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

A Hallucination-Resistant Automatic Evaluation Metric for Image Captioning

松田 一起　和田 唯我　杉浦 孔明
Kazuki Matsuda　Yuiga Wada　Komei Sugiura

慶應義塾大学
Keio University

In the field of image captioning, constructing automatic evaluation metrics that align closely with human judgment is crucial for effective model development. A key challenge in this field is addressing hallucinations, which are instances where models generate words unrelated to the image, a frequent issue in image captioning. Existing metrics often fail to manage hallucinations, primarily due to their limited capability in contrasting candidate captions against a diverse range of reference captions. To overcome this, we propose DENEB, a novel metric for image captioning, specifically robust to hallucinations. DENEB incorporates the Sim-Vec Transformer, a mechanism capable of processing multiple references and extracting similarity vectors effectively. Additionally, to train DENEB, we have expanded the Polaris dataset to create Polaris2.0, significantly enhancing supervised automatic evaluation metrics. Our dataset comprises 32,978 images and 32,978 human judgments from 805 annotators. Our approach achieved state-of-the-art performance on Composite, Flickr8K-Expert, Flickr8K-CF, PASCAL-50S, FOIL, and the Polaris 2.0 dataset, thereby demonstrating its effectiveness and robustness to hallucinations.

1. はじめに

画像キャプション生成は，視覚障害者の補助，医療画像解析，ロボティクスにおける説明生成など様々な社会応用がなされており，幅広く研究が行われている [Dognin 22, Ghandi 23]．画像キャプション生成の社会応用に際して，安全性の観点から画像内に存在しない単語群を出力しないことが望ましい．画像内に存在しない単語を出力する現象はハルシネーションと呼ばれ，画像キャプション生成モデルにおいては度々発生することが知られている [Shekhar 17]．ハルシネーションへの対処は社会応用において重要な課題であるにもかかわらず，本分野における既存の自動評価尺度はハルシネーションに頑健でない．実際に，データ駆動型自動評価尺度 [Hessel 21, Sarto 23] は，人間による評価との相関が高い一方で，ハルシネーションに十分に対処できないことが実験により示されている [Vedantam 15, Hessel 21, Sarto 23, 和田 24]．したがって，本分野では人間による評価との相関が高く，ハルシネーションに頑健な自動評価尺度が望まれる．

画像キャプション生成の自動評価尺度は classic metrics, reference-free metrics, pseudo-multifaceted metrics, multi-faceted metrics の 4 種に大別できる．Classic metrics [Papineni 02, Vedantam 15, Anderson 16, Wada 23] は，n-gram やシーングラフに基づくルールベースの古典的な自動評価尺度である．これらの尺度は人間による評価との相関係数が著しく低いことから，データ駆動型自動評価尺度として reference-free metrics や pseudo-multifaceted metrics が提案された [Hessel 21, Sarto 23, Lee 21, 和田 24]．Reference-free metrics [Hessel 21, Sarto 23, Lee 21] は画像を入力に用いた参照文群を要さない自動評価尺度である．これらの尺度は classic metrics に比べ，人間による評価との高い相関を得ている．しかし，これらの性能は，画像と言語間の接地性能に大きく依存しており，特に画像の局所領域に対する接地性能が不十分であるため，ハルシネーションに頑健でない．実際に，CLIP [Radford 21] を用いた reference-free metrics である CLIP-S は，classic metrics である CIDEr と比較して，ハルシネーションに対処する性能が低いことが知られている [Hessel 21]．Pseudo-multifaceted metrics は，画像エンコーダおよびテキストエンコーダを用いて，画像と複数の参照文を扱う自動評価尺度である [Hessel 21, Sarto 23, 和田 24]．これらの尺度は人間による評価との相関が reference-free metrics よりも高い一方で，複数の参照文を十分効果的に活用していないという問題がある．具体的には，参照文それぞれに対して独立に評価値を計算するため，事実上一つの参照文しか扱っておらず，性能が不十分である．

本研究では，ハルシネーションに頑健な教師あり自動評価尺度 DENEB を提案する．提案尺度は，pseudo-multifaceted metrics と異なり，画像の多角的な説明と候補文を比較可能な構造の導入を行うことで，複数の参照文を十分に活用することができる multifaceted metrics の一つである．具体的には，複数の参照文を効果的に扱うために Sim-Vec Transformer を導入する．また，DENEB を学習するために，画像，参照文群，生成文と対応する人間による評価で構築されたデータセット Polaris2.0 を提案する．本データセットは，本分野において最大のサンプル数を有する Polaris データセット [和田 24] を拡張したものである．Polaris データセットには，画像のバリエーションが人間による評価に比べ極端に少ないという問題点がある．具体的には，画像の総数が全サンプル数の十分の一程度しか存在しない．そこで本研究では，このバリエーションの不均衡を解消するため，約 20,000 枚の画像を追加することで Polaris データセットを拡張し，Polaris2.0 データセットを構築した．

提案手法における貢献は次の通りである．

- ハルシネーションに頑健な画像キャプション生成における自動評価尺度 DENEB を提案する．
- 画像，生成文および参照文群間の類似度を扱う Sim-Vec Transformer モジュールを導入する．
- アダマール積と差分を用いて自動評価に有用な特徴量を抽出する Sim-Vec Extraction (SVE) を導入する．
- 画像のバリエーションがサンプル数に対し極端に少ない Polaris データセットを拡張し，画像を約 20,000 枚追加した新たなデータセット Polaris2.0 を構築する．
- FOIL, Composite, Flickr8K-Expert, Flickr8K-CF, Polaris2.0, PASCAL-50S において提案手法が既存手法を上回る結果を得た．

連絡先: 松田一起，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，k2matsuda0@keio.jp

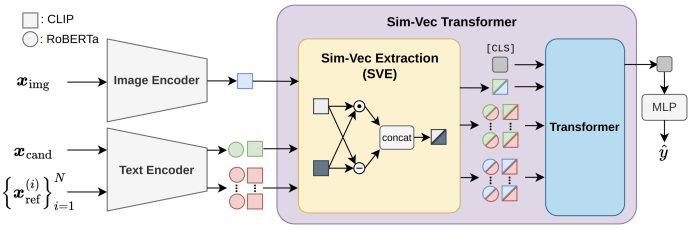

図 1: DENEB のモデル構造

2. 問題設定

本研究では，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尺度を扱う．画像キャプション生成における自動評価尺度は，人間による評価とより近いことが望ましい．特に，画像キャプション生成モデルによるハルシネーションに頑健であることが望ましい．

本タスクでは，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，$\boldsymbol{x}_{\text{cand}}$ 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評価値 $\hat{y}$ を出力する．

画像キャプション生成の自動評価尺度として，参照文群を要さない reference-free metrics が提案されている [Sarto 23, Hessel 21, Lee 21]．しかし，これらの尺度の多くは，参照文群を用いる尺度に比べ人間による評価との相関が低い．また，[和田 24] において指摘されているように，reference-free metrics の性能は，画像と言語における接地の性能に大きく依存している．また，ハルシネーションを含む生成文が与えられたときに自動評価尺度が適切な評価値を出力するためには，人間による正しいキャプションである参照文の活用が不可欠である．実際に，CLIP-S や PAC-S は，n-gram に基づく CIDEr [Vedantam 15] と比較して，ハルシネーションに十分に対処できないことが実験により明らかになっている [Sarto 23, Hessel 21]．したがって，本研究では参照文群を用いる自動評価を扱うことを前提とする．本研究では評価尺度として，人間による評価との相関係数に Kendall's τ を，ハルシネーションへの頑健性の評価にハルシネーションの検出タスクにおける精度を用いた．

3. 提案手法

本研究では，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尺度である DENEB を提案する．提案手法は，Polos [和田 24] をはじめとする学習可能な自動評価尺度と関連が深い．提案手法における，画像の多角的な説明と候補文を比較可能な構造の導入は，CLIPScore [Hessel 21] や PACScore [Sarto 23] をはじめとする画像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．提案手法における新規性は次の通りである．

- 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 DENEB を提案する．
- 画像，生成文および参照文群間の類似度を扱う Sim-Vec Transformer を導入する．
- アダマール積と差分を用いて自動評価に有用な特徴量を抽出する Sim-Vec Extractor (SVE) を導入する．

図 1 に提案尺度 DENEB の全体図を示す．提案手法における Sim-Vec Transformer は Sim-Vec Extraction および Transformer の 2 つのモジュールから構成される．

3.1 入力および埋め込み表現

はじめに，入力 x を以下のように定義する．

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

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，語彙サイズおよびトークン数を表す．

提案手法では，[和田 24] と同様に，$\boldsymbol{x}_{\text{img}}$ から CLIP [Radford 21] および RoBERTa [Liu 19] を用いて特徴量を抽出する．本研究では，CLIP の事前学習に [Sarto 23] を，RoBERTa の事前学習に SimCSE [Gao 21] を用いた．はじめに，CLIP のテキストエンコーダを用いて $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$，$\boldsymbol{x}_{\text{cand}}$ からそれぞれ文埋め込み $\{\boldsymbol{r}_{\text{clip}}^{(i)}\}_{i=1}^{N} \in \mathbb{R}^{N\times d_{\text{clip}}}$，$\boldsymbol{c}_{\text{clip}} \in \mathbb{R}^{d_{\text{clip}}}$ を得る．ここで，d_{clip} は CLIP の出力次元を表す．続いて，CLIP の画像エンコーダを用いて $\boldsymbol{x}_{\text{img}}$ から画像特徴量 $\boldsymbol{v} \in \mathbb{R}^{d_{\text{clip}}}$ を得る．また，RoBERTa を用いて $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$，$\boldsymbol{x}_{\text{cand}}$ からそれぞれ文埋め込み $\{\boldsymbol{r}_{\text{rb}}^{(i)}\}_{i=1}^{N} \in \mathbb{R}^{N\times d_{\text{rb}}}$，$\boldsymbol{c}_{\text{rb}} \in \mathbb{R}^{d_{\text{rb}}}$ を得る．ここで，d_{rb} は RoBERTa の出力次元を表し，文埋め込みは RoBERTa の出力した [CLS] トークンから得た．

3.2 Sim-Vec Extraction モジュール

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，ベクトル間の類似度を捉えた特徴量を抽出する Sim-Vec Extraction (SVE) を導入する．SVE では Parallel Feature Extraction [和田 24] に基づき，アダマール積と差分を用いることで，$\boldsymbol{x}_{\text{img}}, \{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}, \boldsymbol{x}_{\text{cand}}$ 間の類似度を捉えた特徴量を抽出する．

はじめに，次に示す入力から，

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

$\boldsymbol{h}_{\text{clip}}$ と $\boldsymbol{h}_{\text{rb}}$ を次式のように計算する．

$$\boldsymbol{h}_{\text{clip}} = \left[F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{v}); F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{r}_{\text{clip}}^{(1)}); \ldots; F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{r}_{\text{clip}}^{(N)})\right]$$

$$\boldsymbol{h}_{\text{rb}} = \left[F(\boldsymbol{c}_{\text{rb}}, \boldsymbol{r}_{\text{rb}}^{(1)}); \ldots; F(\boldsymbol{c}_{\text{rb}}, \boldsymbol{r}_{\text{rb}}^{(N)})\right]$$

ここで，F はアダマール積 $\odot$ およびベクトルの要素間の差分を用いて $F(\boldsymbol{c}, \boldsymbol{r}) = [|\boldsymbol{c} - \boldsymbol{r}|;\ \boldsymbol{c} \odot \boldsymbol{r}]$ と定義される．ただし，本研究では [和田 24] と異なり，$F(\boldsymbol{c}, \boldsymbol{r})$ において特徴量 c, r の結合は行わない．続いて，$\boldsymbol{h}_{\text{clip}}$ と $\boldsymbol{h}_{\text{rb}}$ を結合することで $\boldsymbol{h}_{\text{inter}} = [\boldsymbol{h}_{\text{clip}}; \boldsymbol{h}_{\text{rb}}]$ を得る．

3.3 Transformer モジュール

先行研究 [和田 24, Sarto 23, Hessel 21] では，N 文の参照文それぞれに対して独立に評価値を計算するため，複数の参照文を十分に活用していないという問題がある．一般に CLIP-S [Hessel 21]，PAC-S [Sarto 23]，および Polos [和田 24] は，

$$\hat{y} = \text{Aggregate}_i f(\boldsymbol{x}^{(i)})$$

と表現することが出来る．ここで，$\boldsymbol{x}^{(i)}$ は i 番目の参照文を含む $\boldsymbol{x}^{(i)} = \{\boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{ref}}^{(i)}, \boldsymbol{x}_{\text{cand}}\}$ を指し，Aggregate$(\cdot)$ は Max 関数をはじめとする任意の写像 $f : \mathbb{R}^N \to \mathbb{R}$ を示す．これらの手法は，$f(\cdot)$ において一文の参照文 $\boldsymbol{x}_{\text{ref}}^{(i)}$ のみを入力として受け取るため，$f(\boldsymbol{x}_{\text{ref}}^{(i)})$ は全ての参照文を効果的に扱っているとはいえない．また，Aggregate 関数は学習可能なパラメタを含まないため，Aggregate 関数は最適化されない．したがって，これらの手法は画像キャプション生成における自動評価において，複数の参照文を十分に活用していないといえる．

そこで提案手法では，自動評価において複数の参照文を効果的に扱うため，N 層の Transformer を用いる．はじめに，SVE から得られた $\boldsymbol{h}_{\text{inter}}$ に [CLS] トークン $\boldsymbol{h}_{[\text{CLS}]}$ を結合することで $\boldsymbol{h} = \left[\boldsymbol{h}_{[\text{CLS}]}; \boldsymbol{h}_{\text{inter}}\right]$ を得る．続いて，N 層の TransformerEncoder を用いて，次式のように $\boldsymbol{g}_{\text{N}}$ を計算する．なお，$\boldsymbol{h}$ には positional encoding を付与する．

$$\boldsymbol{g}_{\text{N}} = \text{TransformerEncoder}\,(\boldsymbol{h})$$

	FOIL 1-ref	FOIL 4-ref	Composite	Flickr8K (Expert)	Flickr8K (CF)	Polaris2.0
Classic metrics						
BLEU	66.5	82.6	30.6	30.8	16.4	40.4
CIDEr	82.5	90.6	37.7	43.9	24.6	48.1
SPICE	75.5	86.1	40.3	44.9	24.4	44.0
Reference-free metrics						
UMIC	–	–	56.1	46.8	30.1	–
CLIP-S	87.2	87.2	53.8	51.2	34.4	46.9
PAC-S	89.9	89.9	55.7	54.3	36.0	47.2
Pseudo-multifaceted metrics						
RefCLIP-S	91.0	92.6	55.4	53.0	36.4	46.9
RefPAC-S	93.7	94.9	57.3	50.6	37.6	50.6
CLAIR	81.4	83.4	55.0	44.6	34.4	52.7
Polos	93.2	95.1	57.6	56.4	37.8	53.9
Multi-faceted metrics						
DENEB (Ours)	**95.4** (+1.7)	**96.5** (+1.4)	**58.2** (+0.6)	**56.8** (+0.4)	**38.3** (+0.5)	**54.3** (+0.4)

表 1: ベースライン手法との定量的比較結果

	HC	HI	HM	MM	Mean
Classic metrics					
BLEU	60.4	90.6	84.9	54.7	72.7
SPICE	63.6	96.3	86.7	68.3	78.7
CIDEr	65.1	98.1	90.5	64.8	79.6
Reference-free metrics					
CLIP-S	56.5	99.3	96.4	70.4	80.7
PAC-S	60.6	99.3	96.9	72.9	82.4
UMIC	66.1	**99.8**	**98.1**	76.2	85.1
Pseudo-multifaceted metrics					
TIGEr	56.0	**99.8**	92.8	74.2	80.7
RefCLIP-S	64.5	99.6	95.4	72.8	83.1
RefPAC-S	67.7	99.6	96.0	75.6	84.7
CLAIR-E	57.7	**99.8**	94.6	75.6	81.9
Polos	70.0	99.6	97.4	**79.0**	86.5
Multifaceted metrics					
DENEB (Ours)	**76.1** (+6.1)	99.7	97.4	77.9	**87.8** (+1.3)

表 2: PASCAL-50S における定量的結果

続いて，g における先頭のトークン $\boldsymbol{g}_{[\mathrm{CLS}]}$ を MLP に入力し，シグモイド関数を適用することで，最終的な評価 $\hat{y}$ を算出する．提案手法では，外れ値に頑健である Huber 損失を損失関数に用いた．ここで，本研究では $N = 3$ とした．

4. 実験

4.1 実験設定

本研究では，FOIL [Shekhar 17], Composite [Aditya 15], Flickr8K-Expert [Hodosh 13]，Flickr8K-CF [Hodosh 13], PASCAL-50S [Anderson 16] および Polaris2.0 [和田 24] において実験を行った．本タスクにおいては，自動評価尺度の評価値と人間による評価との相関が高いことが望ましいため，相関係数に基づき自動評価尺度を評価した．本研究では，標準的な相関係数である Kendall's τ を用いた．評価には，本タスクにおいて標準的である人間による評価との相関係数に加え，ハルシネーションへの頑健性を測るベンチマークである FOIL [Shekhar 17] を用いた．

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不可欠である．[和田 24] では，画像キャプション生成の自動評価における大規模データセット Polaris が提案された．Polaris データセットは本分野における最大規模のデータセットであり，既存のデータセット [Aditya 15, Hodosh 13, Lee 21] と比較して約 10 倍のサンプルを含む．しかし，Polaris データセットには画像のバリエーションが生成文のバリエーションに比べ極端に少ないという問題点があった．そこで本研究では，Polaris データセットに約 2 万枚の画像を追加することで拡張を行い，新たに Polaris2.0 データセットを構築した．ここで，Polaris データセット内のサンプルについては，一つの生成文に対して複数付与された人間による評価の平均値を採用した．

(a) (b)

図 2: Polaris2.0 データセットにおける成功例

Polaris2.0 データセットは 805 人のアノテータから収集された 32,978 個の人間による評価により構成されており，32,978 枚の画像を含む．参照文の総数は 183,472，語彙サイズは 32,870，全単語数は 1,945,956，平均文長は 10.61 語である．生成文の総数は 32,978，語彙サイズは 3,695，全単語数は 288,922，平均文長は 8.76 語である．参照文群，生成文はともに英語で記述されている．我々は Polaris2.0 データセットを，訓練集合，検証集合，テスト集合に分割し，それぞれ 26,382，3,298，3,298 サンプルとした．

本研究では，[Papineni 02, Vedantam 15, Anderson 16] といった画像キャプション生成において標準的な自動評価尺度に加え，[Jiang 19, Lee 21, Hessel 21, Sarto 23, 和田 24, Chan 23], をベースライン手法として用いた．

4.2 実験結果

表 1 に，FOIL データセット [Shekhar 17] におけるベースライン手法との定量的比較結果を示す．本実験において，[和田 24, Sarto 23, Hessel 21] と同様に，1 文および 4 文の参照文が与えられる設定で評価を行った．表 1 に，提案手法およびベースライン手法の，FOIL データセットにおける性能を示す．提案手法は，1 文および 4 文の参照文が与えられる設定において，それぞれ 95.1%，96.3%という性能を得ており，既存手法に比べそれぞれ 1.9，1.2 ポイント上回った．また，表 1 は，画像キャプション生成の自動評価において LLM (GPT-3.5) を使用する CLAIR がハルシネーションを含む生成文とそうでない生成文を十分に区別できていないことを示している．これより，提案手法は，LLM に基づくものを含む既存手法より，ハルシネーションに頑健であると言える．

また，表 1 に，Composite, Flickr8K, Flickr8K-CF, Polaris2.0 におけるベースライン手法との定量的比較結果を示す．表中の “—” は，コードないしはデータが未公開であるため評価実験を行うことができなかった場合を指す．表 1 より，提案手法の性能は Composite，Flickr8K-CF，Polaris2.0 データセットにおいてそれぞれ 58.4，38.3，54.3 であり，既存手法と比較してそれぞれ 0.8, 0.5，0.4 ポイント上回った．

PASCAL-50S は，与えられた 2 文の組のうち，人間による評価がより高い文を特定するタスクである．表 2 に，提案手法およびベースライン手法の，PASCAL-50S [Anderson 16] における性能を示す．表 2 より，提案手法は HC，MM，および Mean においてそれぞれ 76.1%，77.9%，87.8%であり，既存手法を 6.1，0.3，1.9 ポイント上回った．

図 2 に Polaris2.0 データセットにおける提案手法の成功例を示す．図 2-(a) が示すサンプルにおいて，$\boldsymbol{x}_{\mathrm{ref}}^{(1)}$ は “A kitchen with vending machines and a black and white checkered floor.” であり，$\boldsymbol{x}_{\mathrm{cand}}$ は ” black and white checkered floor ” であった．また，本サンプルでは $\boldsymbol{x}_{\mathrm{cand}}$ が画像を部分的に正しく捉えているため，y が 0.5 であった．Polos [和田 24] や RefCLIP-S [Hessel 21] がそれぞれ 0.816, 0.719 と誤って評価したのに対して，提案手法は y に近い値 0.634 と評価した．図 2-(b) が示すにおいて，$\boldsymbol{x}_{\mathrm{ref}}^{(1)}$ “Three traffic signs arranged

Metric	Transformer	SVE	Multifaceted references	FOIL 1-ref	FOIL 4-ref	Flickr8K (CF)	Flickr8K (Expert)	Composite	Polaris2.0
(i)				76.2	76.5	25.1	40.1	37.7	48.1
(ii)	✓		✓	84.3	89.3	24.7	49.6	35.8	45.2
(iii)	✓	✓		94.4	96.1	37.2	55.7	57.4	53.2
(iv)	✓	✓	✓	**95.4**	**96.5**	**38.3**	**56.2**	**58.4**	**54.3**

表 3: ablation studies の定量的結果

on a sign post" であり，$\boldsymbol{x}_{\text{cand}}$ は "a man sitting on a stop sign on a street corner" であった．また，本サンプルでは $\boldsymbol{x}_{\text{cand}}$ がハルシネーション ("a man sitting on") を含んでいるため，y が 0.0 であった．Polos や RefCLIP-S がそれぞれ 0.392, 0.442 と誤って評価したのに対して，提案手法は y に近い値 0.023 と評価した．したがって，我々の提案尺度である DENEB が，人間による評価に近い値を出力しただけでなくハルシネーションに対して頑健であることがわかる．

4.3 Ablation Studies

本研究では，DENEB の有効性を調査するために ablation studies を行った．表 3 に，ablation studies の定量的結果を示す．ablation 条件として以下の 3 つを定めた．

Sim-Vec Transformer ablation. SVE Transformer モジュールを MLP に置き換えることで，SVE Transformer の性能への寄与を調査した．表 3 より，FOIL における Metric(i) の 1-ref，4-ref の精度はそれぞれ 76.2%, 76.5%であり，Metric(iv) に比べ 19.2, 20.0 ポイント減少している．また，Metric(i) における Flickr8K-CF, Flickr8K-Expert, Composite, Polaris2.0 の相関係数はそれぞれ 25.1, 40.1, 37.7, 48.1 であり，提案手法に比べ 13.2, 16.1, 20.7, 6.2 ポイントと大きく減少している．したがって，SVE Transformer が画像キャプション生成の自動評価において有用であることが示された．

SVE ablation. SVE を取り除き，次式で表される系列を Transformer に入力することで，SVE の性能への寄与を調査した．

$$\left[v; c_{\text{clip}}; r^{(1)}_{\text{clip}}; \ldots; r^{(N)}_{\text{clip}}; c_{\text{rb}}; r^{(1)}_{\text{rb}}; \ldots; r^{(N)}_{\text{rb}}\right]$$

表 3 より，FOIL における Metric(ii) の 1-ref，4-ref の精度はそれぞれ 84.3%, 89.3%であり，Metric(iv) に比べ 11.1, 7.2 ポイント減少している．また，Metric(ii) における Flickr8K-CF, Flickr8K-Expert, Composite, Polaris2.0 での相関係数はそれぞれ 24.7, 49.6, 35.8, 45.2 であり，提案手法に比べ，それぞれ 13.6, 6.6, 22.6, 9.1 ポイント減少した．したがって，SVE が画像キャプション生成の自動評価において有益な特徴量を抽出し，性能向上に寄与することが示された．

Multifaceted references ablation. Aggregate 関数を使用せず，複数の多角的な参照文を効果的に扱う機構による性能への寄与を調査した．表 3 より，FOIL における Metric(iii) の 1-ref，4-ref の精度はそれぞれ 93.2%, 94.4%であり，Metric(iv) に比べ 2.2, 1.1 ポイント減少している．また，本機構を取り除いた Metric(iii) における Flickr8K-CF, Flickr8K-Expert, Composite, Polaris2.0 の相関係数はそれぞれ 37.2, 55.7, 57.4, 53.2 であり，提案手法に比べそれぞれ 1.1, 0.5, 1.0, 1.1 ポイント減少した．したがって，Aggregate 関数を使用せず，Transformer を用いて参照文群を扱う本機構が性能向上に寄与することが示された．

5. おわりに

本研究では，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尺度 DENEB を提案した．

本研究の貢献を以下に示す．

- 画像キャプション生成における自動評価尺度 DENEB を提案した．
- 画像，生成文および参照文群間の類似度を扱う Sim-Vec Transformer モジュールを導入した．
- アダマール積と差分を用いて自動評価に有用な特徴量を抽出する Sim-Vec Extraction (SVE) を導入した．
- 画像のバリエーションがサンプル数に対し極端に少ない Polaris データセットを拡張し，画像を約 20,000 枚追加した新たなデータセット Polaris2.0 を構築する．
- Composite, Flickr8K-CF, Polaris2.0, FOIL, PASCAL-50S において提案手法が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

[Aditya 15] Aditya, S., et al.: From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge, *arXiv preprint arXiv:1511.03292* (2015) 3

[Anderson 16] Anderson, P., Fernando, B., Johnson, M., et al.: SPICE: Semantic Propositional Image Caption Evaluation, in *ECCV*, pp. 382–398 (2016) 1, 3

[Chan 23] Chan, D. M., et al.: CLAIR: Evaluating Image Captions with Large Language Models, in *EMNLP* (2023) 3

[Dognin 22] Dognin, P., et al.: Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge, Vol. 73, pp. 437–459 (2022) 1

[Gao 21] Gao, T., et al.: SimCSE: Simple Contrastive Learning of Sentence Embeddings, in *EMNLP*, pp. 6894–6910 (2021) 2

[Ghandi 23] Ghandi, T., Pourreza, H., and Mahyar, H.: Deep Learning Approaches on Image Captioning: A Review, *ACM Computing Surveys*, Vol. 56, No. 3 (2023) 1

[Hessel 21] Hessel, J., Holtzman, A., et al.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning, in *EMNLP*, pp. 7514–7528 (2021) 1, 2, 3

[Hodosh 13] Hodosh, M., et al.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, *JAIR*, Vol. 47, pp. 853–899 (2013) 3

[Jiang 19] Jiang, M., Huang, Q., Zhang, L., Wang, X., et al.: TIGEr: Text-to-image Grounding For Image Caption Evaluation, in *EMNLP*, pp. 2141–2152 (2019) 3

[Lee 21] Lee, H., Yoon, S., et al.: UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning, in *ACL*, pp. 220–226 (2021) 1, 2, 3

[Liu 19] 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) 2

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, in *ACL*, pp. 311–318 (2002) 1, 3

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., et al.: Learning Transferable Visual Models from Natural Language Supervision, in *ICML*, pp. 8748–8763 (2021) 1, 2

[Sarto 23] Sarto, S., et al.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation, in *CVPR*, pp. 6914–6924 (2023) 1, 2, 3

[Shekhar 17] Shekhar, R., Pezzelle, S., Klimovich, Y., et al.: FOIL it! Find One Mismatch Between Image and Language caption, in *ACL*, pp. 255–265 (2017) 1, 3

[Vedantam 15] Vedantam, R., Zitnick, L., and Parikh, D.: CIDEr: Consensus-based Image Description Evaluation, in *CVPR*, pp. 4566–4575 (2015) 1, 2, 3

[Wada 23] Wada, Y., Kaneda, K., et al.: JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models, in *CoNLL*, pp. 424–435 (2023) 1

[和田 24] 和田 唯我, 兼田 寛大, 齋藤 大地, 杉浦 孔明：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, 言語処理学会第 30 回年次大会 (2024) 1, 2, 3