マルチモーダル OCR 特徴を用いた Dynamic Pointer Network によるテキスト付き画像説明文生成

植田有咲　Wei Yang　杉浦孔明
慶應義塾大学
{arinko31,wei.yang,komei.sugiura}@keio.jp

概要

テキスト情報を含む画像の説明文生成は視覚のバリアフリー化を実現するための重要な課題の一つである. 本研究では, テキスト情報を含む画像に対して説明文を生成するタスクに対して, マルチモーダル OCR 特徴を含む複数のモダリティを利用した画像説明文生成モデルを提案する. 提案手法では画像中のテキスト領域を複数のモダリティに分割するマルチモーダル OCR 特徴を導入する. さらに, 画像, 物体領域, マルチモーダル OCR 特徴を含む複数モダリティ間の関係をモデル化するための相互注意を導入する. 提案手法は TextCaps データセットにおいて既存手法を上回る結果を得た.

1　はじめに

テキスト情報を含む標識や看板などは日常生活に多く存在する. テキスト情報を含む画像の説明文生成は, 日常生活における視覚のバリアフリー化を促進する一つの手段である. また, 画像の代替テキスト (alt 属性など) の生成における品質向上にとって有益である.

以上の社会的背景から, 本研究では TextCaps (Text-based Image Captioning) タスクを扱う. TextCaps タスクはテキスト情報を含む画像に対して OCR (Optical Character Recognition) を利用して説明文を生成するタスクである. 図 1 に提案手法の概要図を示す. 図の例では, “a store called del’s sells soft frozen lemonade” という説明文を生成することが望ましい.

本タスクでは OCR トークンの多面的な視覚言語特徴や, OCR トークンと物体間の関係を考慮することが重要である. しかし, 先行研究 [1], [2] ではそれらの考慮が不十分であり, 十分な性能が得られなかった.

そこで本研究では, マルチモーダル OCR 特徴を含む複数のモダリティを利用した画像説明文生成モデルを提案する. ここでマルチモーダル OCR 特徴とは, OCR により得られたテキストとその画像領域に関する視覚言語特徴を指すものとする. 提案手法では, 事前学習済みの CLIP (Contrastive Language-Image Pre-training) [3] モデルを全体画像と OCR トークンの言語特徴量を抽出するために用いる. さらに複数のモダリティ間の関係をモデル化する相互注意を導入することで, 性能の向上が期待できる.

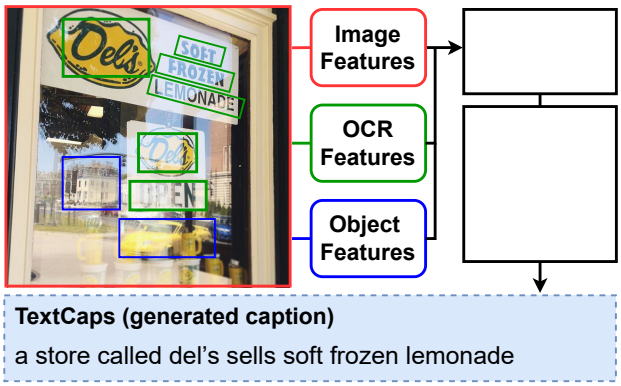


図 1　提案手法の概要図

本研究の新規性は以下の通りである.

- 既存手法と異なり, 画像中のテキスト領域を複数のモダリティに分割するマルチモーダル OCR 特徴を導入する.
- 画像全体, 物体領域, マルチモーダル OCR 特徴を含む複数モダリティ間の関係をモデル化するための相互注意を導入する.

2　問題設定

本研究では, 画像内のテキスト情報を認識し, それらに関連する説明文を自動的に生成する TextCaps タスクを扱う. テキスト情報を含まない画像に対する画像説明文生成については考慮しないものとする. 本タスクの入出力を以下のように定義する.

- **入力**：画像
- **出力**：画像内のテキスト情報に関連する説明文

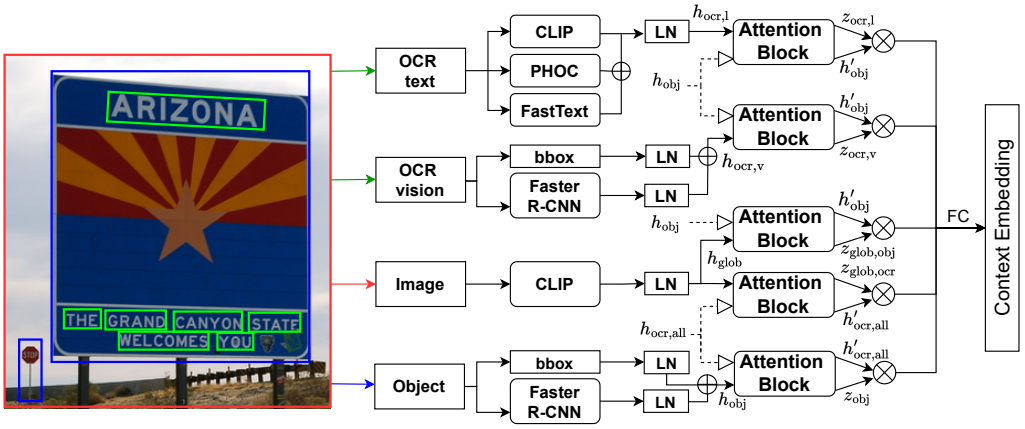


図 2　提案手法のネットワーク図

3　提案モデル

本モデルは 5 つの相互注意と説明文を生成する復号器から構成されている．図 2 に提案手法のネットワーク構造を示す．入力は $\boldsymbol{X} = \{\boldsymbol{X}_{\text{img}}, \boldsymbol{X}_{\text{obj}}, \boldsymbol{X}_{\text{ocr}}\}$ であり，$\boldsymbol{X}_{\text{img}}, \boldsymbol{X}_{\text{obj}}, \boldsymbol{X}_{\text{ocr}}$ はそれぞれ画像全体，物体領域，OCR で検出された領域と検出結果を表す．

3.1　特徴抽出

画像特徴量 $\boldsymbol{X}_{\text{img}}$ は事前学習済みの CLIP を用いて特徴量抽出される．CLIP の RN50x4 モデルを用いて画像全体を符号化し，画像特徴量 $\boldsymbol{h}_{\text{glob}} \in \mathbb{R}^{640}$ を得る．

さらに，Faster R-CNN [4] を用いて画像から l 個の物体を検出し，$\boldsymbol{X}_{\text{obj}}$ を抽出し，TextCaps タスクで fine-tuning を行う．特徴量の次元数は 2048 であり，抽出された特徴量を $\boldsymbol{h}^{(l)}_{\text{obj,fr}}, \boldsymbol{h}^{(l)}_{\text{obj,bb}}$ と定義する．$\boldsymbol{h}^{(l)}_{\text{obj,bb}}$ は物体領域の正規化された座標特徴量を表す．これらに正規化層を適用し，最終的な物体特徴量 $\boldsymbol{h}^{(l)}_{\text{obj}}$ を得る．

j 個の OCR トークンの言語特徴量 $\boldsymbol{h}^{(j)}_{\text{ocr,l}}$ は FastText 特徴量 $\boldsymbol{h}^{(j)}_{\text{ocr,ft}}$ と文字レベルの PHOC [5] 特徴量 $\boldsymbol{h}^{(j)}_{\text{ocr,ph}}$，CLIP (RN50x4) モデルで OCR トークン特徴量を抽出した $\boldsymbol{h}^{(j)}_{\text{ocr,c}}$ から構成される．物体特徴量抽出と同様に Faster R-CNN を用いて OCR トークンの視覚的特徴量 $\boldsymbol{h}^{(j)}_{\text{ocr,fr}}, \boldsymbol{h}^{(j)}_{\text{ocr,bb}}$ を抽出し，TextCaps タスクで fine-tuning を行う．$\boldsymbol{h}^{(j)}_{\text{ocr,bb}}$ は OCR の画像領域の正規化された座標特徴量を表す．最終的な OCR の視覚的特徴量を $\boldsymbol{h}^{(j)}_{\text{ocr,v}}$ と定義する．マルチモーダル OCR 特徴 $\boldsymbol{h}^{(j)}_{\text{ocr,all}}$ を以下のように定義する．

$$\begin{aligned}\boldsymbol{h}^{(j)}_{\text{ocr,all}} = f_{\text{LN}}(&\boldsymbol{W}_{\text{ft}}\boldsymbol{h}^{(j)}_{\text{ocr,ft}} + \boldsymbol{W}_{\text{ph}}\boldsymbol{h}^{(j)}_{\text{ocr,ph}} + \boldsymbol{W}_{\text{c}}\boldsymbol{h}^{(j)}_{\text{ocr,c}} \\ &+ \boldsymbol{W}_{\text{ocr,fr}}\boldsymbol{h}^{(j)}_{\text{ocr,fr}} + \boldsymbol{W}_{\text{glob}}\boldsymbol{h}_{\text{glob}}) + f_{\text{LN}}(\boldsymbol{W}_{\text{ocr,bb}}\boldsymbol{h}^{(j)}_{\text{ocr,bb}})\end{aligned}$$

f_{LN} は正規化層，$\boldsymbol{W}.$ は重み行列を表す．また，画像全体，物体領域，マルチモーダル OCR 特徴を相互注意の入力として用いる．

3.2　相互注意

提案手法は画像全体，物体領域，マルチモーダル OCR 特徴から構成される複数のモダリティを融合させる 5 つの相互注意から構成される．相互注意を以下のように定義する．

$$\begin{aligned}\boldsymbol{h}'_1 &= \text{softmax}\left(\frac{\boldsymbol{W}_Q\boldsymbol{h}_1(\boldsymbol{W}_K\boldsymbol{h}_1)^\top}{\sqrt{d_k}}\right)\boldsymbol{W}_V\boldsymbol{h}_1 \\ \boldsymbol{u}_n &= \text{ReLU}(\boldsymbol{W}_u\boldsymbol{h}'_n), \qquad n = 1, 2 \\ \boldsymbol{\alpha}_m &= \text{softmax}(\boldsymbol{W}_\alpha(\boldsymbol{u}_1 \odot \boldsymbol{u}_2^{(m)})), \quad \boldsymbol{z} = \sum_{m=1}^{M}\boldsymbol{\alpha}_m\boldsymbol{h}'_2,\end{aligned}$$

ここで $\odot$ はアダマール積を表す．$d_k = H/A$, H, A はそれぞれ Transformer モデルの隠れ層の次元数，ヘッド数を表す．$\boldsymbol{W}.$ は学習可能な重みを表す．例として，図 2 における 1 つ目の相互注意では $\boldsymbol{h}_1$ は $\boldsymbol{h}^{(l)}_{\text{obj}}$ として，$\boldsymbol{h}'_2$ は $\boldsymbol{h}^{(j)}_{\text{ocr,l}}$ として定義される．同様に 2 つ目では $\boldsymbol{h}_1, \boldsymbol{h}'_2$ はそれぞれ $\boldsymbol{h}^{(l)}_{\text{obj}}, \boldsymbol{h}^{(j)}_{\text{ocr,v}}$ として定義される．5 つ目の相互注意では $\boldsymbol{h}_1$ は $\boldsymbol{h}^{(j)}_{\text{ocr,all}}$, $\boldsymbol{h}'_2$ は $\boldsymbol{h}^{(l)}_{\text{obj}}$ を表す．

提案手法では画像全体，物体領域，マルチモーダル OCR 特徴を含む複数モダリティ間の関係をモデル化するための 2 つの相互注意を導入する．これらの 2 つの相互注意は入力として $\boldsymbol{h}'_2 = \boldsymbol{h}_{\text{glob}}$ を用いる．図 2 の 3 つ目の相互注意では $\boldsymbol{h}_1$ は $\boldsymbol{h}^{(l)}_{\text{obj}}$ として定義される．同様に 4 つ目の相互注意では $\boldsymbol{h}_1$ は $\boldsymbol{h}^{(j)}_{\text{ocr,all}}$ として定義される．

最終的に，$\boldsymbol{h}'_1 \odot \boldsymbol{z}$ によって 5 つの相互注意から複数モダリティ間の関係をモデル化するための埋め込み表現を計算する．得られた埋め込み表現に全結合層を適用し，相互注意の最終出力とする．

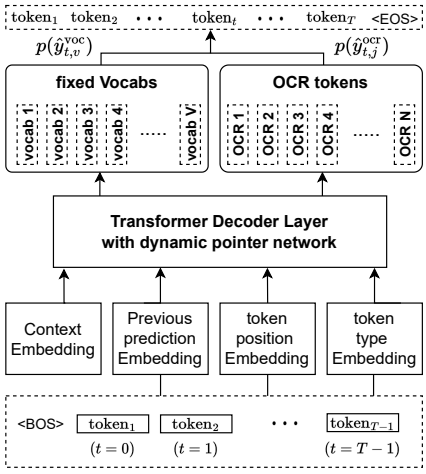


図 3　提案手法の復号器の概要図

3.3　説明文生成

図 3 に画像から説明文を生成するための復号器の概要図を示す. 提案手法では Transformer ベースの自己回帰型の復号器を用いて時刻 t ($t = 0, 1, \cdots, T$) のトークンを予測する. 復号器の特徴はポインタネットワーク (Dynamic Pointer Network) を導入することにより, 予測トークンが固定語彙 ($v = 1, \cdots, V$) または画像中の OCR トークン ($j = 1, \cdots, N$) から選択されることである.

最初の予測トークン（$t = 0$）は, 図 2 の Context Embedding に基づき, 得られる. その後のステップでは, 時刻 t のトークンを予測するために, 時刻 $t-1$ の埋め込み $\boldsymbol{x}_t^{\mathrm{dec}}$, および時刻 $t-1$ に予測されたトークンのタイプ埋め込み（0：固定語彙, 1：OCR トークン）が用いられる. 時刻 $t-1$ の予測として OCR トークンが選択された場合, $\boldsymbol{h}_{\mathrm{ocr,all}}^{(j)}$ を前時刻で予測されたトークンの $\boldsymbol{x}_t^{\mathrm{dec}}$ として入力する. 一方, 前時刻の予測が固定語彙から選択された場合, 対応する重みベクトル $\boldsymbol{w}_v^{\mathrm{voc}}$ を入力として用いる. 各時刻 t で, Transformer モデルは現時刻の予測のための入力 $\boldsymbol{x}_t^{\mathrm{dec}}$ に対応する d 次元のベクトル $\boldsymbol{z}_t^{\mathrm{dec}}$ を出力する. $\hat{\boldsymbol{y}}_{t,v}^{\mathrm{voc}}$ と $\hat{\boldsymbol{y}}_{t,j}^{\mathrm{ocr}}$ は時刻 t での固定語彙または OCR トークンから得られる予測を表す. 時刻 t におけるポインタネットワークを用いた固定語彙の予測確率 $p(\hat{\boldsymbol{y}}_{t,v}^{\mathrm{voc}})$ または OCR トークンの予測確率 $p(\hat{\boldsymbol{y}}_{t,j}^{\mathrm{ocr}})$ は以下のように得られる.

$$p(\hat{\boldsymbol{y}}_{t,v}^{\mathrm{voc}} | S = S_{\mathrm{voc}}) = \mathrm{softmax}((\boldsymbol{w}_v^{\mathrm{voc}})^\top \boldsymbol{z}_t^{\mathrm{dec}})$$

$$p(\hat{\boldsymbol{y}}_{t,j}^{\mathrm{ocr}} | S = S_{\mathrm{ocr}}) = \mathrm{softmax}((\boldsymbol{W}_{\mathrm{ocr}} \boldsymbol{z}_j^{\mathrm{ocr}})^\top (\boldsymbol{W}_{\mathrm{dec}} \boldsymbol{z}_t^{\mathrm{dec}}))$$

ここで, $\boldsymbol{w}_v^{\mathrm{voc}}$ は固定語彙内の v 番目の d-次元のパラメータを表す. $\boldsymbol{W}_{\mathrm{ocr}}$ と $\boldsymbol{W}_{\mathrm{dec}}$ は d × d の重み行列を表す.

$$[p(\hat{\boldsymbol{y}}_{t,v}^{\mathrm{voc}}); p(\hat{\boldsymbol{y}}_{t,j}^{\mathrm{ocr}})] = \{p(\hat{\boldsymbol{y}}_{t,i}^{\mathrm{all}}) | i = 1, \cdots, V+N\}$$

双方の予測確率 $[p(\hat{\boldsymbol{y}}_{t,v}^{\mathrm{voc}}); p(\hat{\boldsymbol{y}}_{t,j}^{\mathrm{ocr}})]$ を考慮し, 固定語彙, 画像内の OCR トークンからなる $V+N$ 個の候補から最も予測確率が高いトークンを出力とする.

4　実験

4.1　実験設定

TextCaps データセットを用いて提案手法の性能評価を行った. 28,408 枚のテキスト付き画像に対する 142,040 文の説明文を含む. 固定語彙は 6,736 語利用した. 説明文の平均語数は 13.47 語である. 我々は [1] に基づきデータセットを分割した. TextCaps データセットの訓練集合, 検証集合, テスト集合はそれぞれ 21,953 枚, 3,166 枚, 3,289 枚の画像から構成されている. TextCaps タスクではテスト集合の正解文が公開されていないため, 訓練集合と検証集合のみを用いた. 評価実験は検証集合を用いて行った. 実験は標準的な手順に従っており, 検証集合はモデルの学習やパラメータの調整には使用されない. 実験に使用した OCR 特徴は, SBD-Trans [6, 7] によって認識されたものである.

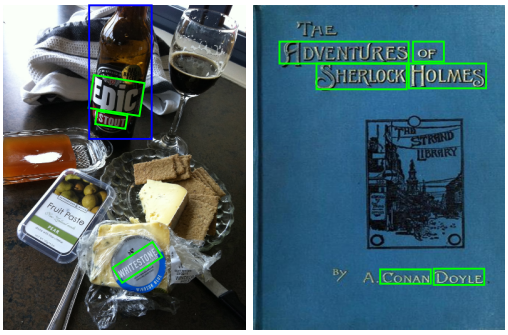
提案手法の学習は 32GB のメモリを搭載した 4 台の Tesla V100 GPU で行った. ハイパーパラメータ設定は [2] と同様である. このモデルは 170,561,578 (171M) の学習可能なパラメータを持ち, 学習には 12 時間必要であった. BLEU-4 [8], METEOR [9], ROUGE-L [10], CIDEr [11], SPICE [12] という 5 つの標準的評価手法を用いて評価を行った.

4.2　定量的結果

M4C-Captioner [1] と SSbaseline [2] をベースライン手法として用いた. ベースライン手法と提案手法の比較実験を行った結果, 提案手法 (条件 (6)) は BLEU-4, SPICE, CIDEr, それぞれ 25.58, 16.60, 99.74 ポイントであり, SSbaseline (BLEU-4: 24.54, SPICE: 15.74, CIDEr: 97.76) に比べて 1, 0.9, 2 ポイント向上する結果が得られた. さらに, M4C-Captioner との比較ではより顕著な向上が見られた. その上, 2 つの画像に関連する相互注意を導入することによって, 条件 (7) の Ours (full) では, BLEU-4, CIDEr は 0.42 (25.58 → 26.00), 0.7 (99.74 → 100.44) ポイントそれぞれ性能が向上した. これらの結果は画像全体, 物体, マルチモーダル OCR 特徴を含む複数の特徴量間の関係をモデル化する相互注意の導入が性能向上に寄与することを示唆している.

表 1　TextCaps タスクにおける定量的結果

	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
(1) M4C-Captioner (paper) [1]	23.30	22.00	46.20	89.60	15.60
(2) SSbaseline (paper) [2]	24.89	22.71	47.24	98.83	15.71
(3) SSbaseline (reproduced)	24.54±0.17	22.68±0.07	47.12±0.07	97.76±0.19	15.74±0.08
(4) Ours (w/o CLIP for Image)	24.88±0.17	22.78±0.07	47.28±0.07	98.66±0.37	15.88±0.10
(5) Ours (w/o CLIP for OCR tokens)	25.58±0.22	23.18±0.15	47.88±0.13	99.40±0.06	16.54±0.05
(6) Ours (w/o image-related attention blocks)	25.58±0.12	23.24±0.05	47.90±0.13	99.74±0.50	16.60±0.04
(7) **Ours (full)**	**26.00±0.01**	**23.40±0.06**	**48.16±0.14**	**100.44±0.50**	**16.70±0.09**



(a)　(b)

(a) SSbaseline: a bottle of whitestone stout next to a glass of it
Ours: a bottle of [epic] [stout] next to a glass of beer
GT: a bottle of epic stout next to a quarter full glass and some snacks

(b) SSbaseline: a book by conan holmes called adventures of sherlock holmes
Ours: a book by [conan] [doyle] called the [adventures] [of] [sherlock] [holmes]
GT: a book cover of the adventures of sherlock holmes by a . conan doyle

図 4　定性的結果

4.3　Ablation Study

画像とマルチモーダル OCR 特徴, および画像に関連する相互注意の有効性を調査するために, ablation study（CLIP-RN50x4 モデルを使用) を 3 つの条件で行った．表 1 に示すように, SSbaseline と条件 (4) を比較して, CLIP に基づくマルチモーダル OCR 特徴を導入するだけで CIDEr スコアは約 1 ポイント向上した（97.76 → 98.66). また, CLIP を利用した画像特徴量を削除することで性能は大きく低下した（条件 (4) と条件 (7)). しかし, 画像全体の CLIP 特徴を導入することでベースライン手法に比べ, 全ての評価指標で性能が向上した (条件 (3) と条件 (5)). 例えば, BLEU-4 は 1 (24.54 → 25.58) ポイント, SPICE は 0.8 (15.74 → 16.54) ポイント, CIDEr は 1.6 (97.76 → 99.40) ポイント, それぞれ向上した. 提案手法で導入した CLIP に基づく画像モダリティは TextCaps タスクでの性能を向上させることがわかった.

条件 (3) と条件 (6) を比較すると, 複数モダリティ間の関係をモデル化するための相互注意を取り除くことで, BLEU-4 と CIDEr のスコアは減少した. この結果は, 画像全体, 物体領域, マルチモーダル OCR 特徴を含む複数特徴量間の関係をモデル化する相互注意が TextCaps タスクにおいて有効であることを示唆している.

4.4　定性的結果

図 4 に定性的結果を示す．図 4 (a) ではベースライン手法は OCR トークンの “whitestone” と物体の “bottle” 間の関係を適切に表現できていない. 一方, 提案手法では正しい OCR トークン “epic” と “stout” を物体 “bottle” に対して適切に選択した．図 4 (b) のベースライン手法は誤った OCR トークンを生成文で用いた. 例として, “a book by conan holmes” は “a book by conan doyle” となるべきである. 一方で, 提案手法は著者と本のタイトルを適切な OCR トークンを用いて表現できている.

5　おわりに

本論文では，テキスト情報を含む画像の説明文生成を行う TextCaps タスクを扱った. 提案手法による新規性は以下である.

- 既存手法と異なり, 画像中のテキスト領域を複数のモダリティに分割するマルチモーダル OCR 特徴を導入した.
- 画像全体, 物体領域, マルチモーダル OCR 特徴を含む複数モダリティ間の関係をモデル化するための相互注意を導入した.
- 全ての評価指標においてベースライン手法を上回る性能を得た.

謝辞

本研究は JSPS 科研費 20H04269, JST Moonshot, NEDO の助成を受けて実施されたものである．

参考文献

[1] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: A dataset for image captioning with reading comprehension. In **ECCV**, pp. 742–758, 2020.

[2] Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. Simple is not easy: A simple strong baseline for TextVQA and TextCaps. In **AAAI**, Vol. 35, pp. 3608–3615, 2021.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **ICML**, pp. 8748–8763, 2021.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. **NeurIPS**, Vol. 28, , 2015.

[5] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. **IEEE Trans. PAMI**, Vol. 36, No. 12, pp. 2552–2566, 2014.

[6] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. In **IJCAI**, pp. 3052–3058, 2019.

[7] Peng Wang, L. Yang, Hui Li, Yuyang Deng, Chunhua Shen, and Yanning Zhang. A simple and robust convolutional-attention network for irregular text recognition. **arXiv preprint arXiv: 1904.01375**, 2019.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, 2002.

[9] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In **ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, 2005.

[10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, 2004.

[11] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In **CVPR**, pp. 4566–4575, 2015.

[12] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In **ECCV**, pp. 382–398, 2016.

[13] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In **AAAI**, Vol. 35, pp. 2286–2293, 2021.

[14] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In **CVPR**, pp. 17990–17999, 2022.

[15] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In **CVPR**, pp. 17980–17989, 2022.

[16] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. **IEEE Trans. PAMI**, Vol. ISSN 0162-8828, pp. 1–10, 2022.

[17] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In **ECCV**, pp. 121–137, 2020.

[18] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In **CVPR**, pp. 17980–17989, 2022.

[19] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. TAP: Text-aware pre-training for Text-VQA and Text-Caption. In **CVPR**, pp. 8747–8757, 2021.

A 付録

A.1 関連研究

画像説明文生成の研究は広く行われている [13–16]. 従来の画像説明文生成タスクでは, 画像内のテキスト情報の理解が必要となるタスクは少ない [17, 18].

テキストに基づいた画像説明文生成 (TextCaps) は, テキスト情報を含む画像に対して説明文を生成することを目的とする. そのため, 従来の画像説明文生成タスクで用いられているような画像ではなく, テキスト情報を含む画像に特化したタスクである. 故に, 一般的な画像理解を目指した従来モデルは本タスクへの適用が難しい. テキスト情報が重要となるため, TextCaps タスクでは OCR 特徴と物体領域を入力として用いるモデルが主流である [1, 2, 19].