テキストと視覚的に表現された情報の融合理解に基づくインフォグラフィック質問応答

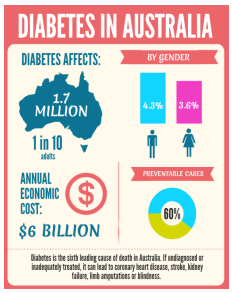
田中涼太 [1]　西田京介 [1]　許俊杰 [2*]　西岡秀一 [1]
[1] 日本電信電話株式会社　NTT 人間情報研究所　[2] 筑波大学大学院　人間総合科学学術院
{ryouta.tanaka.rg,kyosuke.nishida.rx,shuichi.nishioka.gd}@hco.ntt.co.jp
s2021705@s.tsukuba.ac.jp

概要

情報・データ・知識を視覚的に表現した文書画像であるインフォグラフィックを読み解いて，質問に対して回答を行うモデル IG-BERT を提案する．提案モデルではテキストと視覚物体の配置関係の学習および算術演算理解のためのデータ拡張を行う．IG-BERT は，ICDAR 2021 Competition の InfographicVQA タスクにおいて，事前学習データ量を従来モデルの 1/22 に抑えつつ同程度のサイズのモデルの中で最も高い性能を達成し 2 位に入賞した．本研究は，実世界に多数存在する視覚的に表現された文書を知識源として質問応答を行う知的エージェントの発展に貢献できる．

1　はじめに

質問に対して文書を読み解いて人間の様に回答を行う質問応答技術の実現は，AI 分野における重要な課題の一つである．従来はテキストのみで記述された文書に関する質問応答 [1, 2] が活発に取り組まれてきた．一方で，従来研究では実サービスで扱われる HTML や PDF 形式の文書が持つテキスト情報しか理解できないことから，近年では文書を画像として扱い質問応答を行う文書画像質問応答 [3, 4] が注目を集めている．本研究はその中でも情報・データ・知識を視覚的に表現した文書画像を扱う**インフォグラフィック質問応答** [5] に取り組む．本タスクは図 1 で示す様に，テキストに加えてアイコンや図表といった視覚情報の理解，テキストと視覚情報を併せた配置関係の理解，算術演算など様々な能力を必要とする．これらの能力は，オフィス作業や日常生活を支援する AI の発展に向けて必要不可欠である．一方で，文書画像理解を目指した従来モデルの多く [4, 6, 7, 8, 9] はテキストの配置の理解に焦点が置かれており，本タスクへの適用は難しい．



Q: How many females are affected by diabetes?
A: 3.6%

Q: What total percent of B2B and B2C markets use Google+?
A: 79% (40% + 39%)

図 1　InfographicVQA [5] のサンプル例．左の例: 画像中のテキストを抽出して回答する．右の例: 画像中のテキスト (40%，39%) を用いて，算術演算を行い回答する．

本研究では，新たなインフォグラフィック質問応答モデル IG-BERT を提案する．文書画像中のテキストと視覚物体との配置関係を学習する新たなタスク SRP (Spatial Relationship Prediction) と，演算の過程を生成させ算術演算の理解を深めることを可能とする新たなデータ拡張 ADA (Arithmetic operation Data Augmentation) を行う点が新規の貢献である．IG-BERT は，ICDAR 2021 Competition の InfographicVQA [5] において，様々な文書理解タスク [10, 11, 12] で最高スコアを達成した LayoutLMv2 [8] の事前学習データ量を 1/22 に抑えつつ，同程度のサイズのモデルの中で最も高い性能を達成して 18 チーム 337 投稿中 2 位に入賞した．

2　提案モデル

インフォグラフィック質問応答は，質問 w^{q} および文書画像 I が与えられ回答系列 w^{out} を出力するタスクである．回答は，画像中のテキストから回答箇所を抜き出す回答（図 1 左）と数値の演算を必要とする回答（図 1 右）の 2 つのタイプに大別できる．

* NTT におけるインターンシップ期間中の貢献．

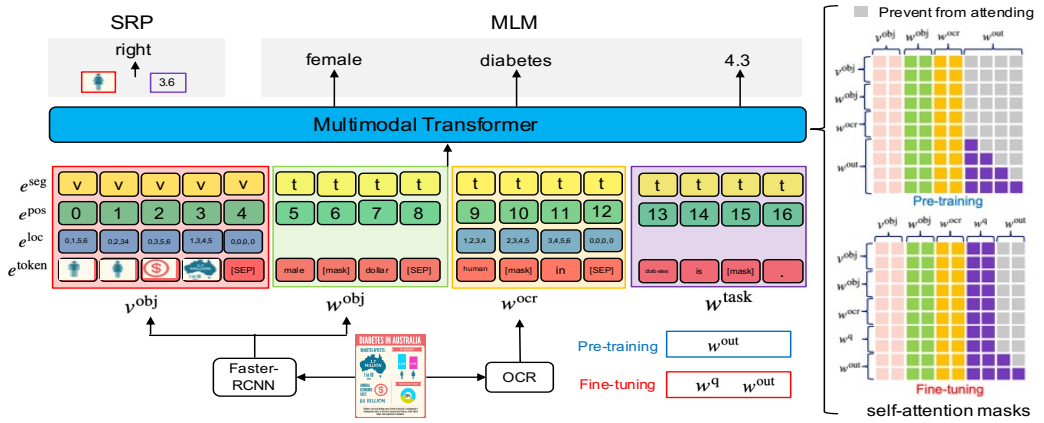


図 2　左図: IG-BERT のアーキテクチャと損失関数 (SRP，MLM)．右図: 自己注意計算時に用いる自己注意マスク M.

IG-BERT は図 2 に示すように，画像からテキストを抽出する OCR，画像から物体特徴を抽出する Faster-RCNN [13]，抽出された特徴量を基に回答を生成するマルチモーダル Transformer [14] から成る．

2.1　入力埋め込み層

画像特徴　Visually29k [15] で事前学習を行った Faster-RCNN を用いて，文書画像 I から固定個の物体を検出して特徴表現 $v_i^{\text{obj}} \in \mathbb{R}^{2048}$ の系列を得る．

テキスト特徴　画像 I から取得される物体ラベル系列 w^{obj} と OCR 単語系列 w^{ocr}，質問文 w^{q}，出力文 w^{out} を WordPiece [16] を用いてトークナイズする．

入力系列　画像・テキスト特徴に基づくモデル入力系列を x^{token} = ([CLS] v^{obj} [SEP] w^{obj} [SEP] w^{ocr} [SEP] w^{task}) とする．事前学習時の w^{task} は w^{out}（キャプション），Fine-tuning 時は w^{q} w^{out} (QA) である．

レイアウト特徴　k 番目のトークンの文書画像中における配置を理解するために，$x_k^{loc} = [x_k^{\min}/W_{\text{im}}, y_k^{\min}/H_{\text{im}}, x_k^{\max}/W_{\text{im}}, y_k^{\max}/H_{\text{im}}]$ を取得する．ここで，$(x_k^{\min}, y_k^{\min})$，$(x_k^{\max}, y_k^{\max})$ はトークン（画像物体あるいは文字）を囲む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．

マルチモーダル入力埋め込み　入力系列中の k 番目の埋め込み $e_k \in \mathbb{R}^{d_k}$ を画像・テキストの特徴およびそのレイアウトに基づき以下のように求める．

$$e_k = \text{LN}(e_k^{\text{token}} + e_k^{\text{seg}} + e_k^{\text{pos}} + e_k^{\text{loc}}),$$

LN は Layer Normalization [17] を示す．e^{token} は，テキストトークン w_k を d 次元に埋め込み，画像（検出物体領域）トークン v_k^{obj} に対しては ReLU を適用後 1 層の FFN にて埋め込む．e_k^{seg} は 2 種のモーダル，e_k^{pos} はトークン系列の絶対位置を表す埋め込みである．e_k^{loc} は，x_k^{loc} を 1 層の FFN に渡して埋め込む．

2.2　マルチモーダル Transformer

入力埋め込み系列 $H^0 = [e_1^0, ..., e_U^0] \in \mathbb{R}^{d \times U}$ を L 層の Transformer に入力し，$H^L = [e_1^L, ..., e_U^L] \in \mathbb{R}^{d \times U}$ を獲得する．$l-1$ 層目の埋め込み系列 H^{l-1} を用いると，l 層目の自己注意 A^l は以下で与えられる．

$$A^l = \text{softmax}(\frac{QK}{\sqrt{d}} + M)V, \quad M_{ij} = \begin{cases} 0, & \text{allow to attend,} \\ -\infty, & \text{others} \end{cases}$$

$$V = W_V^l H^{l-1}, Q = W_Q^l H^{l-1}, K = W_K^l H^{l-1},$$

ここで，W_Q^l，W_K^l，$W_V^l \in \mathbb{R}^{d \times d}$ は学習可能な重みである．$M \in \mathbb{R}^{U \times U}$ は図 2 の右図で示す様に，系列中で出力文 w^{out} のみ left-to-right，他部分は bidirectional となる prefix 付の causal masking [18] を行う．

2.3　演算過程を考慮したデータ拡張 (ADA)

算術演算を含む質問応答を生成タスクとして解くにあたり，演算の過程をモデルに明示的に学習させる．提案手法では，演算対象を A と B，演算結果を C とした時，「$A+B=C$」，「$A-B=C$」，「$100-A=C$」の 3 つのテンプレートを用意する．この時，演算結果 C を出力するタスクは，演算対象（A，B）を抽出する項抽出タスクと，演算子 (+，−) を選択する演算子選択タスクのサブタスクに分割可能である．

具体的には，まず，文書画像から演算対象となる任意の二つの数値を抽出する．次に，テンプレートに抽出した数値を代入し演算を行う．演算結果が正解データ w^{out} と一致する場合，回答の先頭に prompt トークン [NUM] を付与し，新たな回答として学習データに追加する．推論時に [NUM] が出力系列の先頭に存在する場合，後処理として演算を実施した．

2.4 学習タスク

提案モデルは事前学習と Fine-tuning の二段階の学習において，MLM および SRP の損失を線形結合した $L = L_{\text{MLM}} + \lambda_{\text{SRP}} L_{\text{SRP}}$ を最小化する．事前学習時には $\lambda_{SRP} = 0$，Fine-tuning には $\lambda_{SRP} = 1$ とする．

Masked Language Modeling (MLM) BERT [19] と同様に，入力系列のテキストトークン w に対して，トークン系列の 15%をランダムに選択し，[MASK] トークンに置き換える．エンコーダの最終層の出力を用いて，置き換えられたトークンを復元する．L_{MLM} は，生成テキストから計算される損失である．

Spatial Relationship Prediction (SRP) 画像物体と OCR の配置関係を学習する．まず，画像物体系列 v^{obj} と OCR 系列 w^{ocr} に対応するエンコーダの最終出力からそれぞれ 1 つずつランダムに選択し，e^{obj} と e^{ocr} を獲得する．そして，[20] で定義された配置関係を基にクラス分類 (12 クラス) を行う．

$$P^{\text{srp}} = \text{softmax}(W^{\text{srp}}[e^{\text{obj}}; e^{\text{ocr}}]),$$

$W^{\text{srp}} \in \mathbb{R}^{12\times 2d}$ は学習可能な重み，[;] はベクトル連結を表す．L_{SRP} は正解クラスと P^{srp} との交差エントロピー損失である．文書画像への SRP の適用，単語・視覚物体の位置関係学習は本研究が初である．

3 評価実験

事前学習データ インフォグラフィックを扱うウェブページ[1)]から，画像-キャプションペアを 505,868 ペア収集した．HTML タグの alt もしくは title を用いて，キャプションを抽出した．フィルタリングとして，3 単語未満のキャプションの削除と，評価データと同様の画像を URL を基に削除した．

評価データ InfographicVQA [5] を用いた．学習/開発/テストデータの質問数は 23,946/2,801/3,288 であり，画像数は 4,406/500/579 である．テストデータでの評価はリーダーボード[2)]上で実施した．

実験設定 事前学習と Fine-tuning の両方でバッチサイズ 64 とし 30 エポック学習した．Adam [23] を用いて最適化し学習率は $3e-5$ とした．v^{obj}，w^{obj}，w^{ocr}，w^{task} の最大長を 36，20，430，40 とした．開発，テストデータにおける評価は BERT-{base,large} を事前学習時の重みの初期値とした．また，large は base の実験を基にハイパーパラメータを設定した．OCR と Faster-RCNN の詳細は付録に示す．

1) https://infographicplaza.com
2) https://rrc.cvc.uab.es/?ch=17&com=evaluation

評価指標 ICDAR 2021 Competition で採用された ANLS [24] (予測文と正解文集合との平均編集距離) を用いる．また，算術演算を必要とする例に絞った (開発データの 17.4%) 際の ANLS を ANUM と呼ぶ．

比較手法 ベースラインとして，IG-BERT と同じモーダル情報を入力可能な M4C [21] と，事前学習の効果を測るため BERT と LayoutLM [6] を用いた．リーダボード上の主なモデルとして，LayoutLM に視覚情報を取り入れた LayoutLMv2 [8]，複数の教師あり事前学習を行なった TILT [9] がある．

3.1 評価結果と分析

提案モデルはベースラインの性能を上回るか？ 表 1 に示す様に，提案モデルは全ての指標でベースラインモデルを上回った．また，提案モデルと同じモーダル情報を利用しているにも関わらず，事前学習を行っていない M4C-TLV は低い性能であった．

事前学習の効果はあるか？ 表 2 に示す様に，提案モデルから事前学習を除くことで，精度が大きく低下することが確認できる．また，事前学習を除いた IG-BERT は BERT と比べて精度が低いことが分かることから，Fine-tuning のみでは BERT の言語理解能力の忘却 [25] が発生していると考える．

データ拡張手法 ADA は算術演算の理解を改善するか？ 表 2 に示す様に，BERT と IG-BERT のどちらにおいても，提案したデータ拡張により ANUM の向上が確認できる．テンプレートの拡充および複雑化により，更なる演算の高度化が期待できる．

学習タスク SRP は精度向上に寄与するか？ 表 2 に示す様に，提案モデルに SRP を加えることで，大きく性能が向上することが確認できる．

SRP におけるサンプリング対象の選択は精度に影響するか？ 表 3 に結果を示す様に，v^{obj} と w^{ocr} からサンプリングし，配置関係を学習したモデルの精度が最も高いことが分かる．同モーダル内の配置関係よりも，画像物体と OCR テキストの配置関係を理解することが重要であることが示唆される．

リーダボード上のモデルと比較してモデルの性能はどうか？ 表 4 に示す様に，同程度のモデルサイズのモデルの中で，IG-BERT は最も高い精度を達成した．特筆すべきは，同じモーダル情報を利用している LayoutLMv2 に比べて，IG-BERT は事前学習の文書画像数は 1/22 でありパラメータ数も少ないにも関わらず，大きく性能が上回っている．IG-BERT が行う視覚物体の検出・位置関係理解および算術演

表 1 開発データにおけるベースラインモデルとの比較．T (テキスト)，L (レイアウト)，V (画像特徴) を表す．

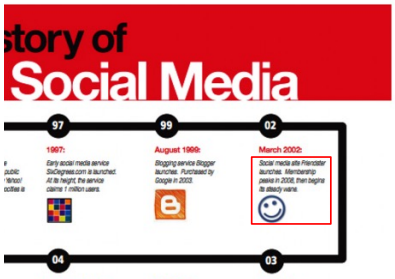
Model	Modal	Pre-train	ANLS
M4C-TL [21]	TL		0.140
M4C-TLV [21]	TLV		0.142
BERT [19]	T	✓	0.206
LayoutLM [6]	TL	✓	0.212
IG-BERT	TLV	✓	**0.292**

表 2 開発データにおける Ablation 評価．

Model	ANLS	ANUM
BERT	**0.206**	**0.161**
BERT w/o ADA.	0.199	0.156
IG-BERT	**0.292**	**0.195**
IG-BERT w/o pretrain	0.176	0.132
IG-BERT w/o SRP	0.275	0.166
IG-BERT w/o SRP and ADA.	0.271	0.159

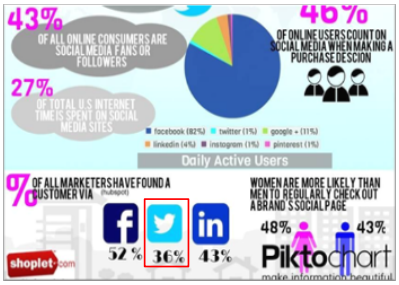
表 3 SRP におけるサンプリング対象ごとの性能評価．k, k' はサンプリング対象のインデクスを表す．

Target	ANLS
$v_k^{\text{obj}} \leftrightarrow v_{k'}^{\text{obj}}$	0.284
$w_k^{\text{ocr}} \leftrightarrow w_{k'}^{\text{ocr}}$	0.271
$v_k^{\text{obj}} \leftrightarrow w_{k'}^{\text{ocr}}$	**0.292**



Q: What is the name of the social media site with a smiley face icon?
GT: friendster **LayoutLM**: twitter
BERT: facebook **IG-BERT**: friendster

Q: How many patients out of 3, does not use social media to seek out health information?
GT: 2 **LayoutLM**: 3
BERT: 1 **IG-BERT**: 2 (3 - 1)



Q: What percentage of all markets have found a customer through Twitter?
GT: 36% **LayoutLM**: 43%
BERT: 43% **IG-BERT**: 82%

図 3 出力例．掲載上，画像の一部をクロップした．GT は正解文，赤枠は回答根拠，() は予測された演算過程である．

表 4 リーダボード上での比較 (2022/01/14 付)．#Docs は事前学習用文書数，#Sups は追加学習用 QA ペア数．

Model	Modal	#Docs	#Sups	Params.	ANLS
BERT fuzzy [19]	T	-	NA	340M	0.208
LayoutLMv2 [8]	TLV	11M	NA	426M	0.283
Ensemble [19, 22]	TLV	NA	0.2M	NA	0.285
BROS [7]	TL	11M	0.12M	NA	0.322
IG-BERT	TLV	0.5M	0	342M	**0.385**
TILT [9]	TLV	1.0M+	0.22M	780M	0.612

算の工夫が性能向上に貢献していると考える．一方で，TILT と比較すると精度が劣っている．TILT では，IG-BERT が行わなかった DocVQA [3] 等の類似タスク [1, 26] での教師あり事前学習が性能向上に大きく寄与することが報告されている．しかし，TILT には IG-BERT で提案した SRP や ADA を代替する学習タスクや算術演算に関する新規点は含まれていない．パラメータ数や事前学習データ量の増加に関しては，今後の課題とする．

出力例・エラー分析 出力例を図 3 に示す．アイコンとテキストの配置関係 (smiley face icon と friendster) を理解する必要がある左の例や，画像中のテキストを用いて演算 (3 - 1) を行う必要のある中央の例では，提案モデルは適切な回答を出力している．一方で，右の例の様に Faster-RCNN で検出が困難な物体 (Twitter ロゴ) に関する質問には，全てのモデルで誤った回答を行なっており，より汎用的な物体認識を基にした文書理解は今後の課題である．

4 関連研究

InfographicVQA は，他の類似タスク [3, 4] と比べてテキストと視覚の融合理解が最も必要となるタスクである [27]．DocVQA [3] に代表される従来のタスクを対象としたモデルの多く [4, 6, 7, 8, 9] はテキストや配置の理解に重点を置いており，視覚情報を含む配置関係や意味の理解，算術演算の理解を行うことは難しい．ICDAR 2021 Competition で高スコアであった LayoutLMv2 [8] や TILT [9] では，事前学習において文字以外の視覚情報に関して考慮されていないが，我々の手法は物体検出により考慮可能である．さらに，我々が提案する SRP や ADA を代替する学習タスクや算術演算は実施されていない．

5 おわりに

インフォグラフィック質問応答モデル IG-BERT および，配置関係学習手法 SRP，算術演算データ拡張手法 ADA を提案し，ICDAR 2021 Competition の InfographicVQA タスクにて，同程度のサイズのモデルの中で最も高い性能を達成し 2 位に入賞した．SRP や ADA は他モデルにも導入可能な汎用性が高い手法である．本研究は，実世界に多数存在する視覚的に表現された文書を知識源として質問応答を行う人工知能の開発に寄与し，Web 検索や対話システムなど産業上重要なサービスの発展に貢献できる．

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] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In **WACV**, pp. 2200–2209, 2021.

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

[5] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In **WACV**, pp. 1697–1706, 2022.

[6] 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 **KDD**, pp. 1192–1200, 2020.

[7] Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. **arXiv preprint arXiv:2108.04539**, 2021.

[8] 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: Multi-modal pre-training for visually-rich document understanding. In **ACL/IJCNLP**, pp. 2579–2591, 2021.

[9] Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. **arXiv preprint arXiv:2102.09550**, 2021.

[10] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In **ICDAR**, pp. 991–995. IEEE, 2015.

[11] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In **ICDARW**, Vol. 2, pp. 1–6. IEEE, 2019.

[12] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In **ICADR**, pp. 1516–1520. IEEE, 2019.

[13] 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.

[14] 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.

[15] Spandan Madan, Zoya Bylinskii, Matthew Tancik, Adrià Recasens, Kimberli Zhong, Sami Alsheikh, Hanspeter Pfister, Aude Oliva, and Fredo Durand. Synthetically trained icon proposals for parsing and summarizing infographics. **arXiv preprint arXiv:1807.10441**, 2018.

[16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. **arXiv preprint arXiv:1609.08144**, 2016.

[17] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. **arXiv preprint arXiv:1607.06450**, 2016.

[18] 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.

[19] 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.

[20] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In **ECCV**, pp. 684–699, 2018.

[21] 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.

[22] Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. Simple is not easy: A simple strong baseline for textvqa and textcaps. **arXiv preprint arXiv:2012.05153**, 2020.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **ICLR**, 2015.

[24] 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.

[25] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In **Psychology of learning and motivation**, Vol. 24, pp. 109–165. Elsevier, 1989.

[26] Minseok Cho, Reinald Kim Amplayo, Seung-won Hwang, and Jonghyuck Park. Adversarial tableqa: Attention supervision for question answering on tables. In **ACML**, pp. 391–406. PMLR, 2018.

[27] Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. Due: End-to-end document understanding benchmark. In **NeurIPS**, 2021.

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **Int. J. Comput. Vis.**, Vol. 123, No. 1, pp. 32–73, 2017.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **CVPR**, pp. 770–778, 2016.

A 付録

A.1 OCR と Faster-RCNN に関する実験設定

OCR には Google Vision API [3] を用いた．Faster-RCNN のバックボーンは Visual Genome [28] で事前学習を行った ResNet-101 [29] を使用した．また，Adam を用いて最適化に用い，バッチサイズを 16，学習率 1e-3 とし，5 epoch 学習を行った．アンカーボックスのスケールを [8, 16, 32]，アスペクト比を [0.5, 1.0, 2.0] とした．

A.2 画像 URL 一覧

図 1，図 3 で用いた画像の URL を以下に示す．

図 1 左 https://www.adelaidebariatriccentre.com.au/diabetes-in-australia
図 1 右 https://www.business2community.com/social-media/3-infographics-100-social-media-statistics-need-0712061
図 3 左 https://visual.ly/community/Infographics/social-media/short-history-social-media
図 3 中央 https://hcsmmonitor.com/2017/02/23/how-modern-health-care-is-being-revolutionized-by-social-media-infographic/
図 3 右 https://visual.ly/community/Infographics/social-media/social-media-marketing-18

3） https://cloud.google.com/vision