AI連携とオーケストレーションのための基盤ビジョンLLM

KHAN Zaid　KUMAR B G Vijay　SCHULTER Samuel　CHANDRAKER Manmohan

要　旨

本稿では、事前定義あるいはカスタム定義されたタスクに対するコンピュータビジョンソリューションの開発とデプロイを自動化するビジョンLLMのフレームワークを提案します。基盤レイヤは、既存の特定タスク向けAIモデルのAPI、技術文書や利用方法とともに、新たなユーザー定義のタスクの理解に基づいたPythonコードを生成するための、強化学習によって自己学習されたコードLLM AIオーケストレータを提供します。特定ドメインでのゼロショット能力は、既存のコンピュータビジョンモデルとデータセットを利用して低い計算コストで学習された視覚言語モデルから得られます。エンジンレイヤは、複数の特定タスク向け視覚言語エンジンで構成され、組み合わせて使用できます。アプリケーション固有レイヤは、標準のファインチューニングツール以外に、LLMによる新たなデータ増強と質問分解を使用してお客様固有のシナリオの性能を改善します。また、AI視覚アシスタントや視覚的対話、法執行、モビリティ、医療画像推論、リモートセンシングを含む広範囲なアプリケーションに適用可能です。

KeyWords　コンピュータビジョン／基盤モデル／エージェント型LLM／オーケストレーション／自己学習／視覚支援

1. はじめに

NECにとって、コンピュータビジョンはヘルスケアや金融、小売、モビリティ、リモートセンシング、セーフティなど広範囲にわたる適用分野において重要な技術です。NECの大目標は、次の２つの戦略を可能にすることです。

・戦略的に重要な基盤モデルを通じてAIビジネスの周辺に防御濠を構築する

・ターゲットとなるアプリケーションドメインへの影響力を加速し広げる、挑戦的なツールキットを構築する

本稿では、これらの目標を実現するための基盤ビジョンLLMアーキテクチャの概要を紹介します。

まず、次のようなワークフローを通じたカスタマイズが必要となる典型的なコンピュータビジョンソリューションを考えてみます。

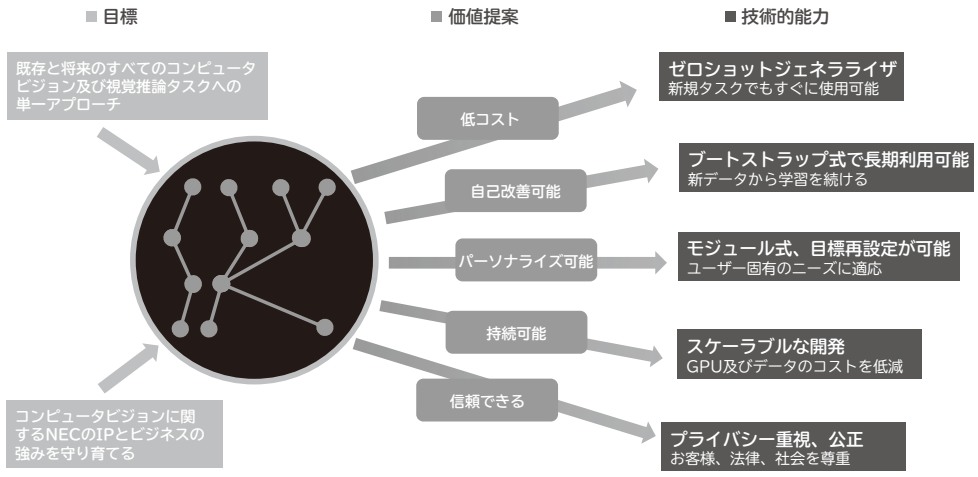

図１　基盤ビジョンLLM創作のためのミッションステートメント

(1) お客様が自然言語または例題を用いて自らのニーズを説明する
(2) 技術者が、利用できるモデルやライブラリ、資料に基づいてタスクを解決するためのコードを書く
(3) デプロイチームがそのソリューションをお客様環境に合わせてチューニングする

これらに対し、新しいタスクを理解して適切なコードを生成し、その後既存のエンジンとAPIを呼び出して与えられたタスクを解決するエージェントとして動作するビジョンLLMを提案します。この設計により、前にあるいは個別に定義されたあらゆる視覚タスクを解決するために、コンピュータビジョンに関するNECのすべてのノウハウを統一的に利用できます。

視覚データはロングレンジの推論には適しておらず、自己教師あり学習は困難で言語とのアラインメントは簡単ではないため、ビジョンLLMは従来のLLMとは異なった検討が必要になります。NECのゴールは、低コストで自己改善可能、パーソナライズ可能、持続可能で信頼できるビジョンLLMを開発することです（**図1**）。これらのビジョンLLMで鍵となる考え方は、エージェントとエンジンレイヤなどの設計上の選択、更にはコードLLMや事前学習されたコンピュータビジョンモデルの利用により、最小限の学習コストで高水準な物理的グラウンディングを達成することです。NECのビジョンLLMは、GPT-4V[1)]のようなマルチモーダルLLMとは一線を画し、NECのレイヤアプローチはよりモジュール化され効率的です。

2. アーキテクチャ概要

アーキテクチャの概要を**図2**に示します。このフレームワークは3層のレイヤで構成されます。1層目の基盤レイヤは、利用可能なコードと技術文書を使用して新規のタスクの解決方法を立案するLLMオーケストレータを備えています。このレイヤはまた、ゼロショット視覚言語（VL）汎化能力を備えた非常に大規模なデータセットで学習された特定ドメイン向け基盤モデルも含んでいます。2層目のエンジンレイヤは画像検索や物体検出などの特定タスク向けのVLモデルを備えています。3層目のレイヤであるアプリケーションレイヤでは、お客様固有のデータや使用法に適応するためのデータ増強、質問分解、プロンプトやファインチューニングなどのツールを利用できます。このようなアプローチは競争上の差別化や市場への浸透に役立つものと考えています。

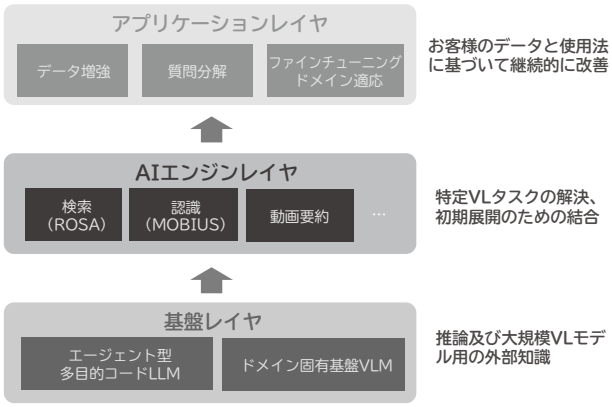

図2　ビジョンLLMのアーキテクチャの概要

3. 技術の詳細と結果

前述したアーキテクチャの実現方法及び主要なベンチマークテスト結果について要点を説明します。

3.1 基盤レイヤ

このフレームワークは、大規模な事前学習済みの特定ドメイン向け基盤モデルだけではなく、利用可能なツールを使用して新規タスクを完遂するための計画を生成するLLMオーケストレータに基礎を置いています。

3.1.1 エージェント型ビジョンLLMオーケストレータ

複雑でまったく新しいタスクを解決するには推論と知覚のステップを何度も繰り返す必要があり、これはプランニング、バックトラッキング、逐次決定で構成されていなければなりません。このようなタスクは知的システムにとって次のフロンティアへの挑戦であり、筆者らは基盤モデル駆動の「AIエージェント」によってこれに取り組もうとしています。ここでは、人間のエンジニアのように利用可能な一連のツールや事前学習済みモデルにアクセスできるLLMオーケストレータを提案します。これにより、エージェントは基盤となるLLMの能力を超えてタスクを解決できるようになります。例えば、LLMは論理的に矛盾が生じやすく演算能力が低い可能性がありますが、LLMで駆動されたエージェントは論理推論であれば論理エンジンに、算術

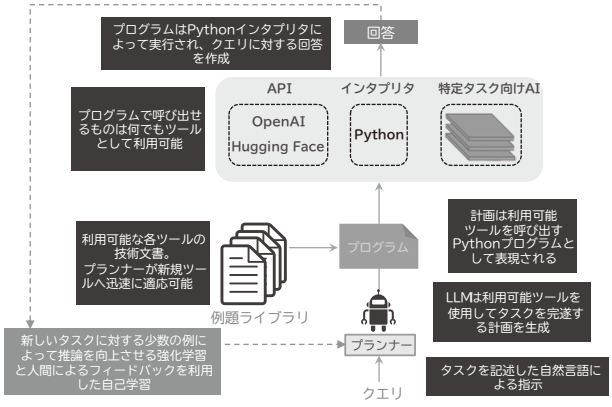

図3　エージェント型アーキテクチャ

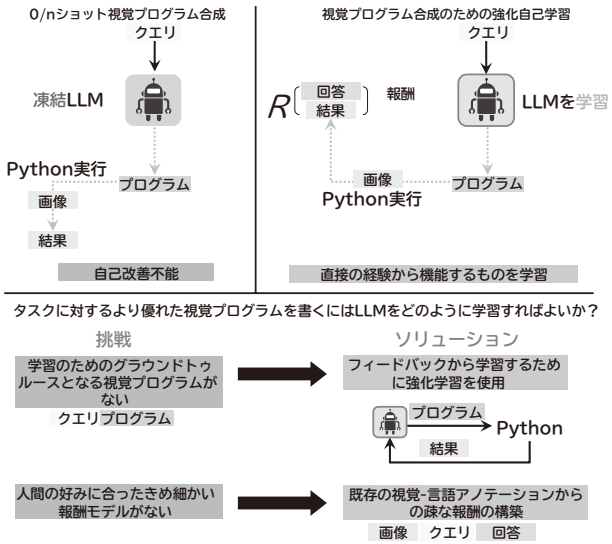

図4　強化自己学習を適用したより能力の高いLLMプランナー

計算を計算機にデリゲートすることができます。また、エージェントにプログラム合成とAPI呼び出しを許可すれば、まったく新しいタスクを解決するために任意の利用可能ツールを組み合わせることが可能になります。

エージェント型アーキテクチャ（**図3**）の中核には、プランナーとして動作するLLMが存在します。自然言語で指示を与えられると、このプランナーは、利用可能なツールを使用してタスクを完遂できる計画を書くことを目指します。この計画は利用可能ツールを呼び出す正式なプログラムとして表現されます。プランナーは、どのツールが利用可能でどう利用できるかを理解するため、技術文書と例題で構成されるライブラリを参照します。これによってプランナーは、新しいツールが追加された場合でも、技術文書を読むことで迅速に適応できます。原理上、プランナーはコードインタプリタを使用することでプログラム的に呼び出せるものはすべてツールとして利用することが可能です。スタート地点として、タスク固有のAI及びサードパーティAPIにアクセス可能な環境を用意しています。続いて、計画（プログラムに相当）をこの環境で実行してクエリに対する回答を作成します。

既存の凍結LLMのプランナーの問題点は、作成した計画に関する経験を欠くために、技術文書だけを基にするとツール利用において微妙な意味合いの理解に失敗する可能性があることです。しかし、LLMをプランナーとして動作するように学習するには学習データが必要なのに、視覚タスクを解決するために作成したプログラムに対し大規模な学習データがありません。

NECの重要な知見は、強化学習を利用してフィードバックから学習を行うことです。そのために、まずプランナーがプログラムを書き実行できる環境を設計しました。プランナーには、最新の特定タスク向けモデルを呼び出せるAPIを提供しています。次に、既存の視覚限度タスク用アノテーションを利用して繰り返し強化自己学習を適用します（**図4**）。例えば、画像データセットv、グラウンドトゥルースy、クエリqがある場合、まずqをプランナーへ入力した後に、画像vに対し生成されたプログラムpを実行します。そして、プログラム実行結果$\hat{y}$をグラウンドトゥルースyと比較して疎な報酬信号を取得し、続いて報酬で重み付けされた行動クローニング損失を適用します。学習済みプランナーは、ChatGPTに基づく凍結プランナーと比較して、質問回答、物体検出、画像テキストマッチングの合成バリエーションにおいて10%、4%、10%も性能が優れています。

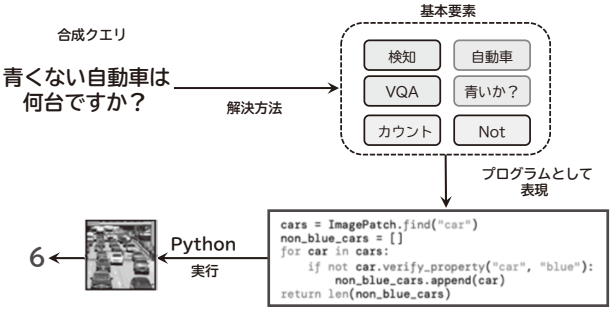

図5　複雑で新しい視覚タスクの解決のために生成されたプログラム例

コンピュータビジョンの一例として、End-to-Endシステムにとって困難な新しい視覚タスクを考えてみます（**図5**）。これは複数の基本的な視覚タスク（物体検出、画像テキストマッチング）とロジックに分解して解決でき、そのための特定タスク向けエンジンが存在しています。プランナーは提供されたAPIによって特定タスク向けAIを制御するPythonプログラムを書いて必要な画像に関する中間情報を取得し、続いて獲得した情報をコードで表現されたロジックと組み合わせ回答にたどり着きます。

これらのAIオーケストレータは、多数の利点を備えています。

- 既存または新しいタスクの仕様を、自動生成されたPythonコードとともに利用可能な視覚モジュールの組み合わせとして解決
- 凍結プランナーと比べ、利用可能なツールの使用法についての理解を改善するように強化学習で訓練されたプランナー
- プログラム例が非常に少ない場合でもタスク推論を改善する自己学習を目的とした人間によるフィードバック
- 一般的なLLMより少ないパラメータで動作するコードLLMを効率的に利用
- パラメータチューニングとデータ増強の操作を自動化

3.1.2 ドメイン基盤VLモデル

これらのアーキテクチャは既存のいかなる視覚モデルでも組み込むことができるものの、時には視覚モデルがライブラリに存在しないことがあります。このため、新規タスクの一般化を容易にする、大量データで学習したドメイン固有の基盤VLモデル（FVLM）も提案しています。重要な知見としては、特定タスク向けのモデルやデータセットは既に数多く存在するため、それらを利用すればFVLMを低い計算コストで学習できるということです。

既に開発したFVLMのドメイン例として、モビリティとヒューマンアナリシスがあります。モビリティFVLMは、いくつかの大規模自動運転データセットに加え、それらに使用される物体検出やセグメンテーション、キャプショニングなどのモデルからの出力も利用して開発しています。ヒューマンFVLMは、人間属性解析や動作認識、人間と物体間の関係性に関するデータセットとモデルの集合体を利用して学習しています。このドメインFVLMを学習するパイプラインの全体像を**図6**に示します。この新モデルは、サイズとしてかなり小さいにもかかわらず（7Bパラメータ）、既存の175Bのクローズドモデルと比較して、性能が0.5%向上しています。

3.2 エンジンレイヤ

このエンジンレイヤは、さまざまなAIモデルやファインチューニング、ドメイン適応などの標準ツールに加え、画像検索や物体検出、医療イメージングやリモートセンシングなどの特定タスク用に開発された技術文書や使用例など、実に多様なもので構成されています。

3.2.1 視覚―言語検索

ここで提案するROSA[2)]は画像モダリティとテキストモダリティを効果的に調整するデータ効率に優れたニューラ

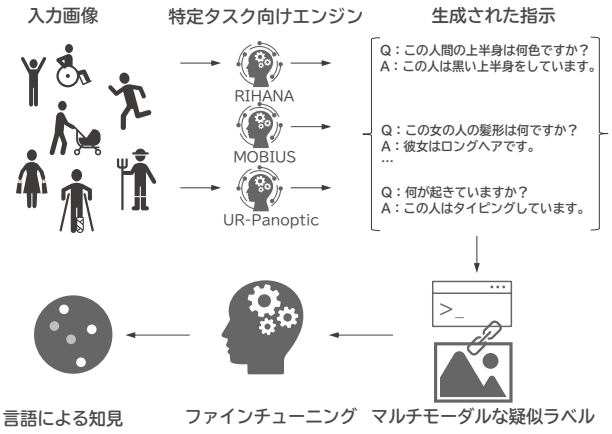

図6　既存の視覚データとモデルを利用して低コストに開発した特定ドメイン向けFVLM

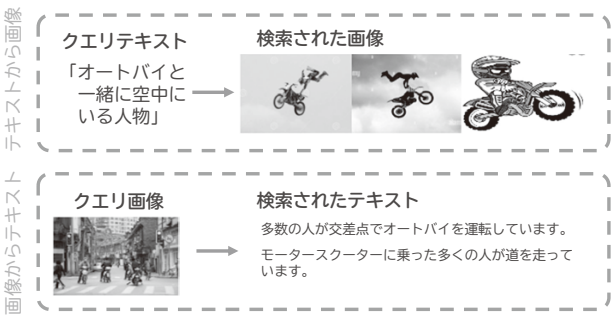

図7　ROSAの緊密な画像－テキスト整合で正確なマルチモーダル検索が可能

図8　オープンボキャブラリな検出器・MOBIUSでの稀な物体カテゴリの正確な位置の言語記述

ルネットワークであり、画像からテキスト及びテキストから画像を正確に検索できるようにします。ゼロショット評価ベンチマーク（**図7**）において、このモデルはRank-1テキスト検索性能で最高水準品を3%上回った一方、競合モデルは30倍多い計算量と100倍大きな学習用データセットが必要でした。

3.2.2 オープンワールドのシーン理解

もう1つのエンジンはオープンボキャブラリで物体検出を行うMOBIUS[3) 4)]（**図8**）であり、自由形式テキストで記述された、あまり見かけないカテゴリや物体の位置特定が可能です。公開されているあるオープンボキャブラリのベンチマークにおいて、学習中のボックスアノテーションなしに初見のカテゴリを検出するというタスクが検出器に与えられたとき、MOBIUSは平均適合率（AP）ポイントで競合性能を4.3%上回りました。

3.3 アプリケーションレイヤ

これらの基盤レイヤとエンジンレイヤは既にお客様領域でのソリューション展開が可能ですが、アプリケーションレイヤではターゲットのデータと使用法に基づいてカスタマイズしたソリューションを実現するためのツールの提供を予定しています。具体的には、少量のターゲットデータを利用するアプローチと、アプリケーションに特化した使用法を、推論がより容易になるように極めて小さなセグメントに分解するというアプローチを提案します。

3.3.1 データ増強

専用タスクやドメインでは利用可能なデータが充分にないことがしばしばあります。より多くのアノテーションを収集することが困難である一方、ラベル付けされていない画像なら利用可能なことが多いです。そこでここでは生成画像―言語モデルを使用して合成データを生産するという戦略のSeITDA[5)]（Self-Taught Data Augmentation、自己学習データ増強）を提案します。SeITDAで学習を行うことにより、安定性、汎化、推論についてそれぞれ9.87%、6.38%、29.81%の改善がみられました。

3.3.2 質問分解

特定ドメイン向けタスクや特殊な推論パターンのあるタスクは、汎用モデルにとって、特に現実的なデータが不十分な環境では困難な課題です。このようなタスクに対して、余分なデータなしに適切なコンテキストを浮き彫りにする対話を通じて汎用モデルを改善する選択的質問分解[6)]を提案します。この分解によって医療データセットに関して最大26%、及び予測誤差を最大20%改善しました。

4. 適用例

ここでは、基盤ビジョンLLMの適用例についていくつか説明します。

AI視覚アシスタント：ビジョンLLMは画像あるいは動画に基づいた出力を発生しますが、コードによるオーケストレーションでは透明性のある段階的な推論が行えます（**図9**）。これによって、視覚障碍者の方々の探し物や安全な移動をはじめとするさまざまなタスクを言語インタフェースで援助できる可能性があります。このアプリ

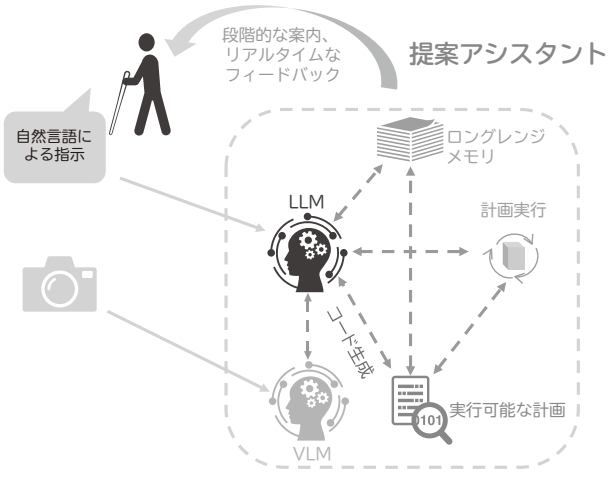

図9　提案したNEC-X向け視覚AIアシスタント

ケーションはNEC-Xにて妥当性の確認を進めています。

視覚的対話：ビジョンLLMは画像とテキストを外部知識と結び付けた推論が可能で、これによって質問回答やマルチモーダルデータによる対話を実現します。このアプリケーションレイヤツールは、外部知識ベースによる画像での言語推論向けに公開されているOK-VQAベンチマークテストで13%の改善を確認できました。

ユースケースとして、NEC-Xにおいて、いくつかの関係性を自動化してSNSインフルエンサーとその視聴者との関わり合い方を改善しようとする例があります（**図10**）。

法執行：法執行のためのタトゥー識別にはVL検索エンジンが使用されています。これにより人間の運用者がタトゥーを解釈し、視覚的類似性を超えて意味に基づいて同一性を判断できます（**図11**）。

モビリティ：モビリティシナリオ用の特定ドメイン向けFVLMによって保険に関する知見と要約を自動化したデモンストレーションを行いました（**図12**）。

リモートセンシング：このアプリケーションレイヤのデータ増強手法により、データが少量であっても衛星画像についての質問にFVLMが回答できるようになります。これによりRS-VS-VQAベンチマークテストにおいてBLIPを上回る改善がみられました（**図13**）。

医療画像：このアプリケーションレイヤの質問分解戦略により、医療画像のQAという、データが不足するドメイン固有のタスクにおいて汎用VLLMの性能が改善します。公開されているPathVQA、SLAKE、VQA-Radの各ベンチマークにおいてそれぞれ22%、10%、26%の改善効果を得ています（**図14**）。

5. 結論と次のステップ

本稿では、お客様のタスクを理解し、外部知識と利用可

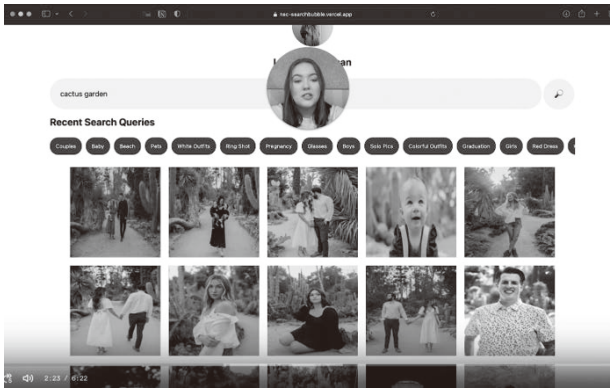

図10　NEC-XのVLサーチアプリケーションでの使用

図11　言語ベースのタトゥー分析システム

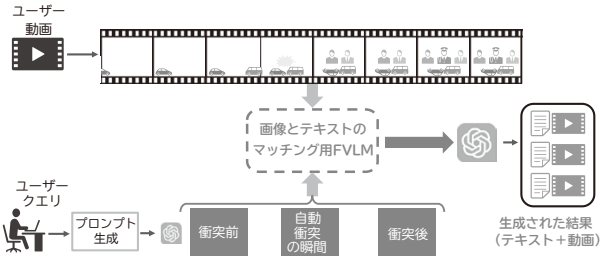

図12　VIR Labアプリケーションの動画要約に使用された筆者らの特定領域向けFVLM

図13　リモートセンシングの例

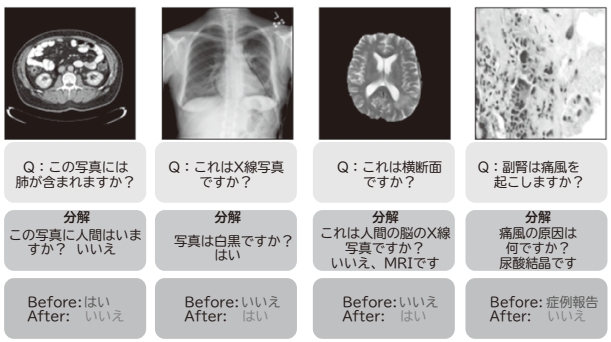

図14　医療用画像の推論の例

能なリソースを用いてタスクを解決するコードを開発することにより、コンピュータビジョンによるソリューションの開発と展開を自動化するという、基盤ビジョンLLMのアーキテクチャを紹介しました。これは、特定ドメイン向けのタスクを解決する新しいFVLMや、特定のアプリケーションで迅速にカスタマイズできるツールを開発することでサポートされています。現在、次のステップとなるいくつかの機能として、(a) デプロイ用パラメータの自動チューニング、(b) アプリケーション報酬に基づく特定タスク向けモデルを更新するための自己学習、(c) ハルシネーションとバイアスの削減、などを開発中です。

* ChatGPTは、米国OpenAI社の商標です。
* その他記述された社名、製品名などは、該当する各社の商標または登録商標です。

参考文献

1) OpenAI: GPT-4 Technical Report, 2023
https://arxiv.org/abs/2303.08774
2) Zaid Khan, Vijay Kumar BG, Xiang Yu, Samuel Schulter, Manmohan Chandraker and Yun Fu: Single-Stream Multi-Level Alignment for Vision-Language Pretraining, European Conference on Computer Vision, 2022
https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136960725.pdf
3) Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Vijay Kumar B.G, Yumin Suh, Manmohan Chandraker and Dimitris N. Metaxas: Taming Self-Training for Open-Vocabulary Object Detection, 2023
https://arxiv.org/abs/2308.06412
4) Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar B.G, Anastasis Stathopoulos, Manmohan Chandraker, Dimitris Metaxas: Exploiting Unlabeled Data with Vision and Language Models for Object Detection, 2022, European Conference on Computer Vision
https://arxiv.org/abs/2207.08954
5) Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu and Manmohan Chandraker: Q: How to Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: Self-Train on Unlabeled Images!, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.15005-15015, 2023
https://www.computer.org/csdl/proceedings-article/cvpr/2023/012900p5005/1POQzQuyw1O
6) Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker and Yun Fu: Exploring Question Decomposition for Zero-Shot VQA, Conference on Neural Information Processing Systems, 2023
https://arxiv.org/abs/2310.17050

執筆者プロフィール

KHAN Zaid
Northeastern University

KUMAR B G Vijay
NEC Laboratories America
Researcher

SCHULTER Samuel
NEC Laboratories America
Senior Researcher

CHANDRAKER Manmohan
NEC Laboratories America
Department Head
University of California San Diego
Professor

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報（日本語）　NEC Technical Journal（英語）

Vol.75 No.2　ビジネスの常識を変える生成AI特集
～社会実装に向けた取り組みと、それを支える生成AI技術～

ビジネスの常識を変える生成AI特集によせて
生成AI技術への取り組み ～基盤から応用、ルール作りまで～

◇ 特集論文

急速に広まる生成AIの市場適用
NECにおける生成AIの取り組みについて
電子カルテと医療文書の作成支援による医師業務効率化
映像×LLMによる報告書作成業務の自動化
映像分析と生成AIによるリアル世界の行動理解
サイバー脅威インテリジェンス生成自動化
生成AIの社内活用を進めるNEC Generative AI Service（NGS）
ソフトウェア・システム開発への生成AIの活用
LLMとMIで革新する素材開発プラットフォーム
LLMと画像分析を活用した被災状況の把握

生成AIの可能性を高める基盤技術
日本語性能に優れたNECの大規模言語モデル（LLM）
NECの生成AIを支える国内企業で最大規模のAIスーパーコンピュータ
より安全な大規模言語モデル（LLM）を目指して
データを秘匿したまま連携できる連合学習技術とLLMへの適用可能性
大規模言語モデル（LLM）によるFew-shotクラスタリングの可能性
オープンドメイン常識推論のための知識拡張型プロンプト学習
AI連携とオーケストレーションのための基盤ビジョンLLM
クエリを考慮した新規手法により関連する企業データを減らしLLM APIの使用コストを最適化

AI技術が社会へ浸透するために
AI標準化・ルールメイクに関する動向とNECの取り組み
人権尊重に向けたNECのAIガバナンスの取り組み
RCModelを用いたAIリスクマネジメントのための人材育成事例

◇ NEC Information
2023年度C&C賞表彰式典開催

Vol.75 No.2
（2024年3月）

特集TOP