複数粒度のマルチモ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

楊 巍 Wei Yang　植田 有咲 Arisa Ueda　杉浦 孔明 Komei Sugiura

慶應義塾大学 Keio University

Visual scene understanding, such as image captioning, can be considered one of the essential topics in the artificial intelligence (AI) field. Image captioning with reading comprehension tasks as an extension of traditional image captioning is more challenging because the generated caption must be related to the text information in the image, and how to read and comprehend text in the context of an image needs to be studied. In this work, we propose multiple image-related attention blocks with multimodal Optical Character Recognition (OCR) information to model the relationship among the global image, multi-level recognized text, and the detected objects in the image. Our model is validated on the standard dataset TextCaps, and the results show that our model outperforms the baseline methods in terms of all evaluation matrices.

1. はじめに

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（alt 属性）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, 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.

本研究の新規性は以下の通りである.

- 画像中のテキスト情報に関して文字から単語まで複数レベルかつマルチモーダル情報として統合する.
- 画像全体の大域的な視覚情報, マルチモーダル OCR 特徴, 画像中の物体群, の三者を扱うマルチモーダル注意機構を導入する.

図 1: TextCaps タスクに対する具体例.

2. 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

連絡先: Wei Yang，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，wei.yang@keio.jp

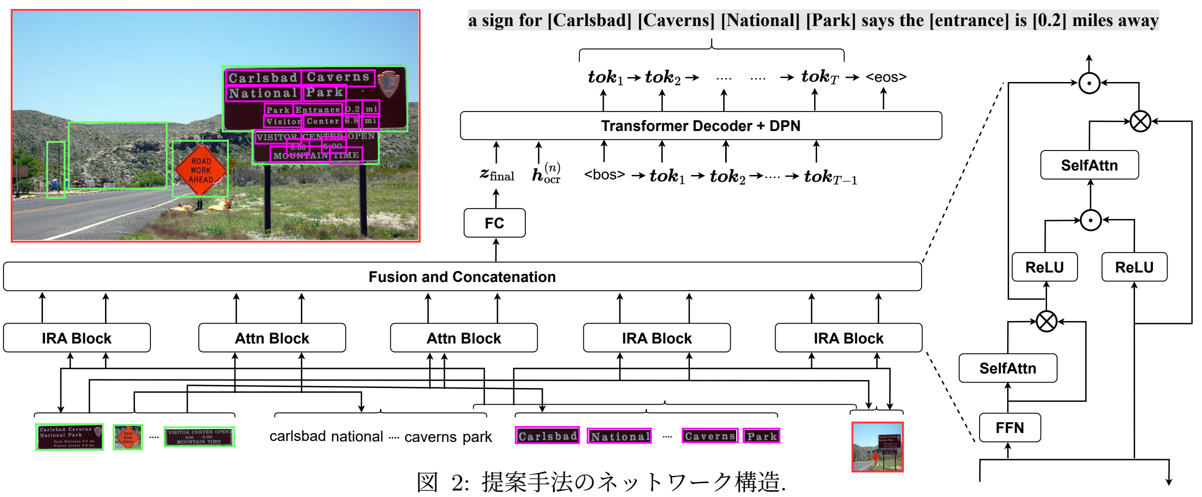


図 2: 提案手法のネットワーク構造.

3. 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. 一つ目は 3.2 節に示す注意機構に基づき, 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, 3.3 節).

3.1 入力

モデルの入力 X を以下のように定義する.

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

$\boldsymbol{X}_{\text{img}}$ は画像全体, $\boldsymbol{X}_{\text{ocr,s}}$ と $\boldsymbol{X}_{\text{ocr,v}}$ 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, $\boldsymbol{X}_{\text{obj}}$ 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 $\boldsymbol{x}_{\text{glob}} \in \mathbb{R}^{640}$ を抽出する. さらに, $\boldsymbol{h}_{\text{glob}} = f_{\text{LN}}(\boldsymbol{W}_{\text{glob}}\boldsymbol{x}_{\text{glob}})$ を得る. ここで, $f_{\text{LN}}(\cdot)$ は Layer normalization を表し, $\boldsymbol{W}_{\text{glob}}$ 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 $(n = 1, \cdots, N)$ 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, sub-word と単語, 様々な粒度 (レベル) を考慮して 3 種類の特徴量 $\boldsymbol{x}^{(n)}_{\text{ocr,phoc}} \in \mathbb{R}^{604}$, $\boldsymbol{x}^{(n)}_{\text{ocr,clip}} \in \mathbb{R}^{640}$, $\boldsymbol{x}^{(n)}_{\text{ocr,ft}} \in \mathbb{R}^{300}$ とを抽出する. それにより, $\boldsymbol{h}^{(n)}_{\text{ocr,s}}$ は下記の様に求める.

$$\boldsymbol{h}^{(n)}_{\text{ocr,s}} = f_{\text{LN}}(\boldsymbol{W}_{\text{ph}}\boldsymbol{x}^{(n)}_{\text{ocr,phoc}} + \boldsymbol{W}_{\text{cl}}\boldsymbol{x}^{(n)}_{\text{ocr,clip}} + \boldsymbol{W}_{\text{ft}}\boldsymbol{x}^{(n)}_{\text{ocr,ft}}) \quad (1)$$

ここで, $\boldsymbol{W}_{.}$ は重み行列を表す.

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 $\boldsymbol{x}^{(n)}_{\text{ocr,bx}}{=}[x^{(n)}_{\min}/W, y^{(n)}_{\min}/H, x^{(n)}_{\max}/W, y^{(n)}_{\max}/H]$ を利用する. $(x^{(n)}_{\min}, y^{(n)}_{\min})$, $(x^{(n)}_{\max}, y^{(n)}_{\max})$ はそれぞれ領域の左上および右下の座標, W, H は画像の幅および高さを表す. 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 $\boldsymbol{x}^{(n)}_{\text{ocr,fr}}$ 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. $\boldsymbol{h}^{(n)}_{\text{ocr,v}}$ は次のようにして得る.

$$\boldsymbol{h}^{(n)}_{\text{ocr,v}} = f_{\text{LN}}(\boldsymbol{W}_{\text{ocr,fr}}\boldsymbol{x}^{(n)}_{\text{ocr,fr}}) + f_{\text{LN}}(\boldsymbol{W}_{\text{ocr,bx}}\boldsymbol{x}^{(n)}_{\text{ocr,bx}}) \quad (2)$$

m $(m = 1, \cdots, M)$ 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.

$$\boldsymbol{h}^{(m)}_{\text{obj}} = f_{\text{LN}}(\boldsymbol{W}_{\text{obj,fr}}\boldsymbol{x}^{(m)}_{\text{obj,fr}}) + f_{\text{LN}}(\boldsymbol{W}_{\text{obj,bx}}\boldsymbol{x}^{(m)}_{\text{obj,bx}}) \quad (3)$$

さらに, OCR と関連がある特徴（画像全体を含む）を考慮して, $\boldsymbol{h}^{(n)}_{\text{ocr}}$ を獲得する.

$$\begin{aligned}\boldsymbol{h}^{(n)}_{\text{ocr}} = f_{\text{LN}}(&\boldsymbol{W}_{\text{p}}\boldsymbol{x}^{(n)}_{\text{ocr,phoc}} + \boldsymbol{W}_{\text{c}}\boldsymbol{x}^{(n)}_{\text{ocr,clip}} + \boldsymbol{W}_{\text{f}}\boldsymbol{x}^{(n)}_{\text{ocr,ft}} \\ &+ \boldsymbol{W}_{\text{fr}}\boldsymbol{x}^{(n)}_{\text{ocr,fr}} + \boldsymbol{W}_{\text{gl}}\boldsymbol{x}_{\text{glob}}) + f_{\text{LN}}(\boldsymbol{W}_{\text{bx}}\boldsymbol{x}^{(n)}_{\text{ocr,bx}}) \quad (4)\end{aligned}$$

3.2 IRA を含むエンコーダモジュール

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 $\boldsymbol{h}_{\text{glob}}$, $\boldsymbol{h}^{(n)}_{\text{ocr,s}}$, $\boldsymbol{h}^{(n)}_{\text{ocr,v}}$, $\boldsymbol{h}^{(n)}_{\text{ocr}}$ と $\boldsymbol{h}^{(m)}_{\text{obj}}$ から構成される. ここでは, 画像中のマルチモダル OCR 特徴 $\boldsymbol{h}^{(n)}_{\text{ocr}}$ と物体 $\boldsymbol{h}^{(m)}_{\text{obj}}$ の間の IRA ブロックを例として定義する. 出力はそれぞれ画像中のマルチモダル OCR 特徴の自己注意に基づく重み付き和 $\boldsymbol{X}'_{\text{ocr}}$ と, 物体特徴 $\boldsymbol{X}'_{\text{obj}}$ である. これ以外の 4 つの Attention ブロックを用いて, $\boldsymbol{h}^{(m)}_{\text{obj}}$ と $\boldsymbol{h}^{(n)}_{\text{ocr,v}}$, $\boldsymbol{h}^{(m)}_{\text{obj}}$ と $\boldsymbol{h}^{(n)}_{\text{ocr,s}}$, $\boldsymbol{h}^{(m)}_{\text{obj}}$ と $\boldsymbol{h}_{\text{glob}}$, $\boldsymbol{h}^{(n)}_{\text{ocr}}$ と $\boldsymbol{h}_{\text{glob}}$ を同様にして得る. まず, 入力埋め込み系列をカーネルサイズが 1 である 2 つの畳み込み層と全結合 Feed-Forward Network （FFN）に入力し, $\tilde{\boldsymbol{h}}^{(n)}_{\text{ocr}}$ を獲得する.

$$\tilde{\boldsymbol{h}}^{(n)}_{\text{ocr}} = \text{Conv1D}\left(\text{ReLU}\left[\text{Conv1D}(\boldsymbol{h}^{(n)}_{\text{ocr}}, 1)\right]\right), \; n = 1, \ldots, N$$

さらに, 自己注意 (SelfAttn) を適用する. SelfAttn は以下で与えられる.

$$\text{SelfAttn}(\boldsymbol{h}) = \text{softmax}\left(\frac{\boldsymbol{h}\boldsymbol{W}_Q\boldsymbol{W}_K^{\top}\boldsymbol{h}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{h}\boldsymbol{W}_V \quad (5)$$

ここで, $\boldsymbol{W}$. は学習可能な重みである．d_k=H/heads, H は隠れ層のサイズ, heads は自己注意機構の Head 数を表す. ここで, 例えば, $\boldsymbol{h}_{\mathrm{ocr}}^{(n)}$ は複数 Attention ブロックに入力するため, 独立したパラメータで自己注意を並列に実行する.

$$\boldsymbol{\alpha}_n = \mathrm{SelfAttn}(\tilde{\boldsymbol{h}}_{\mathrm{ocr}}^{(n)}) \tag{6}$$

Attention ブロックの 1 つ目の出力 $\boldsymbol{X}'_{\mathrm{ocr}}$ は以下の式で得られる.

$$\boldsymbol{X'}_{\mathrm{ocr}} = \sum_{i=1}^{N} \boldsymbol{\alpha}_n \, \tilde{\boldsymbol{h}}_{\mathrm{ocr}}^{(n)} \tag{7}$$

さらに, $\boldsymbol{X}'_{\mathrm{ocr}}$ はモダリティの間の Attention 重みを算出するための誘導として使われる.

$$\boldsymbol{u}_m = \boldsymbol{W}[\mathrm{ReLU}(\boldsymbol{W}_{\mathrm{oc}}\boldsymbol{X'}_{\mathrm{ocr}}) \odot \mathrm{ReLU}(\boldsymbol{W}_{\mathrm{ob}}\boldsymbol{h}_{\mathrm{obj}}^{(m)})] \tag{8}$$

$$\boldsymbol{\beta}_m = \mathrm{SelfAttn}(\boldsymbol{u}_m), m = 1, \ldots, M \tag{9}$$

ここで, M は画像中の検出された物体数である．$\boldsymbol{h}_{\mathrm{obj}}^{(m)}$ と $\boldsymbol{h}_{\mathrm{ocr}}^{(n)}$ を加算する場合, 当該値は画像中に認識されたの OCR トークン数 N となる。最後に, $\boldsymbol{X}'_{\mathrm{obj}}$ に関する加重合計は以下のように求める．これは Attention ブロックの 2 つ目の出力となる。

$$\boldsymbol{X}'_{\mathrm{obj}} = \sum_{m=1}^{M} \boldsymbol{\beta}_m \boldsymbol{h}_{\mathrm{obj}}^{(m)} \tag{10}$$

そして, 各 Attention ブロックの出力に対して要素ごとの乗算 (アダマール積) が適用され, それぞれ得られた 5 つの, $\boldsymbol{z}_{\mathrm{ocr}\to\mathrm{obj}}$(例), $\boldsymbol{z}_{\mathrm{obj}\to\mathrm{ocr,v}}$, $\boldsymbol{z}_{\mathrm{obj}\to\mathrm{ocr,s}}$, $\boldsymbol{z}_{\mathrm{obj}\to\mathrm{img}}$ と $\boldsymbol{z}_{\mathrm{ocr}\to\mathrm{img}}$ を連結した後, さらに, 全結合層に入力することで, 最終的な融合が行われるコンテクスト出力 $\boldsymbol{z}_{\mathrm{final}}$ を得る.

$$\boldsymbol{z}_{\mathrm{ocr}\to\mathrm{obj}} = \boldsymbol{X'}_{\mathrm{ocr}} \odot \boldsymbol{X}'_{\mathrm{obj}} \tag{11}$$

3.3 DPN デコーダモジュール

固定語彙 ($v = 1, \cdots, V$) からまたは画像中の OCR トークン ($n = 1, \cdots, N$) からのコピーで代わりに説明文中のトークンを生成するために, dynamic pointer network (DPN) が含まれる Transformer デコーダを導入する．固定語彙または OCR トークンから選択される予測確率は以下で与えられる.

$$p(\hat{\boldsymbol{y}}_{t,v}^{\mathrm{voc}}) = \mathrm{softmax}\big((\boldsymbol{w}_{\mathrm{voc}}^{(v)})^\top \boldsymbol{z}_{\mathrm{dec}}^{(t)} + \boldsymbol{b}_{\mathrm{voc}}^{(v)}\big)$$

上記の線形演算において, $\boldsymbol{w}_{\mathrm{voc}}^{(v)}$ は固定語彙内の v 番目の単語の d 次元パラメータ, $\boldsymbol{b}_{\mathrm{voc}}^{(v)}$ はスカラーパラメータである.

$$p(\hat{\boldsymbol{y}}_{t,n}^{\mathrm{ocr}}) = \mathrm{softmax}\big((\boldsymbol{W}_{\mathrm{ocr}}\boldsymbol{z}_{\mathrm{ocr}}^{(n)} + b_{\mathrm{ocr}})^\top(\boldsymbol{W}_{\mathrm{dec}}\boldsymbol{z}_{\mathrm{dec}}^{(t)} + b_{\mathrm{dec}})\big)$$

同様に上記において, $\boldsymbol{W}_{\mathrm{ocr}}$ および $\boldsymbol{W}_{\mathrm{dec}}$ は $d \times d$ 重みの行列を表す．$\boldsymbol{z}_{\mathrm{ocr}}^{(n)}(n = 1, \ldots, N)$ は画像中の n 番目の OCR トークンの d 次元の Transformer 出力を表す．b_{ocr} と b_{dec} は d 次元のベクトルとなる.

この二つの式に対して, 時刻 t におけるデコーダ埋め込み (d 次元の $\boldsymbol{z}_{\mathrm{dec}}^{(t)}$) は, 直前時刻 t-1 で予測されたトークンとその表現 ($\boldsymbol{x}_{\mathrm{dec}}^{(t)}$) によって決定される．すなわち, 時刻 t-1 における予測が固定語彙からの単語の場合, その対応する重みベクトル $\boldsymbol{w}_{\mathrm{voc}}^{(v)}$ を Transformer 入力として用いる. 一方, 時刻 t-1 における予測が OCR トークンからの場合, OCR 表現 $\boldsymbol{h}_{\mathrm{ocr}}^{(n)}$ が時刻 t のデコーダ埋め込みになり, Transformer 入力となる. それに加えて, 時刻 t-1 に予測されたトークンの d 次元の位置埋め込みとタイプ埋め込み (0: 固定語彙, 1: OCR トークン) も考慮する．最初の時刻 t=0 の予測に対しての入力はエンコーダモジュールから最終的なコンテクスト出力となる. 最後に, V 次元の $p(\hat{\boldsymbol{y}}_{t,v}^{\mathrm{voc}})$ ($v = 1, \cdots, V$) と N 次元の $p(\hat{\boldsymbol{y}}_{t,n}^{\mathrm{ocr}})$ ($n = 1, \ldots, N$) を連結して, V+N 個 (固定語彙 + OCR トークン) の候補から最も予測確率が高い要素を擬似ラベルとして出力する．損失関数として, 単語ごとの交差エントロピーを用いた.

4. 実験

4.1 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.

4.2 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 GPU 1 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 22ms/iteration と 79.4 ± 1.90 ms であった.

4.3 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.

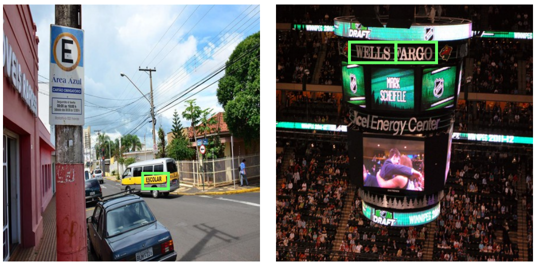
表 1 に示すように, BLEU-1 から 4 について, SSbaseline はそれぞれ 69.56%, 49.78%,, 34.96%,, 24.54%, であり, 提案手法はこのベースライン手法をそれぞれ 1.86, 1.96, 1.76, 1.46 ポイントで上回ったことが確認できる．METEOR, ROUGE-L と SPICE スコアについては, SSbaseline と比較して, それぞれ 0.72, 1.04, 0.96 ポイントで改善した. 特に, 主要尺度として用いられた CIDEr スコアは 2.68 ポイントで大きく上回った. M4C を含むベースラインのいずれにおいても, 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”) が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

表 1: ベースライン手法との定量的比較結果 [%].

Method		$\boldsymbol{x}^{(n)}_{\mathrm{ocr,clip}}$	$\boldsymbol{x}_{\mathrm{glob}}$	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±0.12	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).

(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.

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.

5. おわりに

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.

謝辞

本研究の一部は, JSPS 科研費 20H04269, 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

[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)