

ローカル大規模マルチモーダルモデルによる画像 ・PDF形式文書データの構造化の取り組み

BonBon株式会社 Medical AI R&D Engineer 林 樹永

主題

電子カルテ内の非構造化データの構造化に取り組んでいる。

- 電子カルテが医療文書の構造化を可能にした。
- 一方で、電子カルテは非構造化データも連携される。
 - PDFや画像の形式
 - 規定がない文書
- これらは自由記述形式で非構造化データである。
- 主に検査結果等経過記録で構造化することで医療情報の活用が発展する。

過去の取り組み：OCR

PDF/画像形式の医療文書にOCRを行い、文字列とその座標から構造解析を実施した。

文字認識

画像



文字領域検出



```
{
  "description": "WAITING",
  "boundingPoly": {
    "vertices": [
      {
        "x": 344,
        "y": 821
      },
      {
        "x": 2025,
        "y": 879
      },
      {
        "x": 2016,
        "y": 1127
      },
      {
        "x": 335,
        "y": 1069
      }
    ]
  }
},
```

OCRの場合の課題

OCR結果からテキストマイニングと構造解析により傷病名抽出や事実性判定が実現できたがルールベースの構造解析は汎用性を欠く。

- 大量の辞書を用意する。
- 様々な文書構造に合わせた座標解析を用意する。
- 文書画像以外の形式に対応できない。

大規模マルチモーダルモデルで画像・PDF形式の文書画像の汎用的構造化に取り組む

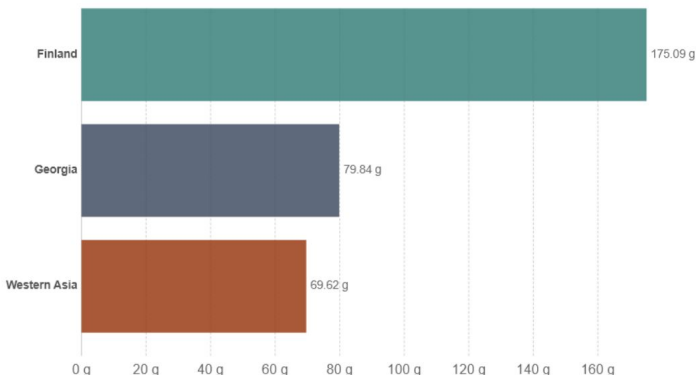
過去の取り組み：大規模マルチモーダルモデル

画像や自然言語等の様々な形式のデータセットを用いた機械学習タスクで訓練したモデルで文書画像データから内容の抽出やグラフの説明をする等の推論能力を持つ。

User What is the sum of average daily meat consumption for Georgia and Western Asia?
Provide a step-by-step reasoning before providing your answer.

Daily meat consumption per person, 1997

Average daily meat consumption per person, measured in grams per person per day. Countries with daily meat consumption greater than the expected EU average of 165g per person are shown in red; yellow are those countries below 165g but exceeding the more ambitious limit of 40g per person; and in blue are those below 40g per person.



Source: UN Food and Agricultural Organization (FAO)

CC BY

GPT-4 To find the sum of average daily meat consumption for Georgia and Western Asia, we will follow these steps:

1. Identify the average daily meat consumption for Georgia.
2. Identify the average daily meat consumption for Western Asia.
3. Add the values from steps 1 and 2.

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.

Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.

Step 3: Add the values from steps 1 and 2.

$$79.84 \text{ g (Georgia)} + 69.62 \text{ g (Western Asia)} = 149.46 \text{ g}$$

So, the sum of average daily meat consumption for Georgia and Western Asia is 149.46 grams per person per day.

グラフのGeorgiaとWestern Asiaの食肉消費量の平均を合算してください。

Georgiaの食肉消費量の平均は79.84g, Western Asiaは69.62gなので合算して149.46gです。

補足: Google Cloud PlatformのVertex AI

Google Cloud PlatformのVertex AIは機械学習に特化した様々なサービスが使える。Vertex AI の Model Garden によって様々な学習済みモデルを API形式で利用可能である。

Model Garden

EXPLORE GENERATIVE AI

VIEW MY MODELS

DEPLOY FROM HUGGING FACE

Modalities

Language48

Vision86

Tabular6

Document5

Speech1

Video4

Multimodal6

Tasks

Generation56

Classification51

Detection33

Extraction17

Recognition17

Translation12

Embedding3

Segmentation7

Retrieval1

Open vocabulary detection2

Open vocabulary segmentation2

Search models

Browse, customize, and deploy machine learning models with Model Garden. Choose from models created by Google and other providers.

Gemini

Imagen 2
Generate images with text prompts

Gemma

Chirp
Generate speech-to-text

Sort by: TrendingNewestLast UpdateAlphabetical

Foundation models

SHOW ALL (70)

Pre-trained multi-task models that can be further tuned or customized for specific tasks. Models marked with + are available in Vertex AI Studio.

Gemini 1.0 Pro

The best performing Gemini model with features for a wide range of tasks

Gemini 1.5 Pro

Created from the ground up to be multimodal (text, images, videos) and to scale across a wide range of tasks

Gemini 1.0 Pro Vision

Created from the ground up to be multimodal (text, images, videos) and to scale across a wide range of tasks

Imagen 2 for Generation and Editing

Use text prompts to generative novel images, edit existing ones, edit parts of an image with a mask and more.

Claude 3 Opus (Pre Anthropic)

it power text model, navigate tasks and open-ended with remarkable fluency

Fine-tunable models

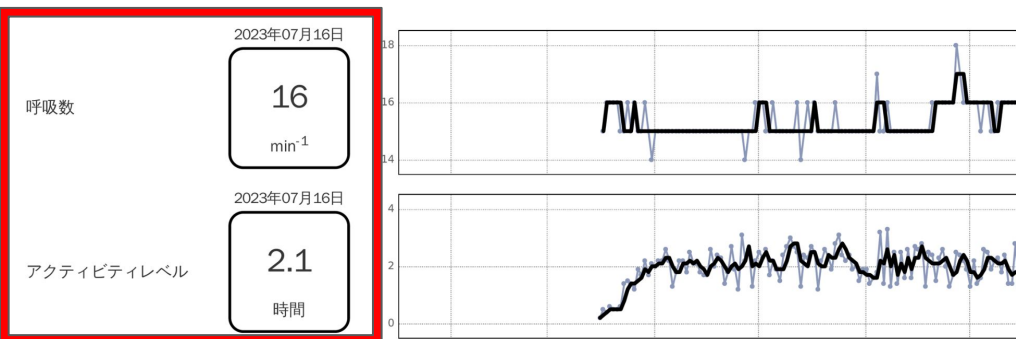
SHOW ALL (44)

Models that data scientists can further fine-tune through a custom notebook or pipeline.

過去の取り組み：大規模マルチモーダルモデルによる構造化

Google Cloud PlatformのVertex AI サービスで管理されている大規模マルチモーダルモデルAPIを用いて入力に画像とプロンプトを与え構造化解析を行い一定の成果を得た。

この画像に記載されている測定値と単位と項目名を json形式にしてください



```
{  
  '呼吸数': {'value': 16, 'unit': 'min-1'},  
  'アクティビティレベル': {'value': 2.1,  
    'unit': '時間'},  
  ....}
```

大規模マルチモーダルモデルによる構造化を踏まえて

大規模マルチモーダルモデルAPIによる高精度な構造化が可能になった。さらに下記背景を踏まえてローカルモデルの検証を行うことは有益と判断した。

- 公開モデルがAPIモデルに並ぶ結果を出している。
 - 公開モデルとはソースコードや学習結果を計算機環境にダウンロードして動かせるモデルのこと。ローカルモデルと呼称される。
- APIの場合リクエストに文書画像データを含めることになり、外部にデータを渡すことになる。ローカルモデルは計算機環境でデータ処理できる。
- APIの場合提供会社がシステム制限を定めている。ローカルモデルは計算機環境を拡張することができる (※コストを無視する場合)。

画像・PDF形式の自由記述文書データ構造化をローカルモデルで検証する

精度検証方法

- データ受領
 - 日本医療ネットワーク協会(現 LDI)、宮崎市郡医師会病院 より匿名化済みデータを受領。PDFの画像化及びデータの選定を行いおよそ200枚の画像データを得た。
- 採用モデル
 - 下記の種類のAPI提供モデル及びローカルモデルを採用する。APIモデルはGoogle Cloud PlatformのVertex AI サービスを用いる。ローカルモデルはHugging Faceより入手する。
- 評価方法
 - 5種類の画像を選定し目視で文書構造化の正解率を判定 (X個所のうちY箇所について構造化できた場合 Y/X で構造化できていると計算する)。

モデル名	概要	提供形態	合計パラメータ数
Claude 3.7 Sonnet	Anthropicが開発するモデル	API	不明
Gemini 2.0	Googleが開発するモデル	API	不明
Qwen 2 VL	Alibabaが開発するモデル	ローカルモデル	2B
Qwen 2.5 VL	Alibabaが開発するモデル	ローカルモデル	3B
Gemma 3	Googleが開発するモデル	ローカルモデル	4B
Janus pro	DeepSeekが開発するモデル	ローカルモデル	1B

補足: Hugging Face

Hugging Faceとは機械学習モデルやデータセットを公開・利用することができる GitHubのようなプラットフォームである。利用者はHugging Faceが提供するライブラリを用いて公開されているモデルを各自の計算機環境にダウンロードし利用することができる。

The screenshot displays the Hugging Face Models interface. On the left, there are navigation tabs for 'Tasks' and 'Libraries', with 'Tasks' selected. Below these, a search bar 'Filter Tasks by name' is present. The main content area is divided into two sections: 'Multimodal' and 'Computer Vision'. The 'Multimodal' section includes buttons for 'Audio-Text-to-Text', 'Image-Text-to-Text', 'Visual Question Answering', 'Document Question Answering', 'Video-Text-to-Text', 'Visual Document Retrieval', and 'Any-to-Any'. The 'Computer Vision' section includes buttons for 'Depth Estimation', 'Image Classification', 'Object Detection', 'Image Segmentation', 'Text-to-Image', 'Image-to-Text', 'Image-to-Image', 'Image-to-Video', 'Unconditional Image Generation', 'Video Classification', 'Text-to-Video', 'Zero-Shot Image Classification', 'Mask Generation', 'Zero-Shot Object Detection', 'Text-to-3D', 'Image-to-3D', 'Image Feature Extraction', and 'Keypoint Detection'. The main content area on the right shows a list of models under the 'Models' tab, with a total of 1,688,079 models. The list includes models like 'nvidia/parakeet-tdt-0.6b-v2', 'ACE-Step/ACE-Step-v1-3.5B', 'Lightricks/LTX-Video', 'nari-labs/Dia-1.6B', 'lodestones/Chroma', 'tencent/HunyuanCustom', 'PrimeIntellect/INTELLECT-2', 'cognition-ai/Kevin-32B', 'JetBrains/Mellum-4b-base', 'black-forest-labs/FLUX.1-dev', 'Qwen/Qwen3-235B-A22B', 'deepseek-ai/DeepSeek-Prover-V2-671B', 'Qwen/Qwen3-30B-A3B', and 'hexgrad/Kokoro-82M'. Each model entry shows its name, a brief description, and statistics like 'Updated', 'Downloads', and 'Likes'.

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

Audio-Text-to-Text Image-Text-to-Text

Visual Question Answering

Document Question Answering Video-Text-to-Text

Visual Document Retrieval Any-to-Any

Computer Vision

Depth Estimation Image Classification

Object Detection Image Segmentation

Text-to-Image Image-to-Text Image-to-Image

Image-to-Video Unconditional Image Generation

Video Classification Text-to-Video

Zero-Shot Image Classification Mask Generation

Zero-Shot Object Detection Text-to-3D

Image-to-3D Image Feature Extraction

Keypoint Detection

Models 1,688,079 Filter by name

Full-text search Sort: Trending

nvidia/parakeet-tdt-0.6b-v2
Automatic Speech Recognition · Updated 12 days ago · 91.8k · 761

ACE-Step/ACE-Step-v1-3.5B
Text-to-Audio · Updated 3 days ago · 372

Lightricks/LTX-Video
Text-to-Video · Updated 7 days ago · 241k · 1.42k

nari-labs/Dia-1.6B
Text-to-Speech · Updated 7 days ago · 154k · 2.08k

lodestones/Chroma
Text-to-Image · Updated 1 day ago · 447

tencent/HunyuanCustom
Image-to-Video · Updated 4 days ago · 111

PrimeIntellect/INTELLECT-2
Updated about 7 hours ago · 25 · 107

cognition-ai/Kevin-32B
Updated 7 days ago · 568 · 106

JetBrains/Mellum-4b-base
Text Generation · Updated 6 days ago · 3.31k · 329

black-forest-labs/FLUX.1-dev
Text-to-Image · Updated Aug 16, 2024 · 2.64M · 10.2k

Qwen/Qwen3-235B-A22B
Text Generation · Updated 12 days ago · 94.3k · 783

deepseek-ai/DeepSeek-Prover-V2-671B
Text Generation · Updated 13 days ago · 7.93k · 760

Qwen/Qwen3-30B-A3B
Text Generation · Updated 13 days ago · 163k · 530

hexgrad/Kokoro-82M
Text-to-Speech · Updated Apr 11 · 1.85M · 4.3k

文書画像構造化結果

- API提供モデルの Claude 3.7 SonnetおよびGemini 2.0はいずれも高い精度の構造化を実現した。
- ローカルモデルの中では Qwen2.5-VLおよびQwen2-VLの構造化精度が高かった。
- Gemma 3やJanus proの構造化精度は低かった。多くの推論ミスがあり構造化もできていなかった。

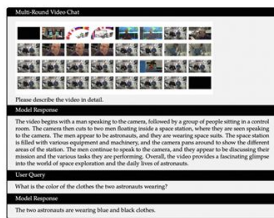
モデル名	Claude 3.7 Sonnet	Gemini 2.0	Qwen2.5-VL	Qwen2-VL	Gemma 3	Janus-Pro
画像1	6/7	6/7	5/7	3/7	0/7	0/7
画像2	6/6	6/6	4/6	3/6	2/6	0/6
画像3	2/2	2/2	0/2	0/2	0/2	0/2
画像4	5/5	5/5	3/5	3/5	0/5	0/5
画像5	8/10	10/10	6/10	7/10	0/10	0/10

参考:Qwen-VLシリーズ

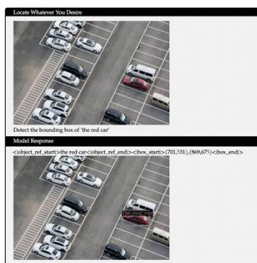
Qwen-VLシリーズはAlibaba Cloudによって開発された公開マルチモーダルモデルであり、様々な画像・言語データの推論が可能である。



General Chat



Video Understanding



Grounding



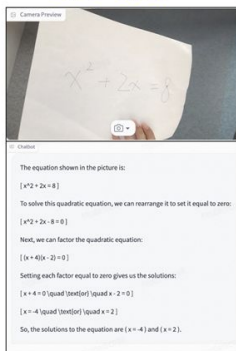
Multilingual OCR



Long Document Understanding



Math & Code



Live Chat

Formula Recognition

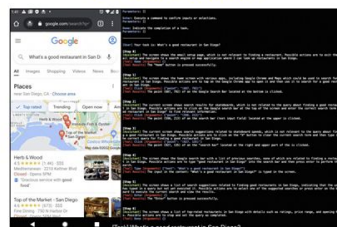


Qwen2-VL

Function Calling



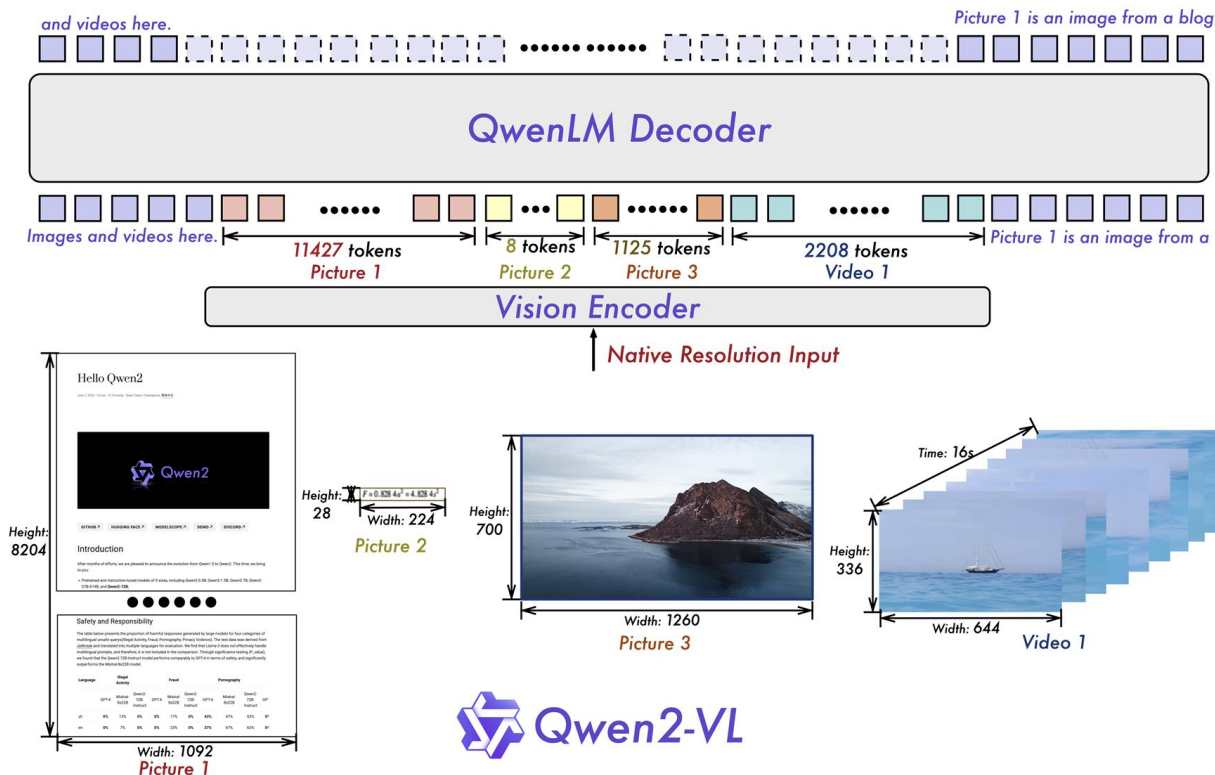
UI Interaction



Beyond

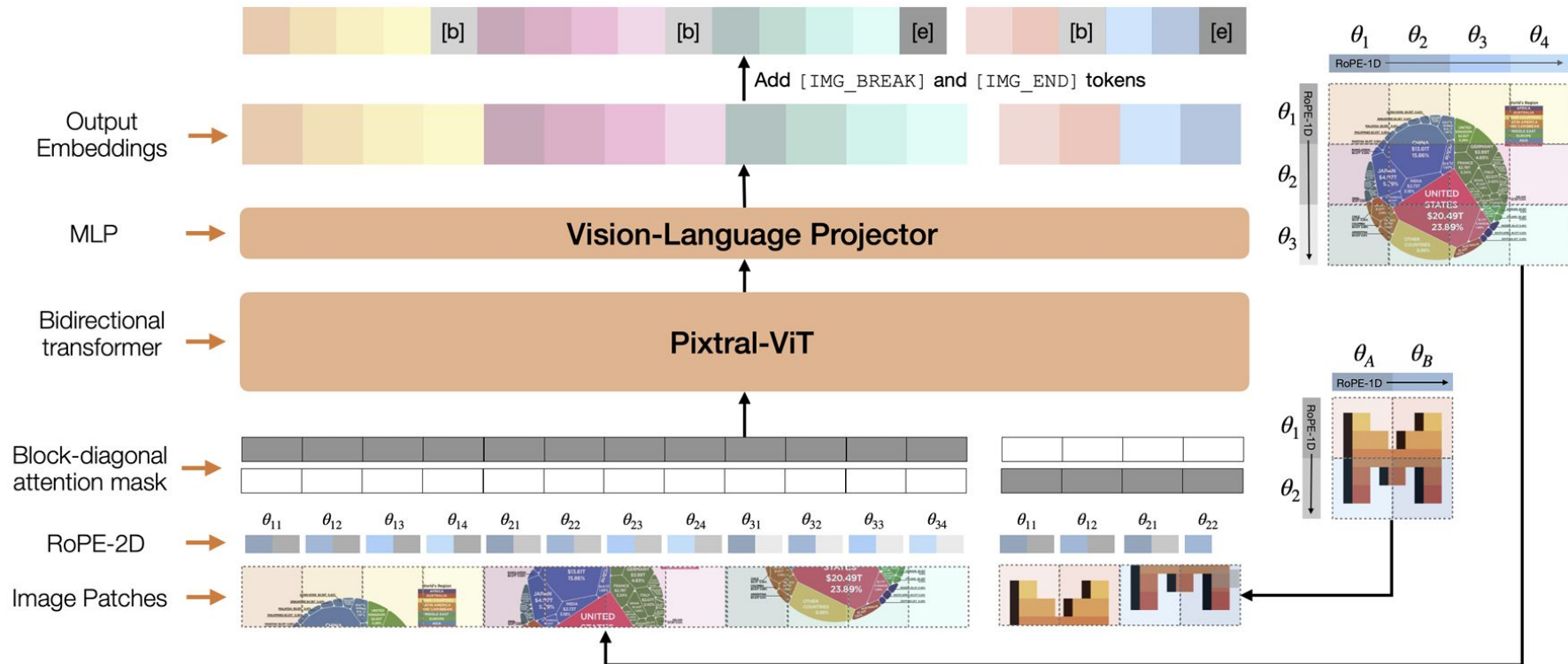
参考:Qwen2-VL以降で採用されるモデル構造

Qwenシリーズの言語モデルの Decoderと画像および動画解析用の Vision Encoderを採用しVision Encoderに工夫がなされている。



参考: Qwen2-VL以降の Vision Encoder

RoPE-2DというVision Encoderの技術を採用し、画像とその位置情報を言語モデルに使うことができる。これによって様々な解像度の画像を元のサイズで推論することができる。



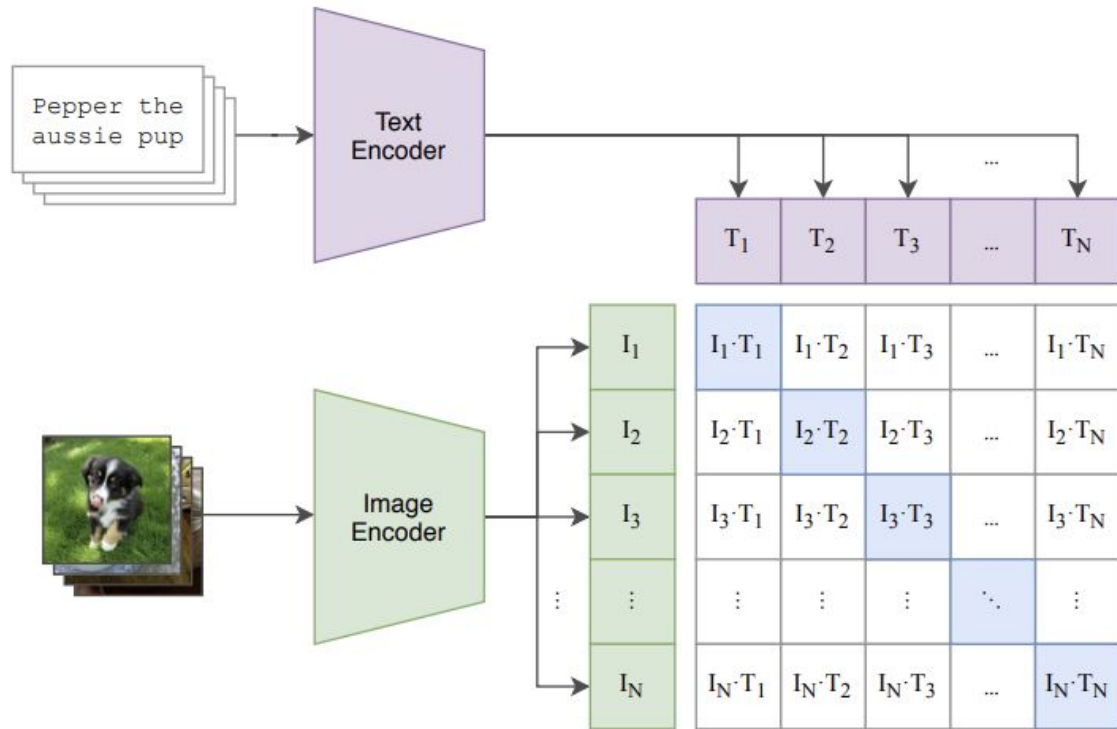
参考: Qwen 2.5-VLの文書画像関連のベンチマークの精度比較

Qwen 2.5-VLは様々な文書画像関連のベンチマークデータセットで API提供モデルに並ぶ精度を出している。

Datasets	Claude-3.5 Sonnet	Gemini 1.5 Pro	GPT 4o	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
OCR-related Parsing Tasks							
CC-OCR	62.5	73.0	66.9	64.7	79.8	77.8	74.5
OmniDocBench _{edit en/zh} ↓	0.330/0.381	0.230/ 0.281	0.265/0.435	0.275/0.324	0.226 /0.324	0.308/0.398	0.409/0.543
OCR-related Understanding Tasks							
AI2D _{w. M.}	81.2	88.4	84.6	89.1	88.7	83.9	81.6
TextVQA _{val}	76.5	78.8	77.4	83.4	83.5	84.9	79.3
DocVQA _{test}	95.2	93.1	91.1	95.1	96.4	95.7	93.9
InfoVQA _{test}	74.3	81.0	80.7	84.1	87.3	82.6	77.1
ChartQA _{test Avg.}	90.8	87.2	86.7	88.3	89.5	87.3	84.0
CharXiv _{RQ/DQ}	60.2 /84.3	43.3/72.0	47.1/84.5	42.4/82.3	49.7/ 87.4	42.5/73.9	31.3/58.6
SEED-Bench-2-Plus	71.7	70.8	72.0	71.3	73.0	70.4	67.6
OCRBench	788	754	736	854	885	864	797
VCR _{En-Hard-EM}	41.7	28.1	73.2	-	79.8	80.5	37.5
OCR-related Comprehensive Tasks							
OCRBench_v2 _{en/zh}	45.2/39.6	51.9/43.1	46.5/32.2	49.8/52.1	61.5 /63.7	56.3/57.2	54.3/52.1

参考: Gemma 3, Janus-ProのVision Encoder

CLIPのVision Encoderは入力画像を固定長にリサイズする。文書画像のような文字が多い画像の場合リサイズによって情報が落ち推論精度が低くなることが考えられる。



Qwen2.5-VLモデルの推論誤りの考察

培養・同定	1	Staphylococcus epidermidis	陽性
	2	Candida glabrata	1+
	3	嫌気性菌	陰性
	4		
	5		
	6		
	7		
	8		
	9		
	10		

出力結果

「4. **培養・同定結果**:\n - 培養・同定の結果が記載されており、陽性菌や陰性菌の種類がリストアップされています。\\n - 例として、「Staphylococcus epidermidis」や「Candida glabrata」などの菌が記載されています。」

- 上記が代表的な推論誤りの例で、内容は読み取れているが検査結果が抽出できていないように見える。プロンプトの調整やモデルサイズを上げて推論を試みる。それでも構造化ができないなら Qwen2.5-VLモデルの学習を試みる。
- この例以外には構造化が途中で止まるものがある。出力する文字列長を調整する。

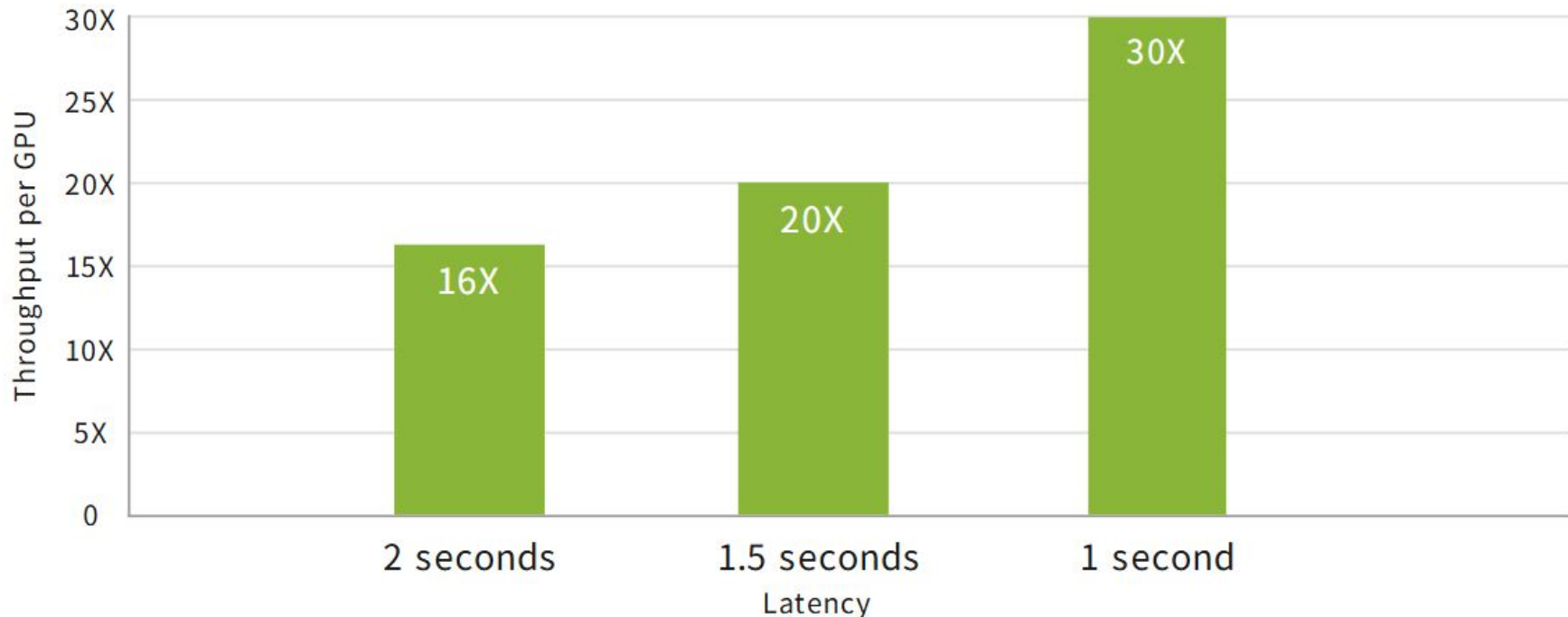
速度計測方法と結果

- NVIDIA A100 40GBでランダムに画像 200枚程度を用いた平均速度計測を実施した。
- 計測対象のローカルモデルは構造化精度の高い Qwen2.5-VL、Qwen2-VLおよびそれらの量子化済みローカルモデルとした。
- 平均時間はAPIモデルのほうがローカルモデルより早かった。
- 千年カルテに1日に連携される医療画像は 1~3万枚とされている。それらを 24時間以内に処理したいなら 1枚あたり2.8 ~ 8.6秒で推論する必要がある。ローカルモデルはその条件を満たさなかったが、高速に推論するハードウェアの採用や複数 GPUによる推論により処理時間は短縮される。

モデル名	GPUメモリ(GB)	提供形態	合計パラメータ数	1枚当たり処理時間平均(秒)
Claude 3.7 Sonnet	-	API	不明	18.0
Gemini 2.0	-	API	不明	8.2
Qwen2.5-VL	16.0	ローカルモデル	3B	29.3
Qwen2-VL	15.1	ローカルモデル	2B	26.2
Qwen2.5-VL-AWQ	12.0	ローカルモデル	3B	39.9
Qwen2-VL-AWQ	3.9	ローカルモデル	2B	23.1

参考：速度向上が見込める GPUの比較

今回用いたNVIDIA A100と比較してNVIDIA H100は言語モデルの推論速度の点で優れている。H100を採用し文書画像推論の高速化は十分見込める。



H100 to A100 Comparison - Relative Performance

まとめ

- ローカルマルチモーダルモデルで画像、PDF形式の文書データの構造解析が一部実現できた。
- 指示文の調整・大きなモデルの検討やGPU等計算機の条件を変えて構造化の精度や速度を確認していく。

謝辞

今回の取り組みのテーマ設定およびデータ提供してくださった、宮崎市郡医師会病院及びライフデータイニシアティブ・日本医療ネットワーク協会・京都大学 吉原博幸先生に感謝申し上げます

引用

- 吉原博幸. "1. 千年カルテ: 次世代医療基盤法に基づく診療データの研究開発への二次利用 薬剤疫学 27.1 (2022): 3-10.
- Google Cloud Vision API <https://cloud.google.com/vision?hl=ja>
- OpenAI. "GPT-4 Technical Report." *ArXiv* abs/2303.08774 (2023): n. pag.
- Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." *arXiv preprint arXiv:2312.11805* (2023).
- <https://huggingface.co/tasks/document-question-answering>
- Huang, Yupan, et al. "Layoutlmv3: Pre-training for document ai with unified text and image masking." *Proceedings of the 30th ACM International Conference on Multimedia*. 2022.
- Mathew, Minesh, Dimosthenis Karatzas, and C. V. Jawahar. "Docvqa: A dataset for vqa on document images." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021.
- Jaume, Guillaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. "Funsd: A dataset for form understanding in noisy scanned documents." *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2. IEEE, 2019.
- Hugging Face <https://huggingface.co/models>
- NVIDIA H100 Tensor コア GPU <https://www.nvidia.com/ja-jp/data-center/h100/>
- Claude 3.7 Sonnet and Claude Code <https://www.anthropic.com/news/claude-3-7-sonnet>
- Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmlR, 2021.
- Agrawal, Pravesh, et al. "Pixtral 12B." *arXiv preprint arXiv:2410.07073* (2024).
- Wang, Peng, et al. "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution." *arXiv preprint arXiv:2409.12191* (2024).
- Bai, Shuai, et al. "Qwen2. 5-vl technical report." *arXiv preprint arXiv:2502.13923* (2025).
- Team, Gemma, et al. "Gemma 3 technical report." *arXiv preprint arXiv:2503.19786* (2025).
- Team, Gemma, et al. "Gemma: Open models based on gemini research and technology." *arXiv preprint arXiv:2403.08295* (2024).
- Chen, Xiaokang, et al. "Janus-pro: Unified multimodal understanding and generation with data and model scaling." *arXiv preprint arXiv:2501.17811* (2025).