VisualMRC: 文書画像に対する機械読解

田中涼太 *　　西田京介 *　　吉田仙
日本電信電話株式会社　NTT メディアインテリジェンス研究所
{ ryouta.tanaka.rg, kyosuke.nishida.rx, sen.yoshida.tu }@hco.ntt.co.jp

1　はじめに

近年，知識源となるテキストを読み解いて質問に応える**機械読解**（Machine Reading Comprehension; MRC）が自然言語理解の主要な課題として注目を集めている [1, 2]．機械読解は Web 検索の高度化に加えて，e-commerce 向けのチャットボット [3] や専門文献の読書補助 [4] などへの応用が期待されているが，従来の研究では，実サービスで扱われる HTML や PDF 形式の文書が持つ視覚的な情報を自然言語と併せて理解することが出来ていない．

そこで本研究では，言語と視覚の融合理解に向けて，**VisualMRC** タスクおよびデータセットを提案する[1)]．図 1 に示す様に，本タスクにおいてモデルは，質問に対して文書画像中のテキストを読み解き，自然文で回答する．本タスクを解くためには，ベースとなる自然言語理解・生成に加えて，文書レイアウト（図 1 の画像で色付けされた矩形領域）の理解，テキスト認識（OCR）など様々な能力が必要となる．

本研究ではさらに，VisualMRC タスクに適用可能な新たなモデルを提案する．テキストコーパスで事前学習されたエンコーダデコーダを文書中のテキスト・非テキストオブジェクトの位置・外観を理解できるように拡張することで，最先端の text-based VQA モデル [6] およびベースとした T5 [7], BART [8]（テキスト情報のみ入力）を上回る性能を達成した．

2　関連研究

Text-based VQA　機械読解における入力テキストを画像としたタスクである VQA はこれまで広く研究されてきた [11, 12]．近年では，TextVQA [9] に代表される，日常シーンの画像に含まれるテキストの理解が必要となるデータセットが複数個公開されている [13, 14, 15]．これらのデータセットの画像中の

Q: Who were the winners of the Ig Nobel prize for Biology and Chemistry?
A: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.

図 1　VisualMRC データセットの例．

単語は少量であるのに対して，VisualMRC は自然文で書かれた複数の文章が視覚的に配置された文書の理解能力の開発に重点を置いている．

文書画像に対する VQA　VisualMRC と同時期に作成されたデータセットとして，文書画像の理解が必要な DocVQA [10] がある．VisualMRC との主な違いを以下に挙げる．**(i)** VisualMRC は 35 のドメインから画像を収集したのに対し，DocVQA の情報源は単一ドメインである．**(ii)** VisualMRC は現代の Web ページから文書画像を収集したのに対し，DocVQA の大半の画像は 1960–2000 年頃の古い文書であり，手書き・タイプ文字を含む．**(iii)** VisualMRC の画像は最低 3 つの自然文を含むのに対し，DocVQA ではその保証が無い．**(iv)** VisualMRC は回答生成を必要とするが，DocVQA は多くの場合 SQuAD [1] の様な単一の短いスパンが回答となる．

マルチモーダル QA　画像中のテキストではなく，プレーンテキストと画像の組を入力として用いるデータセットとして，TQA [16] や RecipeQA [17] などがある．例えば，TQA では中学校の教科書からテキストと図を分けて抽出しているが，我々の研究は人間が実際の文書を読むときと同じ視覚情報を機械が扱えるようにすることを目指している．

* Equal contribution.
1)　データセットのさらなる詳細情報は https://github.com/nttmdlab-nlp/VisualMRC および [5] に示す．

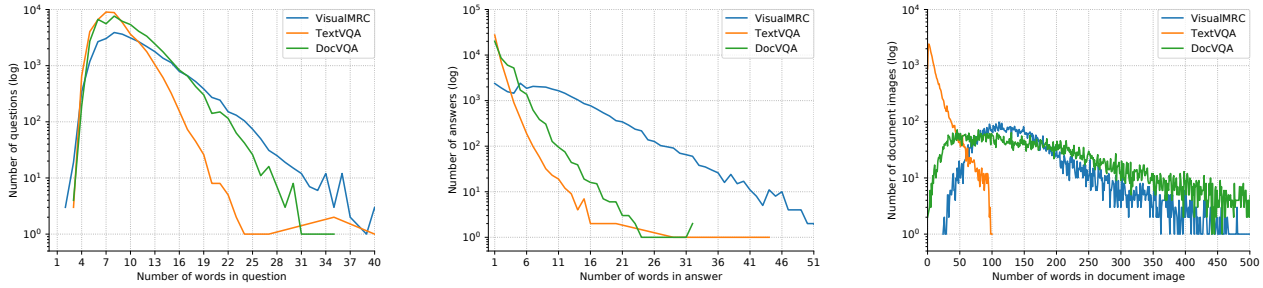


図 2　VisualMRC，TextVQA [9]，DocVQA [10] の単語数に関する分布．**左**: 質問，**中**：回答，**右**：文書画像．

3　VisualMRC

本章では最初にタスクを定義し，その後にデータセットの収集方法および分析結果を示す．

3.1　タスク定義

VisualMRC のタスクを下記の様に定義する．

Task 1 (End-to-end VisualMRC)**.** 質問 q および文書画像 I が与えられた時，モデルは回答 a を生成する．

本タスクは，NarrativeQA [18] や MS MARCO [19] と同じ生成型機械読解であり，回答は入力テキスト中のスパンに限定されない．

次に，文書画像の理解は，以下の 2 つのサブタスクに分解可能である．

Subtask 1 (ROI **検出**)**.** 画像 I が与えられた時，モデルは ROI（Region-Of-Interest）の集合を検出する．各 ROI r_i は矩形領域 b_i および意味クラス l_i を持つ．

Subtask 2 (OCR)**.** ROI r_i が与えられた時，モデルは単語認識を行う．各単語は矩形領域 $b_{i,j}$ および表層 $w_{i,j}$ を持つ．

3.2　データ収集方法

文書画像収集　94 名のワーカに依頼し，2020 年 1-3 月に 35 のドメインの Web ページ（英語）の 10,197 枚のスクリーンショット画像 I を収集した．

正解 ROI アノテーション　45 名のワーカが文書画像中の ROI（$\{r_i\}$; Subtask 1 の正解) のアノテーションを実施した．9 つの意味クラス（Heading/Title, Subtitle/Byline, Paragraph/Body, Picture, Caption, List, Data, Sub-Data, Other）の定義は付録に示す．

QA 作成／関連 ROI 選択　495 名のワーカが，各文書画像に対して 3 つの異なる質問および回答の組を作成した．その際，各ワーカは回答に必要となる**関連 ROI** を付与済みの正解 ROI 集合から選択した．

データ分割　URL ドメインに基づき訓練・開発・テストセットを 21,015，2,839，6,708 質問とした．

表 1　データセットの統計．

	TextVQA	DocVQA	VisualMRC
Image type	daily scenes	industry documents	webpages
Num. images	28,472	12,767	10,197
Num. questions	45,536	50,000	30,562
Uniq. num. questions	36,593	36,170	29,419
Perc. uniq. answers	51.74	64.29	91.82
Avg. len. questions	8.12	9.49	10.55
Avg. len. documents	12.17	182.75	151.46
Avg. len. answers	1.51	2.43	9.53

3.3　統計および分析

VisualMRC を代表的な関連データセットである TextVQA [9] と DocVQA [10] と比較して分析する．

質問　表 1 に示す様に，VisualMRC のユニークな質問の割合（96.3%）は TextVQA（80.7%）や DocVQA（72.3%）よりも高い．また，質問長の分布は TextVQA や DocVQA に比べてロングテールである（図 2 左）．さらに，VisualMRC の “what” や “what is the” で始まる質問の割合（42.0%，9.5%）は，TextVQA（78.2%，22.9%）や DocVQA（68.1%，58.2%）と比べて低く，多様な質問が含まれている．

回答　表 1 に示す様に，VisualMRC のユニークな回答の割合（91.82%）は TextVQA （51.74%）や DocVQA（64.29%）よりも高い．VisualMRC は生成型の機械読解タスクであるため，回答の平均長（9.53 単語）は TextVQA（1.51）や DocVQA（2.43）に比べて長い（図 2 中）．また，“yes”，“no” から始まる回答（10.04%，2.67%）は TextVQA（4.90%，0.97%）や DocVQA （0.12%，0.15%）に比べて多い．

文書画像　画像の正解 ROI 集合から OCR（Tesseract [20]）で抽出された平均単語数は， VisualMRC（151.46）や DocVQA（182.75）が TextVQA（12.17）よりも多い．また，VisualMRC を LDA [21] により分析すると，科学，旅行，健康，教育，ニュース，政治など多様なトピックが抽出された．一方で，DocVQA のトピックは食料と栄養が大半であった [10].

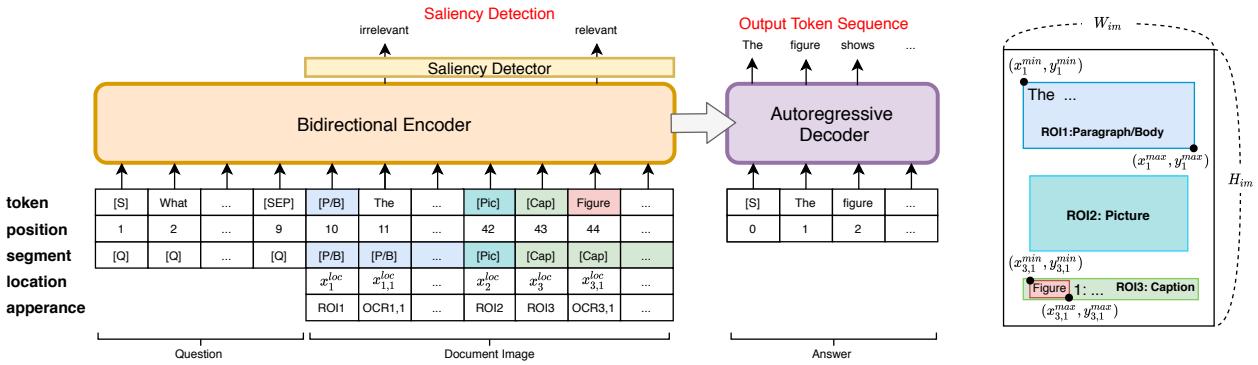


図 3　**左**: 提案モデルアーキテクチャ．**右**: ROI および OCR トークンの例．

4　提案モデル

提案モデルはメインとサブモジュールから構成される．サブモジュールは 3.1 節に示す SUBTASK 1（ROI 検出）用の Faster R-CNN [22] と，SUBTASK 2（OCR）用の Tesseract [20] を用いる．メインモジュールは，エンコーダデコーダ型の Transformer [23] アーキテクチャに基づき，質問，文書画像，およびサブモジュールの出力に基づいて動作する．

本章ではメインモジュールについて詳細を述べる．提案モデルの貢献は，T5 [7] や BART [8] の入力系列の拡張およびマルチタスク学習の導入により，事前学習により得られた自然言語理解・生成能力を転移した上で視覚情報を考慮可能な点である．

4.1　入力トークン系列

質問 q のトークナイズ結果と各 ROI r_i から得られた OCR 単語系列 $\{w_{i,j}\}$ を連結する．ROI の意味的な役割を考慮するため，各 ROI の意味クラスに対応するクラスラベルトークン $[\mathrm{L}_i]$（図 3 の [P/B] など）を各 OCR 単語系列の前に挿入する．

$$x^{\text{token}} = \left\{ \begin{matrix} [\mathrm{S}], q_1, ..., q_m, [\mathrm{SEP}], [\mathrm{L}_1], w_{1,1}, ..., w_{1,M}, \\ [\mathrm{L}_2], \ldots, [\mathrm{L}_N], w_{N,1}, ..., w_{N,M} \end{matrix} \right\},$$

ベースモデルとして T5（BART）を用いる場合，[S] は’question:’ (<s>)，[SEP] は’context:’ (</s>) である．

4.2　入力 Embedding 系列

トークン系列 $\{x^{\text{token}}\}$ を embedding 系列 $\{z\}$ に変換してエンコーダに渡す．系列中の k 番目の embedding である $z_k \in \mathbb{R}^H$ は下記の様に求める．

$$z_k = \mathrm{LN}(z_k^{\text{token}} + z_k^{\text{pos}} + z_k^{\text{seg}} + z_k^{\text{loc}} + z_k^{\text{app}})$$

ここで，$\mathrm{LN}(\cdot)$ は layer normalization [24] を表す．z^{token} および z^{pos} はベースモデルが用いるトークンおよびトークン系列中の位置 z^{pos} の embeddings である．

以下に，3 つの追加 embeddings について説明する．

セグメント　文書構造の意味を明確に捉えるため，k 番目のトークンが属する ROI クラスに対して学習可能なベクトル $z_k^{\text{seg}} \in \mathbb{R}^H$ を導入する．

画像中の位置　k 番目のトークンに対応する矩形領域（ROI あるいは OCR 単語）の画像中の相対位置

$$x_k^{loc} = [x_k^{\text{min}}/W_{\text{im}}, y_k^{\text{min}}/H_{\text{im}}, x_k^{\text{max}}/W_{\text{im}}, y_k^{\text{max}}/H_{\text{im}}],$$

を，1 層の FFN により $z_k^{\text{loc}} \in \mathbb{R}^H$ に変換する．$(x_k^{\text{min}}, y_k^{\text{min}})$，$(x_k^{\text{max}}, y_k^{\text{max}})$ は矩形領域の左上および右下の座標，W_{im}，H_{im} は画像の幅および高さを表す．

外観　ROI と OCR 単語の視覚的表現を利用するため，k 番目のトークンの矩形領域の画像を Faster R-CNN [22] に渡し，2048 次元の fc7 特徴ベクトルを獲得する．さらに，ReLU 活性化関数を適用後，1 層の FFN に渡し $z_k^{\text{app}} \in \mathbb{R}^H$ に変換する．

4.3　Saliency Detection

質問に関連する OCR 単語をより正確に発見するため，エンコーダの出力を用いて適合度を求める．

$$P_{i,j} = \mathrm{sigmoid}(w^{s\top} h_{w_{i,j}} + b^s),$$

$h_{w_{i,j}}$ は単語 $w_{i,j}$ に対するエンコーダの最終層の出力，$w^s \in \mathbb{R}^H$ と $b^s \in \mathbb{R}$ は学習可能な重みである．

損失関数　Saliency detection を教師有り学習するための正解ラベルは与えられていないため，各 OCR 単語に対して疑似ラベル $s_{i,j}$ を割当て学習を行う．

$$L_{\text{sal}} = -\frac{1}{NM} \sum_i^N \sum_j^M \left(\begin{matrix} s_{i,j} \log P_{i,j} + \\ (1 - s_{i,j}) \log(1 - P_{i,j}) \end{matrix} \right)$$

ここで，$s_{i,j}$ は OCR 単語が回答文かつ関連 ROI に含まれている場合に 1，それ以外は 0 となる．

マルチタスク学習　L_{nll} を sequence-to-sequence 学習における負の対数尤度とし，メインモジュールの学習を $L_{\text{multi}} = L_{\text{nll}} + L_{\text{sal}}$ の最小化により行う．

表 2　main 設定における性能．

Model	B-4	M	R-L	Bs
M4C	10.3	12.8	28.1	86.1
T5	41.5	31.7	53.0	90.5
LayoutT5	**43.4**	34.6	**54.6**	**90.8**
w/o SD	43.3	**34.9**	54.4	90.7
BART	36.4	28.8	48.7	90.1
LayoutBART	**38.7**	**31.9**	**52.8**	**90.7**
w/o SD	37.7	31.3	**52.8**	90.6
$\text{LayoutT5}_{\text{LARGE}}$	**44.9**	**37.3**	**57.1**	91.3
$\text{LayoutBART}_{\text{LARGE}}$	43.0	36.1	57.0	**91.5**

表 3　end-to-end 設定における性能．

Model	B-4	M	R-L	Bs
M4C	10.2	12.7	28.0	86.1
T5	38.6	29.8	50.2	90.0
LayoutT5	**41.0**	**33.2**	**52.2**	**90.3**
BART	34.6	27.5	47.3	90.0
LayoutBART	**36.4**	**30.5**	**50.5**	**90.4**
$\text{LayoutT5}_{\text{LARGE}}$	**42.1**	35.6	54.5	90.9
$\text{LayoutBART}_{\text{LARGE}}$	40.6	34.6	55.2	91.2
Human	39.6	**41.0**	**57.9**	**91.9**

表 4　Ablation 評価（main 設定）．

Model	B-4	M	R-L	Bs
T5	41.5	31.7	53.0	90.5
+ lbl	42.9	32.5	53.2	90.5
+ seg	43.6	32.8	53.3	90.5
+ loc	**44.1**	33.5	53.7	90.5
+ app	43.3	**34.9**	**54.4**	**90.7**
BART	36.4	28.8	48.7	90.1
+ lbl	37.6	30.3	50.7	90.3
+ seg	37.8	30.3	50.9	90.4
+ loc	**38.1**	30.3	51.4	90.5
+ app	37.7	**31.3**	**52.8**	**90.6**

5　評価実験

VisualMRC を用いて評価実験を行う．提案モデルの初期化に T5_{BASE} [7]，$\text{BART}_{\text{BASE}}$ [8] を用いたものをそれぞれ LayoutT5，LayoutBART と呼ぶ．

実験設定　Task 1 に対応する end-to-end 設定と，人手で付与された正解 ROI が与えられる main 設定の 2 種類を用いる．学習時は main 設定で行う．

ベースライン　最先端の text-based VQA モデル M4C [6] と，ベースとした T5_{BASE}，$\text{BART}_{\text{BASE}}$（テキスト情報のみを入力）を用いる．

評価指標　生成タスクで広く用いられる指標である BLEU-4（B-4）[25], METEOR（M）[26], ROUGE-L（R-L） [27]，BERTScore（Bs） [28] を用いる．

5.1　評価結果

提案モデルはベースラインの性能を上回るか？　表 2 に示す様に，提案モデルは視覚的な位置・外観を考慮することにより全ての指標でベースラインの性能を上回った．M4C は視覚情報を考慮するモデルであるが事前学習されていないため性能が低かった．また，ベースとしたモデルのサイズを LARGE に変更することで，全ての指標で性能が改善した．

Saliency detection とのマルチタスク学習は効果があるか？　表 2 に示す様に，saliency detection（SD）とのマルチタスク学習により性能が改善した．LayoutT5 では効果が小さかったが，これは T5 の事前学習に saliency detection に類似したタスクである抽出型の機械読解 [1] が含まれているためと考える．

提案モデルは end-to-end 設定で高い性能を達成できるか？　表 3 に示す様に，実世界の問題設定と同じく ROI 検出を自動的に行う必要がある end-to-end 設定においても，提案手法は全ての指標でベースラインの性能を上回り，main 設定と比べても大きく性能が低下することは無かった．ROI 検出の mAP は 7.86%であり，文書構造解析データ [29] における Faster R-CNN の mAP（5.1%）[30] と同程度であった．

提案モデルは人間の性能を上回るか？　表 3 に end-to-end 設定・ランダムサンプリングした 3000 件の QA における人間の性能と提案モデル（LARGE サイズ）の性能の比較結果を示す．BLEU-4 では提案モデルが人間を上回ったが，他の指標は人間が大きく提案モデルを上回っていた．

入力トークン・embedding 系列の変更は効果があるか？　表 4 に ablation 評価の結果を示す．クラスラベルトークンの挿入（lbl）によるトークン系列の変更，セグメント（seg），画像中の位置（loc），外観（app）の追加 embeddings の全てで効果があった．外観の embeddings のみ BLEU-4 で性能が低下したが，これは動画にグラウンディングされた対話の従来研究 [31] でも同様の傾向であった．

提案モデルは全ての ROI クラスの理解に効果があるか？　LayoutT5 は全クラス・全指標で T5 の性能を上回った．（詳細は付録に示す）．相対的には写真や図表の理解が不十分であり今後の課題としたい．

6　おわりに

本研究では，VisualMRC を新たな視覚・言語融合理解の課題として提起し，データセットの作成およびモデルの提案を行った．本データセットは，実世界の文書を視覚的に読解可能な知的エージェントの開発に寄与し，Web 検索やチャットボットなど産業上重要なサービスの発展に貢献できる．また，入力トークン・embeddings 系列の変更により，事前学習により獲得された自然言語理解・生成の能力を転移した上で文書の視覚情報を考慮する方法の有効性を示した．提案モデルは汎用性が高く様々な事前学習モデルに導入可能である．

参考文献

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, pp. 2383–2392, 2016.

[2] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *ACL*, pp. 784–789, 2018.

[3] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. Superagent: A customer service chatbot for e-commerce websites. In *ACL*, pp. 97–102, 2017.

[4] Yining Hong, Jialu Wang, Yuting Jia, Weinan Zhang, and Xinbing Wang. Academic reader: An interactive question answering system on academic literatures. In *AAAI*, pp. 9855–9856, 2019.

[5] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, 2021.

[6] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA. In *CVPR*, pp. 9992–10002, 2020.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, Vol. 21, No. 140, pp. 1–67, 2020.

[8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pp. 7871–7880, 2020.

[9] 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.

[10] Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. *arXiv*, Vol. 2007.00398, , 2020.

[11] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pp. 2425–2433, 2015.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6325–6334, 2017.

[13] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, 2018.

[14] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pp. 4290–4300, 2019.

[15] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pp. 10126–10135, 2020.

[16] Aniruddha Kembhavi, Min Joon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pp. 5376–5384, 2017.

[17] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *EMNLP*, pp. 1358–1368, 2018.

[18] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, Vol. 6, pp. 317–328, 2018.

[19] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated machine reading comprehension dataset. *arXiv*, Vol. 1611.09268v3, , 2018.

[20] R. Smith. An overview of the tesseract OCR engine. In *ICDAR*, pp. 629–633, 2007.

[21] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022, 2003.

[22] Ren, Shaoqing, He, Kaiming, Girshicka, Ross, Sun, and Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pp. 91–99, 2015.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 6000–6010, 2017.

[24] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv*, Vol. 1607.06450, , 2016.

[25] 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.

[26] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, pp. 376–380, 2014.

[27] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out@ACL*, pp. 74–81, 2004.

[28] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020.

[29] Dominika Tkaczyk, Pawel Szostek, and Lukasz Bolikowski. GROTOAP2 - the methodology of creating a large ground truth dataset of scientific articles. *D Lib Mag.*, Vol. 20, No. 11/12, 2014.

[30] Carlos Soto and Shinjae Yoo. Visual detection with context for document layout analysis. In *EMNLP-IJCNLP*, pp. 3462–3468, 2019.

[31] Hung Le and Steven C. H. Hoi. Video-grounded dialogues with pretrained generation language models. In *ACL*, pp. 5842–5848, 2020.

A 付録

ROI クラスの定義 下記に，クラウドワーカに対するインストラクションで用いた ROI クラスの定義を示す．

- **Heading/Title** The title or caption of a page, chapter, etc.
- **Subtitle/Byline** The secondary or subordinate title of a page or a line of text giving the author's name.
- **Paragraph/Body** The "normal" or "main" text that would be read.
- **Picture** Pictures or images without any text or data within.
- **Caption** Text placed next to an image, data, quote, etc. that provides or explains information about an image or data.
- **List** Typically bulleted lists, where each bullet is not a full sentence.
- **Data** Tables, charts, graphs, infographic, or other figures with data or information.
- **Sub-data** Text present inside of tables, charts, graphs, infographic, or other figures.
- **Other** Any other text that does not fit in the other categories.

質問文の多様性 図 4 に質問文の最初の 3 単語の分布を示す．“what” 質問に偏っている TextVQA [9] や DocVQA [10] に比べて多様な質問を含むことが分かる．

実験設定 学習には 8 枚の NVIDIA Quadro RTX 8000 GPUs を用いた．バッチサイズ 32 とし，7 エポック学習（12 時間程度）した．オプティマイザには Adam を用い，学習率は 3e-5 とした．開発セットの ROUGE-L スコアを用いてモデルを選択した．

出力例 図 5 に提案モデルおよびベースラインの出力例を示す．提案モデルは視覚的なデータ表の最初の行に回答に必要な情報（77.3%）が含まれることを発見できたが，視覚情報を用いない T5 では，他の割合の表現（less than 1 percent）を誤って回答していた．

ROI の意味クラス別の性能 表 5 に示す通り，LayoutT5 は全 ROI クラスの理解で T5 の性能を上回っていた．

実行時間 表 6 に 1 つの質問への回答に要する平均時間を示す．GPU は 1 枚の NVIDIA Quadro RTX 8000 を用いた．LayoutT5 は多数の領域・OCR トークンに対する外観の embeddings を計算する必要があるが，大幅な速度低下は起きていない．Faster R-CNN が高速に動作する理由は，ROI の探索を行わず画像表現を獲得するためだけに使っているためである．モデルサイズが速度低下には寄与しており，LARGE サイズは BASE サイズの約 3.5 倍の動作時間を要した．

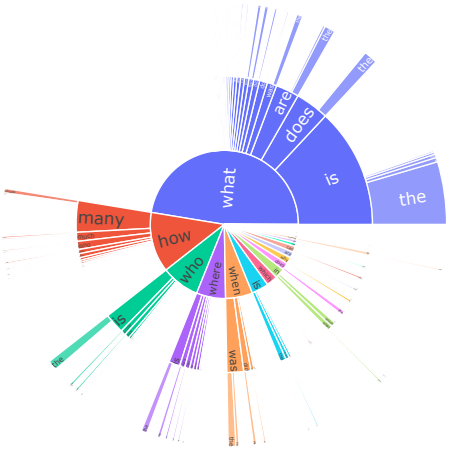
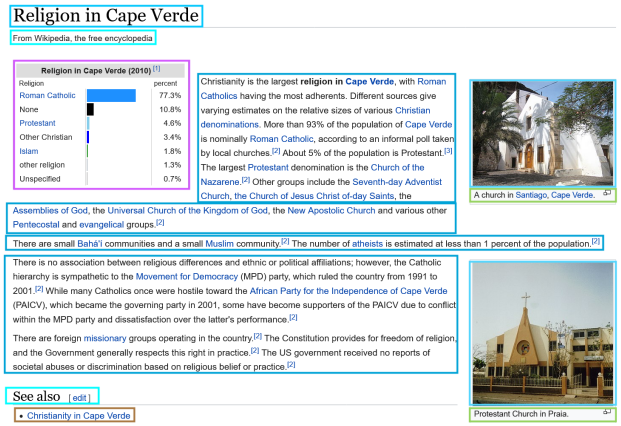


図 4 質問文の最初の 3 単語の分布．



Q: what is the percentage of roman catholics in cape verde?
Ground-Truth: the percentage of roman catholics in cape verde is 77.3%.
M4C: the percentage of young women in cape town are about 54% of western somalia
T5: percentage of roman catholics in cape verde is less than 1 percent.
LayoutT5: the percentage of roman catholics in cape verde is 77.3%.

図 5 LayoutT5 による出力例．

表 5 意味クラス別の T5/LayoutT5 の性能．

ROI Class	BLEU-4	METOR	RUGE-L	BERTscore
Heading/Title	37.9/**42.8**	29.8/**32.5**	49.9/**52.6**	89.9/**90.3**
Paragraph/Body	42.7/**44.1**	32.3/**35.0**	54.0/**55.1**	90.6/**90.8**
Subtitle/Byline	39.6/**46.3**	29.9/**33.8**	48.0/**52.6**	90.0/**90.8**
Picture	25.9/**32.0**	24.8/**29.8**	44.9/**49.0**	89.4/**90.3**
Caption	31.2/**41.1**	28.0/**33.1**	50.3/**55.5**	89.6/**91.0**
List	35.7/**39.0**	30.4/**33.1**	48.1/**50.4**	90.0/**90.7**
Data	31.8/**32.7**	26.1/**29.3**	42.2/**46.4**	88.9/**89.6**
Sub-Data	30.1/**41.4**	26.4/**32.4**	42.8/**50.6**	88.9/**90.6**
Other	34.1/**41.5**	28.1/**32.5**	48.4/**51.7**	89.8/**90.5**

表 6 1 つの質問への回答に要する平均時間 (秒)．

Model	Avg. Gen Time
T5	0.1812
LayoutT5	0.2253
$\text{LayoutT5}_{\text{LARGE}}$	0.4489