文字情報を考慮したシーン画像検索

宮脇峻平 [1]　長谷川拓 [2]　西田京介 [2]　加藤拓真 [1]　鈴木潤 [1,3]
[1] 東北大学　[2]NTT 人間情報研究所　[3] 理化学研究所
{shumpei.miyawaki.q1,takuma.kato.r3}@dc.tohoku.ac.jp　jun.suzuki@tohoku.ac.jp
{taku.hasegawa.ps,kyosuke.nishida.rx}@hco.ntt.co.jp

概要

本研究では，Vision and Language の分野で，テキストと視覚領域の意味表現を紐づけるための手法として注目される，画像とテキストを独立したモデルでエンコードするデュアルエンコーダ型の検索モデルを用いて画像検索に取り組む．デュアルエンコーダでは，特に画像エンコーダが画像中の文字情報であるシーンテキストと視覚情報を融合して理解できているかどうかが明らかでないことから，TextCaps [1] を用いて画像検索を行い，シーンテキストの影響について調査を行う．実験の結果より，シーンテキストと周辺視覚情報の融合的な理解を助けるモデル化が効果的であることを示す．

1　はじめに

計算機によるテキストや視覚情報の適切な理解を実現することは，人工知能研究における最終目的の一つである．Vision and Language の分野では，この目的の実現に向けてテキストと視覚領域の意味表現の紐付けを行うための方法論が研究されている．一般的に，1) テキストと視覚情報を連結して一つの Transformer [2] ベースのモデルに入力し，異なるモダリティ間でアラインメントを学習するクロスエンコーダ，2) 画像・テキストに関する二つの独立したエンコーダを用いて対照学習の枠組みを導入してアラインメントを学習するデュアルエンコーダ，の二つの方法論が広く用いられる．クロスエンコーダは異なるモダリティの理解を行いやすいが，画像検索などの高速かつ大規模な推論が必要なタスクには適していないことが知られている [3]．

本研究では，高速に推論可能なデュアルエンコーダを用いてシーンテキストと視覚領域の意味表現を適切に紐づけることを目的として，画像中の文字情報を考慮する画像検索モデルを提案する．本研究の貢献は以下の通りである．

- デュアルエンコーダを対象に，画像中の文字情報と視覚情報を融合して理解するための拡張を行い，シーンテキストを周辺視覚情報から読むための事前学習を提案する．
- クロスエンコーダを用いた先行研究と同様に，実験結果から画像中の文字情報と周辺視覚情報の融合的な理解がモデルの検索性能に有効な影響を示すことを明らかにした．

2　関連研究

2.1　デュアルエンコーダによる検索手法

オープンドメイン質問応答タスクでは，Karpukhin ら [4] が提案したデュアルエンコーダを用いた密なベクトル意味空間での類似度に基づく検索手法である Dense Passage Retrieval（DPR）が広く用いられており，デュアルエンコーダにおけるモダリティの関係性についても研究されている [5, 6]．

デュアルエンコーダを用いた検索手法は画像検索でも研究されており [3, 7, 8, 9]，Jia ら [8] は WEB から収集した画像とテキストのペアからなる大規模な学習データを用いることで，線形演算可能なテキストと視覚情報の埋め込み表現を学習している．

またクロスエンコーダについても研究されており [10, 11, 12, 13]，UNITER [11] は Transformer [2] の注意機構を用いてテキストと視覚領域のクロスモーダルなアラインメントを学習している．

さらに多言語を考慮したモデルも注目されており [14, 15]，Huang ら [16] はデュアルエンコーダの枠組みを用いて多言語マルチモーダル検索における事前学習法を提案している．

2.2　画像中の文字情報のモデル化

標識や商品名など我々の身の回りにはシーンを説明するために有効なテキスト情報が溢れており，

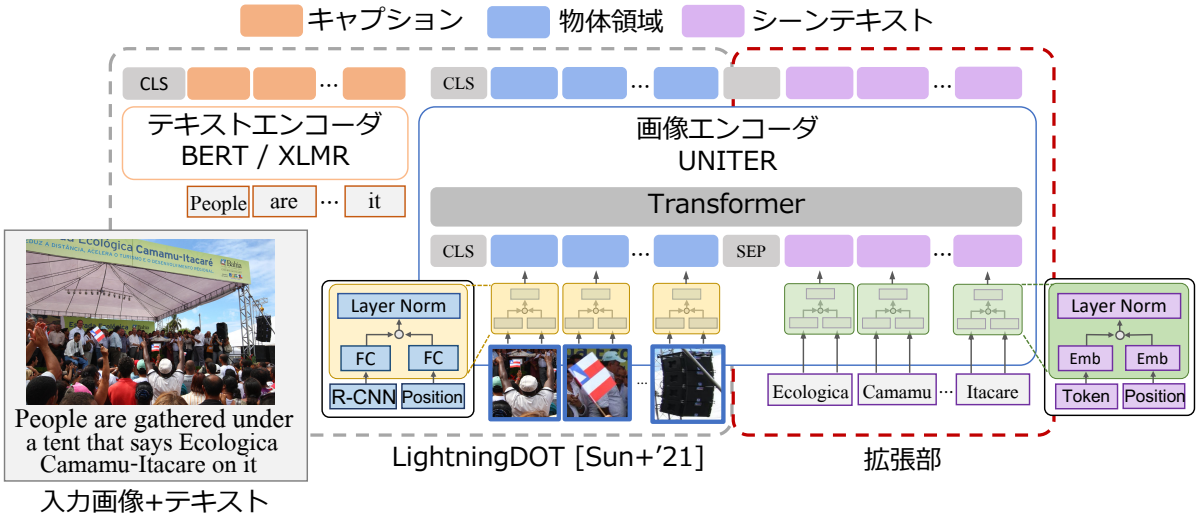

図 1　ベースラインモデルの概要

シーンテキストは類似するシーン画像に対してより詳細な意味情報を提供する [17, 18]．先行研究ではシーンテキストの意味情報として，テキストの言語情報や視覚情報だけでなく，文字単位の特徴量（PHOC） [19] やレイアウト [20, 21, 22, 23]，シーンテキスト間の位置情報 [24] などが使用される．しかしデュアルエンコーダによる画像中の文字情報と周辺視覚情報の融合した理解については明らかでなく，特にシーンテキストをモダリティの対象としない CLIP [7] では画像中の文字情報と物体の視覚情報の意味的な紐付けが難しい [25, 26]．

また撮影場所や生産地などに依存するシーンテキストはゼロショットの語彙となる場合があるため [1]，先行研究におけるシーンテキストはテキストエンコーダが持つ語彙とは異なるものとして扱われる [27, 28, 24]．また大規模なテキストコーパスを用いて事前学習を行うことでシーンテキストの語彙を広くカバーするモデルも提案されている [23]．

その他，シーンテキスト検出・検索タスクでは，従来より注意機構 [29] や Connectionist Temporal Classification（CTC）ロス [30] などが使用される．Wang ら [31] は CTC ロスに加え，シーンテキストおよび視覚表現のクロスモーダルな類似関係をモデル化するシーンテキスト検索モデルを提案している．

3　文字情報を考慮した画像検索

本章では，高速に推論可能なデュアルエンコーダを用いて画像中の文字情報を考慮した画像検索モデルを提案する．

3.1　ベースラインモデル

本研究ではデュアルエンコーダにおける詳細なアラインメントを学習することを目的として，画像全体を入力する CLIP [7] ではなく，Faster R-CNN を用いた物体検出器 [32] により分割された物体領域を扱う LightningDOT [3] をベースラインモデルとして用いる（図 1）．画像・テキストエンコーダには，UNITER [11] および BERT [33] を用いる．

LightningDOT では，画像 $\boldsymbol{v}$ とキャプション $\boldsymbol{w}$ のペアからなるデータセット D をそれぞれ画像エンコーダ E_v およびテキストエンコーダ E_w に入力する．そして，ベクトル化された二つの CLS ベクトルについて，対応するペアに対して内積値 $sim(\boldsymbol{w}, \boldsymbol{v}) = E_w(\boldsymbol{w}_{\mathrm{CLS}})^\top E_v(\boldsymbol{v}_{\mathrm{CLS}})$ が高くなるようモデルパラメータを最適化する Cross Modal Retrieval（CMR）を行う（式 1）．なお負例の選択についてはインバッチネガティブサンプリングを採用し，ミニバッチ B 内の他の画像とキャプションをそれぞれ負例として扱うこととする．

$$
\begin{aligned}
\mathcal{L}_{CMR}(B) &= \frac{1}{2B}\sum_{b=1}^{B} \mathcal{L}_{TR}^{(v_b)} + \mathcal{L}_{IR}^{(w_b)} \\
\mathcal{L}_{IR}^{(w_b)} &= -\log \frac{\exp^{sim(\boldsymbol{w}_b, \boldsymbol{v}_b)}}{\sum_{k=1}^{B} \exp^{sim(\boldsymbol{w}_b, \boldsymbol{v}_k)}} \\
\mathcal{L}_{TR}^{(v_b)} &= -\log \frac{\exp^{sim(\boldsymbol{v}_b, \boldsymbol{w}_b)}}{\sum_{k=1}^{B} \exp^{sim(\boldsymbol{v}_b, \boldsymbol{w}_k)}}
\end{aligned}
\tag{1}
$$

CMR の他，[MASK] に置き換えられたトークン $\boldsymbol{w}_m$ を周辺テキスト $\boldsymbol{w}_{\setminus m}$ および視覚情報 $\boldsymbol{v}_{CLS}$ から予測する Visual Masked Language Modeling（VMLM）

[1)]（式 2），また **0** ベクトルに置き換えられた物体領域のベクトル表現 $\boldsymbol{v}_m$ を周辺視覚情報 $\boldsymbol{v}_{\setminus m}$ およびテキスト情報 $\boldsymbol{w}_{CLS}$ から予測する Semantic Masked Region Modeling（SMRM）[2)]（式 3）を行う [3)]．

$$\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{(\boldsymbol{w},\boldsymbol{v})\sim D} \log P_\theta(\boldsymbol{w}_m|\boldsymbol{w}_{\setminus m}, \boldsymbol{v}_{\mathrm{CLS}}) \quad (2)$$

$$\mathcal{L}_{MRM}(\theta) = -\mathbb{E}_{(\boldsymbol{w},\boldsymbol{v})\sim D} f_\theta(\boldsymbol{v}_m|\boldsymbol{v}_{\setminus m}, \boldsymbol{w}_{\mathrm{CLS}}) \quad (3)$$

3.2 シーンテキストと周辺視覚情報の理解

先行研究における画像中の文字情報を組み込んだクロスモーダルモデルについては，クロスエンコーダを用いた手法が広く研究される [17, 23, 20, 21, 22] 一方で，画像中の文字情報がデュアルエンコーダに与える影響については明らかでない．そこで本研究では，LightningDOT [3] の画像エンコーダである UNITER [11] を対象に，シーンテキストを新たなモダリティの対象として考慮する（図 1）．

Masked Scene-Text Modeling (MSM)

シーンテキストに対して物体の視覚情報やキャプションとの適切なアラインメントをデュアルエンコーダに学習させるため，3.1 節で説明した VMLM を拡張し，シーンテキストに対してマスク予測に基づく最適化を行う [4)]．ここではシーンテキストのテキスト情報のみをマスクすることにより，モデルが「周辺の視覚情報からシーンテキストを読む」ことが期待できる．具体的には t をマスクするトークン数として，マスクするインデックスを $\boldsymbol{m} \in \mathbb{N}^t$ とし，マスクされたシーンテキストのトークン列を $\boldsymbol{s}_m$，マスクされていない周辺文脈としてのシーンテキストのトークン列を $\boldsymbol{s}_{\setminus m}$ とし，式 4 の negative log-likelihood を最小化する．

$$\mathcal{L}_{MSM}(\theta) = -\mathbb{E}_{(\boldsymbol{w},\boldsymbol{v},\boldsymbol{s})\sim D} \log P_\theta(\boldsymbol{s}_m|\boldsymbol{s}_{\setminus m}, \boldsymbol{v}, \boldsymbol{w}_{\mathrm{CLS}}) \quad (4)$$

4 実験設定

画像中の文字情報を考慮した画像検索を行うため，アノテータによって画像中のテキストを読解するように生成された [5)] TextCaps [1] を利用する（表 1）．TextCaps [1] では，キャプションと画像のペアに加え，Rosseta-en [34] を用いたシーンテキストの情報がデータとして与えられる [6)]．学習時のパラメータについては Appendix. A.1 を参照されたい．

表 1　学習データセット

データセット	訓練		開発	
	# images	# captions	# images	# captions
TextCaps [1]	21,953	109,764	3,166	15,830

表 2　TextCaps の開発セットにおける検索性能

	BERT	IR@k			TR@k		
		k=1	k=5	k=10	k=1	k=5	k=10
STARNet [18]	-	19.8	40.1	51.6	28.7	53.7	65.1
LightningDOT [3]	en	15.7	35.0	45.8	19.7	42.1	53.0
w/ OCR	en	**34.4**	**57.5**	**66.5**	45.7	67.8	**77.3**
w/ OCR+MSM	en	**34.4**	57.4	66.0	**45.8**	**68.4**	76.4
LightningDOT [3]	mul	13.8	33.1	43.8	14.5	35.1	46.6
w/ OCR	mul	21.9	45.0	56.4	30.0	54.0	65.5
w/ OCR+MSM	mul	**22.8**	**46.0**	**57.3**	**30.1**	**55.1**	**66.2**

5 実験

5.1 文字情報を考慮する検索モデルの評価

3.2 節で説明した，画像中の文字情報を視覚情報として考慮するための手法について評価した（表 2）．評価指標は，キャプションをクエリとした画像検索（IR）および画像をクエリとしたキャプション検索（TR）の両方について，Recall@k を用いた．また表 2 には，TextCaps [1] で学習された STARNet [18] の検索性能を同時に示す．さらに 2.2 節で述べたように，テキストエンコーダが持つ語彙ではシーンテキストを多くカバーすることが難しい [27, 28, 24] ため，テキストエンコーダの語彙数を増やした多言語 BERT [33] の評価も行った [7)]．

表 2 より，シーンテキストを新たなモダリティとして画像エンコーダに入力した場合に，英語 BERT および多言語 BERT の両設定で検索性能が大幅に向上した．これは TextCaps [1] がシーンテキストに

1）Devlin ら [33] の設定に従い，入力トークン列のうち 15% のトークンに対して，うち 80% をマスクトークン [MASK] に置換，10% をランダムなトークンに置換，残り 10% は置換なしとして処理を行う．
2）SMRM では，**0** にマスクされた物体領域のベクトル値を予測する Masked Region Feature Regression，およびマスクされた物体領域の物体クラスを予測する Masked Region Classification の二つのサブタスクからなる [3, 11]．
3）各イテレーションで無作為にタスクを一つ選択する．
4）VMLM と同様に Devlin ら [33] と同様の設定で行った．
5）Mafla ら [18] は，TextCaps における少なくとも一つのシーンテキストに言及するというバイアスについて言及している．
6）TextCaps [1] は，シーンテキストを含む画像およびシーンテキストに言及するキャプションが，それぞれ 96.9% および 81.3% と高く（COCO [35] では，それぞれ 12.7% および 2.7%），評価セットには訓練および開発データ中に含まれない OCR トークンが 2901/6329 個と多く含まれる．
7）なお英語 BERT および多言語 BERT の語彙数は，それぞれ 28,996，119,547 である．

表 3　視覚情報に対するアブレーション

MSM	modality		IR@k			TR@k		
	IMG	ST	k=1	k=5	k=10	k=1	k=5	k=10
-	✓	-	15.7	35.0	45.8	19.7	42.1	53.0
なし	✓	✓	34.4	57.5	66.5	45.7	67.8	77.3
	✓	-	11.0	28.8	39.6	12.4	28.0	38.9
	-	✓	13.9	27.5	34.1	11.2	26.3	33.3
あり	✓	✓	34.4	57.4	66.0	45.8	68.4	76.4
	✓	-	9.7	26.3	36.6	11.2	26.5	36.8
	-	✓	14.8	29.1	35.5	14.9	29.7	37.1

言及するキャプションデータであるというバイアス [18] が要因であると考えられるが，少なくともシーンテキストを含むデータセットに対してシーンテキストと周辺視覚情報を融合するモデル化が効果的であることを示している．また 3.2 節で述べた MSM については，多言語 BERT を用いた場合に検索性能が向上した．MSM による画像中の文字情報のモデル化について 5.2 節で詳細な評価を行う．

5.2　視覚情報に対するアブレーション

5.1 節では，視覚情報として画像中の文字情報を考慮することで検索性能が改善することを示した．本節では，視覚情報として画像中の文字情報（ST）および物体の視覚情報（IMG）の融合的な理解ができているか調査するため，視覚情報のモダリティ選択についてアブレーションを行った（表 3）．

表 3 より，画像エンコーダに入力するモダリティを，IMG・ST のみとした場合にモデルの検索性能が大きく低下したことから，画像エンコーダから画像中の文字情報および周辺視覚情報が共存した視覚情報がエンコードされることが分かる．また ST のみを対象とする場合に，MSM ありの検索性能が MSM なしに比べて高い値を示したことから，MSM による事前学習がシーンテキストをモデル化するために有効な手法であることが分かる．一方 IMG のみを対象とする場合に，MSM ありの検索性能が低くなった結果は，テキストエンコーダから出力されるキャプションに対して，テキストとしての情報を持つ ST との親和性が IMG よりも高くなったのが原因と考えられる．このことから，キャプションと物体の視覚情報におけるクロスモーダルな関係を損なうことなく，シーンテキストを視覚情報として融合できるような工夫が今後必要となる．

表 4　シーンテキストとキャプションのトークン重複数に対する検索結果（MSM あり）

	重複数（IR）			重複数（TR）		
	0	1	2	0	1	2
トークン数	2,302	512	212	11,785	2,484	1,004
- acc@1	46.1	45.3	44.3	31.9	36.6	46.3
- acc@5	69.2	66.8	64.2	54.9	59.0	69.2
- acc@10	77.0	75.4	74.5	63.9	67.1	76.7

5.3　トークン重複数別の検索性能の変化

シーンテキストが TR・IR に与える影響について更なる調査を行うため，キャプションとの親和性が高くなる要因の一つであるトークンの重複数に対する検索性能を，TextCaps [1] の開発セットを用いて Accuracy@k で評価した[8]（表 4）．

表 4 から，TR ではキャプションとのトークン重複数が多くなるほど検索性能が高い値を示したことから，シーンテキストとキャプション間で重複するトークンが，画像中の文字情報とキャプションの親和性を高める要因の一つであることが分かる．また IR においてトークン重複数に依存しない結果を示したのは，トークン重複数と検索対象である視覚情報のモダリティ間の差異や，シーンテキストには権限性の低い文字情報が多く含まれることが原因と考えられる．

6　おわりに

本研究では TextCaps [1] を対象に，画像中の文字情報をデュアルエンコーダに組み込んだ際の影響について調査を行った．実験の結果より，デュアルエンコーダにおいてもシーンテキストが有効であることを示すと同時に，シーンテキストのモデル化について MSM が効果的であることを示した．

今後の展望としては，クロスモーダルな意味関係についてより詳細な紐付けを行う工夫を模索したい．また Jia ら [8] のような大規模な学習データを用いた影響や，TextCaps 以外の評価データについても同様効果があるかどうかデータ横断的な調査を行いたい．本研究で得られた知見は，文字情報の理解に強い Vision and Language の事前学習モデルの構築や，事前学習モデルに基づく大規模かつ高速な画像検索の重要な基礎・応用課題に貢献できる．

8）なおシーンテキストは spaCy を用いて（“ADJ”,“ADV”,“NOUN”,“PROPN”,“VERB”）の内容語からなる品詞に限定した．

参考文献

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

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NeurIPS**. Curran Associates, Inc., 2017.

[3] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. LightningDOT: Pre-training visual-semantic embeddings for real-time image-text retrieval. In **NAACL-HLT**, pp. 982–997, 2021.

[4] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In **EMNLP**, pp. 6769–6781, 2020.

[5] Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. In **ACL**, pp. 2173–2183, 2021.

[6] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In **NAACL-HLT**, pp. 5835–5847, 2021.

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

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In **ICML**, pp. 4904–4916, 2021.

[9] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. **CoRR**.

[10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In **NeurIPS**, pp. 13–23, 2019.

[11] Chen Yen-Chun, Li Linjie, Yu Licheng, El Kholy Ahmed, Ahmed Faisal, Gan Zhe, Cheng Yu, and Liu Jingjing. UNITER: universal image-text representation learning. In **ECCV**, pp. 104–120, 2020.

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

[13] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In **CVPR**, pp. 5579–5588, 2021.

[14] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In **CVPR**, pp. 3977–3986, 2021.

[15] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In **CVPR**, pp. 4155–4165, 2021.

[16] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In **NAACL-HLT**, pp. 2443–2459, 2021.

[17] 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. 8751–8761, 2021.

[18] Andres Mafla, Rafael S. Rezende, Lluis Gomez, Diane Larlus, and Dimosthenis Karatzas. Stacmr: Scene-text aware cross-modal retrieval. In **WACV**, pp. 2220–2230, 2021.

[19] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. **IEEE**, pp. 2552–2566, 2014.

[20] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In **SIGKDD**, pp. 1192–1200, 2020.

[21] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multimodal pre-training for visually-rich document understanding. In **ACL/IJCNLP**, pp. 2579–2591, 2021.

[22] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In **AAAI**, pp. 13878–13888, 2021.

[23] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. Latr: Layout-aware transformer for scene-text VQA. **CoRR**, 2021.

[24] Jing Wang, Jinhui Tang, Mingkun Yang, Xiang Bai, and Jiebo Luo. Improving ocr-based image captioning by incorporating geometrical relationship. In **CVPR**, pp. 1306–1315, 2021.

[25] Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, Chris Olah. Multimodal neurons in artificial neural networks. **Distill**, 2021.

[26] David A. Noever and Samantha E. Miller Noever. Reading isn't believing: Adversarial attacks on multi-modal neurons. **CoRR**, 2021.

[27] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In **CVPR**, pp. 8317–8326, 2019.

[28] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In **CVPR**, 2020.

[29] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In **ICLR**, 2015.

[30] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In **ICML**, pp. 369–376, 2006.

[31] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In **CVPR**, pp. 4558–4567, 2021.

[32] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In **CVPR**, pp. 6077–6086, 2018.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **NAACL-HLT**, pp. 4171–4186, 2019.

[34] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In **SIGKDD**, pp. 71–79, 2018.

[35] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. **CoRR**, Vol. abs/1504.00325, , 2015.

A　参考情報

A.1　学習パラメータ

4 章で説明したモデルの学習設定を図 5 に示す.

表 5　実験設定

モデルアーキテクチャ	LightningDOT [3]
隠れ層次元数	768
ミニバッチサイズ	4096
gradient accumulation steps	6
学習率	$5e-5$
学習ステップ数	150, 000
画像エンコーダ	UNITER [11]
隠れ層次元数	768
Transformer 層数	12
attention head 数	12
最適化アルゴリズム	AdamW $\beta_1 = 0.9,\ \beta_2 = 0.98$
Warmup Steps	10, 000
dropout	0.1
事前学習タスク比	16 （CMR） 8 （MLM） 4 （MRFR） 4 （MRC-kl） 8 （MSM）
テキストエンコーダ	BERT [33]
アーキテクチャ	BertForMaskedLM
隠れ層次元数	768
Transformer 層数	12
attention head 数	12
語彙数	28, 996 （英語 BERT） 119, 547 （多言語 BERT）