マルチモーダル情報に基づく画像説明文の教師あり自動評価

Supervised Automatic Evaluation for Image Captioning Based on Multimodality

齋藤 大地 * 和田 唯我 兼田 寛大 杉浦 孔明
Daichi Saito Yuiga Wada Kanta Kaneda Komei Sugiura

慶應義塾大学
Keio University

画像キャプション生成では，モデルが出力した生成文の品質を適切に評価することが重要である．しかし，n-gram に基づく自動評価尺度は人間による評価との相関が低いことが報告されている．日本語の画像キャプション生成では JaSPICE がそれらに代わる自動評価尺度として提案されているものの，表層表現の不一致に対して適切に評価を行うことができない．また，COMET をはじめとする学習可能な自動評価尺度は，機械翻訳における自動評価タスクに最適化されており画像を考慮しないため，画像キャプション生成には適していない．そこで本論文では，画像キャプション生成に対する自動評価尺度 SuiSei を提案する．SuiSei は，画像特徴量と言語特徴量を扱うマルチモーダル特徴抽出機構および idf を考慮した文埋め込み機構を用いて人間による評価を回帰する．実験の結果，SuiSei はベースライン尺度と比較して人間による評価との相関係数が高いことを示した．

1 はじめに

画像キャプション生成は，視覚障害者支援 [Gurari 20, Ahsan 21, Dognin 22, Ghandi 23]，医療画像解析 [Pavlopoulos 19, Huang 21, Ayesha 21]，ロボティクスにおける説明生成 [Kambara 22, Magassouba 20, Ogura 20] など，幅広い分野に応用されている．本研究分野では，画像キャプション生成モデルが出力した生成文の品質を適切に評価することが重要である．

先行研究では，n-gram に基づく自動評価尺度は人間による評価との相関が低いことが指摘されている [Anderson 16]．そのため，JaSPICE [Wada 23] などの人間による評価との相関が高い自動評価尺度が提案されているものの，表層表現の不一致に対して適切に評価を行うことができない．また，COMET [Rei 20] をはじめとする学習可能な自動評価尺度は，機械翻訳における自動評価タスクに最適化されており，画像を考慮しないため画像キャプション生成における性能は不十分である．したがって，画像キャプション生成において人間による評価との相関が十分に高い自動評価尺度が構築されることが望ましい．

画像キャプション生成では，画像のどこに着目するか，画像をどのように自然言語で表現するかなどによって無数の正解文が存在し得る．そのため，生成文が正解文と大きく異なっていても画像に対して適切である場合があり，単なる文の類似度比較だけでは不十分な点において困難なタスクである [Yi 20]．

そこで本論文では，画像キャプション生成モデルに対する自動評価尺度 SuiSei[1]を提案する．SuiSei は，画像特徴量と言語特徴量を扱うマルチモーダル特徴抽出機構および idf (inverse document frequency, 逆文書頻度) を考慮した文埋め込み機構を用いた回帰を行うことで，人間による評価との高い相関を実現する．

例えば，正解文が {「少年がサッカーをしている」,「男の子がサッカーの試合をしている」} であるような画像に対し，画像キャプション生成モデルが「子供がサッカーボールを蹴っている」という生成文を出力したとする．このとき，提案手法は画像と正解文に対して生成文がどの程度適切であるかを示す評価値を計算する．

提案手法が既存手法と異なる点は，画像特徴量と言語特徴量を扱うマルチモーダル特徴量抽出機構を自動評価尺度に導入する点，および idf を考慮した文埋め込み機構を導入する点である．マルチモーダル特徴量抽出機構の導入により，正解文だけでなく画像に対しても生成文が適切であるかを考慮することができる．また，idf を考慮した文埋め込み機構の導入により，より重要な単語に注目して評価値を予測することができる．

提案手法の新規性は以下の通りである．

*連絡先：慶應義塾大学理工学部情報工学科
〒 223-8522 神奈川県横浜市港北区日吉 3-14-1
E-mail: daichi-s@keio.jp

[1]**S**upervised m**U**ltimodal evaluat**I**on **S**ystem for imag**E** capt**I**oning

- 画像キャプション生成に対する自動評価尺度に画像特徴量と言語特徴量を扱うマルチモーダル特徴抽出機構を導入した SuiSei を提案する．
- 画像キャプション生成に対する自動評価尺度に idf を考慮した文埋め込み機構を導入する．

2　関連研究

画像キャプション生成の研究は広く行われており [Li 20, Alayrac 22, Yu 22, Li 23]，鑑賞者の感情に基づいた絵画の説明文生成 [Ishikawa 23] や Transformer による生活支援ロボットの指示文生成 [Kambara 22] など，さまざまな分野に応用されている．画像キャプション生成に関するサーベイ論文である [Ming 22] では，画像キャプション生成モデル，標準データセット，評価尺度などについての包括的な総括がなされている．

画像キャプション生成における標準的な評価尺度としては，BLEU [Papineni 02], ROUGE [Lin 04], METEOR [Banerjee 05], CIDEr [Vedantam 15], BERTScore [Zhang 20] などが挙げられる．また COMET は，深層学習に基づいて人間による評価を回帰する学習可能な自動評価尺度であり，ルールベースの自動評価尺度と比較して人間による評価との相関が高いことが示されている [Rei 20]．日本語の画像キャプション生成では，シーングラフに基づいて評価を行う JaSPICE も既存手法に比べ人間による評価と相関が高いことが報告されている [Wada 23]．近年では，BERT [Devlin 19] に基づいて対照学習を行う UMIC [Lee 21] や，CLIP [Radford 21] を用いて画像と生成文の類似度を計算する CLIP-Score [Hessel 21] など，正解文を使用しない自動評価尺度も登場している．

3　問題設定

本論文では，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，人間による評価に近いことが望ましい．具体的には，自動評価尺度の評価値と人間による評価との相関係数が高いことが望ましい．

本タスクでは，画像 $\boldsymbol{x}_{\mathrm{img}}$，生成文 $\boldsymbol{x}_{\mathrm{cand}}$，$N_t$ 個の正解文 $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ を入力とし，$\boldsymbol{x}_{\mathrm{img}}$ と $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ に対して $\boldsymbol{x}_{\mathrm{cand}}$ がどの程度適切であるかを示す評価値を計算する．また，本論文では日本語の画像キャプション生成における自動評価を前提とし，提案手法の評価には人間による評価との相関係数（Pearson/Spearman/Kendall の相関係数）を使用する．

4　提案手法

本論文では，人間による評価を回帰させる COMET を拡張した，画像キャプション生成に対する自動評価尺度 SuiSei を提案する．本手法における拡張は，画像特徴量と言語特徴量を用いた人間による評価の回帰であり，学習可能なパラメータをもつ自動評価尺度一般に適用可能であると考えられる．

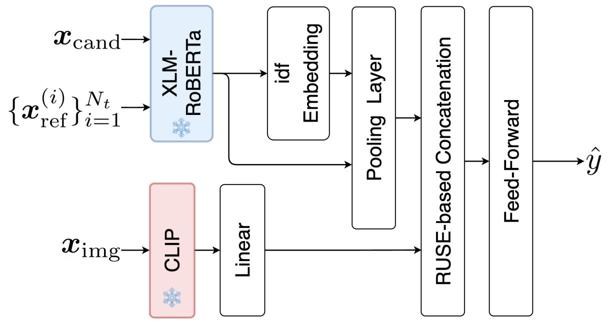

図 1: 提案手法のモデル構造．

提案手法の新規性は以下の通りである．

- 画像キャプション生成に対する自動評価尺度に画像特徴量と言語特徴量を扱うマルチモーダル特徴抽出機構を導入した SuiSei を提案する．
- 画像キャプション生成に対する自動評価尺度に idf を考慮した文埋め込み機構を導入する．

4.1　入力

図 1 に提案手法のモデル構造を示す． ネットワークの入力は以下の $\boldsymbol{x}$ であり，出力は $\boldsymbol{x}$ に対する評価値の予測 $\hat{y}$ である．ここで，$\boldsymbol{x}$ は以下のように定義される．

$$\boldsymbol{x} = \left\{\boldsymbol{x}_{\mathrm{cand}}, \left\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\right\}_{i=1}^{N_t}, \boldsymbol{x}_{\mathrm{img}}\right\} \tag{1}$$

ただし，$\boldsymbol{x}_{\mathrm{cand}} \in \{1,0\}^{V\times L}$, $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t} \in \{1,0\}^{N_t\times V\times L}$, $\boldsymbol{x}_{\mathrm{img}} \in \mathbb{R}^{3\times H\times W}$ である．ここで，V は語彙サイズ，L は最大トークン長，N_t は 1 サンプルにおける正解文の数，H, W はそれぞれ $\boldsymbol{x}_{img}$ の高さと幅を表す．

本モデルではまず，XLM-RoBERTa [Conneau 20] を使用して $\boldsymbol{x}_{\mathrm{cand}}$ および $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ からそれぞれ言語特徴量 $\boldsymbol{h}_{\mathrm{cand}} \in \mathbb{R}^{L\times d_{\mathrm{XLM}}}$ および $\{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t} \in \mathbb{R}^{N_t\times L\times d_{\mathrm{XLM}}}$ を得る．ここで，d_{XLM} は XLM-RoBERTa の出力次元数である．また，事前学習済みの CLIP 画像エンコーダ（ViT-B/16）[Radford 21] を用い，$\boldsymbol{x}_{\mathrm{img}}$ から画像特徴量 $\boldsymbol{h}_{\mathrm{img}} \in \mathbb{R}^{d_{\mathrm{CLIP}}}$ を得る．ここで，d_{CLIP} は CLIP 画像エンコーダの出力次元数である．

4.2　文埋め込み機構

idf embedding では，$\boldsymbol{h}_{\mathrm{cand}}$ および $\{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ から idf を考慮した文埋め込みを作成する．文章の類似性の計算において，文書集合の中でその語を含む文書数，すなわち文書頻度（document frequency）が低い単語は，文書頻度が高い単語よりも重要であることが知られて

いる [Zhang 20]．そこで本モジュールでは，文書頻度の低いトークンをより重要なトークンとして評価値の予測に反映させるために，トークンごとに idf を計算し，$\boldsymbol{h}_{\mathrm{cand}}$ および $\{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ に重みとして掛け合わせる．ここで，N 個の正解文が与えられたとき，トークン w に対する文書頻度を $\mathrm{df}(w)$ とすると，w に対する idf は以下のように計算される．

$$\mathrm{idf}(w) = \log \frac{N}{\mathrm{df}(w)} \tag{2}$$

これを用いて，以下のように $\boldsymbol{h}_{cand}$ から $\boldsymbol{h}'_{cand}$ を得る．

$$\boldsymbol{h}'_{\mathrm{cand}} = \left\{ \mathrm{idf}(\boldsymbol{x}_{\mathrm{cand}}^{(l)}) \cdot \boldsymbol{h}_{\mathrm{cand}}^{(l)} \mid l = 1, \ldots, L \right\} \tag{3}$$

ただし，L は最大系列長を表す．同様に，$\{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ から idf を考慮した文埋め込み $\{\boldsymbol{h'}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ を作成し，$\{\boldsymbol{h'}_{\mathrm{cand}}$, $\{\boldsymbol{h'}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}\}$ を得る．ここで，term frequency (tf) を使用しない理由は，処理する対象が単一の文であり，多くの場合 tf の値が 1 になるためである．

4.3 マルチモーダル特徴抽出機構

RUSE-based concatenation では，COMET と同様に RUSE [Shimanaka 18] に基づく手法を用いて各特徴量を結合し，Feed-Forward Network (FFN) により $\boldsymbol{x}$ に対する評価値の予測 $\hat{y}$ を得る．入力は $\{\boldsymbol{h}_{\mathrm{cand}}, \{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}, \boldsymbol{h}'_{\mathrm{cand}}, \{\boldsymbol{h'}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}, \boldsymbol{h}_{\mathrm{img}}\}$，出力は $\hat{y}$ である．本モジュールでは，RUSE に基づく手法を用いて $\boldsymbol{h}_{\mathrm{cand}}$ および $\{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ から以下のように $\boldsymbol{h}_{\mathrm{RUSE}}$ を得る．

$$\boldsymbol{h}_{\mathrm{RUSE}} = \left\{ \left\{ \boldsymbol{h}_{\mathrm{cand}} \odot \boldsymbol{h}_{\mathrm{ref}}^{(i)} \right\}_{i=1}^{N_t}, \left\{ \left| \boldsymbol{h}_{\mathrm{cand}} - \boldsymbol{h}_{\mathrm{ref}}^{(i)} \right| \right\}_{i=1}^{N_t}, \boldsymbol{h}_{\mathrm{cand}} \odot \boldsymbol{h}_{\mathrm{img}}, |\boldsymbol{h}_{\mathrm{cand}} - \boldsymbol{h}_{\mathrm{img}}| \right\} \tag{4}$$

ここで，$\odot$ はアダマール積を表し，$\boldsymbol{h}_{\mathrm{cand}}$ および $\{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ には Pooling 層を適用した．最後に，$\boldsymbol{h}_{\mathrm{RUSE}}$，$\boldsymbol{h}_{\mathrm{cand}}$，$\{\boldsymbol{h}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ を結合し，$\boldsymbol{h}$ を得る．同様に，$\boldsymbol{h'}_{\mathrm{cand}}$，$\{\boldsymbol{h'}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ についても RUSE に基づく変換を行い，$\boldsymbol{h}$ と結合して FFN により出力 $\hat{y}$ を得る．損失関数には，平均二乗誤差を用いた．

5 実験設定

本研究では，画像キャプション生成における人間による評価を学習するため，*Shichimi* データセット [Wada 23] および PFN-PIC-gen データセットを用いた．*Shichimi* データセットおよび PFN-PIC-gen は，日本語の画像キャプション生成に対する自動評価タスクにおいて最大規模のコーパスである．これらのデータセットにおける人間による評価は，画像に対して生成文が適切であるかを 5 段階で評価したものである．*Shichimi* データセットは，画像，生成文，正解文，および生成文に対する人間による評価で構成された大規模コーパスであり，500 人のアノテータによる評価値を含む．サンプル数は 103,170，語彙サイズは 6,269，平均文長は 12.4 文字である．また，PFN-PIC-gen は [Wada 23] で収集されたものであり，PFN-PIC [Hatori 18] の画像から生成したキャプションに対して人間による評価が付与されたデータセットを指す．サンプル数は 1,920，語彙サイズは 647，平均文長は 20.2 文字である．

表 1: 提案手法の設定．

最適化手法	Adam ($\beta_1 = 0.9$，$\beta_2 = 0.999$)
学習率	3.0×10^5
バッチサイズ	32
エポック数	9

提案手法の評価には，SuiSei が出力した評価値 $\{\hat{y}_i\}_{i=1}^{N}$ と人間による評価値 $\{y_i\}_{i=1}^{N}$ に対する相関係数を用いた．ここで，N はサンプル数を表す．相関係数には，Pearson，Spearman，Kendall の相関係数を使用した．

本実験では，*Shichimi* データセットを訓練集合とテスト集合に分割し，それぞれ 51,988 サンプル，51,182 サンプルとした．また，PFN-PIC-gen は，ゼロショット性能を検証するために全てテスト集合として使用した．ただし，訓練集合をモデルの学習に用い，テスト集合を評価に用いて実験を行った．

表 1 に提案手法の設定を示す． ここで，提案手法における訓練可能パラメータ数は 1.84×10^8 であった．また，モデルの学習にはメモリ 24GB 搭載の GeForce RTX 3090 および Intel Core i9 12900K を使用し，訓練時間および 1 サンプルあたりの推論時間は，それぞれ約 1.2 時間および約 7.6ms であった．

6 実験結果

6.1 定量的結果

表 2, 表 3 に *Shichimi* データセットおよび PFN-PIC-gen における定量的結果を示す．本タスクでは，自動評価尺度の評価値と人間による評価との相関係数に基づいて自動評価尺度を評価する．ベースライン尺度には，日本語の画像キャプション生成において標準的な尺度である BLEU, ROUGE, METEOR, CIDEr, JaSPICE を用いた．また，提案手法は COMET を拡張したものであるため，COMET もベースライン尺度に採用した．

表 2 より，*Shichimi* データセットにおける提案手法は，Pearson, Spearman, Kendall の相関係数において，それぞれ 0.672, 0.644, 0.504 であり，JaSPICE と比較して 0.173, 0.113, 0.091 ポイント上回った．また，COMET と比較して，0.048, 0.071, 0.062 ポイント上回った．*Shichimi* データセットにおいて，人間による評価に対する Pearson, Spearman, Kendall の相関係数

表 2: *Shichimi* データセットにおける各自動評価尺度と人間による評価との相関係数.

	Pearson	Spearman	Kendall
BLEU	0.296	0.343	0.260
ROUGE	0.366	0.340	0.258
METEOR	0.345	0.366	0.279
CIDEr	0.312	0.355	0.269
JaSPICE	0.499	0.531	0.413
COMET	0.624	0.573	0.442
SuiSei	**0.672**	**0.644**	**0.504**
Human	0.759	0.750	0.669

(a)

(b)

図 2: 成功例における画像.

はそれぞれ，0.759, 0.750, 0.669 であった．人間による評価に対する相関係数が 1.0 よりも小さい理由は，人間による評価に完全な一貫性がなく，同一サンプルに対する評価値が必ずしも一致しないためである．また，人間による評価に対する相関係数は，自動評価尺度の性能における上限値であると考えられる．

同様に，表 3 より PFN-PIC-gen における提案手法は，Pearson, Spearman, Kendall の相関係数において，それぞれ 0.576, 0.590, 0.443 であり，JaSPICE と比較して 0.030, 0.017, 0.005 ポイント，COMET と比較して，0.137, 0.155, 0.118 ポイント上回った．

6.2 定性的結果

図 2 に定性的結果における成功例を示す．図 2 (a) は *Shichimi* データセットにおける結果の一つであり，$\boldsymbol{x}_{\mathrm{cand}}$ は「デニムパンツをはいた少女がサッカーボールを蹴ろうとしている」，$\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ は {「少女がサッカーボールと戯れている」,「ボーダー柄のシャツを着た少女がサッカーボールで遊んでいる」} である．図 2 (a) における人間による評価値 y と提案手法の予測値 $\hat{y}$ はそれぞれ，$y = 5, \hat{y} = 0.974$ であった．テスト集合において，この入力に対する SuiSei の値は上位 3%の値であるため，提案手法は図 2 (a) の例において人間による評価に近い評価値を出力していると言える．

表 3: PFN-PIC-gen における各自動評価尺度と人間による評価との相関係数.

	Pearson	Spearman	Kendall
BLEU	0.484	0.466	0.352
ROUGE	0.500	0.474	0.365
METEOR	0.423	0.457	0.352
CIDEr	0.416	0.462	0.353
JaSPICE	0.547	0.573	0.438
COMET	0.439	0.435	0.325
SuiSei	**0.576**	**0.590**	**0.443**

表 4: Ablation study の結果.

Model	$\boldsymbol{x}_{\mathrm{img}}$	idf を考慮した文埋め込み	Pearson	Spearman	Kendall
(i)		✓	0.532	0.536	0.402
(ii)	✓		0.517	0.515	0.386
(iii)	✓	✓	0.576	0.590	0.443

同様に図 2 (b) も *Shichimi* データセットにおける定性的結果の一つであり，$\boldsymbol{x}_{\mathrm{cand}}$ は「時計塔の前に道路標識が立っている」，$\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N_t}$ は {「ビッグベンの前に地下鉄のサインが出ている」,「高い時計台と地下と表示された看板がある」} である．図 2 (b) における人間による評価値 y と提案手法の予測値 $\hat{y}$ はそれぞれ，$y = 5, \hat{y} = 0.950$ であった．テスト集合において，この入力に対する SuiSei の値は上位 6%の値であるため，提案手法は図 2 (b) の例においても人間による評価に近い評価値を出力していると言える．

6.3 Ablation Study

次に，以下の二つの条件を Ablation study に定めた．その結果を表 4 に示す．

Image Ablation 入力から $\boldsymbol{x}_{\mathrm{img}}$ を取り除くことによる性能への影響を調査した．その結果，Model (i) は Pearson, Spearman, Kendall の相関係数においてそれぞれ 0.532, 0.536, 0.402 であり，Model (iii) と比較して 0.044, 0.054, 0.041 ポイント下回った．この結果から，$\boldsymbol{x}_{\mathrm{img}}$ の自動評価尺度への導入が提案手法の性能向上に寄与していることが確認できた．

Sentence Embedding Ablation idf を考慮した文埋め込みを FFN の入力から取り除くことによる性能への影響を調査した．その結果，Model (ii) は Pearson, Spearman, Kendall の相関係数においてそれぞれ 0.517, 0.515, 0.386 であり，Model (iii) と比較して 0.059, 0.075, 0.057 ポイント下回った．この結果から，idf を考慮した文埋め込みの自動評価尺度への導入が提案手法の性能向上に寄与していることが確認できた．

7 結論

本論文では，画像キャプション生成に対する自動評価を扱った．本研究の貢献を以下に示す．

- 画像キャプション生成に対する自動評価尺度に，画像特徴量と言語特徴量を扱うマルチモーダル特徴抽出機構を導入した SuiSei を提案した．
- 画像キャプション生成に対する自動評価尺度に idf を考慮した文埋め込み機構を導入した．
- SuiSei はベースライン尺度と比較して，人間による評価との相関係数が高いことを示した．

謝辞

本研究の一部は，JSPS 科研費 23H03478，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参考文献

[Ahsan 21] Ahsan, H., Bhalla, N., Bhatt, D., and Shah, K.: Multi-Modal Image Captioning for the Visually Impaired, in *NAACL*, pp. 53–60 (2021)

[Alayrac 22] Alayrac, J.-B., Donahue, J., Luc, P., et al.: Flamingo: a Visual Language Model for Few-Shot Learning, *NeurIPS*, Vol. 35, pp. 23716–23736 (2022)

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

[Ayesha 21] Ayesha, H., et al.: Automatic Medical Image Interpretation: State of the Art and Future Directions, *Pattern Recognition*, Vol. 114, p. 107856 (2021)

[Banerjee 05] Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, in *ACL*, pp. 65–72 (2005)

[Conneau 20] Conneau, A., Khandelwal, K., et al.: Unsupervised Cross-lingual Representation Learning at Scale, in *ACL*, pp. 8440–8451 (2020)

[Devlin 19] Devlin, J., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *NAACL-HLT*, pp. 4171–4186 (2019)

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

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

[Gurari 20] Gurari, D., Zhao, Y., Zhang, M., and Bhattacharya, N.: Captioning Images Taken by People Who Are Blind, in *ECCV*, pp. 417–434 (2020)

[Hatori 18] Hatori, J., et al.: Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions, in *ICRA*, pp. 3774–3781 (2018)

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

[Huang 21] Huang, J.-H., Wu, T.-W., et al.: Contextualized Keyword Representations for Multi-modal Retinal Image Captioning, in *ICMR*, pp. 645–652 (2021)

[Ishikawa 23] Ishikawa, S. and Sugiura, K.: Affective Image Captioning for Visual Artworks Using Emotion-Based Cross-Attention Mechanisms, *IEEE Access*, Vol. 11, pp. 24527–24534 (2023)

[Kambara 22] Kambara, M., et al.: Relational Future Captioning Model for Explaining Likely Collisions in Daily Tasks, in *IEEE ICIP*, pp. 2601–2605 (2022)

[Lee 21] Lee, H., Yoon, S., Dernoncourt, F., and Jung, K.: UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning, in *ACL*, pp. 220–226 (2021)

[Li 20] Li, X., Yin, X., Li, C., Zhang, P., and Hu, X. o.: Oscar: Object-Semantics Aligned Pre-training for Vision-language Tasks, in *ECCV*, pp. 121–137 (2020)

[Li 23] Li, J., Li, D., et al.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, *ICML* (2023)

[Lin 04] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, in *Text summarization branches out*, pp. 74–81 (2004)

[Magassouba 20] Magassouba, A., Sugiura, K., et al.: Multimodal Attention Branch Network for Perspective-Free Sentence Generation, in *CoRL*, pp. 76–85 (2020)

[Ming 22] Ming, Y., et al.: Visuals to Text: A Comprehensive Review on Automatic Image Captioning, *IEEE/CAA JAS*, Vol. 9, No. 8, pp. 1339–1365 (2022)

[Ogura 20] Ogura, T., Magassouba, A., Sugiura, K., et al.: Alleviating the Burden of Labeling: Sentence Generation by Attention Branch Encoder–Decoder Network, *IEEE RAL*, Vol. 5, No. 4, pp. 5945–5952 (2020)

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

[Pavlopoulos 19] Pavlopoulos, J., Kougia, V., and Androutsopoulos, I.: A Survey on Biomedical Image Captioning, in *SiVL*, pp. 26–36 (2019)

[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)

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

[Shimanaka 18] Shimanaka, H., et al.: RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation, in *WMT18*, pp. 751–758 (2018)

[Vedantam 15] Vedantam, R., Lawrence Zitnick, C., and Parikh, D.: CIDEr: Consensus-Based Image Description Evaluation, in *CVPR* (2015)

[Wada 23] Wada, Y., Kaneda, K., and Sugiura, K.: JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models, in *CoNLL* (2023)

[Yi 20] Yi, Y., Deng, H., and Hu, J.: Improving Image Captioning Evaluation by Considering Inter References Variance, in *ACL*, pp. 985–994 (2020)

[Yu 22] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., et al.: CoCa: Contrastive Captioners are Image-Text Foundation Models, *TMLR* (2022)

[Zhang 20] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, in *ICLR* (2020)