対照学習と階層的特徴抽出に基づく画像キャプション生成の自動評価

戸倉 健登[1,a)]　松田 一起[1,b)]　和田 唯我[1,c)]　杉浦 孔明[1,d)]

概要

画像キャプション生成モデルの開発には高品質な自動評価尺度が必要である．本研究では，マルチモーダル LLM による画像説明を用いた対照学習と，局所的な画像特徴量抽出を導入した自動評価尺度 Fawaris を提案する．また，提案手法を訓練するため，本タスクにおいて最大規模の Cygnus データセットを構築した．実験の結果，人間による評価との相関係数において，提案手法は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，視覚障害者の補助や alt-text の自動生成など，様々な研究と社会応用が行われている [8, 3, 9, 33, 35, 22]．画像キャプション生成モデルの開発には，人間による評価に近い自動評価尺度が不可欠である．

一方，古典的な自動評価尺度 [19, 5, 27, 28] は，人間による評価との相関が著しく低いという問題がある [10, 23, 29, 24]．そのため，画像と言語のエンコーダを用いたデータ駆動型の自動評価尺度 [10, 23, 29, 38] が提案されたが，未だ人間による評価との相関が十分に高いとは言えない．実際に，Flickr8K-Expert データセット [11] における，人間による評価間の相関係数が 0.73 であるのに対し [4]，人間による評価と既存尺度 [23, 29, 38] の相関係数は 0.57 程度である．

データ駆動型の自動評価尺度は，教師なし自動評価尺度と教師あり自動評価尺度に大別される．教師あり自動評価尺度は，教師なし自動評価尺度に比べて，人間による評価との相関が高いことが知られている．既存の教師あり自動評価尺度 [29, 38] では，画像と言語の接地を行うエンコーダに，alt-text で学習された CLIP[20] を使用しており，画像キャプション生成モデルが出力する生成文の特徴量抽出に適していない可能性がある．また，画像の局所領域を考慮する機構が導入されていないため，生成文が参照文群に含まれない要素を説明する例において，不当に低い評価を出力する問題がある [38].

そこで，本研究では，対照学習と階層的特徴抽出に基づく教師あり自動評価尺度 Fawaris を提案する．提案手法は，CLIP-DescriptionAlign (CDA) を導入することにより，短文である alt-text を用いて学習された CLIP エンコーダが，本タスクに適した特徴量抽出をすることが期待される．また，Hierarchical Image Encoder (HIE) の導入により，自動評価尺度が画像の局所特徴を考慮した評価を行うことが期待される．さらに，本タスクにおける最大規模の Nebula データセット [38] は，人間による評価に分布の不均衡が存在するため，本研究では Cygnus データセットを提案する．教師あり学習における学習に用いるデータ分布の不均衡は，モデルの汎化性能低下や，学習を不安定化させることが知られている．そのため，Cygnus データセットの構築に際し，本研究では人間による評価の偏りを軽減するよう，Nebula データセットを約 2.3 倍に拡張した．

提案手法における新規性は次の通りである．

- CLIP エンコーダに対してマルチモーダル LLM による画像説明を用いた対照学習を行う CDA を導入する．
- 大域的な画像特徴量に加え，Grounded-SAM[21] と AlphaCLIP[26] を用いて局所的な画像特徴量の抽出を可能とした HIE を導入する．

2. 問題設定

本研究では画像キャプション生成に対する自動評価を扱う．画像キャプション生成における自動評価尺度は，生成文に対する値が人間による評価と近いことが望ましい．具体的には，画像 $\boldsymbol{x}_{\text{img}}$，生成文 $\boldsymbol{x}_{\text{cand}}$，および N 個の参照文 $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ が入力として与えられ，自動評価尺度は $\boldsymbol{x}_{\text{cand}}$ が $\boldsymbol{x}_{\text{img}}$ と $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ に対してどの程度適切であるかの値 $\hat{y}$ を出力する．

図 1 に本タスクにおける代表例を示す．図の例に対して生成文 “A baby sleeping with a stuffed teddy bear, their faces close together.”，および参照文群 “A baby wrapped

[1] 慶應義塾大学
[a)] tkento1985@keio.jp
[b)] k2matsuda0@keio.jp
[c)] yuiga@keio.jp
[d)] komei.sugiura@keio.jp

図 1 入力画像の代表例

in a blanket facing a small teddy bear.”, “A baby is sleeping next to a teddy bear” が与えられるとき，自動評価尺度は画像，生成文，参照文群から，人間による評価と近い値を出力することが期待される．この例では，生成文が流暢かつ詳細に，関連性を考慮した文であるため，自動評価尺度は高い値を出力することが望ましい．

3. 提案手法

本研究では，DENEB[38] を拡張した，画像キャプション生成における教師あり自動評価尺度 Fawaris を提案する．提案手法における新規性は次の通りである．

- CLIP エンコーダ [20] に対してマルチモーダル LLM による画像説明を用いた対照学習を行う CLIP-DescriptionAlign (CDA) を導入する．
- 大域的な画像特徴量に加え，Grounded-SAM[21] と AlphaCLIP[26] を用いて局所的な画像特徴量の抽出を可能とした，Hierarchical Image Encoder (HIE) を導入する．

ここで，画像の局所領域における特徴を抽出する AlphaCLIP 特徴量の導入は，CLIP-S[10] や PAC-S[23] をはじめとする，マルチモーダル特徴量を扱う自動評価尺度に広く適用可能であると考えられる．

図 2 に提案手法 Fawaris のモデル図を示す．提案手法は，図に示す通り，HIE および Sim-Vec Transformer モジュール [38] の 2 つのモジュールで構成される．提案手法は以下に示す入力 $\boldsymbol{x}$ から，生成文に対する評価 $\hat{y}$ を出力する．

$$\boldsymbol{x} = \left\{ \boldsymbol{x}_{\text{img}}, \left\{ \boldsymbol{x}_{\text{ref}}^{(i)} \right\}_{i=1}^{N}, \boldsymbol{x}_{\text{cand}} \right\}$$

ここで，$\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times H\times W}$，$\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N} \in \{0,1\}^{N\times V\times L}$，$\boldsymbol{x}_{\text{cand}} \in \{0,1\}^{V\times L}$ はそれぞれ画像，N 個の参照文群，生成文を表す．また，H，W，N，V，L はそれぞれ画像の高さと幅，参照文の数，語彙サイズ，トークン数を表す．

3.1 CLIP-DescriptionAlign

CLIP エンコーダは短文である alt-text を用いて学習されているため，画像キャプション生成モデルが出力する生成文の特徴量抽出に適していない可能性がある．そのため本研究では，CLIP エンコーダに対してマルチモーダル LLM による画像説明を用いた対照学習を適用した CLIP-DescriptionAlign (CDA) を導入する．

はじめに，$\boldsymbol{x}_{\text{img}}$ に対し，マルチモーダル LLM を用いて画像説明 $\boldsymbol{x}_{\text{desc}}$ を生成する．このとき，マルチモーダル LLM には LLaVA-NeXT[16] を使用し，入力するプロンプトには “Create a short caption of about 20 words for this image, noting the relationship between the objects in the image.” を用いる．次に，$\boldsymbol{x}_{\text{desc}}$，$\boldsymbol{x}_{\text{img}}$ および $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ を用いて対照学習を行う．本研究では，$\boldsymbol{x}_{\text{img}}$ と $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ 間，および $\boldsymbol{x}_{\text{img}}$ と $\boldsymbol{x}_{\text{desc}}$ 間の InfoNCE[18] 損失をそれぞれ計算し，その平均値を最終的な損失とする．

$\boldsymbol{x}_{\text{img}}$ に対する $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ の InfoNCE 損失は，アンカーを $\boldsymbol{x}_{\text{img}}$ とし，対応する $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ を正例，バッチ内に存在する正例以外の $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ を負例として計算する．$\boldsymbol{x}_{\text{img}}$ に対する $\boldsymbol{x}_{\text{desc}}$ の InfoNCE 損失も同様に，アンカーを $\boldsymbol{x}_{\text{img}}$ とし，対応する $\boldsymbol{x}_{\text{desc}}$ を正例，バッチ内に存在する正例以外の $\boldsymbol{x}_{\text{desc}}$ を負例として計算する．

3.2 Hierarchical Image Encoder

画像キャプション生成の自動評価では，画像と言語の関係を詳細に捉えるために，大域的な特徴量と局所的な特徴量の両方を抽出することが重要である．先行研究 [38] では，画像に対する特徴量抽出に CLIP エンコーダのみを用いており，局所的な画像特徴が十分に考慮されていない．そのため，大域的な画像特徴量に加え，Grounded-SAM と AlphaCLIP を用いて局所的な画像特徴量抽出を可能とした Hierarchical Image Encoder (HIE) を導入する．

大域的な画像特徴量抽出では，CDA を適用した CLIP 画像エンコーダを使用することで，本タスクに適した画像特徴量抽出を行う．具体的には，CLIP 画像エンコーダを用いて，$\boldsymbol{x}_{\text{img}}$ から画像特徴量 $\boldsymbol{v}_{\text{clip}} \in \mathbb{R}^{d_{\text{clip}}}$ を取得する．また，局所的な画像特徴量抽出では，はじめに $\boldsymbol{x}_{\text{cand}}$ と $\boldsymbol{x}_{\text{img}}$ を Grounded-SAM に入力し，$\boldsymbol{x}_{\text{cand}}$ と関連する領域のマスク画像 $\boldsymbol{x}_{\text{mask}}$ を取得する．ここで，Grounded-SAM では $\boldsymbol{x}_{\text{img}}$ に対するプロンプトを $\boldsymbol{x}_{\text{cand}}$ とし，$\boldsymbol{x}_{\text{cand}}$ に存在する単語との類似度が 0.25 以上の領域にマスクを生成する．その後，$\boldsymbol{x}_{\text{img}}$ および $\boldsymbol{x}_{\text{mask}}$ を AlphaCLIP に入力し，画像の局所特徴量 $\boldsymbol{v}_{\text{a-clip}}$ を抽出する．

3.3 Sim-Vec Transformer

画像キャプション生成の自動評価尺度では，画像，生成文および参照文群間の類似度を適切に捉えることが重要である．そのため，本研究では先行研究 [38] に従い，画像キャプション生成における画像，生成文，および参照文群間の類似度を正確に捉えるために Sim-Vec Transformer モジュールを導入する．

Sim-Vec Transformer モジュールでは，入力 $\boldsymbol{s}_{\text{in}}$ から生

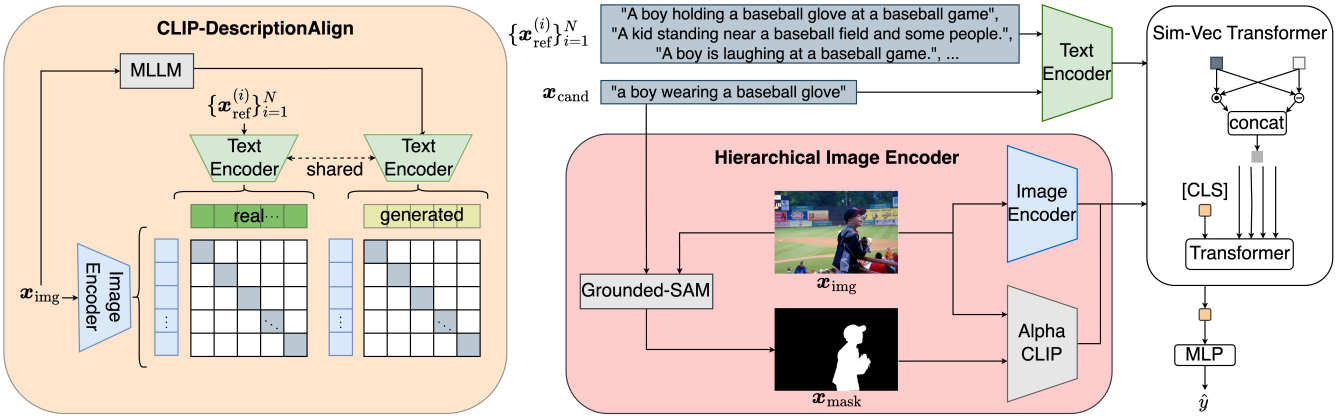


図 2 Fawaris のモデル図

成文の自動評価に有益な特徴量を抽出する．

$$\boldsymbol{s}_{\text{in}} = \left\{ \boldsymbol{c}_{\text{clip}}, \boldsymbol{r}_{\text{clip}}^{(i)}, \boldsymbol{c}_{\text{rb}}, \boldsymbol{r}_{\text{rb}}^{(i)}, \boldsymbol{v}_{\text{clip}}, \boldsymbol{v}_{\text{a-clip}} \right\}$$

ここで，$\boldsymbol{r}_{\text{clip}}^{(i)} \in \mathbb{R}^{N \times d_{\text{clip}}}$ および $\boldsymbol{c}_{\text{clip}} \in \mathbb{R}^{d_{\text{clip}}}$ は，それぞれ $\boldsymbol{x}_{\text{ref}}^{(i)}$ と $\boldsymbol{x}_{\text{cand}}$ から CLIP テキストエンコーダによって抽出した言語特徴量であり，$\boldsymbol{r}_{\text{rb}}^{(i)} \in \mathbb{R}^{N \times d_{\text{rb}}}$ および $\boldsymbol{c}_{\text{rb}} \in \mathbb{R}^{d_{\text{rb}}}$ は，それぞれ $\boldsymbol{x}_{\text{ref}}^{(i)}$, $\boldsymbol{x}_{\text{cand}}$ から RoBERTa[17] を使用して抽出した言語特徴量である．

はじめにベクトル間の類似度を捉えるため，$\boldsymbol{s}_{\text{in}}$ から次に示す特徴量 $\boldsymbol{g}_{\text{inter}}^{(i)}$ を得る．

$$\boldsymbol{g}_{\text{inter}}^{(i)} = \left\{ F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{v}), F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{r}_{\text{clip}}^{(i)}), F(\boldsymbol{c}_{\text{rb}}, \boldsymbol{r}_{\text{rb}}^{(i)}) \right\}$$

ここで，$\boldsymbol{v} = \{\boldsymbol{v}_{\text{clip}}, \boldsymbol{v}_{\text{a-clip}}\}$ であり，$F(\boldsymbol{c}, \boldsymbol{r})$ は [38] と同様に，アダマール積と要素間の差分を用いた $F(\boldsymbol{c}, \boldsymbol{r}) = [\boldsymbol{c} \odot \boldsymbol{r}; |\boldsymbol{c} - \boldsymbol{r}|]$ を指す．次いで $\boldsymbol{g}_{\text{inter}}$ に `[CLS]` トークン $\boldsymbol{g}_{[\text{CLS}]}$ を結合し，$\boldsymbol{g} = \left[\boldsymbol{g}_{[\text{CLS}]}; \boldsymbol{g}_{\text{inter}}\right]$ を得る．その後 $\boldsymbol{g}$ を N 層の transformer エンコーダに入力し $\boldsymbol{h}_{\text{N}}$ を得る．ここで，本研究では N=3 とする．続いて，$\boldsymbol{h}_N$ から `[CLS]` トークンに対応する特徴量 $\boldsymbol{h}_{[\text{CLS}]}$ を取得し，MLP とシグモイド関数を用いて最終的な評価 $\hat{y}$ を算出する．提案手法における損失関数には，外れ値に頑健な Huber 損失を用いる．

4. 実験

4.1 実験設定

画像キャプション生成に対する教師あり自動評価尺度の構築には，大規模なデータセットが必要である．Nebula データセット [38] は，本分野における最大規模のデータセットであるが，人間による評価の分布が不均衡である．具体的には，$y = 0$ であるサンプルは $y = 0.25$ のサンプルと比較して約 11 倍も多い．教師あり学習において，学習データの分布が不均衡である場合，モデルの汎化性能低下や，学習を不安定化させることが知られている．

そこで，本研究では人間による評価の偏りを軽減した Cygnus データセットを構築した．Cygnus データセットは，57,808 枚の画像と 1,085 人のアノテータから収集された 68,148 個の人間による評価により構成されており，Nebula データセットの約 1.8 倍の画像を含む．Cygnus データセットでは，Nebula データセットで使用された標準的な画像キャプション生成モデル [32, 7, 36, 25, 14, 30, 31, 13] に加えて，画像キャプション生成において高い性能を示す GPT-4V[1] および LLaVA-NeXT[16] を用いた．

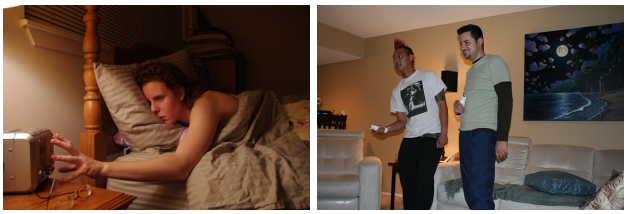
	Composite	Flickr8K (Expert)	Flickr8K (CF)	Nebula	FOIL 1-ref	FOIL 4-ref
Rule-based metrics						
BLEU [19]	30.6	30.8	16.4	40.4	66.5	82.6
ROUGE [15]	32.4	32.3	19.9	42.6	71.7	79.3
CIDEr [27]	37.7	43.9	24.6	48.1	82.5	90.6
METEOR [5]	38.9	41.8	22.2	46.8	78.8	82.6
SPICE [4]	40.3	44.9	24.4	44.0	75.5	86.1
Unsupervised metrics						
BERTScore [37]	30.1	46.7	22.8	47.0	88.6	92.1
BARTScore [34]	43.5	37.8	24.3	43.8	85.3	91.1
CLIP-S [10]	53.8	51.2	34.4	46.9	87.2	87.2
RefCLIP-S [10]	55.4	53.0	36.4	46.9	91.0	92.6
CLAIR [6]	55.0	44.6	34.4	52.4	81.4	83.4
Supervised metrics						
PAC-S [23]	55.7	54.3	36.0	47.2	89.9	89.9
RefPAC-S [23]	57.3	50.6	37.3	50.6	93.7	94.9
Polos [29]	57.6	56.4	37.8	53.9	93.2	95.1
DENEB [38]	58.2	56.8	**38.3**	**54.3**	**95.4**	**96.5**
Ours	**58.6**	**58.1**	38.0	52.7	95.1	95.8

表 1 ベースライン手法との定量的比較結果

4.2 実験結果

本研究では，Composite[2]，Flickr8K-Expert[11]，Flickr8K-CF[11]，Nebula[38]， FOIL[24] を用いて定量的比較を行った．表 1 にベースライン手法との定量的比較結果を示す．先行研究 [10, 12, 4, 29] と同様に，Flickr8K-CF における評価には τ_b (Kendall-B) を用い，Composite，Flickr8K-CF，Nebula の評価には τ_c (Kendall-C) を使用した．提案手法の人間による評価との相関係数は，Composite，Flickr8K-Expert において 59.6，58.1 であり，既存手法と比較してそれぞれ 0.4，1.3 ポイント上回った．また，FOIL で 1 文および 4 文の参照文が与えられる場合において，それぞれの精度は 95.1%，95.8%であり，既存手法と同程度の性能であった．

図 3 に Composite データセット内における提案手法の



(a) (b)

図 3 Composite データセットにおける成功例

Metric	HIE	CDA	Nebula	Flickr8K (Expert)	Flickr8K (CF)	Composite	FOIL 1-ref	FOIL 4-ref
(i)	✓		50.8	58.0	37.4	55.4	**95.2**	95.3
(ii)		✓	**52.9**	56.8	37.9	58.0	94.9	**96.0**
(iii)	✓	✓	52.7	**58.1**	**38.0**	**58.6**	95.1	95.8

表 2 ablation studies の定量的結果

成功例を示す．図 3-(a) の例では，$\boldsymbol{x}_{\text{cand}}$ が “a man sitting on a couch with a laptop and a laptop .” であり，$\boldsymbol{x}_{\text{ref}}^{(1)}$ は “A woman laying in her bed touching the alarm clock” であった．また，この例では $\boldsymbol{x}_{\text{cand}}$ に ‘a laptop’ や ‘a man’, ‘sitting’ といった誤りを含むため，人間による評価 y が 0 であった．この例に対して，既存手法である RefPAC-S[23] が 0.704 と誤った値を出力したのに対して，提案手法は 0.060 と y に近い値を出力した．次に，図 3-(b) が示す例では，$\boldsymbol{x}_{\text{cand}}$ は “a couple of young men standing in a living room holding wii controllers.” であり，$\boldsymbol{x}_{\text{ref}}^{(1)}$ は “two men playing with wii motes by a couch” であった．本サンプルでは $\boldsymbol{x}_{\text{cand}}$ が画像を流暢かつ詳細に説明しているため，人間による評価 y は 1.0 であった．このサンプルにおいて，既存手法である RefPAC-S は 0.898 という値を出力したのに対して提案手法は 0.915 と評価した．したがって，提案手法である Fawaris は，既存手法に比べ人間による評価と近い評価値を出力したと言える．

4.3 Ablation Studies

各モジュールの有効性を調査するために ablation studies を行った．表 2 に ablation studies の結果を示す．

CLIP-DescriptionAlign ablation

CDA を適用した CLIP エンコーダ [20] を, 適用前の CLIP エンコーダに置き換えることで，CDA の性能への寄与を調査した．表 2 より，Flickr8K-Expert および Composite における Metric(i) の相関係数は，それぞれ 58.0，55.4 であり，提案手法に比べ 0.1，3.2 ポイント減少している．したがって，CDA が画像キャプション生成の自動評価において有効であることが示唆される．

Hierarchical Image Encoder ablation

画像特徴量の抽出から HIE を取り除くことで，HIE の性能への寄与を調査した．表 2 より，Flickr8K-Expert および Composite における Metric(ii) の相関係数は，それぞれ 56.8, 58.0 であり，提案手法に比べ 1.3, 0.6 ポイント減少している．したがって，HIE が画像キャプション生成の自動評価において有効な特徴量を抽出することが示唆された．

5. おわりに

本研究では，画像キャプション生成の自動評価尺度を構築した．提案手法における貢献は次の通りである．

- 画像キャプション生成における，対照学習と階層的特徴抽出に基づく教師あり自動評価尺度 Fawaris を提案した．
- CLIP エンコーダ [20] に対してマルチモーダル LLM による画像説明を用いた対照学習を行う CLIP-DescriptionAlign (CDA) を導入した．
- 大域的な画像特徴量に加え，Grounded-SAM[21] と AlphaCLIP[26] を用いて局所的な画像特徴量の抽出を行う Hierarchical Image Encoder (HIE) を導入した．
- 57,808 枚の画像と 1,085 人のアノテータから収集し，人間による評価の偏りを軽減した Cygnus データセットを構築した．
- 提案手法である Fawaris は，Composite[2]，Flickr8K-Expert[11] データセットにおいて，既存手法を上回る結果を得た．

謝辞

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] Achiam, J., Adler, S., Agarwal, S. et al.: GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774* (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] Ahsan, H., Bhatt, D., Shah, K. and Bhalla, N.: Multi-Modal Image Captioning for the Visually Impaired, *NAACL-HLT*, pp. 53–60 (2021).

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

[5] Banerjee, S. et al.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *ACL*, pp. 65–72 (2005).

[6] Chan, D. M. et al.: CLAIR: Evaluating Image Captions with Large Language Models, *EMNLP* (2023).

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

[8] Dognin, P. et al.: Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge,

Vol. 73, pp. 437–459 (2022).

[9] Ghandi, T., Pourreza, H. and Mahyar, H.: Deep Learning Approaches on Image Captioning: A Review, *ACM Computing Surveys*, Vol. 56, No. 3 (2023).

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

[11] Hodosh, M. et al.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, *JAIR*, Vol. 47, pp. 853–899 (2013).

[12] Kim, J. et al.: Mutual Information Divergence: A Unified Metric for Multimodal Generative Models, *NeurIPS*, Vol. 35, pp. 35072–35086 (2022).

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

[14] Li, J., Li, D., Xiong, C. and Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, *ICML* (2022).

[15] Lin, C.: ROUGE: A Package For Automatic Evaluation Of Summaries, *ACL*, pp. 74–81 (2004).

[16] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S. and Lee, J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (2024).

[17] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M. et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint arXiv:1907.11692* (2019).

[18] Oord, A., Li, Y. and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018).

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

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

[21] Ren, T., Liu, S., Zeng, A. et al.: Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, *arXiv preprint arXiv:2401.14159* (2024).

[22] Rotstein, N., Bensaïd, D., Brody, S. et al.: FuseCap: Leveraging large language models for enriched fused image captions, *WACV*, pp. 5689–5700 (2024).

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

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

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

[26] Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D. and Wang, J.: Alpha-CLIP: A CLIP Model Focusing on Wherever You Want, *CVPR* (2024).

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

[28] Wada, Y., Kaneda, K. et al.: JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models, *CoNLL*, pp. 424–435 (2023).

[29] Wada, Y., Kaneda, K., Saito, D. and Sugiura, K.: Polos: Multimodal Metric Learning from Human Feedback for Image Captioning, *CVPR* (2024).

[30] Wang, J. et al.: GIT: A Generative Image-to-text Transformer for Vision and Language, *TMLR* (2022).

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

[32] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. et al.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, *ICML*, pp. 2048–2057 (2015).

[33] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T. and Chen, E.: A Survey on Multimodal Large Language Models, *arXiv preprint arXiv:2306.13549* (2023).

[34] Yuan, W., Neubig, G. and Liu, P.: BARTScore: Evaluating Generated Text as Text Generation, *NeurIPS*, Vol. 34, pp. 27263–27277 (2021).

[35] Yunhao, G., Xiaohui, Z., Jacob, H. et al.: Visual Fact Checker: Enabling High-Fidelity Detailed Caption Generation, *CVPR* (2024).

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

[37] Zhang, T., Wu, F., Weinberger, K. et al.: BERTScore: Evaluating Text Generation with BERT, *ICLR* (2020).

[38] 松田一起，和田唯我，杉浦孔明：ハルシネーションに頑健な画像キャプション生成の自動評価，第 38 回 人工知能学会全国大会 (2024).