TSUBAME 共同利用　令和6年度　産業利用　成果報告書

利用課題名　基盤モデルの進化的構築技術の確立
英文：Establishing an Evolutionary Design Methodology for Foundation Models

三﨑航
Kou Misaki

Sakana AI 株式会社
Sakana AI
https://sakana.ai/

邦文抄録（300 字程度）
大規模言語モデル（LLM）に代表される基盤モデルは、さまざまな分野で活用され始め、工学における問題解決の方法論に大きな変革をもたらしている。しかしながら、基盤モデルを構築および利用するための体系的な方法論は確立されておらず、研究者や開発者は依然として経験や暗黙知に大きく依存しているのが現状である。本プロジェクトでは、事前学習後の基盤モデルの応用において特に重要な役割を果たす、ファインチューニング用データセットの自動構築に取り組む。具体的には、画像と言語の融合分野である Vision and Language（以下、VL）におけるファインチューニングデータセットを自動的に構築する手法を開発する。本プロジェクトでは進化計算をはじめとする進化的手法の概念をヒントに、効率的かつ効果的なデータセットの自動構築を可能にする方法論の確立を目指す。

英文抄録（100 words 程度）
Foundation models based on large language models (LLMs), are increasingly being utilized across various fields, driving a paradigm shift in problem-solving methodologies in engineering. However, there is still no established systematic methodology for building foundation models efficiently, and researchers and developers must largely rely on trial and error. This study focuses on the automatic construction of training datasets, which play a crucial role in developing foundation models. Specifically, we aim to develop a method for automatically constructing training datasets for foundation models in the Vision and Language (VL) domain. By leveraging evolutionary approaches, we seek to establish a methodology for efficiently and effectively automating the construction of foundation models.

Keywords: LLMs, Foundation models, Vison and language models, Evolutionary approach

1. 背景と目的

大規模言語モデル（LLM）に代表される基盤モデルは、近年の人工知能（AI）研究を牽引しており、今後もその発展が重要な役割を果たすと考えられる。深層学習の基盤技術が成熟した結果として基盤モデルが登場したように、基盤モデルのさらなる発展が次世代の技術革新に不可欠である。しかしながら、基盤モデルの構築および応用に関しては、依然として体系的な方法論が確立されておらず、研究開発の進展を阻む要因となっている。

本プロジェクトでは、基盤モデルの応用において不可欠なファインチューニング用データセットの自動構築を目的とする。特に、画像と言語の融合領域である Vision and Language（以下、VL）に焦点を当て、ファインチューニングデータセットを自動生成する手法の開発を行う。

本プロジェクトでは、進化計算を代表とする進化的アプローチを活用し、ファインチューニングデータセットを自動構築する手法を提案する。最終的に、本手法によって生成されたデータセットが、人手で構築されたデータセットと同等の性能を達成可能であることを実証する。

2. 概要

本プロジェクトでは、進化的アプローチを用いたデータセット構築の手法として、以下の二つのアプローチを検証した。

2.1 VL タスクの自動設計

第一の手法は、ファインチューニング用タスクそのも

Task Name	Instruction
Contextual text interpretation	Interpret the meaning of text within the context of the surrounding visual elements. For example, explain what a label or sign refers to based on nearby objects or scenes.
Text classification	Classify the identified text in the image into predefined categories such as titles, publication information, and volume details.
Hierarchical text reasoning	Determine the importance and hierarchy of multiple text elements within the image, such as distinguishing between main and supporting text in complex visual settings.

表 1. GPT-4o によって自動生成されたタスク例

のを自動設計するものである。従来の研究では、画像キャプション生成、画像を用いた対話、視覚推論の三つのタスクが一般的に利用されている。しかし、これら以外にも例えば背景領域の詳細記述、物体の動作に関する質疑応答、画像内の文字認識といった多様なタスクの可能性が考えられる。次トークン予測の最適化が LLM の成功をもたらしたように、ファインチューニングにおいても最適なタスクが存在するはずである。

本研究では、タスクの進化的生成手法を提案する（図 1 参照）。まず、任意のデータセット D を用意し、対象の視覚言語モデル（VLM）をファインチューニングする。その後、検証用データセットを用いてモデルの性能を評価し、誤答サンプルを収集する。これらの誤答データを基に、新たなタスク（タスク名と指示文）を自動生成し、アノテーションモデルを用いて適切な指示文および回答文を作成する。このプロセスを繰り返すことで、モデルが苦手とする領域を補完しつつ、最適なタスクを設計することが可能となる。本手法では、タスク設計およびアノテーションに GPT-4o を活用し、各ステップで 10 種類のタスクを生成する。実際に生成されたタスク例を表 1 に示す。

2.2 質疑応答文の自動設計

第二の手法は、詳細な画像キャプションを基に質疑応答データを自動生成するものである。まず、GPT-4o を用いて学習画像ごとに詳細なキャプションを生成する。次に、生成されたキャプションを基に、GPT-4o を活用して画像に関する質問文および回答例を自動生成する。さらに、これらの質問文を発展させ、より簡単な質問とより難しい質問を生成し、それぞれに対する回答を作成する。このプロセスを反復することで、多様な質疑応答データを自動生成し、データセットの質を向上させる。

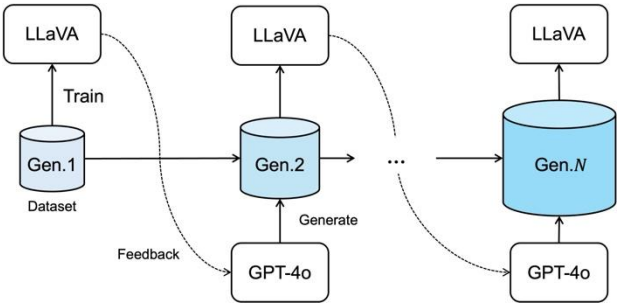

図 1. タスク自動設計の概要図

3. 結果および考察

3.1 実験設計

本研究では、LLaVA-1.5 [1]を学習対象の VLM として採用した。Flamingo [2]などのクロスアテンション機構を用いた VLM も検証したが、LLaVA-1.5 がより高い性能を示したためである。

また、性能検証には TextVQA [3]と呼ばれるベンチマークを採用した。TextVQA は、画像内に写っているテキストに関する認識能力を測るベンチマークである。高い正答率を得るためには、正確なテキスト認識と物体認識、そしてそれらの情報を用いた推論能力が問われる。画像と言語の双方にわたる複数の認識能力が必要となることから、本プロジェクトの評価対象に適していると考え、採用した。

3.2 実験結果

図 2 に、VL タスク自動設計のアプローチ（2.1 節）の

TextVQA での評価結果を示す。図 2 の横軸はタスク設計の回数を示し、縦軸は TextVQA での正答率を表している。図 2 から明らかなように、タスク設計を繰り返すことで、モデルの性能が向上していることがわかる。全体的なコストの都合で合計 5 回のタスク設計を行ったが、最終的に人手でタスク設計を行っているデータセット（LLaVA-1.5[1]で 58.2%）と同程度の性能（57.4%）を達成可能なことを示せた。なお、LLaVA-1.5 による結果は我々が手元で再現した結果である。

また、我々は 2.2 節のアプローチで構築したデータセットを用いて VLM をファインチューニングし、同様に TextVQA で性能評価を行った。質疑応答生成のラウンドを繰り返したが、最終的な回答精度は 56.1%に留まった。

3.3 現状の課題

本プロジェクトで目標とした VLM のファインチューニング用データセットの自動構築について、以下の二つの課題が明らかになった。

アノテーションモデルの性能限界：例えば、TextVQA には画像内に写っているアナログ時計の時刻を回答させるものがあるが、GPT-4o はほとんどのサンプルに対して誤答してしまう。同様の現象がほかのサンプルにも生じうる。そのため、誤ったアノテーションがファインチューニング用データに含まれてしまい、VLM の性能向上を阻害する要因となった。

タスクの多様性の低下：タスク設計を繰り返すにつれて、異なる指示文であっても実質的に類似したタスクが生成される傾向が見られた。この結果、データセットの多様性が損なわれ、性能の向上が限定的になっていく。

3.4 マルチエージェントによるタスク自動設計

タスクの多様性の低下を解決するために、マルチエージェント手法を導入した。具体的には、我々は GPT-4o と Claude Sonnet 3.5（Claude-3.5）を用い、それぞれのモデルの表現力を相互補完することを試みた。

具体的には、GPT-4o で生成した 10 個のタスクを Claude-3.5 に渡し、内容を洗練化させた後、GPT-4o に戻し最終的に 5 個のタスクを選択させる。このプロセスを繰り返すことで、データセットの進化を図った。

しかし、図 3 に示すように、GPT-4o 単独で設計したタスクの方が高い性能を示した。エージェント間の議論を促進する実験も行ったが、建設的な議論の形跡は見られなかった。今後は、異なるエージェントの知識を効果的に統合する手法が求められる。

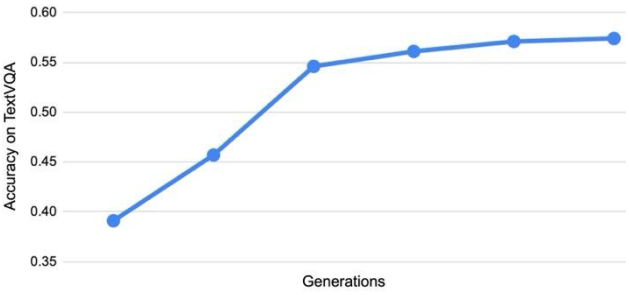

図 2. TextVQA における評価結果

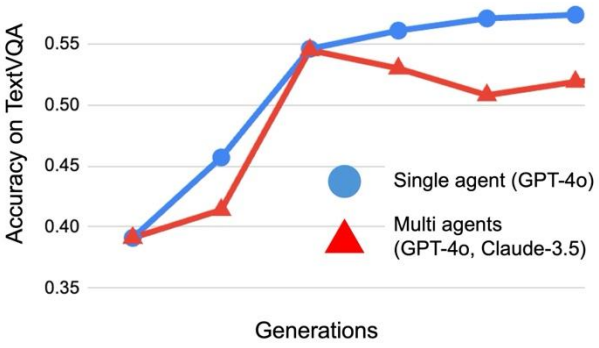

図 3. マルチエージェントによる評価結果

4. まとめ、今後の課題

本プロジェクトでは、基盤モデルのファインチューニングデータセットを自動構築する手法を提案し、人手構築のデータセットと同等の性能を達成できることを示した。しかしながら、より効果的な自動構築手法の確立にはさらなる研究が必要である。特に、マルチエージェントシステムの活用を通じて、個々のモデルの限界を克服する枠組みの開発が今後の課題となる。

謝辞

本プロジェクトの遂行にあたり、計算資源および環境を提供してくださった東京科学大学 情報基盤センターの皆様ならびに関係者の皆様に深く感謝申し上げます。

参考文献

[1] H. Liu+, Improved Baselines with Visual Instruction Tuning, CVPR, 2024.

[2] J. B. Alayrac+, Flamingo: a Visual Language Model for Few-Shot Learning, NeurIPS, 2022.

[3] A. Singh+, Towards VQA Models That Can Read, CVPR, 2019.