Polos: 画像キャプション生成における教師あり自動評価尺度

和田唯我　兼田寛大　齋藤大地　杉浦孔明
慶應義塾大学
{yuiga,k.kaneda,daichi-s,komei.sugiura}@keio.jp

概要

画像キャプション生成タスクでは，生成文の品質が適切に評価されることが重要である．しかし，近年のデータ駆動型自動評価尺度は，多様な画像および言語に対する汎化性能が低いという問題が指摘されている．この問題は，これらの尺度が画像キャプション生成とは無関係なタスクで学習された埋め込み表現を用いており，また古典的手法によって類似度を計算しているだけに過ぎないためであると考えられる．そこで，本研究では，画像キャプション生成タスクにおける教師あり自動評価尺度 Polos を提案する．Polos は画像と言語を入力とし，大規模対照学習によって学習された埋め込みを用いた並列クロスモーダル特徴抽出機構により評価値を計算する．また，本研究では人間のフィードバックに基づき自動評価尺度を学習するフレームワーク M^{2}LHF を提案する．さらに，Polos を学習するため，550 人の被験者から 13 万サンプルの人間による評価を収集した最大規模のデータセット Polaris を構築した．実験の結果，Polos は Composite, Flickr8K-Expert, Flickr8K-CF, PASCAL-50S, FOIL, Polaris において，既存手法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，視覚障害者の補助，画像に関する対話生成，画像に基づく質問応答など，多くの用途で社会応用されている．本分野において効率的なモデル開発を行うには，人間による評価に近い自動評価尺度の構築が不可欠である．先行研究では，古典的な自動評価尺度 [1–5] が人間による評価と十分に相関していないことが報告されており [5,6]，その結果様々なデータ駆動型自動評価尺度が提案された [6–9]．しかし，これらの手法は画像キャプション生成とは無関係なタスクで学習された埋め込み表現を用いており，さらにコサイン類似度や最適輸送等の古典的手法によって類似度を計算しているだけに過ぎない．そのため，これらの手法は多様な画像および言語における汎化性能が低く，またこれらのモデルは hallucination を適切に扱うことが出来ない点も指摘されている [6]．

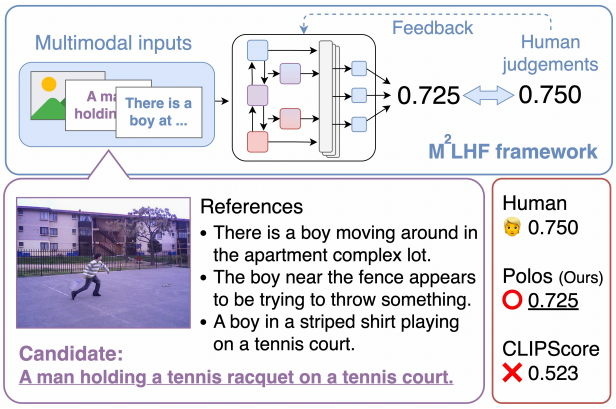

図 1: 提案手法 Polos および M^{2}LHF.

そこで，本研究では，画像キャプション生成タスクにおける教師あり自動評価尺度 Polos を提案する．さらに，人間のフィードバックに基づく自動評価尺度を構築するためのフレームワーク Multimodal Metric Learning from Human Feedback (M^{2}LHF) も提案する．図 1 に提案手法 Polos および M^{2}LHF の概要図を示す．提案手法 Polos は M^{2}LHF に基づき，人間による評価を学習する．また，提案手法は SimCSE [10] で事前学習された RoBERTa [11] や CLIP [12] を用いた並列クロスモーダル特徴抽出機構を用いて有用な特徴量を抽出する．既存手法は古典的手法によってスカラの類似度を計算しているのに対して，提案手法は本機構によりベクトル空間における複雑な関係を捉えることができる．

さらに，M^{2}LHF に基づく Polos を学習するため，本研究では多様な人間による評価によって構成された新たなデータセット Polaris を構築した．Polaris は現時点で最大規模のデータセット [13] の約 10 倍のサンプルが含まれており，13 万サンプルの人間による評価および多様なキャプションで構成される．

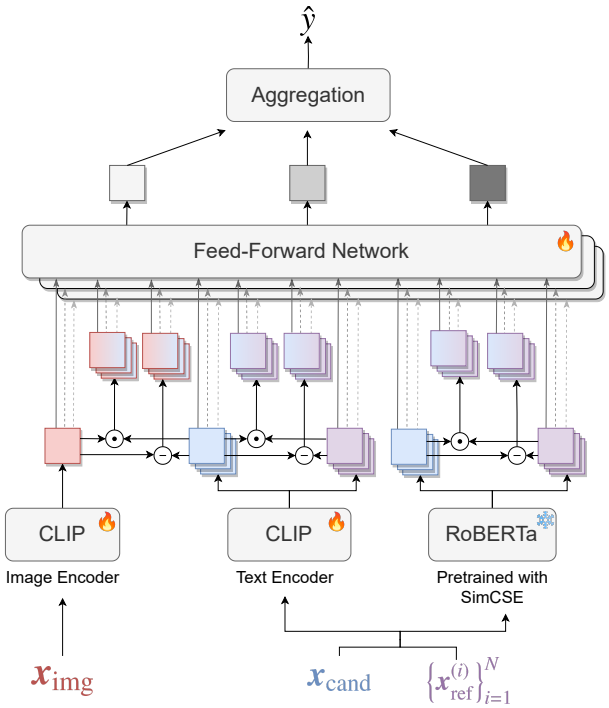

図 2: 提案手法 Polos のモデル図.

提案手法における新規性は以下の通りである.

- 画像キャプション生成タスクにおける教師あり自動評価尺度 Polos を提案する.
- 人間のフィードバックに基づく自動評価尺度を構築するフレームワーク M^{2}LHF を提案する.
- SimCSE および CLIP に基づく並列クロスモーダル特徴抽出機構を導入する.
- Polos を学習するため，13 万の人間による評価を含むデータセット Polaris を構築する.
- Composite, PASCAL-50S, FOIL, Flickr8K-Expert, Flickr8K-CF, Polaris において既存手法を上回る結果を得た.

2 提案手法

本研究では，画像キャプション生成タスクにおける教師あり自動評価尺度 Polos を提案する．図 2 に提案手法のモデル図を示す．また，人間のフィードバックに基づく自動評価尺度を構築するためのフレームワーク M^{2}LHF を提案する．本手法は，機械翻訳における自動評価尺度である COMET [14] や RUSE [15] をマルチモダリティへと拡張した手法である．提案手法は M^{2}LHF に基づき，入力 $\boldsymbol{x}$ から評価値 $\hat{y}$ を計算し，人間による評価 y を直接回帰する．

ここで，文埋め込みの生成方法には十分留意する必要がある．前述の通り，データ駆動型自動評価尺度の多くは，多様な画像および言語に対する汎化性能が低い．これは多くの尺度が CLIP [12] のテキストエンコーダのみを使用している点に起因すると考えられる．Sarto らが指摘しているように，CLIP は比較的短く簡潔な alt-text によって学習されているため，本研究で扱う長いキャプションを評価するには不十分である可能性がある [16]．そのため，本研究では CLIP の文埋め込みと比較して，教師あり SimCSE [10] に基づく文埋め込みがより有益であると考える．SimCSE は，文埋め込み生成における対照学習手法であり，STS タスクにより有用性が実証されている [10]．以上より，本研究では文埋め込みを得るため，CLIP のテキストエンコーダおよび SimCSE で学習された RoBERTa [11] を使用する.

2.1 入力および埋め込み表現

本手法における入力 $\boldsymbol{x}$ は生成文 $\boldsymbol{x}_{\text{cand}} \in \{1,0\}^{V\times L}$, i 番目の参照文 $\{\boldsymbol{x}_{\text{ref}}^{(i)}\} \in \{1,0\}^{N\times V\times L}$，画像 $\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times H\times W}$ を用いて，$\boldsymbol{x} = \{\boldsymbol{x}_{\text{cand}}, \{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}, \boldsymbol{x}_{\text{img}}\}$ と定義される．ここで，V, L, N, H, W はそれぞれ，語彙サイズ，最大トークン長，参照文の数，および画像の幅と高さを指す.

はじめに，SimCSE で学習された RoBERTa [11] を用いて，$\boldsymbol{x}_{\text{cand}}, \{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ からそれぞれ，文埋め込み $\boldsymbol{c}_{\text{rb}} \in \mathbb{R}^{L\times d_R}$ および $\{\boldsymbol{r}_{\text{rb}}^{(i)}\}_{i=1}^{N} \in \mathbb{R}^{N\times L\times d_R}$ を得る．ここで，d_R は RoBERTa の出力次元を指し，文埋め込みは RoBERTa の出力した `[CLS]` トークンから得た．続いて，CLIP のテキストエンコーダを用いて，$\boldsymbol{x}_{\text{cand}}, \boldsymbol{x}_{\text{ref}}^{(i)}$ から，それぞれ文埋め込み $\boldsymbol{c}_{\text{clip}} \in \mathbb{R}^{d_{\text{CLIP}}}$ および $\boldsymbol{r}_{\text{clip}}^{(i)} \in \mathbb{R}^{d_{\text{CLIP}}}$ を得る．ここで，d_{CLIP} は CLIP の出力次元を指す．また，CLIP の画像エンコーダ (ViT-B/16) より，$\boldsymbol{x}_{\text{img}}$ から $\boldsymbol{v} \in \mathbb{R}^{d_{\text{CLIP}}}$ を抽出する.

2.2 並列クロスモーダル特徴抽出機構

本研究では，M^{2}LHF に効果的な特徴抽出器として，並列クロスモーダル特徴抽出機構を導入する．本機構は，アダマール積と差分を使用する RUSE [14, 15] をマルチモダリティへと拡張した機構である．CLIP の学習において，対応する画像と文のコサイン類似度を最小化するよう設計されている点を考えると，アダマール積を CLIP 特徴量に適用することは有用であると考えられる．また，差分やアダマール積はベクトル同士における要素間の増幅や減衰を捉えることが出来るため，類似度をスカ

ラではなくベクトルで表現することができる．したがって，本機構では，アダマール積と差分を使用することで，画像キャプション生成の評価において有用な特徴量を CLIP および RoBERTa から抽出できると考えられる．

はじめに，本機構では次に示す入力

$$\left\{\boldsymbol{c}_{\text{clip}},\ \boldsymbol{r}_{\text{clip}}^{(i)},\ \boldsymbol{c}_{\text{rb}},\ \boldsymbol{r}_{\text{rb}}^{(i)},\ \boldsymbol{v}\right\} \tag{1}$$

より，$\boldsymbol{h}_{\text{inter}}^{(i)}$ を次式のように計算する．

$$\boldsymbol{h}_{\text{inter}}^{(i)} = [F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{r}_{\text{clip}}^{(i)});\ F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{v});\ F(\boldsymbol{c}_{\text{rb}}, \boldsymbol{r}_{\text{rb}}^{(i)})] \tag{2}$$

ここで，F はアダマール積 $\odot$ を用いて以下のように定義された関数である．

$$F(\boldsymbol{c}, \boldsymbol{r}) = [\boldsymbol{c};\ \boldsymbol{r};\ |\boldsymbol{c} - \boldsymbol{r}|;\ \boldsymbol{c} \odot \boldsymbol{r}] \tag{3}$$

続いて，$\boldsymbol{h}_{\text{inter}}^{(i)}$ より，MLP を用いて i 番目の参照文における類似度 $\boldsymbol{h}^{(i)}$ を計算する．ここで，本研究では事前実験により，MLP が Transformer [17] よりも高い性能が得られたため MLP を採用した．

最後に，評価値 $\hat{y}$ を次のように計算する．

$$\hat{y} = \underset{i}{\text{Aggregate}}(\sigma(\text{MLP}(\boldsymbol{h}^{(i)}))) \tag{4}$$

ここで，σ はシグモイド関数を指し，Aggregate は任意の写像 $f : \mathbb{R}^N \to \mathbb{R}$ を指す．Aggregate 関数には，Max 関数や Mean 関数などが考えられ，本研究では Max 関数を採用した．さらに損失関数には，回帰問題において標準的な損失である平均二乗誤差を用いた．

3 実験設定

本研究では，画像，キャプション，および人間による評価で構成された Polaris データセットを構築した．教師あり自動評価尺度の学習には多様なキャプションおよび人間による評価が含まれた大規模コーパスが不可欠である．しかし，我々の知る限り，そのようなデータセットは限られる．したがって，我々は 550 人の被験者から 131,020 の人間による評価を収集し，Polaris データセットを構築した．Polaris は現時点で最大規模のデータセット [13] の約 10 倍のサンプルが含まれており，13 万サンプルの人間による評価および多様なキャプションで構成される．

Polaris データセットは次に示す 10 個の標準的なモデルの推論結果が含まれている: SAT [18], $\mathcal{M}^2$-Transformer [19], VinVL [20], GRIT [21], $\text{BLIP}_{\text{base}}$, $\text{BLIP}_{\text{large}}$ [22], GIT [23], OFA [24], $\text{BLIP-2}_{\text{flan}}$, $\text{BLIP-2}_{\text{opt}}$ [25]．ここで，$\text{BLIP}_{\text{base}}, \text{BLIP}_{\text{large}}$ はそれぞれ，ViT-B および ViT-L を画像エンコーダに使用した BLIP であり，$\text{BLIP-2}_{\text{flan}}, \text{BLIP-2}_{\text{opt}}$ はそれぞれ大規模言語モデルとして Flan-T5 [26] および OPT [27] を使用した BLIP-2 を指す．我々は，キャプションの多様性を確保するため，最新のモデルだけでなく比較的性能の低いモデルも採用した．また，Polaris データセットには，前述の 10 個のモデルにおける MS-COCO [28] および nocaps [29] の推論結果が含まれる．本研究では，キャプションの多様性を考慮し，MS-COCO に加え，MS-COCO よりも多様なクラスが含まれている nocaps も採用した．

	Composite	Flickr8K (Expert)	Flickr8K (CF)	Polaris
Classic metrics				
BLEU [1]	30.6	30.8	16.4	46.3
ROUGE [3]	32.4	32.3	19.9	46.3
CIDEr [4]	37.7	43.9	24.6	52.1
METEOR [2]	38.9	41.8	22.2	51.2
SPICE [5]	40.3	44.9	24.4	51.0
SPARCS [30]	43.1	48.1	10.4	43.3
Similarity-based metrics				
MoverScore [8]	30.1	46.7	22.8	46.4
BERTScore [7]	30.1	46.7	22.8	51.6
BARTScore [31]	43.5	37.8	24.3	47.3
LEIC [32]	–	–	29.5	–
TIGEr [33]	45.4	–	–	–
ViLBERTScore [9]	52.4	50.1	–	–
CLIP-S [6]	53.8	51.2	34.4	52.3
RefCLIP-S [6]	55.4	53.0	36.4	52.3
MID [34]	55.7	54.9	37.3	51.3
Learning-based metrics				
PAC-S [16]	55.7	54.3	36.0	52.5
UMIC [35]	56.1	46.8	30.1	49.8
RefPAC-S [16]	57.3	55.9	37.6	56.0
Polos (Ours)	**57.6** (+0.3)	**56.4** (+0.5)	**37.8** (+0.2)	**57.8** (+1.8)

表 1: 定量的結果 (Kendall's τ)

4 実験結果

本研究では，Composite [13], Flickr8K [36], Flickr8K-CF [36], PASCAL50-S [5], FOIL [37]，および Polaris データセットによる実験を行った．（PASCAL50-S および FOIL における実験結果は付録 A.1 に示す．）表 1 に，Composite, Flickr8K, Flickr8K-CF, Polaris にお

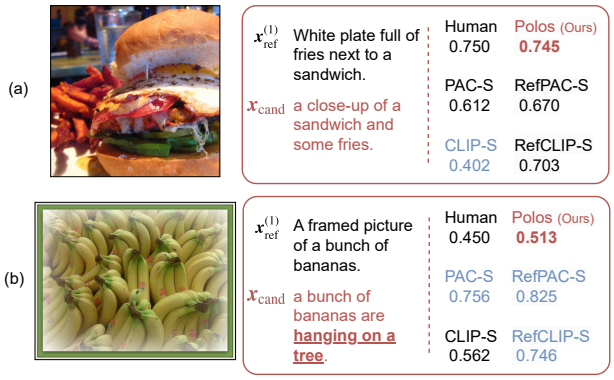

図 3: 提案手法の成功例.

ける実験結果を示す．本実験では，先行研究 [5,6,34] と同様に，Flickr8K-CF においては τ_b (Kendall-B) を，その他のデータセットにおいては τ_c (Kendall-C) を使用した．表より，提案手法における相関係数は Composite, Flickr8K-Expert, Flickr8K-CF, Polaris において，それぞれ 57.6, 56.4, 37.8, 57.8 であり，RefPAC-S と比較して，0.3, 0.5, 0.2, 1.8 ポイント上回った．

図 3 に Polaris データセットにおける提案手法の成功例を示す．ここで，赤色と青色で示した尺度はそれぞれ，y に最も近い尺度および y と大きく乖離した尺度を指す．また，下線は $\boldsymbol{x}_{\text{cand}}$ における重大な誤りを表す．図 3 (a) は，$\boldsymbol{x}_{\text{cand}}$ が画像の内容を適切に捉えていたため，y が 0.750 であったサンプルである．本例では，CLIP-S が 0.402 と誤って評価していたのに対して，提案手法は 0.745 と適切な評価値を出力した．図 3 (b) は $\boldsymbol{x}_{\text{cand}}$ が画像の内容を部分的に正しく捉えていた例である．本サンプルでは，画像から句 “hanging on a tree” が適切であるか不明瞭であるため，y は 0.450 であった．この例では，RefPAC-S と RefCLIP-S がそれぞれ 0.825, 0.746 と $\boldsymbol{x}_{\text{cand}}$ を過大評価していたのに対して，提案手法は 0.513 と適切な評価値を出力した．以上より，提案手法は人間による評価に近い値を出力したといえる．

4.1 Ablation Study

本研究では，提案手法の有効性を調査するため，ablation studies を行った．表 2 に ablation studies の結果を示す．ここで，‘P’ は並列クロスモーダル特徴抽出機構の有無を指す．

Parallel feature extraction ablation. 並列クロスモーダル特徴抽出機構の性能への寄与を調べるため，アダマール積と差分による処理を取り除いた．具体的には，$F(\mathbf{c}, \mathbf{r}) \leftarrow [\mathbf{c};\ \mathbf{r}]$ へと変更した．表 2 より，Metric (i) は Metric (vi) と比較して，Composite, Flickr8K, Polaris において，それぞれ 18.3, 15.4, 6.4 ポイント減少した．したがって，並列クロスモーダル特徴抽出機構は性能向上に寄与したといえる．

Metric	P	$\boldsymbol{x}_{\text{img}}$	CLIP	RoBERTa	Aggregate	Composite	Flickr8K	Polaris
(i)		✓	✓	✓	Max	39.3	41.0	51.4
(ii)	✓		✓	✓	Max	56.8	55.5	57.1
(iii)	✓			✓	Max	55.0	53.2	55.4
(iv)	✓	✓	✓		Max	56.0	55.0	57.4
(v)	✓	✓	✓	✓	Mean	55.1	55.4	52.1
(vi)	✓	✓	✓	✓	Max	**57.6**	**56.4**	**57.8**

表 2: Ablation studies の結果.

M^{2}LHF ablation. M^{2}LHF の性能向上への寄与を調査するため，$\boldsymbol{x}_{\text{img}}$, $(\boldsymbol{c}_{\text{clip}}, \{\boldsymbol{r}_{\text{clip}}^{(i)}\}_{i=1}^{N})$ および $(\boldsymbol{c}_{\text{rb}}, \{\boldsymbol{r}_{\text{rb}}^{(i)}\}_{i=1}^{N})$ を除去した．まず，画像特徴量の性能への寄与を調べるために $\boldsymbol{x}_{\text{img}}$ を除去した結果，Metric (ii) は Metric (vi) と比較して，Composite, Flickr8K, Polaris において，それぞれ 0.4, 1.4, 0.7 ポイント減少した．このことから，$\boldsymbol{x}_{\text{img}}$ は性能向上に寄与したといえる．続いて，各言語特徴量の性能への寄与を調べるため，CLIP および RoBERTa を除去した結果，Metric (iii) は，それぞれ 2.6, 3.2, 2.4 ポイント減少した．同様に，RoBERTa を取り除くことでも，性能は低下した．これらの結果より，M^{2}LHF は確かに性能へ寄与したといえる．

Aggregation mechanism ablation. Aggregate 関数の性能への寄与を調べるため，Max 関数を Mean 関数へと変更した．表 2 より，Metric (v) は Metric (vi) と比較して，Composite, Flickr8K, Polaris において，それぞれ 2.5, 1.0, 5.7 ポイント減少した．このことから，Max 関数の性能への寄与が確認できた．

5 おわりに

本研究では，画像キャプション生成タスクにおける教師あり自動評価尺度 Polos を提案した．提案手法の新規性は次の通りである．(i) 人間のフィードバックに基づく自動評価尺度を構築するためのフレームワーク M^{2}LHF を提案した．(ii) SimCSE および CLIP に基づく並列クロスモーダル特徴抽出機構を導入した．(iii) Polos を学習するため，13 万の人間による評価で構成されたデータセット Polaris を構築した．(iv) Composite, PASCAL-50S, FOIL, Flickr8K-Expert, Flickr8K-CF, Polaris において既存手法を上回る結果を得た．

謝辞

本研究は，Apple 社の助成を受けて実施された．本研究で述べられた見解，意見，発見，結論および推奨は全て著者らのものであり，明示的または暗黙的を問わず，Apple 社の見解，方針または立場を反映するものではない．また，本研究の一部は，JSPS 科研費 23H03478，JST CREST，NEDO の助成を受けて実施されたものである．

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **ACL**, pp. 311–318, 2002. 1, 3, 6

[2] Satanjeev Banerjee, et al. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In **ACL**, pp. 65–72, 2005. 1, 3, 6

[3] Chin Lin. ROUGE: A Package For Automatic Evaluation Of Summaries. In **ACL**, pp. 74–81, 2004. 1, 3, 6

[4] Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In **CVPR**, pp. 4566–4575, 2015. 1, 3, 6

[5] Peter Anderson, Basura Fernando, Mark Johnson, et al. SPICE: Semantic Propositional Image Caption Evaluation. In **ECCV**, pp. 382–398, 2016. 1, 3, 4, 6

[6] Jack Hessel, Ari Holtzman, et al. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In **EMNLP**, pp. 7514–7528, 2021. 1, 3, 4, 6

[7] Tianyi Zhang, et al. BERTScore: Evaluating Text Generation with BERT. In **ICLR**, 2020. 1, 3, 6

[8] Wei Zhao, et al. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In **EMNLP-IJCNLP**, pp. 563–578, 2019. 1, 3, 6

[9] Hwanhee Lee, Seunghyun Yoon, et al. ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In **Eval4NLP**, pp. 34–39. 1, 3, 6

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **EMNLP**, pp. 6894–6910, 2021. 1, 2

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv preprint arXiv:1907.11692**, 2019. 1, 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning Transferable Visual Models from Natural Language Supervision. In **ICML**, pp. 8748–8763, 2021. 1, 2

[13] Somak Aditya, Yezhou Yang, Chitta Baral, et al. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. **arXiv preprint arXiv:1511.03292**, 2015. 1, 3, 6

[14] Ricardo Rei, et al. COMET: A Neural Framework for MT Evaluation. In **EMNLP**, pp. 2685–2702, 2020. 2

[15] Hiroki Shimanaka, et al. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In **WMT18**, pp. 751–758, 2018. 2

[16] Sara Sarto, et al. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In **CVPR**, pp. 6914–6924, 2023. 2, 3, 6

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. Attention Is All You Need. In **NIPS**, Vol. 30, pp. 5998–6008, 2017. 3

[18] Kelvin Xu, Jimmy Ba, Ryan Kiros, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In **ICML**, pp. 2048–2057, 2015. 3

[19] Marcella Cornia, et al. Meshed-Memory Transformer for Image Captioning. In **CVPR**, pp. 10578–10587, 2020. 3

[20] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, et al. VinVL: Revisiting Visual Representations in Vision-language Models. In **CVPR**, pp. 5579–5588, 2021. 3

[21] Masanori Suganuma, Takayuki Okatani, et al. GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features. In **ECCV**, pp. 167–184, 2022. 3

[22] Junnan Li, et al. BLIP: Bootstrapping Language-image Pre-training for Unified Vision-language Understanding and Generation. In **ICML**, pp. 12888–12900, 2022. 3

[23] Jianfeng Wang, et al. GIT: A Generative Image-to-text Transformer for Vision and Language. **TMLR**, 2022. 3

[24] Peng Wang, et al. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-sequence Learning Framework. In **ICML**. 3

[25] Junnan Li, Dongxu Li, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In **ICML**, 2023. 3

[26] Hyung Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, et al. Scaling Instruction-finetuned Language Models. **arXiv preprint arXiv:2210.11416**, 2022. 3

[27] Susan Zhang, Stephen Roller, et al. OPT: Open Pre-trained Transformer Language Models. **arXiv preprint arXiv:2205.01068**, 2022. 3

[28] Tsung Lin, et al. Microsoft COCO: Common Objects in Context. In **ECCV**, pp. 740–755, 2014. 3

[29] Harsh Agrawal, Karan Desai, et al. nocaps: Novel Object Captioning at Scale. In **ICCV**, pp. 8948–8957, 2019. 3

[30] Joshua Feinglass, et al. SMURF: SeMantic and linguistic UndeRstanding Fusion for Caption Evaluation via Typicality Analysis. In **IJCNLP**, pp. 2250–2260, 2021. 3

[31] Weizhe Yuan, Graham Neubig, et al. BARTScore: Evaluating Generated Text as Text Generation. In **NeurIPS**, Vol. 34, pp. 27263–27277, 2021. 3, 6

[32] Yin Cui, Guandao Yang, et al. Learning to Evaluate Image Captioning. In **CVPR**, pp. 5804–5812, 2018. 3

[33] Ming Jiang, et al. TIGEr: Text-to-image Grounding For Image Caption Evaluation. In **EMNLP**, 2019. 3, 6

[34] Jin Kim, et al. Mutual Information Divergence: A Unified Metric for Multimodal Generative Models. In **NeurIPS**, Vol. 35, pp. 35072–35086, 2022. 3, 4, 6

[35] Hwanhee Lee, Seunghyun Yoon, et al. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In **ACL**, pp. 220–226, 2021. 3, 6

[36] Micah Hodosh, et al. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. **JAIR**, Vol. 47, pp. 853–899, 2013. 3, 6

[37] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, et al. FOIL it! Find One Mismatch Between Image and Language caption. In **ACL**, pp. 255–265, 2017. 3, 6

A 付録

A.1 実験結果

本研究では，Composite [13], Flickr8K [36], Flickr8K-CF [36], Polaris に加えて，PASCAL50-S [5], FOIL [37] においても実験を行った．PASCAL-50S は画像キャプション生成の自動評価尺度構築における標準的なベンチマークであり，FOIL は自動評価尺度が hallucination を適切に対処可能であることを検証するための標準的なデータセットである．

A.1.1 PASCAL-50S における実験結果

	HC	HI	HM	MM	Mean
Classic metrics					
BLEU [1]	60.4	90.6	84.9	54.7	72.7
METEOR [2]	63.8	97.7	93.7	65.4	80.2
ROUGE [3]	63.7	95.3	92.3	61.2	78.1
SPICE [5]	63.6	96.3	86.7	68.3	78.7
CIDEr [4]	65.1	98.1	90.5	64.8	79.6
Similarity-based metrics					
ViLBERTScore [9]	49.9	99.6	93.1	75.8	79.6
BERTScore [7]	65.4	98.1	96.4	60.3	80.1
MoverScore [8]	65.1	97.1	93.2	65.6	80.3
TIGEr [33]	56.0	**99.8**	92.8	74.2	80.7
CLIP-S [6]	56.5	99.3	96.4	70.4	80.7
RefCLIP-S [6]	64.5	99.6	95.4	72.8	83.1
MID [34]	67.0	99.7	97.4	76.8	85.2
Learning-based metrics					
PAC-S [16]	60.6	99.3	96.9	72.9	82.4
RefPAC-S [16]	67.7	99.6	96.0	75.6	84.7
UMIC [35]	66.1	**99.8**	**98.1**	76.2	85.1
Polos (Ours)	**70.0** (+3.0)	99.6	97.4	**79.0** (+1.2)	**86.5** (+1.3)

表 3: Pascal50-S における実験結果.

PASCAL-50S [4] は，与えられた 2 文の組のうち，人間による評価が高い文を特定するタスクである．具体的には，HC (正しい人間による文の組), HI (正しい人間による文と誤った文の組), HM (人間による文と自動生成された文の組), MM (両者とも自動生成された文の組) の 4 つのカテゴリそれぞれにおいて，人間による評価の高い文を特定するタスクである．本研究では，[6] と同様に，48 文のキャプションから無作為に選択された 5 文を用いて評価を行った．

表 3 に PASCAL-50S における実験結果を示す．表より，提案手法の性能は HC，MM，および Mean において，それぞれ 70.0%, 79.0%, 86.5% であり，既存手法と比較して，それぞれ 3.0, 1.2, 1.3 ポイント上回った．以上より，提案手法が PASCAL-50S の 3 カテゴリにおいて既存手法を上回る結果を得たことを確認できた．

A.1.2 FOIL における実験結果

	1-ref	4-ref
BLEU [1]	66.5	82.6
ROUGE [3]	71.7	79.3
METEOR [2]	78.8	82.6
CIDEr [4]	82.5	90.6
SPICE [5]	75.5	86.1
BARTScore [31]	85.3	91.1
MoverScore [8]	88.4	88.4
BERTScore [7]	88.6	92.1
CLIP-S [6]	87.2	87.2
MID [34]	90.5	90.5
PAC-S [16]	89.9	89.9
RefCLIP-S [6]	91.0	92.6
RefPAC-S [16]	**93.7**	94.9
Polos (Ours)	93.3	**95.4**

表 4: FOIL における実験結果.

本研究では，提案尺度が hallucination を適切に対処可能か確認するため，FOIL データセットによる実験を行った．本実験においても，先行研究 [6] と同様に，それぞれ 1 文および 4 文の参照文群が付与された 3 万枚の画像を用いて評価を行った．

表 4 に FOIL データセットにおける定量的結果を示す．提案手法の性能は，1 文および 4 文の参照文群が付与された設定において，それぞれ 93.3%と 95.4%であり，特に後者の設定において，RefPAC-S を 0.8 ポイント上回った．したがって，表 1, 3, 4 より，提案手法が既存手法を上回る結果を得たこと，および hallucination を適切に対処可能であることを確認できた．