



大規模マルチモーダルモデル (LMM) の展開

相澤清晴

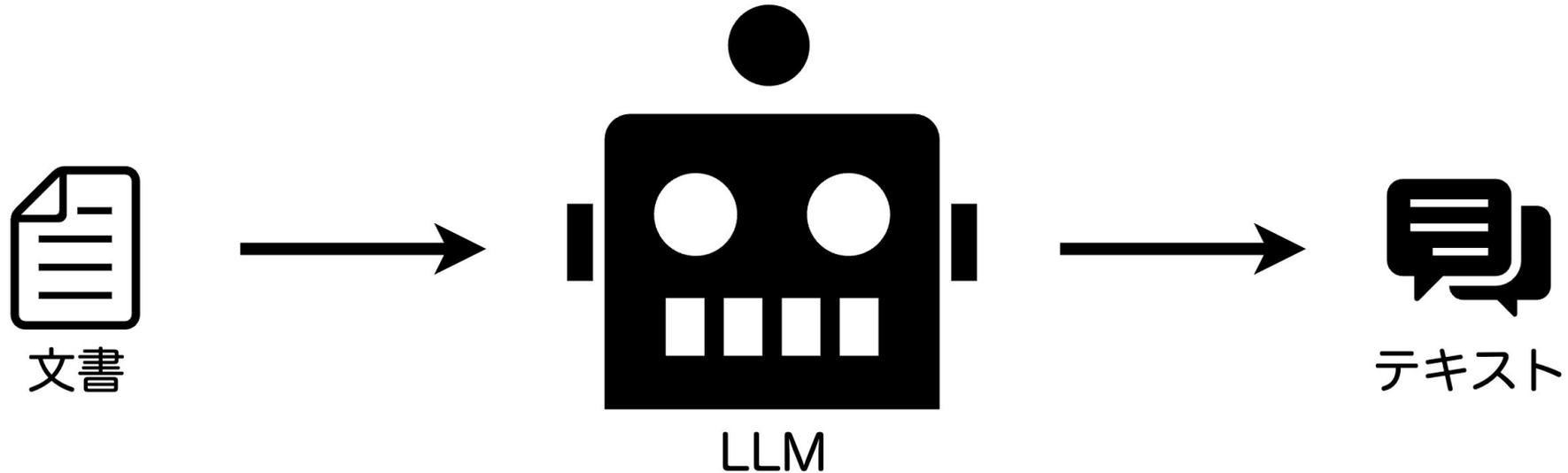
東京大学 大学院情報理工学系研究科

VRセンター

目次

- LMMについて
- 画像を交えた指示を行うLMMについて
- LMMの技術の要
- LLaVAについての概要
- その事例
- まとめ

LLM: 大規模言語モデル



言語処理的タスク

テキスト生成

翻訳

要約

質問応答(Q&A)

会話

コード生成

感情分析

文法やスタイルの修正

キーワード抽出

分類

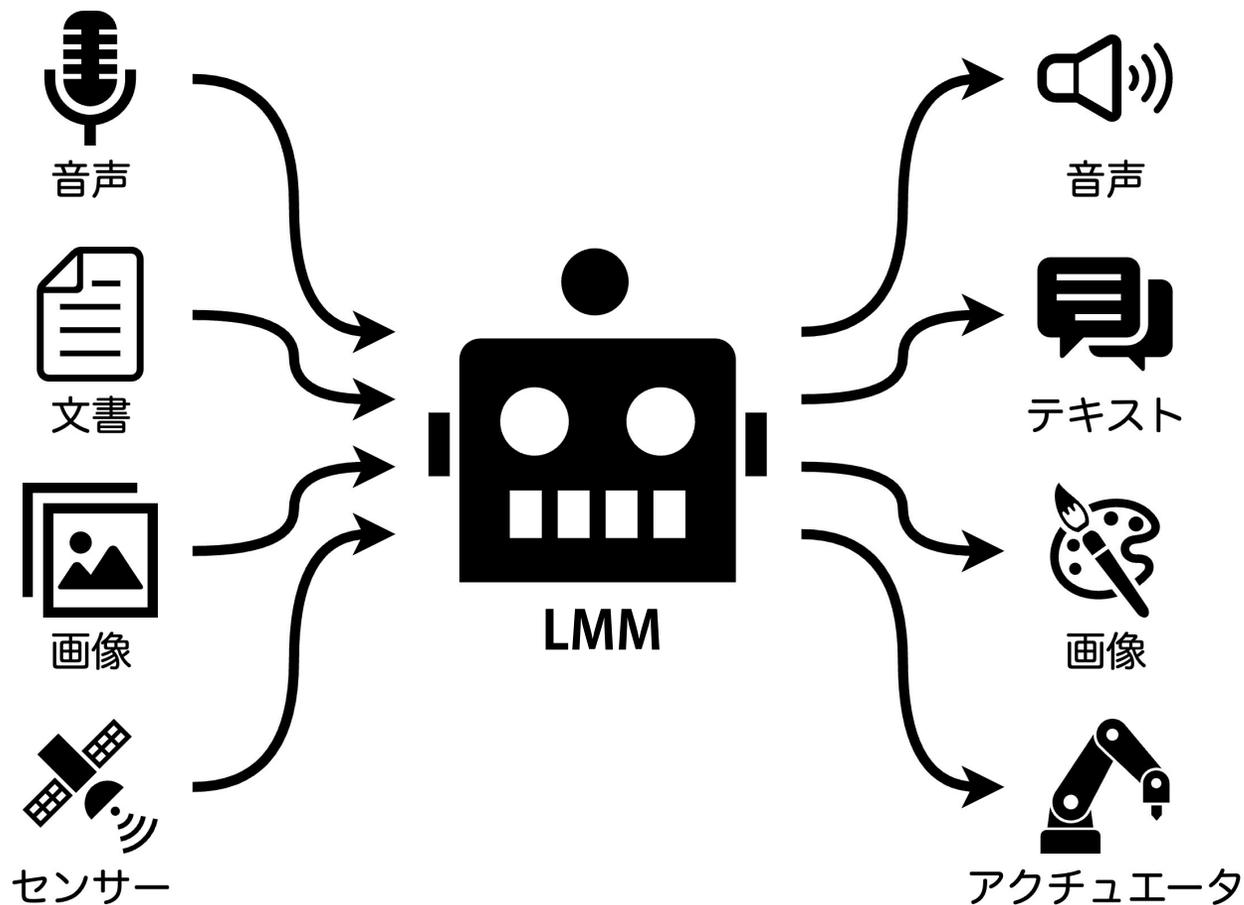
マルチモーダルの必要性



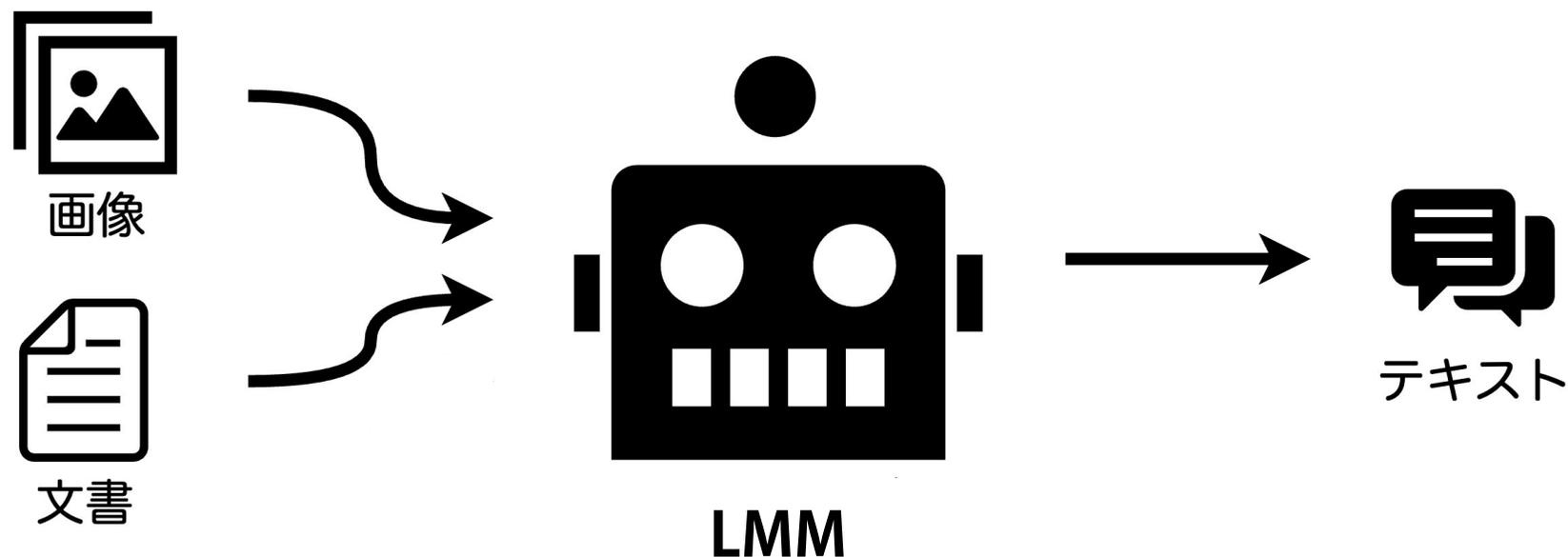
(DALL-E3による生成画像)

LMM: 大規模マルチモーダルモデル

画像に対しての問い合わせを行うなど、
入出力のモダリティが広がり、さらに進化し始めている。



画像を用いた対話やQAでは、



以下、画像に限定して話を進める。

デモ

- GPT-4

LMMの展開

(注) 網羅的ではない。

• 一般サービス

2023.10 GPT-4 (OpenAI)

2023.12 Bard Gemini
(Google)

データ, コード, モデルが
公開されている。

• 学術発表

2022.04 Flamingo (DeepMind)

2023.01 BLIP2 (SalesForce)

2023.03 OpenFlamingo (LAION)

2023.04 Mini-GPT4 (KAUST)

2023.08 Qwen VL (Alibaba)

2023.04 LLaVA (Microsoft Research, UW)

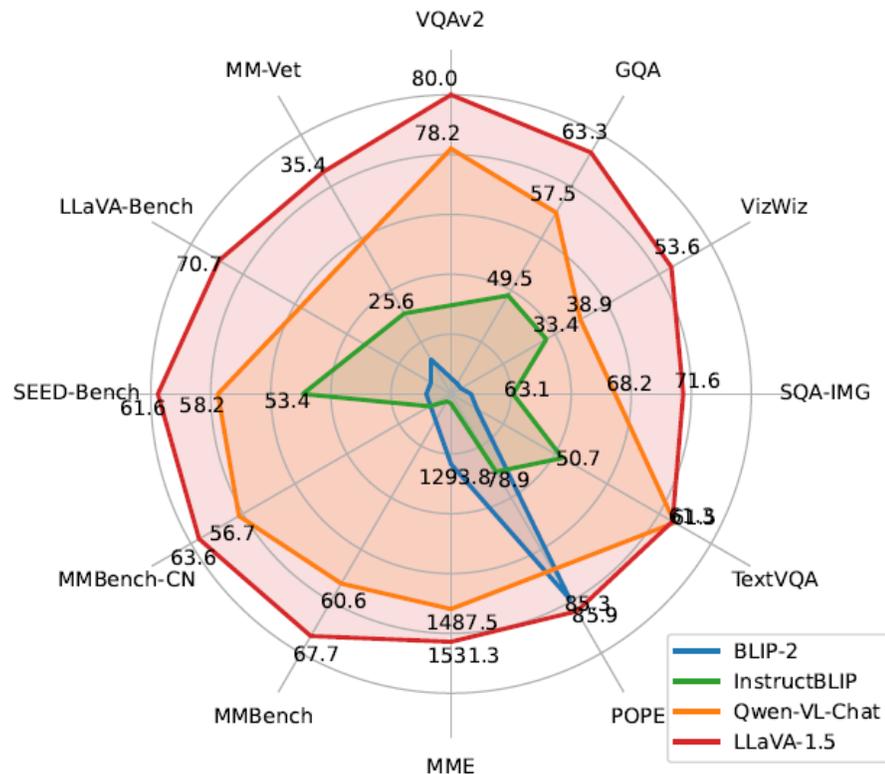
2023.10 LLaVA 1.5

LMMの技術の要

- (1) 大規模言語モデル (LLM) の基盤
- (2) 画像を言語と同様にLLMで扱うセマンティクス
- (3) 画像をもちいた会話や質疑応答性能

LLaVA (LLaVA 1.5, 23.10)

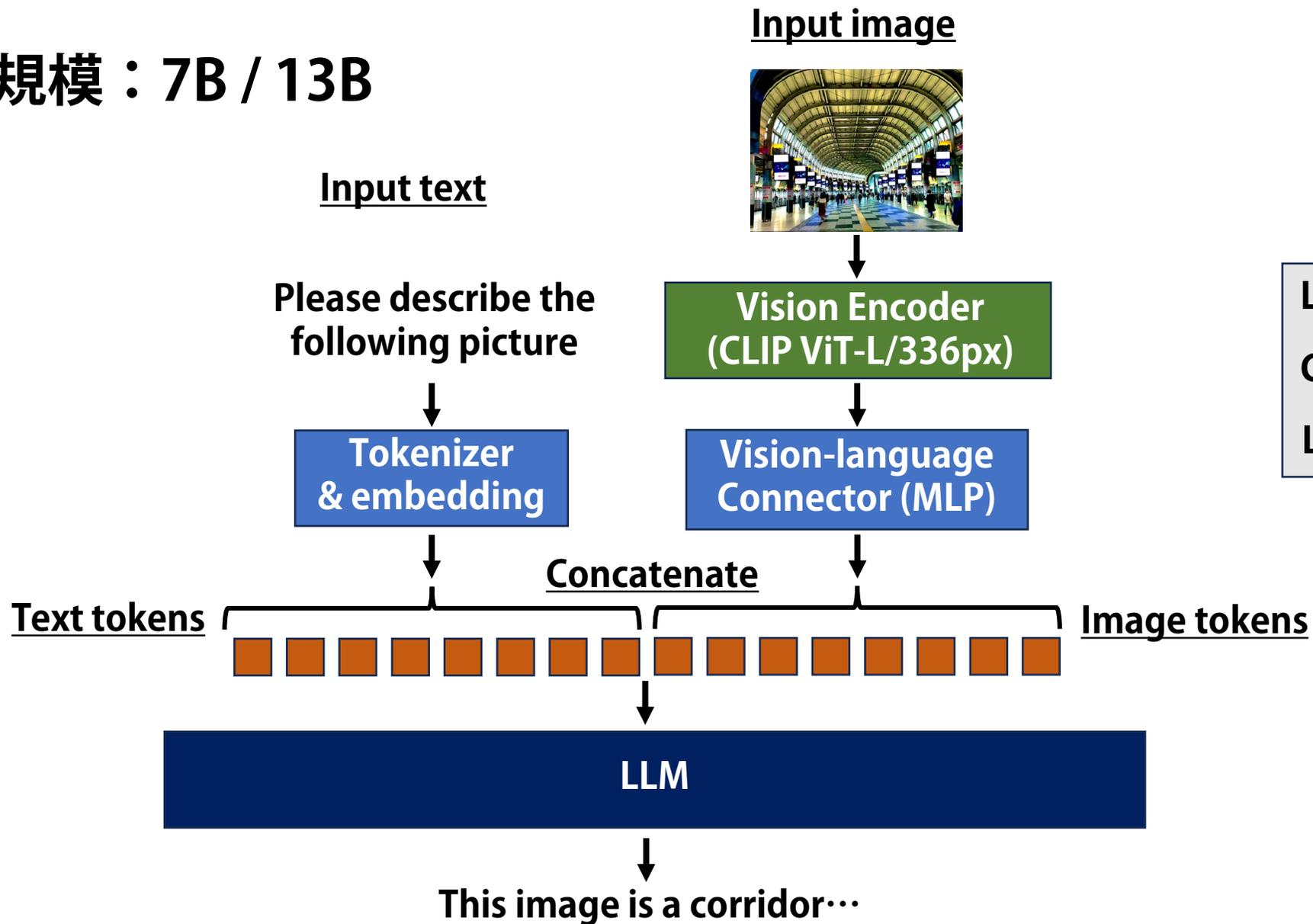
- シンプルで明解なアーキテクチャであり， 23.10時点で，ほとんどのベンチマークデータセットに対して，公開モデルにてもっともよい性能を示している。
- モデル，コード，データが公開されている。



H.Liu, et. al, Visual instruction tuning, arXiv 2304.08485
H.Liu, et. al, Improved baseline with visual instruction tuning, arXiv 2310.03744

LLaVA アーキテクチャ

モデル規模：7B / 13B



LLaVA: Microsoft, UW

CLIP: OpenAI

LLM: Meta, LMSYS org

LLaVA

(1) 大規模言語モデル (LLM)

Llama(Meta) の派生モデルであるVicuna(LMSYS org)を大規模言語モデルに利用. Llama, Vicunaいずれもコード, データ, モデルが公開されている. 7B, 13Bのサイズのモデルが用いられた.

Vicunaの対話性能は, ChatGPTの90%と評価されている.

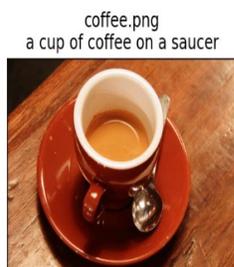
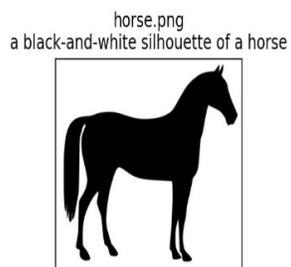
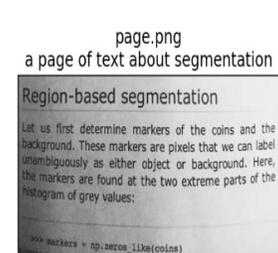
(2) 画像をLLMで言語と同様に扱うセマンティクス

- * CLIPの画像エンコーダの利用

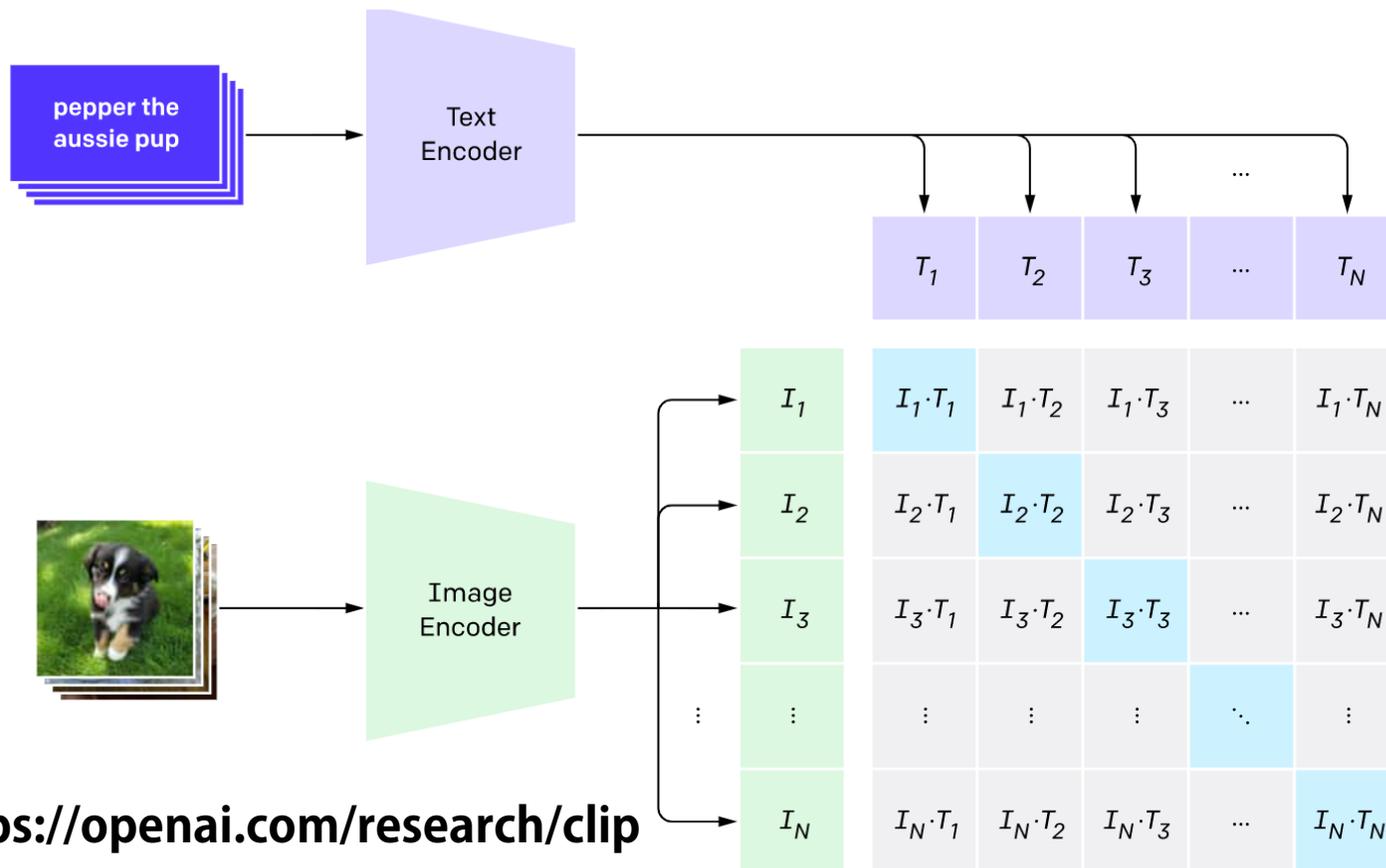
- * 画像特徴のLLMへのアライメント

CLIP 画像言語(VL)モデル (OpenAI 2021)

- インターネットから収集した膨大な画像テキスト対(4億)を用いる。
- 膨大な量のデータを用いたシンプルな学習アルゴリズム。N個のペアに対して、本来のペアが近づくように、テキストと画像のエンコーダーを学習する。
- 画像特徴と言語特徴を近づけることができる。



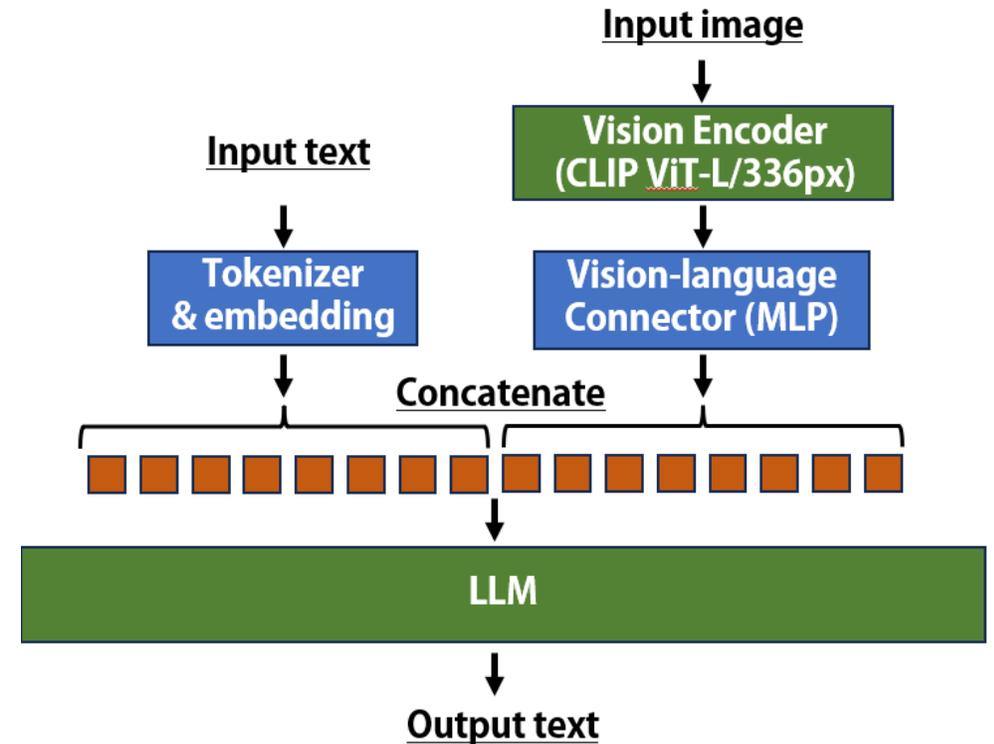
Interacting with CLIP.ipynb



<https://openai.com/research/clip>

LLaVA: 画像特徴のアライメント

- CLIPの画像エンコーダーはそのまま用いて、MLPをコネクタとして挟み、その学習をすることで、画像特徴をLLMに適合させる。
- コネクタの事前学習では、LLMも固定し、比較的小規模なimage-text pairのデータ (595K) を用いる。



LLaVA

(3) Visual Instruction Tuning

一般に、Instruction tuningとは、LLMが指示に従った応答を行うようにするため、様々な指示と応答からなるデータセットでの学習であり、LLMの汎化性能を効率よく向上できる。それに倣い、画像を用いたVQA(Visual Question Answering)データを用いてInstruction Tuningを行う。

LLMとコネクタMLPのファインチューニング

*テキストだけでGPT-4で生成したチャットデータによるVQAデータセット

LLaVA -Instruct158K dataset

-会話: 58K

-詳細な記述: 23K

-複雑な推論: 77K

*マルチモーダルな複数のVQAデータセットの利用。

- 学術タスク向けVQAである

Open-knowledge VQA,

- OCRのVQA,

-領域レベルVQA 等々

すべて合わせて、LLaVA1.5では、665K 程度

Visual Instructions

画像のデータセットから、その画像のキャプションとオブジェクトの位置を示す B B (バウンディングボックス) の情報を用い、テキストだけのGPT-4に与えて、Q Aを作成。

Context type 1: Captions

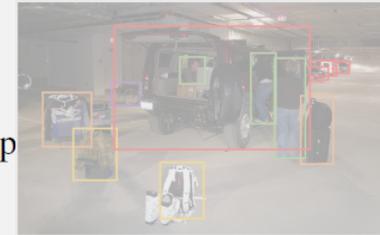
A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

LLaVAの出力



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

User

LLaVA

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

[Start a new conversation, and clear the history]

User

LLaVA

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention **due to his unconventional choice of ironing his clothes on top of a moving car**. The city street around him is bustling with activity, adding to the unique nature of the scene.

ユーザー：この画像のどこが普通ではないのか？

LLaVA: この画像の**普通でない点**は、ミニバンかバンの荷台でアイロンをかけている男性である。ここはアイロンをかける場所ではなく、通常、アイロンがけは自宅のようなもっと静止した安全な場所で、**アイロン台を使って行うものだからだ**。画像に描かれている光景は、車の上でその場しのぎのアイロンがけをするという奇妙なものだ。これは安全でなく、型破りである。さらに、どうやってこのような不安定な環境でアイロンをかけている間、男性がどのようにバランスを保ち、安定しているのかは不明である。

LLaVAの出力



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

User

LLaVA

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

[Start a new conversation, and clear the history]

User

LLaVA

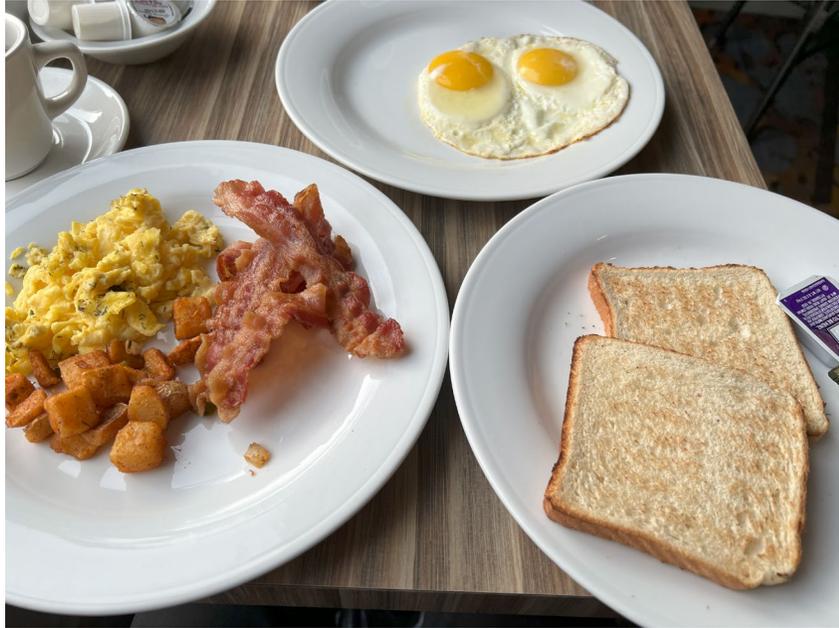
What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention **due to his unconventional choice of ironing his clothes on top of a moving car**. The city street around him is bustling with activity, adding to the unique nature of the scene.

ユーザー：このシーンで何が起
こっているのか？

LLaVA: 画像は、賑やかな市街地
で黄色いSUVの上に立つ男を描い
ている。彼は携帯用のはしごを持
ち、車の上に立ちながらアイロ
ンをかけているようだ。周囲には他
の車、信号機、消火栓がある。さ
らに車の上にいる男の近くに1人、
離れたところにもう1人いる。ハ
シゴの上の男性は、アイロンをか
けるという型破りな方法で注目を
集めているようだ。走行中の車の
上でアイロンをかけるという型破
りな方法で注目を集めているよう
だ。周囲の街路は賑やかだ。この
シーンのユニークさをさらに高め
ている。

LLaVAの出力



This image is a photograph. Please describe what is happening in this photo in detail in English.

The image features a dining table with three white plates filled with breakfast food. The plates are placed on the table, with one plate on the left side, another plate in the middle, and the third plate on the right side. The breakfast consists of eggs, toast, and bacon, making it a hearty meal.

There are also two cups placed on the table, one near the top left corner and the other near the top right corner. The table setting creates an inviting atmosphere for a delicious breakfast.

画像は、**朝食の料理**が盛られた**3枚の白い皿**が置かれた食卓。左側に1枚、真ん中にもう1枚、右側に3枚目の皿が置かれている。朝食は**卵**、**トースト**、**ベーコン**で、ボリュームのある食事である。このテーブルセッティングが、おいしい朝食を食べるための雰囲気を作り出している。

また、**テーブルの上には2つのカップ**が置かれ、1つは左上の角近くに、もう1つは右上の角近くに置かれている。このテーブルセッティングは、おいしい朝食のための魅力的な雰囲気を作り出している。

LLaVAの出力



最速!! 1巻 (松田尚正提供)

コマの内容についての詳細を問うと

The comic panel features two men, one of them wearing a racing suit and the other wearing sunglasses. They are standing in front of a large Brazilian flag, which suggests that the scene takes place in Brazil. The man in the racing suit appears to be a race car driver, while the other man might be a spectator or a fellow racer.

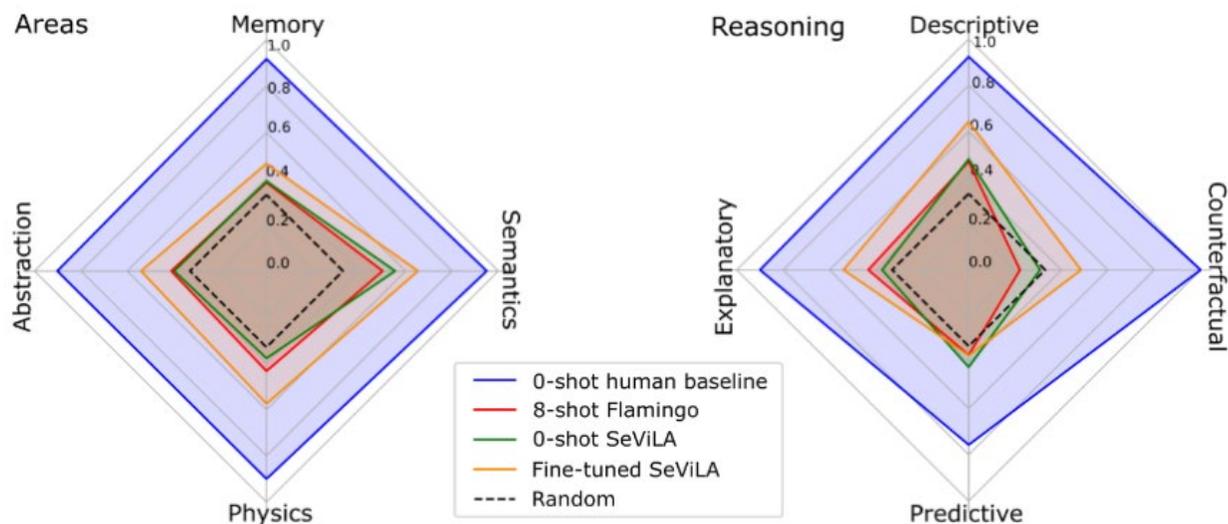
The background of the panel shows a crowd of people, indicating that the event might be a race or a gathering related to racing. The man in the racing suit is holding a trophy, which further supports the idea that he is a race car driver who has just won a race. The man in sunglasses is smiling, possibly celebrating the victory or showing support for the driver. The overall context of the panel seems to be a celebration or a moment of triumph in the world of racing.

この漫画のコマには2人の男性が登場し、1人はレーシングスーツ、もう1人はサングラスをかけている。彼らは大きなブラジル国旗の前に立っており、このシーンがブラジルで起こっていることを示唆している。レーシングスーツを着た男性はレースカーのドライバーのようだが、もう一人の男性は観客かレーサー仲間かもしれない。

コマの背景には群衆が描かれており、レースかレース関連の集まりであることを示している。レーシングスーツの男性はトロフィーを手にしており、彼がレースに勝ったばかりのレーシングドライバーであることをさらに裏付けている。サングラスの男は微笑んでおり、おそらく勝利を祝っているか、ドライバーを応援しているのだろう。このコマの全体的な背景は、レースの世界における祝賀か勝利の瞬間のようだ。

最後に

- 一般画像でなく，特定分野に合わせたLMMの研究開発も進んでいる．バイオメディカル分野の学術論文PubMedのデータを用いて，LLaVAのファインチューンを進めたLLaVA-Med (Microsoft 23.6)などが開発されている．
- ビデオのマルチモーダル解析のための知覚を問うベンチマークが作られ，6つの課題での評価が行われた Perception Test (23.5, DeepMind)では，人との比較があり，まだ差はととても大きく，進展の余地は多い．



まとめ

- LMMの概要
- LLMの発展としてのLMM
- LLaVAの概要
- LLaVAの技術要素
- LLaVAの出力例

ご清聴ありがとうございました。