2. 視覚言語モデルとCT画像診断レポート生成への応用

森　健策　名古屋大学大学院情報学研究科/名古屋大学情報基盤センター/国立情報学研究所医療ビッグデータ研究センター

近年の人工知能（artificial intelligence：AI）技術の進展は目覚ましく，毎日新たな技術が生み出されているといっても過言ではない。「ChatGPT」（OpenAI社）などに代表される生成AI技術は，さまざまな質問に対する回答を生成できることが可能となった。ChatGPT登場当初は，プログラムソースコードを書くことは「まだまだ」感が強かったが，最近では「Claude」（Anthropic社）によるAIプログラミングの能力がきわめて高くなり，ソフトウエア開発のかなりの部分でAIによる自動化も可能となっている[1]。その背後にあるのは，大規模言語モデル（large language model：LLM），視覚言語モデル（vision language model：VLM），AIエージェントなどの新たなAI技術開発の存在である。

一方，医療分野における生成AIの利活用も加速しようとしている段階である。放射線科の領域で見ると，北米放射線学会（RSNA）の企業展示においては，ディクテーションされた放射線レポートの言語モデルを用いた構造化レポートへの変換，胸部X線画像の所見レポートの自動生成などの分野において，生成AIを活用した製品が展示されている。画像を入力すると所見文を生成する技術は，画像情報と言語情報とを接続する技術を用いている。画像所見レポート生成については，二次元画像を対象としたものから，三次元画像を対象としたもの，そして四次元画像を対象としたものへと広がっている。ある時刻で撮影された医用画像への所見文生成のみならず，医用画像の比較読影を対象とした所見文生成へとその応用分野は広がっている。

われわれの研究グループにおいても，経時三次元CT画像を対象として比較読影を行い，日本語の所見文を生成するシステムの開発を進めてきた[2,3]。これは，画像基盤モデルと言語モデルを用いてVLMを構築し，経時三次元CT画像の所見文を生成するものである。国立情報学研究所医療ビッグデータ研究センターの画像データベースとそれに付随する所見文を用いてモデルの訓練を行っている。

本稿では，RSNA 2025におけるわれわれの研究グループのEducation Exhibitの内容に基づき，VLMについて簡単に解説するとともに，VLMに基づく三次元CT画像の所見文レポート生成手法について紹介したい。

VLM

1. VLMとは

VLMは，画像とテキストの両方を処理できるAIモデルである。一般的には，入力された画像に映っている物体やその画像が撮影されたシーンについて，キャプション生成をすることができるものである。放射線科の領域で見ると，CT画像やMR画像の読影レポート生成に利用できるものである。

典型的なVLMは，画像エンコーダ，画像特徴をLLMの埋め込み空間へ写像する射影層（アダプタ），およびテキストの処理と出力の生成を担う大規模言語モデル（LLM）といった要素から構成される。入力画像と「キャプションを生成せよ」といったテキストプロンプトをそれぞれエンコーダで処理し，統合した上でLLMを介して出力を生成する。

このようなVLMの実装には，LLaVA型のアーキテクチャがよく利用される[4]。LLaVA型のアーキテクチャは，画像エンコーダの出力をアダプタでLLMの埋め込み空間へ写像し，テキストはLLM自身がトークン化・埋め込みすることで，両者を同じ形式のトークン列としてLLMに入力し，LLMが最終的な出力を生成する構成である。LLaVA型のアーキテクチャVLMは，画像も言語も同じトークン列として取り扱うものであり，LLMをマルチモーダル拡張したものとも言える。

2. VLMの仕組み

VLMがテキストと画像を統一的に扱えるカギは，両者を同じ形式のトークン（埋め込みベクトル）の列に変換する点にある。テキストのトークン化においては，人間が読めるテキストを，ニューラルネットワークが処理できるベクトル列へと変換する。テキストは，まずトークン（直感的には単語に近い小さな断片）に分割され，各トークンに数値ID（トークンID）が割り当てられる。例えば，「I love my dog.」→["I", "love", "my", "dog", "."]→[23, 65, 2356, 788, 2]のような変換を行う。各トークンは，埋め込み行列Eを通じて実数ベクトルへ写像される。Eは語彙数×埋め込み次元

〈0913-8919/26/￥300/論文/JCOPY〉