自動評価尺度 Pearl を提案した．本研究の主な貢献は次の通りである．(i) 参照文ベースおよびフリーの両設定における評価を可能とするため，Image-Guided Evaluation Module (Img-GEM) および Reference-Guided Evaluation Module (Ref-GEM) を導入した．(ii) 画像キャプション生成向け自動評価尺度のための，新たな大規模データセット Spica を構築した．(iii) 標準ベンチマークにおいて，既存の LLM フリー自動評価尺度を上回る結果を得た．

謝辞

本研究の一部は，JSPS 科研費 23K28168，JST ムーンショットの助成を受けて実施されたものである．

参考文献

[1] *Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks* (2023).

[2] Aditya, S. et al.: From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge, *arXiv preprint arXiv:1511.03292* (2015).

[3] Agrawal, H., Desai, K. et al.: nocaps: Novel Object Captioning at Scale, *ICCV*, pp. 8948–8957 (2019).

[4] Anderson, P., Fernando, B., Johnson, M. et al.: SPICE: Semantic Propositional Image Caption Evaluation, *ECCV*, pp. 382–398 (2016).

[5] Chan, D., Petryk, S. et al.: CLAIR: Evaluating Image Captions with Large Language Models, *EMNLP*, pp. 13638–13646 (2023).

[6] Cornia, M., Stefanini, M., Baraldi, L. et al.: Meshed-Memory Transformer for Image Captioning, *CVPR*, pp. 10578–10587 (2020).

[7] Hessel, J., Holtzman, A. et al.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning, *EMNLP*, pp. 7514–7528 (2021).

[8] Hodosh, M., Young, P. and Hockenmaier, J.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, *JAIR*, Vol. 47, pp. 853–899 (2013).

[9] Lee, H., Yoon, S. et al.: UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning, *ACL*, pp. 220–226 (2021).

[10] Lee, Y. et al.: FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model, *ACL*, pp. 3732–3746 (2024).

[11] Li, J. et al.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, *ICML*, pp. 12888–12900 (2022).

[12] Li, J. et al.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, *ICML*, pp. 19730–19742 (2023).

[13] Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R. et al.: Microsoft COCO: Common Objects in Context, *ECCV*, pp. 740–755 (2014).

[14] Matsuda, K. et al.: DENEB: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning, *ACCV*, pp. 3570–3586 (2024).

[15] Navigli, R., Conia, S. and Ross, B.: Biases in large language models: origins, inventory, and discussion, *ACM*, Vol. 15, No. 2, pp. 1–21 (2023).

[16] Ohi, M. et al.: HarmonicEval: Multi-modal, Multi-task, Multi-criteria Automatic Evaluation Using a Vision Language Model, *arXiv preprint arXiv:2412.14613* (2024).

[17] Panickssery, A., Bowman, S. R. and Feng, S.: LLM Evaluators Recognize and Favor Their Own Generations, *NeurIPS* (2024).

[18] Papineni, K., Roukos, S., Ward, T. and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, *ACL*, pp. 311–318 (2002).

[19] Radford, A., Kim, J. W., Hallacy, C. et al.: Learning Transferable Visual Models from Natural Language Supervision, *ICML*, pp. 8748–8763 (2021).

[20] Rei, R., Stewart, C., Farinha, A. and Lavie, A.: COMET: A Neural Framework for MT Evaluation, *EMNLP*, pp. 2685–2702 (2020).

[21] Sarto, S., Barraco, M. et al.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation, *CVPR*, pp. 6914–6924 (2023).

[22] Sarto, S., Moratelli, N. et al.: Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training, *arXiv preprint arXiv:2410.07336* (2024).

[23] Shekhar, R., Pezzelle, S., Klimovich, Y. et al.: FOIL it! Find One Mismatch Between Image and Language caption, *ACL*, pp. 255–265 (2017).

[24] Shimanaka, H. et al.: RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation, *WMT*, pp. 751–758 (2018).

[25] Sidorov, O., Hu, R. et al.: TextCaps: a Dataset for Image Captioning with Reading Comprehension, *ECCV*, pp. 742–758 (2020).

[26] Suganuma, M., Okatani, T. et al.: GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features, *ECCV*, pp. 167–184 (2022).

[27] Vedantam, R., Zitnick, L. and Parikh, D.: CIDEr: Consensus-based Image Description Evaluation, *CVPR*, pp. 4566–4575 (2015).

[28] Wada, Y., Kanta, K. et al.: Polos: Multimodal Metric Learning from Human Feedback for Image Captioning, *CVPR*, pp. 13559–13568 (2024).

[29] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z. et al.: GIT: A Generative Image-to-text Transformer for Vision and Language, *TMLR* (2022).

[30] Wang, P. et al.: OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-sequence Learning Framework, *ICML*, pp. 23318–23340 (2022).

[31] Yao, Z. et al.: HiFi-Score: Fine-Grained Image Description Evaluation with Hierarchical Parsing Graphs, *ECCV*, pp. 441–458 (2024).

[32] Zeng, Z., Sun, J., Zhang, H., Wen, T., Su, Y., Xie, Y. et al.: HICEScore: A Hierarchical Metric for Image Captioning Evaluation, *ACM*, p. 866–875 (2024).

[33] Zeng, Z., Xie, Y., Zhang, H., Chen, C., Chen, B. et al.: MeaCap: Memory-Augmented Zero-shot Image Captioning, *CVPR*, pp. 14100–14110 (2024).

[34] Zhang, D., Li, J., Zeng, Z. and Wang, F.: Jasper and Stella: distillation of SOTA embedding models, *arXiv preprint arXiv:2412.19048* (2024).

[35] Zhang, P., Li, X., Hu, X. et al.: VinVL: Revisiting Visual Representations in Vision-language Models, *CVPR*, pp. 5579–5588 (2021).