URL　https://journal.ntt.co.jp/article/26653
DOI　https://doi.org/10.60249/24065002

グラフィカルな文書を理解できる「tsuzumi」

大規模言語モデル（LLM：Large Language Models）の応用先として，医療やカスタマーサポート，オフィスDXなどが挙げられます．こうした分野で扱う情報には，テキストのみならず図や表をはじめとしたさまざまな視覚情報が含まれており，LLMをAI（人工知能）技術の中核として発展させていくためには視覚情報を理解できるように拡張する必要があります．本稿では，文書を視覚情報から理解する「視覚読解技術」に関する一連の取り組みについて紹介します．

キーワード：#tsuzumi，#LLM，＃視覚読解

田中　涼太（たなか　りょうた）／壹岐　太一（いき　たいち）
長谷川　拓（はせがわ　たく）／西田　京介（にしだ　きょうすけ）
NTT人間情報研究所

文書を視覚的に理解する「視覚読解技術」

私たちが扱う文書はテキストや視覚要素（アイコンや図表など）を含み，多様な種類・形式が存在します．こうした実世界の文書を読解し理解する技術の実現は，AI（人工知能）分野における重要課題の1つです．近年では，汎用的な言語理解・生成能力を持つ大規模言語モデル（LLM: Large Language Models）をはじめとするAIが数多く登場し，人間の読解能力を超えるなど大きく発展してきましたが，文書中のテキスト情報しか理解できない限界がありました．この問題に対して，NTTでは人の情報理解と同様に，文書を視覚情報から理解する技術として，**図1**で示す「視覚読解技術」を提唱しました．

私たちはこれまでに，本技術の実現をめざしてVisualMRC[1]やSlideVQA[2]といったデータセットを構築してきました．これらのデータセットは，Webページのスクリーンショットやプレゼンテーション資料といった1枚・複数枚の文書画像に対する質問応答データであり，言語情報のみならず，文字の大きさや色，図や表，グラフ，レイアウトの情報といった視覚情報の理解を必要とします．私たちは，物体認識技術を適用して抽出した文書中の領域（タイトルや段落，画像，キャプション，リストなど）と，さらに文字認識技術を適用して抽出した文字の位置・外観情報を追加入力とし，これらを統合して考慮可能な視覚読解モデルであるLayoutT5[1]と，さらに複数の文書画像間の関係性を理解可能なM3D[2]を提案しました．これらの視覚情報を考慮したモデルは，テキストのみを考慮したモデルに比べて高い性能を示しており，人の情報処理から着想を得た本技術の有効性を確認しました．

こうした取り組みで得られた知見をベースに，情報・データ・知識を視覚的に表現したインフォグラフィック文書に対する質問応答の性能を競うICDAR（International Conference on Document Analysis and Recognition）2021 DocVQA competitionに参加しました．本コンペティションにおける質問応答例を**図2**に示します．図2中のQ1に答えるには，文書中の右中段のアイコンが女性を意味することを理解しなければなりません．また，Q2に答えるには，文書中から数値を抽出し，「40％＋39％＝79％」という計算を行う必要があります．このように，テキストに加

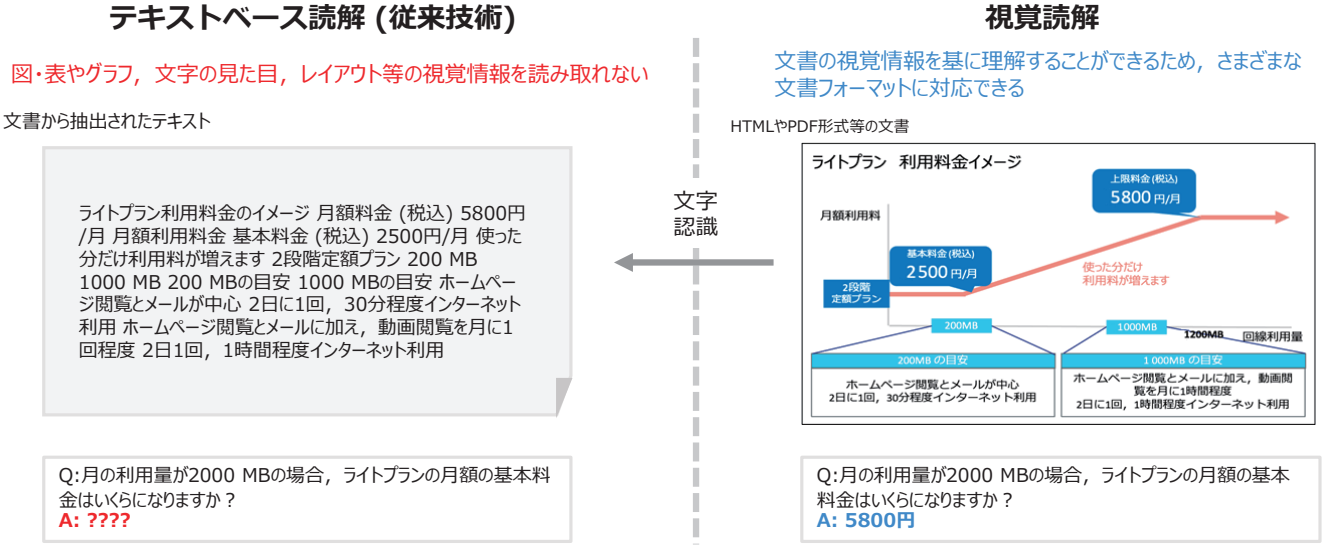
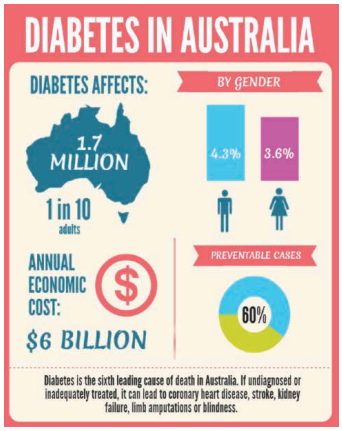


図1　視覚読解技術の概要図



Q1: How many females are affected by diabetes?
A: 3.6%

Q2: What total percent of B2B and B2C markets use Google+?
A: 79%（40% + 39%）

図 2　InfographicVQA における質問例

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，算術演算などさまざまな能力が必要であり，挑戦的な課題です．そこで，私たちは新たなインフォグラフィック質問応答モデルであるIG-BERTを提案しました[3]．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タスクの導入や，演算の過程を生成させる新たなデータ拡張手法の導入を行いました．その結果，私たちが提出したシステムは従来モデルで必要とする事前学習データ量の22分の 1 に抑えつつ，同程度のサイズのモデルの中でもっとも高い性能を達成し，18チーム337投稿中 2 位を獲得することができました．

従来の「視覚読解技術」の課題

これまでの視覚読解技術は任意のタスク（例えば，請求書に対する情報抽出タスク）に対して対応することができませんでした．つまり，目的のタスクごとに，一定数のデータを用意して学習を行わない限り，所望のタスクで高い性能を出すことは難しい状況でした．そのため，従来の技術では，目的に合わせたデータセットの作成およびモデルの学習が必要であり，作成・計算コストが高く，ユーザのニーズに合わせたモデルを構築するうえでの障壁となっていました．そこで，私たちは汎用な言語理解・生成能力を持つLLMを活用し，任意のタスク用に学習を行わなくても応答できる，高い指示遂行能力を視覚読解モデルで実現することをめざしました．具体的には，テキスト情報しか理解できないLLMに対して，LLMの推論能力を壊すことなく，どのように文書画像に含まれる図表などの視覚情報をテキストと融合させてLLMに理解させるかが，解決すべき課題になります．

LLMを活用した「視覚読解技術」

文書画像に含まれる図表などの視覚情報をLLMに理解させるためには，画素（ピクセル）の集まりとして表される視覚情報をLLMが処理しやすいかたちに変換する必要があります．tsuzumiの視覚読解では，**図 3** に示すように，画像エンコーダと軽量なアダプタの組合せにより，テキスト理解を保って視覚理解の追加を実現しました[4]．画像エンコーダは画素の集合を言語的な意味に対応付け，アダプタはその意味をLLM（tsuzumi）が処理できるよう変換します．

次に，技術をもう一段深く解説するとともに，本技術による実現例についても紹介します．

■日本特有の画像を理解できる画像エンコーダ

画像エンコーダは，画像に何が映っているかの視覚情報を処理する役割を果たします．私たちは，画像を入力してベクトルに変換する画像エンコーダと，テキストを入力してベクトルに変換するテキストエンコーダを用意し，画像のベクトルとその画像の内容を表しているテキストのベクトルの距離が近くなるように，そして同時に無関係な画像とテキストのベクトルの距離が遠くなるように，画像エンコーダを学習しています．これにより，画像エンコーダによって得られる視覚情報をテキスト情報に結び付けることができます．画像エンコーダの学習の際には，一般的な画像と英語キャプションだけでなく，日本特有の画像とその日本語のキャプションも収集し，数億ペアのテキストと画像の学習データを構築しました．日本特有の画像には，日本語文字が含まれていたり，日本の独自の風景が映っていたりします．さらに，日本語キャプションを使って学習することにより「青信号」や「真っ赤な太陽」といったような日本語特有の表現も学習できるように工夫しています．私たちは英語のテキストと画像で学習したエンコーダをベースにして，日本語にも適応させる技術を開発し，英語，日本語の両方に強いモデルを実現することにも取り組んでいます[5]．

■視覚理解を追加するアダプタとその学習

アダプタは，いわば，画像エンコーダの「言葉」をLLMの「言葉」に翻訳するような役割を担います．LLMはテキストをいくつかの文字列に区切ったトークンという単位に分解します．そして，トークンをニュー

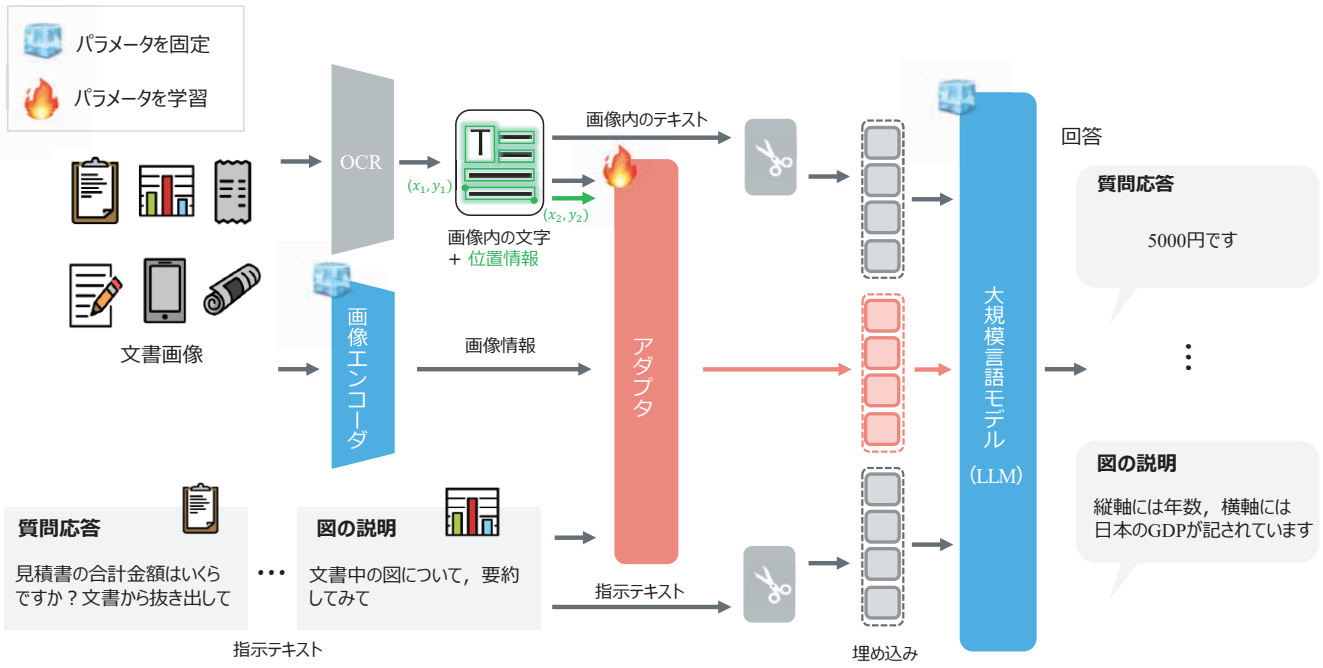


図 3　LLM を活用した視覚読解技術の概念図

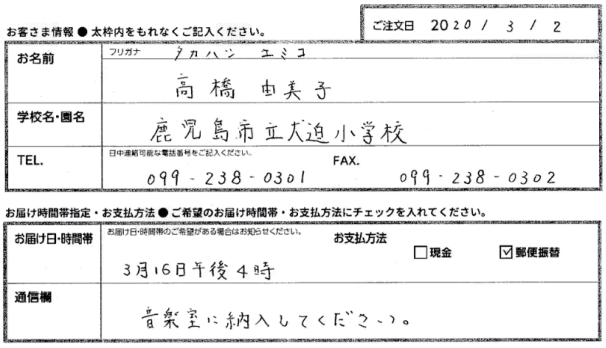
ラルネットワークへと入力するために，事前に学習した対応テーブルに従ってベクトル（数値列）に変換します．このトークンに対応するベクトルこそLLMが入力として受け取る「言葉」であり，埋め込みと呼びます．アダプタは，画像エンコーダの出力を埋め込みに変換することで画像をLLMに伝えます．

アダプタは少量のパラメータを持ったニューラルネットワークであるため学習が必要です．tsuzumiの視覚読解では，画像エンコーダとLLMのパラメータは固定して，アダプタのパラメータだけを学習対象とすることで，LLMの推論能力を保ちます．さらに，独自に収集したデータセットを使用する多段階の学習によって文書画像に適したアダプタを実現します．まず，画像からキャプションテキストを予想するタスクを大量のデータで学習し，物体や風景，位置関係といった一般的な視覚的概念をtsuzumiに伝えられるようにします．続いて，文書画像，画像内の光学文字認識（OCR：Optical Character Recognition）結果，指示テキストを入力とし，タスク遂行の結果を出力するよう学習します．例えば，見積書画像と「合計金額を教えて」という入力に対して「5000円」と回答する情報抽出などを幅広く学習することによって，レイアウトや図表といった文書画像に含まれる情報や回答スタイルを伝えられるようにします．

■LLMを活用した視覚読解による実現例

私たちは，tsuzumiの視覚読解の学習に先行して，世界中の既存の文書画像を対象とした研究を調査し，データセットを網羅的に収集することでInstructDocデータセットを作成しました[4]．また，InstructDocを用いた実験によって，LLMが初見のタスク（学習データには含まれないタスク）に対しても高い成功率を達成できるようになることを実証しました．tsuzumiの視覚読解では，この成果に基づいて，汎用的なタスク遂行に向けた学習データセットを構築し，学習しています．tsuzumiの視覚読解の動作例を**図 4** に示します．左側の例は帳票の読取りタスクです．注文票の画像と「お届け日と納入場所をそれぞれ教えてください」という指示に対して，モデルは「3月16日午後4時 音楽室」と回答しています．表のお届け日・時間帯欄から正しく抽出しています．データセットにさまざまな帳票の画像が含まれているので，標準的なレイアウトを理解して質問に回答できます．さらに，注文票に納入場所の欄はありませんが，通信欄の自由形式テキストに基づいて回答できています．これはtsuzumiのテキスト理解が保たれているためです．次に，右側の例は写真の認識タスクです．猫の写った画像と「画像のメインの物体をtype，color，captionキーでjson化して」という指示を与えています．ここでJSONとは，ソフトウェアの分野でよく利用される情報を構造化するためのテキスト形式です．モデルは「{“type”:“cat”, “color”: “white”, “caption”: “猫がノートパソコンのキーボードに手を置いています”}」と回

指示

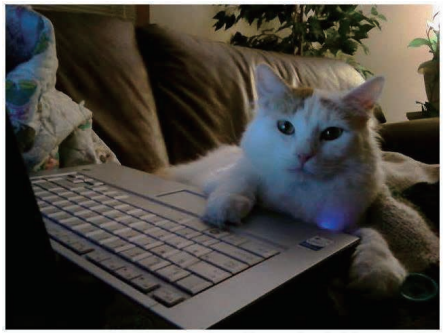

お客さま情報 ● 太枠内をもれなくご記入ください。
ご注文日 2020 / 3 / 2
お名前 フリガナ タカハシ ユミコ 高橋 由美子
学校名・園名 鹿児島市立犬迫小学校
TEL. 099-238-0301 FAX. 099-238-0302
お届け時間帯指定・お支払方法 ● ご希望のお届け時間帯・お支払方法にチェックを入れてください。
お届け日・時間帯 3月16日午後4時
お支払方法 □現金 ☑郵便振替
通信欄 音楽室に納入してください。

お届け日と納入場所をそれぞれ教えてください

モデルの回答

3月16日午後4時 音楽室

指示



画像のメインの物体をtype，color，comment キーでjson化して

モデルの回答

{“type”: “cat”, “color”: “white”, “caption”: “猫がノートパソコンのキーボードに手を置いています”}

図 4　tsuzumi視覚読解の動作例

答しています．モデルは，画像内の主な物体として猫を認識したうえで指示に合うようにtype，color，captionそれぞれをテキスト化し，JSON形式に従ったテキストを出力しています．出力形式の制御は画像のタグ付け等へ応用できます．

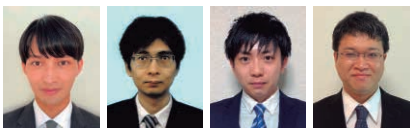
以上のように，tsuzumiの視覚読解は，テキスト理解と画像理解を柔軟に結び付けることによって，ユーザのニーズに合わせたタスク遂行を可能とします．

今後の目標

今後は，現在の文書読解モデルをさらに発展させることをめざして，１つひとつのモジュールの発展に取り組んでいきます．さらに，視覚以外のモーダルともLLMを結び付けていくことでLLMの応用範囲を広げ，最終的には人とAIの共生社会の実現をめざして研究開発と実用化を進めていきます．

■参考文献

(1) R. Tanaka, K. Nishida, and S. Yoshida：“VisualMRC: Machine Reading Comprehension on Document Images,” AAAI 2021, pp. 13878-13888, Feb. 2021.
(2) R. Tanaka, K. Nishida, K. Nishida, T. Hasegawa, I. Saito, and K. Saito：“SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images,” AAAI 2023, pp. 13636-13645, Washington D.C. , U.S.A. Feb. 2023.
(3) 田中・西田・許・西岡：“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,” NLP2022, pp. 52-57, 2022.
(4) R. Tanaka, T. Iki, K. Nishida, K. Saito, J. Suzuki：“InstructDoc: A Dataset for Zero-Shot Generalization of Visual Document Understanding with Instructions,” AAAI 2024, pp. 19071-19079, Vancouver, Canada, Feb. 2024.
(5) T. Hasegawa, K. Nishida, K. Maeda, and K. Saito：“DueT: Image-Text Contrastive Transfer Learning with Dual-adapter Tuning,” EMNLP 2023, pp. 13607-13624, Singapore, Dec. 2023.



（左から）　田中 涼太／ 壹岐 太一／
長谷川 拓／ 西田 京介

人とAIの共生社会の実現をめざして，視覚読解技術の研究開発を推進していきます．

◆問い合わせ先

NTT人間情報研究所
思考処理研究プロジェクト
E-mail　tsuzumi-nttlab @ ntt.com