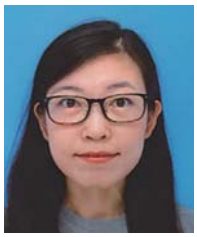
［研究助成（A）］

テキスト付き画像と感情ラベルを付与した画像向けの画像キャプショニングと質問応答

Generating Captions/ Answers with Multi- level Multimodal Encoder on Text- based Images with Reading Comprehension Tasks

2231033



研究代表者　慶応義塾大学 大学院理工学研究科　特任助教　楊　巍

［研究の目的］

実世界の画像を対象として説明文を生成する技術の実現は，人工知能分野における重要な課題の一つである。商品名等のテキストが印刷された商品パッケージや看板は日常生活に多く存在するため，テキストを含む画像に対する説明文生成は広範な応用を有する。また，ウェブサイト等における画像の代替テキスト生成，データ拡張［Magassouba 21］，視覚のバリアフリー化促進の手段としても有益である。

本研究では Image captioning with reading comprehension タスク［Sidorov 20］を扱う。本タスクの対象は，画像に含まれるテキストを含む画像を入力として，説明文を生成することである。本タスクの具体例を図 1 に示す。図において，緑/ピンク/赤矩形で囲っている領域は，

図 1　TextCaps タスクに対する具体例

それぞれ入力として画像中の検出物体，OCR 領域と認識された OCR トークン，画像全体を表す。画像全体中の検出された物体と認識された OCR トークン・領域の配置関係（“sign” と “salzburg/2/km” と “the top one”）を理解した上で，適切な説明文（“a collection of road signs, the top one reads salzburg 2 km”）を生成することが望ましい。

一般的な画像説明文生成に比べ，本タスクは，画像全体・画像中の物体・テキストを統合する点において挑戦的課題である。これまでに本タスクに関する既存研究は数多く行われている［Zhu 21, Sidorov 20］．［Sidorov 20］では OCR 情報と画像中の物体群を入力とする Transformer モデルであり，TextCaps データセットを用いて検証を行った。しかし，OCR 情報と物体間の関係を明示的に対応付けることが不十分であるとともに，bi-modal 特徴量の扱いに関する計算量およびパラメータが多いという問題があった。また，［Zhu 21］では，画像中のテキスト情報と物体間の関係のモデル化が考慮されているが，画像全体・画像中に出現するテキスト・物体間の関係のモデル化が不十分と考えられる。

提案手法では IRA ブロック（image-related attention block）を導入し，画像全体・画像中のテキスト・検出物体間の関係をモデル化する。ゆえに，生成文は画像中のテキストと物体のみ

に依存するものではない。加えて，画像中には標準的な単語だけではなくURLなどの文字列も含まれる可能性があるため，subwordレベルに基づく特徴を考慮する。これにより，単語レベルと文字レベル（低レベル）だけでなく，柔軟性があるsubwordレベルに基づく特徴を利用可能である。

［研究の内容，成果］

1. 問題設定

本タスク，画像内容に合った説明文を自動的に生成するタスクであり，特に，説明文が画像中に含まれるテキスト情報とそれに関する推論を含んでいる必要がある。画像から認識されたOCRトークンの特徴と画像全体の視覚的な情報を統合した上で，説明文を生成することが望ましい。

本タスクの入出力を以下のように定義する．

入力：画像全体，画像中の各OCR領域・トークン，画像中の各物体

出力：画像中に出現するテキストを含む画像に対する説明文

2. 提案手法

提案手法はTextCapsタスクにおいて良好な性能が報告されているSsbaseline［Zhu 21］を拡張したモデルである。複数視覚モダリティおよび言語情報の間の関係を有効的に統合し，モデリング化することは，Visual-and-Language（V&L）関連の問題を解決するための基本戦略と考えられる。したがって，提案手法で行った拡張は，言語（e. g., 質問，OCRトークン，指示文など）および画像関連のモダリティ（e. g., 画像全体，画像中の物体/OCR領域，セグメンテーションされた画像など）を入力として考慮した他の既存のV&Lモデルにも広く適用することができると考えられる。

提案手法ネットワークの構造を図2に示す。提案モデルは二つのモジュールから構成される。

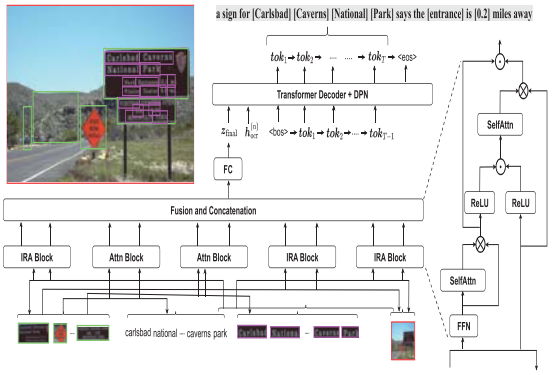


図2　提案手法のネットワーク構造

一つ目は注意機構に基づき，マルチモーダルOCRを含む複数モダリティの融合を目指して構築されたエンコーダモジュールである（図2の下段）。二つ目はTransformerアーキテクチャに基づき，Dynamic Pointer Network（DPN）による固定語彙と画像中のOCRトークンから選択可能となるデコーダモジュールである（図2の上段）。

モデルの入力Xを以下のように定義する。

$$\boldsymbol{X} = \boldsymbol{X}_{img}, \boldsymbol{X}_{ocr,s}, \boldsymbol{X}_{ocr,v}, \boldsymbol{X}_{obj}$$

$\boldsymbol{X}_{img}$は画像全体，$\boldsymbol{X}_{ocr,s}$と $\boldsymbol{X}_{ocr,v}$はそれぞれ認識された各OCRトークンとその領域を表す。また，$\boldsymbol{X}_{obj}$は検出された画像中の各物体領域である。

画像全体の特徴量はCLIP［Radford 21］（バックボーンRN50x4）モデルの全結合層を利用して640次元の特徴量 $\boldsymbol{x}_{glob} \in \mathbb{R}^{604}$を抽出する。さらに $\boldsymbol{h}_{glob} = f_{LN}(\boldsymbol{W}_{glob}\, \boldsymbol{x}_{glob})$ を得る。ここで，$\boldsymbol{f}_{LN}(\cdot)$ は Layer normalization を表し，$\boldsymbol{W}_{glob}$は重み行列を表す。

画像中のOCRトークンの言語的な意味を捉えるため，n（n＝1, …, N）番目のOCRトークンに対して，Pyramidal Histogram of Characters（PHOC）［Almazán 14］，同CLIP（RN50x4）モデルとFastTextに基づき，それぞれ文字，subwordと単語，様々な粒度（レベル）を考慮して3種類の特徴量を抽出する。

一方，画像中のn番目のOCRトークンは画像全体における配置を理解するために，OCR

トークンの相対位置（それぞれ領域の左上および右下の座標，画像の幅および高さ）を利用する。そして，OCR 領域の視覚的表現を利用するため，n 番目の OCR 領域は画像として Faster R-CNN［Ren 15］に渡し，2048 次元の fc6 層の特徴ベクトルを獲得する。さらに，fc7 層の重みは本説明文生成タスクとデータセットに対してファインチューニングされる。

m（m＝1, …, M）番目の物体領域の特徴抽出にも OCR 領域と同様，Faster R-CNN モデルとファインチューニングが適用される。以上の特徴ベクトルもそれぞれ $\boldsymbol{f}_{LN}(\cdot)$を行って，$\boldsymbol{h}^{(n)}_{ocr,s}$, $\boldsymbol{h}^{(n)}_{ocr,v}$, $\boldsymbol{h}^{(m)}_{obj}$を得られる。さらに，OCR と関連がある特徴（画像全体を含む）を考慮して，$\boldsymbol{h}^{(n)}_{ocr}$を獲得する。

Image-related Attention（IRA）Block を含むエンコーダモジュールは，複数モダリティの関係をモデリングするための 5 つの Attention ブロックから構成されている。各 Attention ブロックの入力は $\boldsymbol{h}_{glob}$, $\boldsymbol{h}^{(n)}_{ocr,s}$, $\boldsymbol{h}^{(n)}_{ocr,v}$, $\boldsymbol{h}^{(n)}_{ocr}$ と $\boldsymbol{h}^{(m)}_{obj}$ から構成される。

3. 実験

3.1 データセット

本研究で取り扱う TextCaps データセット［Sidorov 20］は，テキスト付き画像説明文生成タスクにおける標準データセットである。訓練・検証・評価セットは，それぞれ 21,953, 3,166 と 3,289 の画像が含まれている。単一の画像に対して複数（5 つ）のテキストに関する説明文が付与されている。評価セットの正解文が公開されていないため，それぞれ 109,765 と 3,166 の説明文が含まれている訓練と検証セットを用いた。訓練集合を提案モデルのパラメータ更新に使用し，検証セットでモデルの評価を行った。実験は Zhu ら［Zhu 21］と同様，標準的な手順に従っており，検証セットはトレーニングやハイパーパラメータの調整に使用していない。説明文の品質評価では，BLEU-1-4［Papineni 02］，METEOR［Banerjee 05］，ROUGE-L［Lin 04］，CIDEr［Vedantam 15］，SPICE［Anderson 16］を算出した。データセット内と画像の参考文群に出現する頻度を評価可能となる CIDEr が主要尺度として用いられた。

3.2 実験設定

ネットワーク内の Attention block は，層数が 8，隠れ層の次元数が 768，Head 数が 12 であった。最適化には Adam を使用し，学習率は 5×10^{-4}，Max interation 数は 12,000，バッチサイズは 128 であった。なお，固定語彙サイズは 6,736 であり，画像ごと OCR トークン数と物体数は，ぞれぞれ 50 と 100 であった。SBD-Trans［Liu 19, Wang 19］によって，OCR トークンを認識された。

提案手法のパラメータ数は 244M である。学習にはメモリ 16GB 搭載の Tesla V100 GPU 4 枚を使用した。一方で，推論に同 GPU1 枚を用いた。訓練に要した平均時間は 10 時間，1 サンプルあたりの学習と推論に要した平均時間がそれぞれ 22 ms/iteration と 79.4±1.09 ms であった。

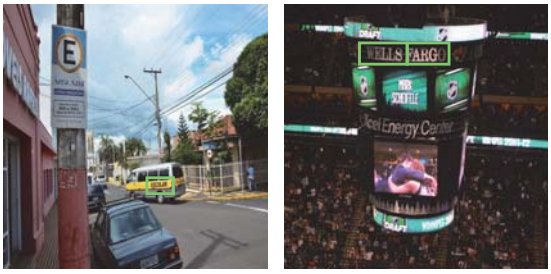
3.3 実験結果

表 1 に，提案手法とベースライン手法の定量的比較結果を示す。ベースライン手法として，M4C［Sidorov 20］と Ssbaseline［Zhu 21］を利用した。ベースライン（SSbaseline）と提案手法についてはそれぞれ 5 回の実験における平均と標準偏差を示してある。

表 1 に示すように，BLEU-1 から 4 について，SSbaseline はそれぞれ 69.56%, 49.78%, 34.96%, 24.54% であり，提案手法はこのベースライン手法をそれぞれ 1.86, 1.96, 1.76, 1.46 ポイントで上回ったことが確認できる。

表 1　ベースライン手法との定量的比較結果　［%］

Method		$\boldsymbol{z}^{(n)}_{ocr,clip}$	$\boldsymbol{z}_{glob}$	IRA	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
M4C					-	-	-	23.30	22.00	46.20	89.60	15.60
SSbaseline					-	-	-	24.89	22.71	47.24	98.83	15.71
SSbaseline (repd.)					69.56±0.30	49.78±0.16	34.96±0.16	24.54±0.17	22.68±0.07	47.12±0.07	97.76±0.19	15.74±0.08
Ours (full)		✓	✓	✓	**71.42**±0.20	**51.74**±0.12	**36.72**±0.10	**26.00**±0.01	**23.40**±0.06	**48.16**±0.14	**100.44**±0.50	**16.70**±0.09
Ours	(i)		✓		70.98±0.13	51.28±0.17	36.32±0.21	25.58±0.22	23.18±0.15	47.88±0.13	99.40±0.06	16.54±0.05
	(ii)	✓			69.88±0.10	50.12±0.12	35.36±0.12	24.88±0.17	22.78±0.07	47.28±0.07	98.66±0.37	15.88±0.10
	(iii)	✓	✓		71.04±0.22	51.28±0.17	36.24±0.14	25.58±0.12	23.24±0.05	47.90±0.13	99.74±0.50	16.60±0.04



(a)
Ours: a yellow van with the word [escolar] on the side
SSbaseline: a yellow car is parked in front of a building that says escolar
GT: yellow and gray van that has the word escolar on it

(b)
Ours: a scoreboard at a stadium that says ' [wells] [fargo] ' on it
SSbaseline: a scoreboard at a sports game that says ' wells ' on it
GT: stadium with a scoreboard that says wells fargo on it

図 3　TextCaps における定性的結果（成功例）

METEOR，ROUGE-L と SPICE スコアについては，SSbaseline と比較して，それぞれ 0.72，1.04，0.96 ポイントで改善した。特に，主要尺度として用いられた CIDEr スコアは 2.68 ポイントで大きく上回った。M4C を含むベースラインのいずれにおいても，すべての評価尺度で提案手法が既存手法を上回るスコアを示しており，提案手法の有効性が確認された。

図 3 に提案モデルおよびベースライン（SSbaseline）の出力例を示す。図 3(a) では，提案モデルは画像中の物体と OCR トークンが説明文に必要な関係（“van” with the word “escolar”）が含まれる適切的な記述を生成できたが，ベースラインでは，物体と OCR トークンの関係の表現（“car” と “escolar”）を誤って説明文を生成していた。そして，図 3 (b) では，提案手法（“says ‘wells fargo’ on it”）に比べてベースライン手法（“says ‘wells’ on it”）は表現が完全ではないという面で劣る。

テキスト付き画像説明文生成タスクにおいて，提案手法の性能をさらに検証するために，TextCaps データセット以外のデータセット ST-VQA（task-1）[Biten 19] を用い，zero-shot 予測を行った。図 4 に，提案モデルに基づく出力例を示す。例えば，図 4(b) の画像に対して，検出された物体（“sign”）と認識された OCR トークン（“salem” and “Portland”）の間の関係を推論することが可能となり，提案

(a)
Pred: a bus that says ' [more] [destinations] ' on the side of it

(b)
Pred: a green sign that says [salem] and [portland] on it

図 4　提案モデルを用い，ST-VQA（task-1）における予測例

モデルは適切に文を生成できている。

Ablation Study には，以下の 3 条件を定めた

（ i ）w/o CLIP for OCR tokens: OCR トークン特徴においては，CLIP を利用する言語特徴を導入する場合としない場合で，性能の比較を行った。

（ ii ）w/o CLIP for Image: 画像全体に関する CLIP 特徴量の有無について，性能への影響を調べた。

（iii）w/o image-related attention （IRA） blocks: 画像全体，OCR 情報，画像中の物体群の間の相互注意の有無に性能への影響を調べた。

表 1 に示すように，条件（iii）vs.（ i ）に関して評価尺度（BLEU-4 から SPICE まで）がそれぞれ 0 %，0.06%，0.02%，0.34%，0.06% 減少していた。一方で，条件（iii）vs.（ ii ）によって各指標がそれぞれ 0.7%，0.46%，0.62%，1.08%，0.72% で減少している。すなわち，画像全体に関する特徴を導入したことで，全指標で説明文生成の性能が向上した。また，特に性能向上に寄与していた要素は画像全体であったと考えられる。さらに，条件（iii）vs. Ours（full）によって，Ours（full）の方，各評価尺度がそれぞれ 0.42%，0.16%，0.26%，0.7%，0.1% を上回った。これより，画像中の OCR トークン情報，物体群の間の統合的処理がモデルの性能向上に有益であったことがわかる。

4. まとめ

本研究ではテキスト付き画像の説明文生成タスクを扱った。提案手法による貢献は以下である。

・既存手法と異なり，画像全体と複数粒度のマルチモーダル OCR 特徴を導入した。
・複数のモダリティの利用可能となり，画像全体と関連がある複数のモダリティの自己注意を導入し，画像全体と複数のモダリティの間の相互注意機構を導入した。
・TextCaps データセットにおいて，提案手法がベースライン手法を全ての評価尺度で上回った。
・質問応答に関する研究は［成果の発表，論文など］の［3］となる。

［成果の発表，論文など］

［1］ Switching Text-based Image Encoders for Captioning Images with Text, Arisa Ueda, Wei Yang (90%), Komei Sugiura, *IEEE Access*, 11: 55706-55715, 2023. 6.

［2］ 複数粒度のマルチモーダル情報を用いたテキスト付き画像の説明文生成 (Generating Captions with Multi-level Multimodal Encoder on Image Captioning with Reading Comprehension Tasks), Wei Yang (90%), Arisa Ueda, Komei Sugiura, 2023 年度人工知能学会全国大会（第 37 回), Kumamoto, 2023. 6.

［3］ Multimodal Encoder with Gated Cross-attention for Text-VQA Tasks, Wei Yang (90%), Arisa Ueda, Komei Sugiura, 言語処理学会第 29 回年次大会，pp. 1580-1585, 2023. 3

［4］ 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，言語処理学会第 29 回年次大会，pp. 425-430, 2023. 3

［参考文献］

［Almazán 14］ Almazán, J., et al.: Word spotting and recognition with embedded attributes, *IEEE Trans. PAMI*, Vol. 36, No. 12, pp. 2552-2566（2014）

［Anderson 16］ Anderson, P., et al.: SPICE: Semantic propositional image caption evaluation, in *ECCV*, pp. 382-398（2016）

［Banerjee 05］ Banerjee, S. and Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in *ACL Workshop*, pp. 65-72（2005）

［Biten 19］ Biten, A. F., et al.: Scene text visual question answering, in *ICCV*, pp. 4291-4301（2019）

［Lin 04］ Lin, C. -Y.: ROUGE: A package for automatic evaluation of summaries, in *Text Summarization Branches Out*, pp. 74-81（2004）

［Liu 19］ Liu, Y., et al.: Omnidirectional scene text detection with sequential-free box discretization, in *IJCAI*, pp. 3052-3058（2019）

［Magassouba 21］ Magassouba, A., Sugiura, K., and Kawai, H.: CrossMap Transformer: A Crossmodal Masked Path Transformer Using Double Back-Translation for Vision-and-Language Navigation, *IEEE RA-L*, Vol. 6, No. 4, pp. 6258-6265（2021）

［Papineni 02］ Papineni, K., et al.: BLEU: A method for automatic evaluation of machine translation, in *ACL*, pp. 311-318（2002）

［Radford 21］ Radford, A., et al.: Learning transferable visual models from natural language supervision, in *ICML*, pp. 8748-8763（2021）

［Ren 15］ Ren, S., He, K., Girshick, R., and Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks, *NeurIPS*, Vol. 28,（2015）

［Sidorov 20］ Sidorov, O., Hu, R., Rohrbach, M., and Singh, A.: TextCaps: A dataset for image captioning with reading comprehension, in *ECCV*, pp. 742-758（2020）

［Vedantam 15］ Vedantam, R., Lawrence Zitnick, C., and Parikh, D.: CIDEr: Consensus-based image description evaluation, in *CVPR*, pp. 4566-4575（2015）

［Wang 19］ Wang, P., et al.: A Simple and Robust Convolutional-Attention Network for Irregular Text Recognition, *arXiv preprint arXiv : 1904. 01375*（2019）

［Zhu 21］ Zhu, Q., Gao, C., Wang, P., and Wu, Q.: Simple is not easy: A simple strong baseline for TextVQA and TextCaps, in *AAAI*, Vol. 35, pp. 3608-3615（2021）