



Technology White Paper

AIの現場適応を加速する アラヤのVision Language Model 活用事例



Contents

はじめに	2
Vision Language Modelの構造	3
各Vision Language Modelの構造の詳細	3
VLMのビジネス上のインパクト	7
事例から見るVLMを活用するメリット	7
VLMの系譜と発展	10
VLMとVision Foundation Modelの違い	12
VLMの選択	13
VLMの評価	15
ケーススタディ	16
事例1：シーングラフによる状況認識と安全管理システムの実用化	16
事例2：データセットマネジメントのデータセットキャプションング	18
事例3：VLMを用いたロボットの遠隔操作	20
VLM導入における課題	22
モデルサイズの大きさ	22
VLMのバイアス	22
VLM認識に対しての敵対的アタック	23
VLMのリアルタイム性	23
結論 ～VLM導入実現に向けて～	23
アラヤについて	25
お問合せ	25

はじめに

現在、人工知能(AI)の発展は急速に進んでおり、特に視覚情報とテキスト情報を組み合わせて処理できるVision Language Model(以下、VLM)は、ビジネスに新たな可能性を提供しています。VLMとは、画像や動画といった視覚データをテキスト情報と関連付けることで、従来の画像認識技術では対応しきれなかった複雑なタスクにも取り組むことができる技術です。例えば、VLMを用いることで、画像の中に「何があるのか」を識別するだけでなく、それが「どのような状況に置かれているか」「他の物体とどのような関係にあるか」といった文脈や意味を理解することが可能になります。

このように、VLMが多様な用途に対応できる理由は、視覚情報とテキスト情報を組み合わせることで、より深い文脈の理解が可能になるためです。画像やテキストのみの単一の情報(モーダルと言います)では、そのデータが示す内容や意味を限定的にしか解釈できません。画像認識技術だけでは、「何が映っているか」はわかっても、その背景にある状況や意図までは把握できないことが多いですが、VLMは、視覚とテキストを融合することで、画像に映る物体の名称や属性だけでなく、物体同士がどのような関係にあるのか、どのような場面が描かれているのか、といったより深いレベルでの情報を引き出すことが可能です。

さらに、VLMは視覚とテキストの情報を相互に補完し合うことで、単純な情報処理を超えた「理解」に近づいています。例えば、製造業においては、機械の異常を示す画像データと、それに関連する説明やエラーコードのテキストデータを結びつけることで、問題を特定し、適切な対応を提示することができます。また、医療の分野では、診断画像と患者の症状や診断データを組み合わせて解析することで、より正確な診断支援が可能となります。このように、視覚とテキストの情報を組み合わせて解釈することは、より包括的なデータ解析を実現し、さまざまな業界での応用を可能にしています。

Vision Language Modelの構造

VLMは、その入力内容と出力内容によって、大きく三つに分類されています。

一つ目が、vision-languageの関係性理解に特化したモデル、二つ目は、画像と言語など複数のモダリティ(マルチモダリティ)の入力を処理してテキスト(単一モダリティ)な出力を生成するモデル、三つ目は、複数のモダリティを入力し複数のモダリティを出力するモデルに分類することができます。

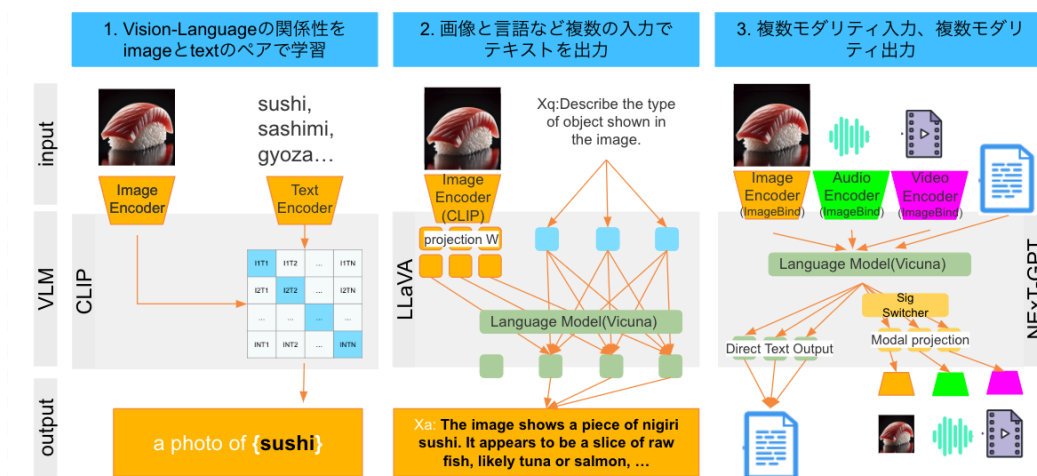


図1 :<https://arxiv.org/pdf/2404.07214>の分類を元に、Vision Language Modelの構造と種類を分類し、VLMの構造は各論文を参考に作成

各Vision Language Modelの構造の詳細

言語とペアで学習された柔軟な物体検出

「Vision-Language の関係性をimageとtextのペアで学習する」ことで、画像を単語と紐付けて理解する潜在空間を構築し両者を類似度として比較することが可能となっています。この時、最も大事な特徴とし

て、事前に特定のクラスで学習されていなくても回答の候補となる単語リストを更新するだけで画像を分類することが可能な能力(ゼロショット学習といいます)を獲得していることです。図1では、寿司の画像と同時に候補となる単語リスト("sushi", "sashimi", "gyoza"...)を候補として入力し、画像と単語の類似度を比較し、最も高かった"Sushi"という回答を得ています(実際の単語リストは、openai/CLIPで公開されているdata/prompts.mdからFood101を利用)。

画像に何が写っているか単語のリストから選択する。人が行うとするとすごく簡単なことのように思えますが、これまでの多くの画像認識モデルでは、固定された単語リスト(クラスといいます)で事前に学習し、固定のクラスにのみ分類することが多く、単語のリスト、つまりクラスやクラス数が更新された時に、再学習が必要でした。VLMの発展は、このゼロショット学習能力を持つCLIPと呼ばれるモデルが開発されたことが大きな役割を果たしています。

画像からの自然なテキストデータへの変換

「画像と言語の複数の入力でテキストを出力」では、言語とペアで学習された柔軟な物体検出能力で画像から言語と関連する特徴を抽出し、質問文と一緒に大規模言語モデルに入力することで、画像に含まれる言語的な特徴と、質問文を合わせて豊かな文脈の理解と自然な回答の生成ができるようになっていきます。

これは、言語とペアで学習されたモデルでは、画像とテキストを共通の潜在空間にマッピングし関連づけられるが、その出力は主に単語や短いフレーズに限定されてしまう、という課題を解決し、VLMがより自然に対話的な説明を行えるようになっています。

例えば、風景の画像を入力すると、その中に含まれる要素(山、川、建物など)を言語的に説明したり、それらに関する質問に答えたりすることができます。これは、ある陸橋の下からの写真ですが、詳細なキャプションとして、下から見た鉄の橋であること、二つの電球があることなどを説明してくれています。また、QA Result では、この画像の天気はという質問に対して、晴と回答することができています。このような形で画像内の情報の構造化などに活用が進んでいます。



Uploaded Image

Generated Caption:

```
{
  "<MORE_DETAILED_CAPTION>" :
  "The underside of a metal bridge. There are two lights on either side of the bridge. The sky above the bridge is blue. There is a building under the bridge that is yellow. "
}
```

QA Result 1:

```
{
  "<QA>" : "sunny"
}
```

複数のモダリティ入出力を行うVLM

「複数モダリティ入力、複数モダリティ出力」することができる、より高度なマルチモーダルAIモデルも登場してきています。

このモデルでは、CLIPを発展させた言語とペアで学習された柔軟な物体検出能力を持つエンコーダ（例えばImageBind）を使用して、画像や音声、ビデオから言語と関連する特徴を抽出します。そして、これらの特徴を質問文や指示文と一緒に大規模言語モデル（Vicunaなど）に入力することで、入力されたマルチモーダルな情報を統合的に理解し、豊かな文脈の中で自然な回答や対応するモダリティの出力を生成することが可能となっています。

これは、従来の言語とペアで学習されたモデルが、画像とテキストを共通の潜在空間にマッピングして関連づけるものの、その出力が主に単語や短いフレーズ、あるいはテキストのみに限定されてしまうという課題を解決しています。これによりVLMがより自然で対話的な説明を行えるだけでなく、画像や音声、ビデオといったより多様なモダリティでの出力も可能にしています。

例えば、風景の画像を入力すると、その中に含まれる要素（山、川、建物など）を言語的に説明したり、それらに関する質問に答えるだけでなく、「この風景に合った音楽を作ってください」というリクエストに対して、適切な音楽を生成することもできます。また、「この風景をもとに短いビデオを作成してください」という依頼に対して、関連するビデオクリップを生成することも可能です。

現在は、複数のモダリティを組み換え可能な形で用意し、任意の入力の組み合わせと出力の組み合わせにできるようにする試みなども行われており、言語を中心とすることでより柔軟な理解と生成ができるようになっていくことが期待されています。

VLMのビジネス上のインパクト

画像認識AIの登場によって、物体認識結果をもとにしたセキュリティ機能や製造業における危険管理、飼育個体数管理などがこれまで達成されてきました。しかしながら、これらの画像認識AIは、事前にデータから検出したい情報を人手で十分に時間をかけて丁寧に準備された状態で学習されたAIによるもので、検出対象の変更などユースケースごとにモデルを作ることが求められていることが多くありました。特に、画像認識AIで検出した後の処理プロセス自体を専用で設計することが求められることが多く、画像認識から物体・領域を検出するAIとしては機能するが、ユースケースを大幅に変えるには再開発が必要でした。VLMの登場は、AIの学習サイクルそのものを変更し、これらのAI開発・運用のサイクルに根底から変えていくような進化をしつつあります。。

VLMの登場によって、学習を行わずにあらゆる場面でロボットやAIと現実世界の共通認識をもとにコミュニケーションを取ること・自然言語によって現実世界を操作することへの関心が急速に高まってきています。

GoogleやMetaなどのビックテックをはじめ多くの企業がVLMを活用した自然なコミュニケーション可能な状態での自動化や、画像やスライドなどへのアクセシビリティの改善、VLMの生成機能と認識機能を活用した物理世界のシミュレーションを行うモデルへの開発などが取り組まれています。

VLMを活用するメリットについて具体例を交えながら紹介します。

事例から見るVLMを活用するメリット

画像認識結果に基づく行動計画と実行の高品質な自動化

VLMを活用することで、ロボティクス分野やコンピューター操作において、画像認識結果に基づく高品質な行動計画と実行の自動化が可能になります。例えば、Anthropic社のClaude 3.5の「Computer Use」の事例では、AIが視覚情報を使ってコンピューターインターフェースを理解・操作し、反復的なプロセスの自動化やソフトウェアの構築・テスト、リサーチなどのオープンエンドなタスクを効率的に行っています。視覚を持つことで、AIは自動フィードバックとシミュレーションを低コストで実現し、環境への適応やタスクの

7

最適化が迅速に進みます。これにより、人間との協調作業や複雑な作業の効率化が達成され、全体的な生産性と柔軟性が向上します。

Computer Useは、現時点では、まだかなり頓珍漢なことを行うこともありますが、ここで思い出さないといけないのは、私たちは、ChatGPT-3.5が出た当時も同じことを感じた人が多かったということです。それは、ちょうど二年前、2022年11月です。まだ、たったの。

画像認識結果のコンテキスト把握能力の獲得

近年のVLM (Vision-Language Model) の発展により、画像認識結果から何が起きているのか、なぜそれが起きているのかを深く理解する能力が向上しています。これを更に強化することを可能にしているのが、グラフ形式での情報管理です。画像内のオブジェクトや人物、動作などをノードとし、それらの関係性や因果関係をエッジとして表現することで、複雑なシーンの理解が可能となりました。

例えば、ある画像において「子供がボールを投げている」というシーンがあった場合、モデルは子供(主体)、ボール(対象)、投げる(動作)という要素とそれらの関係性を理解し、「子供がボールを投げている」という文脈を把握します。さらに、「なぜボールを投げているのか」という質問に対して、「遊んでいるから」や「スポーツをしているから」といった因果関係も推論できます。

グラフ形式で扱うことの大きなメリットは、関係性をより厳密な記述として扱いやすいことです。これにより、それぞれのノードが持っている情報と、エッジの関係性から次に何が起こるのかということを予測することや、グラフの更新をVLMに継続的にさせることにより空間の理解をより容易にするといったメリットが模索されています。詳しくは事例1でご紹介します。

画像認識をもとにした高速なプロダクト開発・資料作成

VLMを利用することで、イラストやスケッチから、3Dやプログラミングコードとして生成することができ、商品説明のウェブページや、プロダクトのユーザーインターフェースなどの開発が従来より圧倒的に早くなっています。また、Microsoft Copilotなどで見ると、スライドそのもののデザインを視覚情報と

一緒にフィードバックすることで、誰でもいつでも高品質なフィードバック・改善を受け取れるようになってきています。

画像認識結果に対しての専門家知識を紐付けた活用

例えば、建設設計図面や特許図面など、読み取りは可能だが、その空間的な配置や、図形との関係性が重要になる書類には、単純な文字読み取りを行うOCR 技術だけでは解決することが難しい状態にあります。調整したVLMを活用することで、認識した文字、図形と合わせて専門家知識を参照しながら回答・解釈をさせることが可能になり、特定業務の効率化や、作業品質の一定化などが期待できます。

プライバシーを考慮した情報の埋め込み分析

VLM (Vision-Language Model) を活用することで、プライバシーを保護しながら情報の埋め込み分析が可能になります。例えば、オフィスビルや街中などにおける人の行動データなども、顔画像などの個人情報を直接取り扱わず、テキスト形式で情報を管理することで、プライバシーを保護しながら必要なデータを取得できます。これは、従来は複雑な後処理が必要とされていたタスクであり、VLMの導入により大幅な効率化が実現しました。

この技術は、実験だけでなく、在宅医療や、公共データの定量化など、さまざまな分野での応用が期待されています。プライバシーに配慮した情報分析が可能になることで、個人の安心・安全を守りつつ、社会全体の利便性を高めることができます。

ユーザーの言語化支援

ユーザーが、検索するためのキーワードが思い浮かばない、そもそも現状がどのような状態なのか説明ができないことで課題解決ができないという状態に対して、ユーザーが提供する画像やイラストをもと

に、課題やキーワードを抽出し、課題解決に必要な情報への早くアプローチできるようにするといったことが期待できます。

VLMのメリットは、画像認識結果の情報・位置関係を言語的に学習されている知識と紐付け出力に繋がれることにあり、多くの業界に革新をもたす可能性があります。特に、VLMの視覚と言語を紐づけた柔軟な出力は、VLMの自律的な支援や行動を可能にしており、これまで人的リソースが制限でサポートされていなかった困りごとが加速度的に解決されるようになっていくと考えられます。

VLMの系譜と発展

ここで改めてVLMがどのように発展してきているか、これからどのように発展していくことで加速度的にユーザーの課題解決につながっていくかを整理したいと思います。

初期のVLMは、**CLIP**と呼ばれる視覚と言語の関係性の理解(Vision Language Understanding)から始まり、見えているもののうち代表的なものは何かを答えるということでした。

次のステップでは、視覚と言語の関係性をより複雑な関係性として理解できるように**CLIP**にさらに中規模な言語モデル(**Vicuna**) が追加されCLIPから伝わる、見えているものの確率はどのような分布かを言語空間に投射することで、より詳細な説明ができるようになってきました。ただこの時点でも視覚と言語の組み込みにCLIPモデルが非常に大きな役割を担っていました。

さらに、画像とテキストだけでなく、パッチワークのように音声や動画など複数のモダリティ(データの形式)に対応してモデルが作成されてきており、マルチモーダル入力だけでなく、マルチモーダル出力への対応がどんどん進んでいます。

2D認識から3D認識に発展するVLM

3D空間認識(LEOなど)として、空間上にどのような物体が配置され、相互作用しているのかを理解し回答することが進んでいます。これは、VLMに対してマルチモーダル入力、テキスト出力を行う、空間のキャプションングを行っていることに近く、3D情報を入力し、テキストを出力することが試みられています。

特に、テキスト出力としてロボットを操作可能なアクション(座標や、コマンドなど)を出力することを可能とすることで、3D世界の中で様々なタスクを実行できる汎用なエージェントとして動かすこともできるようになってきています。



引用 : <https://embodied-generalist.github.io/>

画像認識から動画認識に発展するVLM

破綻の少ない動画生成(Gen-3, Dream Machine, Sora, Movie Genなど)、音楽生成(Sunoなど)、そして、連続した画像空間認識(SAM2)、これまで特定の瞬間に対しての理解だったVLMが、時間方向に認識力を発展させています。

高速に認知するVLM

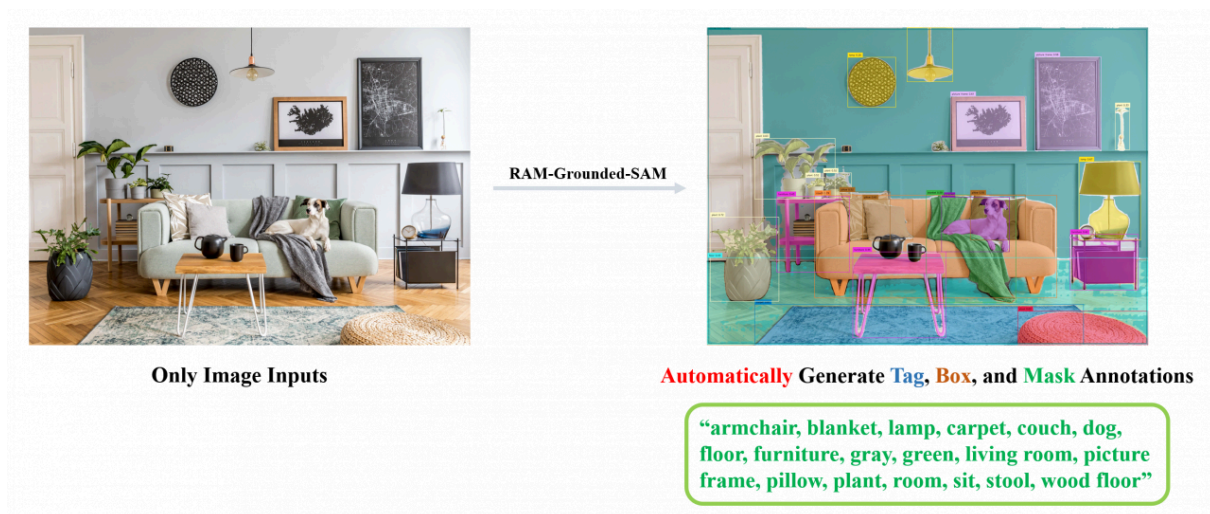
これは、これまでの二つとは属性が異なり、複数のモダリティがより密な結合を行い、認知の高速化、出力の高速化がされているモデルです。特に、今年の5月にOpenAIが発表した感情音声モデル(オムニモデル)は、これまでは複数のモデルによって、段階的に処理されることで達成されていたマルチモーダル入力、マルチモーダル出力と異なり、単一のモデルでエンコーダ処理から、埋め込み、デコーディングまで全てが達成されていることで、大幅な速度の改善がされ、音声の入力から出力までが320ミリ秒に改善したと公表されています。

引用: <https://openai.com/index/hello-gpt-4o/>

VLM とVision Foundation Modelの違い

VLM と似た言葉として、Vision Foundation Modelと呼ばれるものがあります。この二つの違いについては、明確な定義として分割できるものがないのですが、Vision Foundation Modelは、主に、テキストによって画像の空間的な理解を出力する(画像上の物体の位置や、領域など)を目的としているものだと思います。

例えば、画像領域(BBOX)とテキストを紐付けるGrounding-DINOを元に、特定の領域内で単語に関連する領域を塗りつぶすSegment Anythingを連携させることで、図のように分割された画像が実際にどのような単語と関連しているのかを可視化することなどができます。



引用: <https://github.com/IDEA-Research/Grounded-Segment-Anything>

VLMの選択

VLMモデルは、テキストと視覚情報の紐付けによってどのような課題を解決したいかによって、大きく選択肢が変わります。これは、課題解決できることのメリットに対して、VLMのコストがどの程度許容されるかという大きな問題が常に付きまとうためです。

これまでみてきたとおり、VLMで実行できるタスクは、

- 画像に何が含まれているか、それぞれがどのような関係であるのかをテキスト、もしくは構造化されたテキストで説明する。
- マルチモーダルな入力を一度言語的な空間に落とし込み、マルチモーダルな出力として生成する。

という大きく二つのタスクに分類できます。

例えば、"画像からテキストで説明する"は、

- 画像、動画を元にユーザーが調べたいことを画像入力からテキスト化を支援することで画像の類似度だけでなく、周辺情報を含めた状況としての検索を行う
- スライドなどを自動キャプションさせることで、類似のコンテキストを持つスライドを検索できるようにする
- スライド、バナーなど作成したもののわかりやすさ、効果などを予測させる
- 手書きしたUI画像入力を元に、ソースコードとして生成し、インタラクティブなプロトタイピングをすぐに作成する
- 医療用のCT画像など解釈に高い専門性が必要なモノについて、専門に調整したモデルで説明を付与する

などが行われています。それぞれによって、求められる速度・精度・汎用性、そしてコストが異なることが想像いただけると思います。

VLM やLLMの特性として、ある程度小さいモデルにおいても、繰り返しや再起的に情報を検討・生成させることで、確率的に良い精度のものが得られることもあると知られています*1。

一方で、医療用など専用のデータセットでファインチューニングをすることで達成可能な事柄など、従来のAIモデルから変わらない部分もあります*2。

最近、モデル内部での熟考することによって高い精度が出せると話題になっているOpenAI のo1-previewと呼ばれるモデルにおいても、知っているかどうかが決定的な因子になる事柄については、熟考することでは効力がないことが示されるなどされており*3、解くべきタスクをVLM / LLMが得意としているかどうかなど、ある意味、専門人材の採用と同じ難しさがモデル選択時に出てきている部分があります。

*1: Xuezhi Wang, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models.

[arXiv:2203.11171](https://arxiv.org/abs/2203.11171)

*2: Jinhyuk Lee, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. [arXiv:1901.08746](https://arxiv.org/abs/1901.08746)

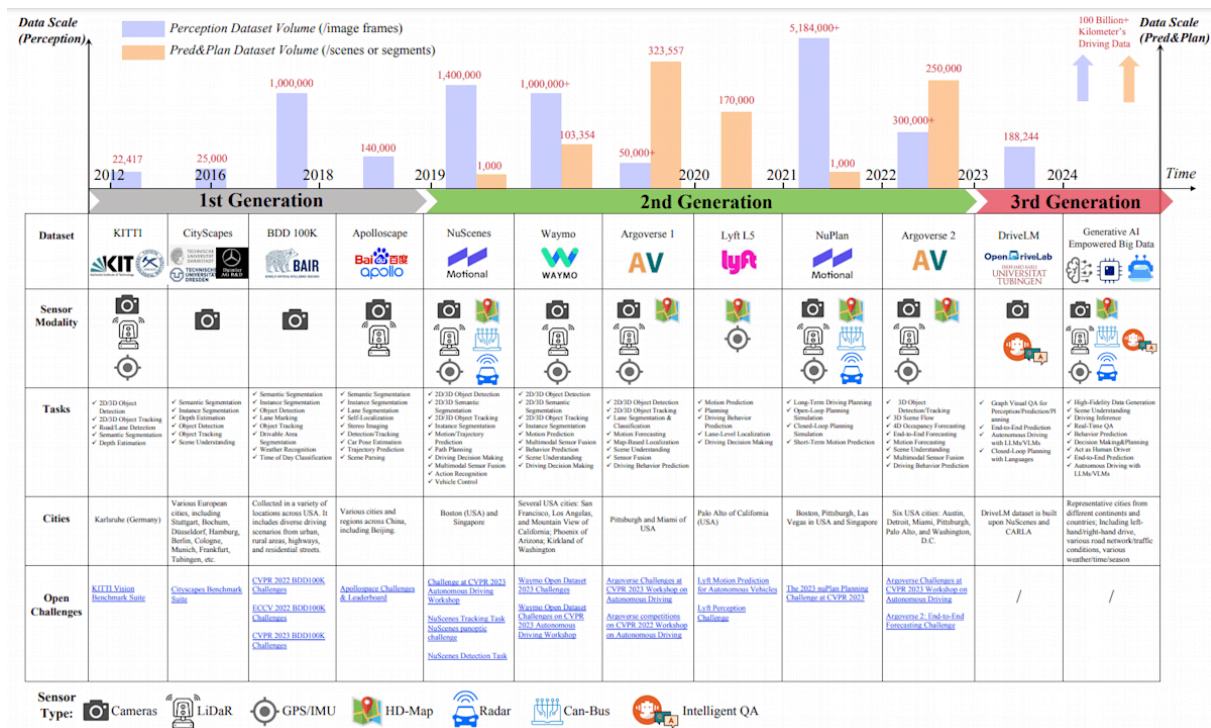
*3: OpenAI, GPT-4 Technical Report. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774)

VLMの評価

データセット

VLMを評価するためのベンチマークは、いくつかの目的別に構成されていることがあります。

例えば、自動運転に関連するようなタスクでは、以下のように、古くから画像認識用のデータセットが作られてきているとともに、3rd Generationとして、VLMが状況の理解や、次の運転行動を予測するためのデータセットなどが公開され始めています。



引用 : <https://arxiv.org/pdf/2401.12888>

特に、これまで、多くのデータセットが日本語に対応していない中、Turingより公開された日本語対応のVLM評価ベンチ(Heron-Bench)は、これまで日本語対応したVLMデータセットがない中で大きな有効性があったと捉えられると考えています。

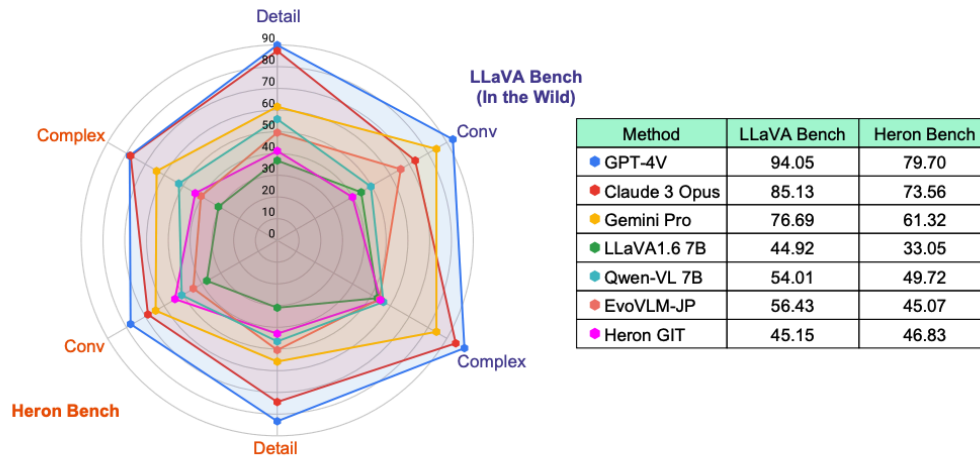


Figure 1: Comparison of evaluation results using the Japanese-translated LLaVA Bench (In the Wild) and the Japanese Heron-Bench.

引用 : Heron-Bench : A Benchmark for Evaluating Vision Language Models in Japanese[Yuichi Inoue, Yu Yamaguchi.,et al., 2024]

ケーススタディ

この章では、VLMの効果的な活用例を具体的に紹介します。

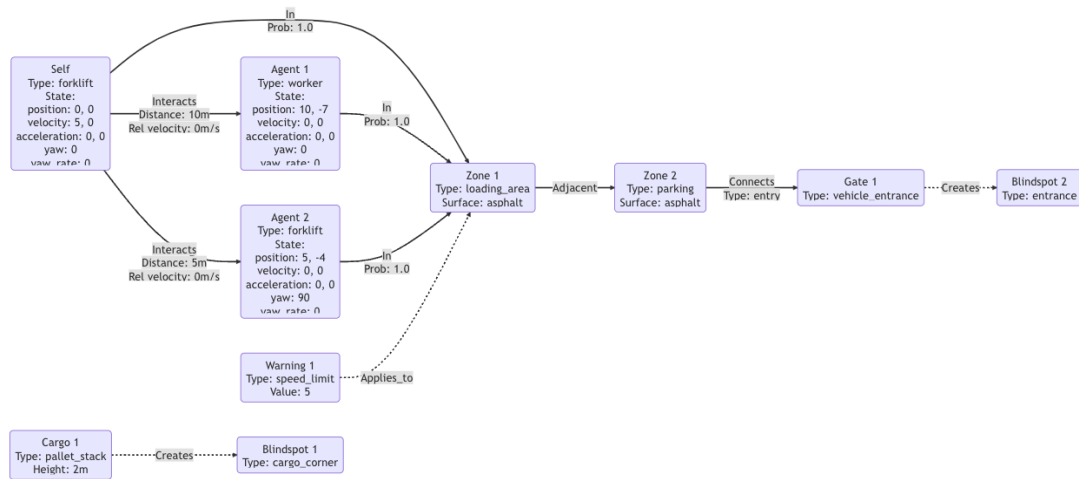
事例1：シーングラフによる状況認識と安全管理システムの実用化

物流現場における人・車両・設備の相互関係を、VLMによって「グラフ構造」として構造化させることで、AIによるリアルタイムな状況理解と予測を可能にする新しいアプローチを検討しています。従来の単純なセ

センサー監視から一歩進み、現場の複雑な関係性を構造化データとして管理・分析することで、事故防止と作業効率の向上を同時に実現への応用などを検討しています。

例えば、以下のようなフォークリフトの一人称視点の画像から、周辺にどのような人がいるのか、どのようなものがあるのかなどをグラフ化することで、グラフ内のそれぞれのオブジェクトの変数がどう変わった場合に何が起こるかなどの予測アルゴリズムの構築などが模索されています。ただし、実際には、現状のVLMで厳密に常に全ての重要なオブジェクトを検出することを担保することは難しく、物体検出AIなどと組み合わせた検証などを進めています。





画像: <https://www.kaggle.com/datasets/shrimantasatpati/inventory-warehouse-object-detection-dataset>

事例2: データセットマネジメントのデータセットキャプションニング

安全管理を行うための小さい画像認識モデルを繰り返し、対象の機器毎に学習しているという現状があります。これには、機器事の測定環境や、測定対象がしっかりと含まれていることをデータセットのバイアスとして管理するとともに、不足しているデータを同程度の品質で生成していくことを目指す必要があります。

入力画像に対して、ローカルで動作可能なVLMモデルを利用し、画像の詳細キャプションと、特定のタグ用のQAを生成を行うことで、データセットの管理を行っています(再掲)。



Uploaded Image

Generated Caption:

```
{
  "<MORE_DETAILED_CAPTION>" :
  "The underside of a metal bridge. There are two lights on either
  side of the bridge. The sky above the bridge is blue. There is a
  building under the bridge that is yellow. "
}
```

QA Result 1:

```
{
  "<QA>" : "sunny"
}
```

また、次の図にあるように不足している領域に対してデータ拡張するために、生成されたキャプションに不足しているシーンの指定を追加し、画像の生成を組みあわすことなどを組み合わせています。



<https://huggingface.co/spaces/black-forest-labs/FLUX.1-schnell> を活用して生成。左は、クレーン先端からの画像。右は、工場内でのヘルメット未着用者を含む画像。

現在アラヤでは、VLMによるデータセットマネジメント・データ生成・自動アノテーション機能と、蒸留と枝刈を組み込んだエッジAIの開発運用を高速に実行可能なエッジAI開発・運用基盤を開発しており、VLMによって半自動的にエッジAI開発を行うことが可能な状態を目指しています。

事例3: VLMを用いたロボットの遠隔操作

アラヤでは、VLMを用いたロボット操作などについて、いくつかの論文を公開しています。本資料では、画像とテキストのゼロショット学習を活用した例を紹介します。

これまでの研究では、3D表現空間、特にボクセルを使用することで3Dマニピュレーションタスクに利点があることが示されてきました。しかし、ボクセルは計算量が N^3 に比例するため、観測サイズに制限がありました。また、ボクセルの構造は空間情報を伝達しますが、解像度が限られているため、意味的に

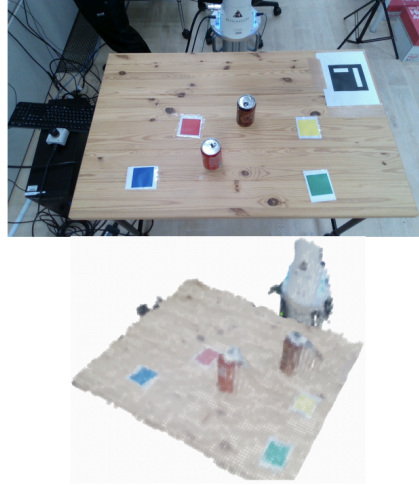


Fig. 1: 720×1280 image view from the RGBD camera (top) vs. 100^3 voxel view (bottom). Fine-grained semantic information somewhat diminished in the voxel view.

重要な情報が不明瞭になる可能性があります。

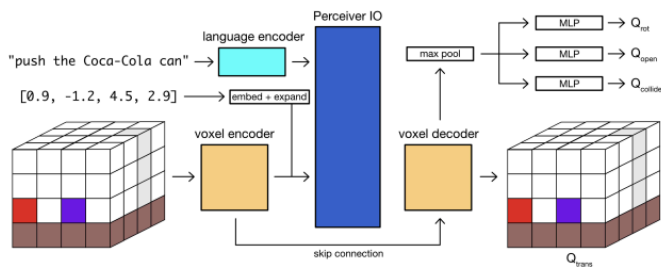


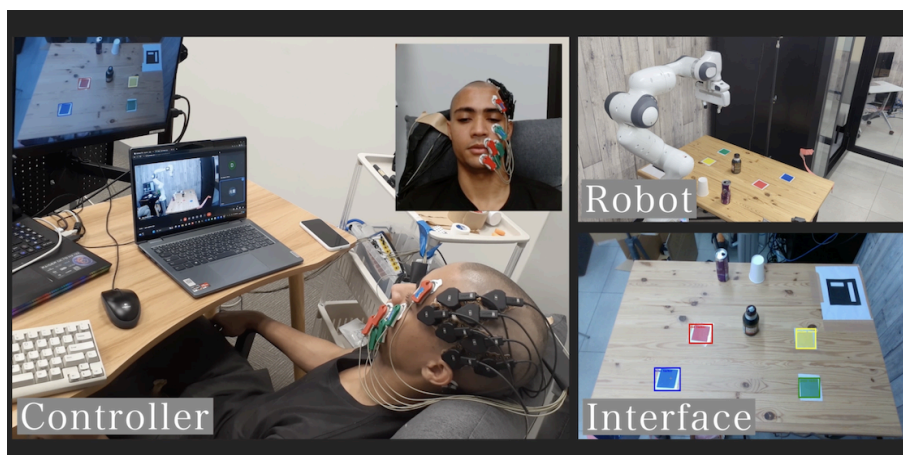
Fig. 2: PerAct architecture. The voxel and text inputs are separately encoded, with proprioceptive inputs embedded and then concatenated with the output of the voxel encoder. These are then processed by a Perceiver IO Transformer, and decoded back to voxels. The voxels are used to directly predict the end-effector translation, and further processed through pooling and fully-connected layers to predict the rotation, gripper state, and motion planner mode.

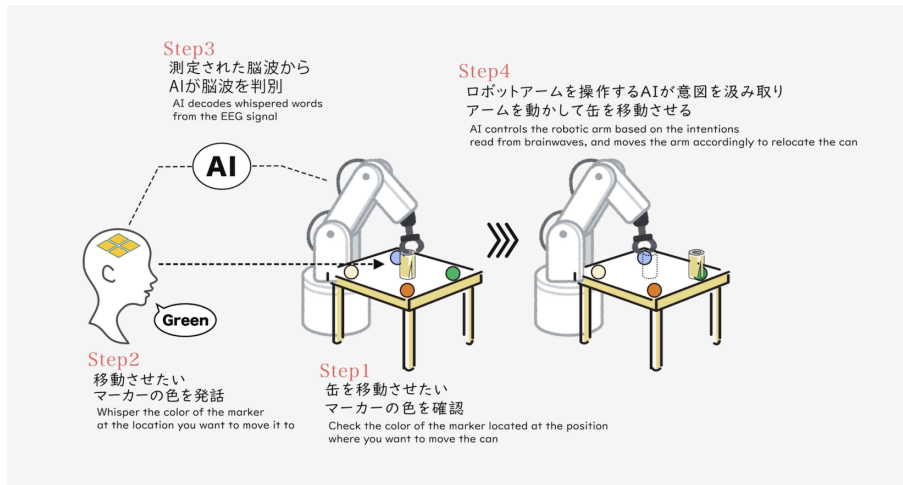
引用 : https://openreview.net/pdf?id=zZVEvf_mT6

この課題を解決するために、本研究では3Dベースのエージェント「Perceiver-Actor」に追加のセグメンテーション情報を条件付けする方法を提案しています。これにより、操作タスクにおいて類似したオブジェクトを正しく区別することが可能になります。

具体的には、事前学習されたセグメンテーションモデルとテキスト・イメージモデルを使用して、関連するオブジェクトのセグメンテーションマスクをゼロショットで抽出します。この方法を実際のロボットで検証した結果、「コーラ」の缶と「ドクターペッパー」の缶といった微細な違いを持つオブジェクトに対しても、正しく操作できることが確認されました。

これらの機能は、アラヤの別プロジェクトで行っている脳波によるロボット操作などに応用を期待しています。単語ベースでロボットを操作するといった研究に応用が可能であり、VLM / AIを駆使することで、Brain Machine Interface の開発を模索しています。





引用：[超高密度脳波計とAIによるロボットアームの遠隔操作実験に成功](#)

VLM導入における課題

VLMの導入は現代のコンピューティングに多くの利点をもたらす一方で、VLMの実装や利用にはいくつかの課題が存在します。本章ではこれらの代表的な課題を示し、VLMの導入成功への影響を探ります。

モデルサイズの大きさ

VLM(Vision-Language Model)は、その高度な機能性ゆえにモデルサイズが非常に大きくなっています。これにより、モデルを動作させるためには高性能なハードウェアスペックが必要となります。具体的には、大容量のGPUメモリや高速な計算能力を持つプロセッサが求められ、小規模な組織や個人では導入が難しい場合があります。このハードウェア要件の高さは、VLMの普及を妨げる要因の一つとなっています。

しかしながら、近年では、sVLM(<7B以下)など高い性能を維持しながら、効率的な圧縮が実行されたモデルなどが出現していること、VLMのコアであるネットワーク構造(Transformer)そのものの処理を高速化できるようなエッジAIアクセラレータの登場がしてきているなど、少しずつ課題として解決されてきています。

VLMのバイアス

VLMは大量のデータを基に学習されますが、そのデータセットには偏りや誤情報が含まれている可能性があります。例えば、音声認識モデルのWhisperにおけるハルシネーション(幻覚)の事例では、モデルが存在しない情報を生成してしまうリスクが報告されており、医学領域での利用などに注意が促されています。同様に、VLMでも全てを正しく認識・理解できるわけではなく、誤った情報や偏見に基づいた出力を行う可能性があります。これらの問題は、モデルの信頼性や公平性に影響を与えるため、慎重なデータ選定とバイアスの検証が必要です。

VLM認識に対しての敵対的アタック

VLMは画像認識技術の一種であるため、敵対的アタックの影響を受けやすいという課題があります。具体的には、特殊なパターンが描かれたTシャツやアクセサリを用いることで、モデルが人を正しく認識できなくなるといった問題が起こり得ます。このような認識防止用の工夫により、セキュリティや安全性に重大な影響を及ぼす可能性があります。例えば、自動運転システムに対して、敵対的アタックになる・なってしまう衣服を着た人が認識されずに事故につながるということが考えられ、認識技術の冗長的な設計などが必要になります。

VLMのリアルタイム性

VLMを自動運転などのリアルタイム性が求められる分野で活用する場合、モデルの処理速度が大きな課題となります。運転状況では、数十ミリ秒以内の応答速度が必要とされていますが、VLMはその複雑さと計算量からリアルタイムでの処理が困難な場合があります。遅延が発生すると、安全性に直結する問題となるため、モデルの軽量化やハードウェアの最適化など、リアルタイム性を確保するための技術的工夫が求められます。

結論 ～VLM導入実現に向けて～

AIが、視覚や、それ以外の複数の感覚を言語的に説明できるようになってくることで、より深い文脈や、事象の変化を理解し、出力を生成できるようになってきています。これは、これまで実社会の中で、人向けに画像や音声で作成されていたフィードバックやシグナルをVLMやマルチモーダルLMが活用できるようになってきていることを示しており、実社会の中に、あなたの隣に、AIが或る社会が駆け足で近づいてきていることを示しています。しかしながら、モデルの大きさやリアルタイム性などの問題から、VLMの浸透はAPI経由でアクセスできる環境から始まり、しだいにヒューマノイドを含むロボットなどに発展していくと考えられます。

VLM導入にあたり、どのような柔軟な解釈が必要なのか、要求が決まった後に却って柔軟性がデメリットにならないか、自律して繰り返しを行わせたいのか、VLMが判断を間違えた場合にどの程度リスクがあるのか、そのような点を多角的にレビューした上で、評価が必要になってきます。

最も大事なポイントは、VLMが本当に必要なのか、これまでの画像認識AIではいけないのか、ルールベースの処理の方が良いのではないかなど、目的に沿った戦略を検討できるかという点です。

アラヤでは、先進的な技術活用を検討するためのPoCから、システムインフラの構築、エッジAI化を通じたプロダクト化支援まで、幅広く皆様の到達したい未来と一緒に伴走することが可能です！ぜひお気軽にご相談ください！！

Mission

ミッション

人類の未来を
圧倒的に面白く！

Vision

ビジョン

すべてのモノに
AIを宿らせる

アラヤについて

株式会社アラヤは、認知神経科学の研究者、金井良太、が率いる、人工知能AIとニューロテックをコア技術とし、「人類の未来を圧倒的に面白く」をミッションに掲げたディープテックベンチャー企業です。製造業、ヘルスケア、建設、アカデミック・リサーチ等の領域において、AIアルゴリズム開発、エッジAI実装、生成AIを活用した先進的なソリューションやDX支援、また、ニューロテック・ブレインテック領域等における高度な研究開発支援をご提供することで、皆様から信頼されるソリューションパートナーであり続ける事を目指します。

先端AI開発支援 : <https://www.araya.org/service/aisupport/>

エッジAIソリューション : <https://www.araya.org/service/edgeaiconsulting/>

お問合せ

株式会社アラヤ

本社所在地

〒101-0025 東京都千代田区神田佐久間町1-11 産報佐久間ビル6F

お問合せフォーム : <https://www.araya.org/contact/>