Title	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
Sub Title	
Author	植田, 有咲
Publisher	慶應義塾大学AI・高度プログラミングコンソーシアム
Publication year	2023
Jtitle	AICカンファレンス予稿集 (2023.) ,p.23- 24
JaLC DOI	
Abstract	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.
Notes	会議名 : AICカンファレンス2023 開催地 : 慶應義塾大学日吉キャンパス 日時 : 2023年3月4日 第2章ポスター発表要旨 ポスター要旨-1
Genre	Conference Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO11003001-20230304-0023

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
慶應義塾大学大学院理工学研究科開放環境科学専攻

Abstract
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.
Keywords: Image Captioning, Multimodal Language Processing, Text-based Image Manipulation

1. 研究背景・目的

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. 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. Fig.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.

2. 方法

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. 提案手法の新規性は以下の通りである.

1. 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.
2. 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.

入力は画像であり, 出力は画像内のテキスト情報に関連する説明文である. 入力画像から特徴量抽出器を用いて全体画像特徴, 物体特徴, OCR 特徴を抽出する. 画像特徴量は事前学習済みの CLIP[1](RN50x4)を用いて特徴量抽出を行った. 物体特徴は Faster R-CNN[2]を用いて特徴量抽出を行った. OCR の言語特徴量は FastText, PHOC[3], 画像特徴量で用いたものと同様の CLIP で特徴量抽出を行った. 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つの相互注意を新しく導入する. これらの2つの相互注意の入力は全体画像の CLIP 特徴量, 物体特徴量, マルチモーダル OCR 特徴量である. それぞれ全体画像と物体間の関係, 全体画像とマルチモーダル OCR 特徴量間の関係モデル化する構造である. 最終的に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算出し, 相互注意の最終出力を得る. 説明文を生成するための復号器では

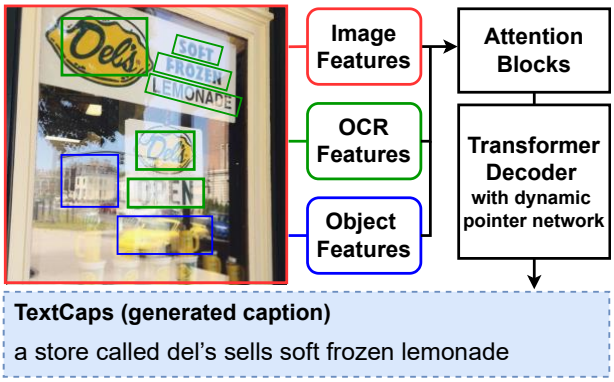

Fig.1 提案手法の概要図

Transformer[4]ベースの自己回帰型の復号器を用いる. 提案手法の復号器はポインタネットワーク(Dynamic Pointer Network)を導入することで予測トークンを固定語彙または画像中の OCR トークンから選択することが可能である.

3. 結果

3.1. 定量的結果

TextCaps データセット[5]を用いて提案手法の性能評価を行った. TextCaps データセットは 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.

Fig.1 に提案手法の定量的結果を示す.提案手法は既存手法に比べて2つの画像に関連する相互注意を導入することで各評価指標で約1ポイント程度性能が向上した.

3.2. 定性的結果

Fig.2 に提案手法と baseline 手法[5],[6]の定性的結果を示す. Fig.2(a)では baseline 手法は OCR トークン”whitestone”と物体“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. Fig.2(b)の baseline 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”

Table.1 TextCaps タスクにおける定量的結果

	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
M4C-Captioner (paper) [5]	23.30	22.00	46.20	89.60	15.60
SSbaseline (paper) [6]	24.89	22.71	47.24	98.83	15.71
Ours (full)	**26.00**$\pm$0.01	**23.40**$\pm$0.06	**48.16**$\pm$0.14	**100.44**$\pm$0.50	**16.70**$\pm$0.09

(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

Fig. 2 提案手法と baseline 手法の定性的結果

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．

4. 考察

Fig. 2 の提案手法と baseline 手法の定性的結果から提案手法は baseline 手法に比べて画像内の物体と OCR トークン間の関係を適切に表現できている．このことは，提案手法の新規性である画像全体，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が TextCaps タスクにおいて有効であることを示唆している．

5. 結論（と今後の展望）

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．提案手法では既存手法と異なり，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．また，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することで全ての評価指標において baseline 手法を上回る性能を得た．

参考文献

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., “Learning transferable visual models from natural language supervision,” in ICML, 2021, pp. 8748–8763.

[2] S.Ren,K.He,R.Girshick,andJ.Sun,“FasterR-CNN: Towards real-time object detection with region proposal networks,” NeurIPS, vol. 28, 2015.

[3] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, “Word spotting and recognition with embedded attributes,” IEEE Trans. PAMI, vol. 36, no. 12, pp. 2552–2566, 2014.

[4] A.Vaswani,N.Shazeer,N.Parmar,J.Uszkoreit,L.Jones, A.N.Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in NeurIPS, 2017, pp. 6000–6010.

[5] O.Sidorov, R.Hu, M.Rohrbach, and A.Singh, “TextCaps: A dataset for image captioning with reading comprehension,” in ECCV, 2020, pp. 742– 758.

[6] Q. Zhu, C. Gao, P. Wang, and Q. Wu, “Simple is not easy: A simple strong baseline for TextVQA and TextCaps,” in AAAI, vol. 35, no. 4, 2021, pp. 3608–3615.