TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　多言語マルチモーダル大規模言語モデルに関する研究

利用課題責任者　小町守

所属　一橋大学大学院　ソーシャル・データサイエンス研究科
URL https://hit.komachi.live

邦文抄録（300 字程度）

多言語マルチモーダル大規模言語モデルの研究について、視覚言語モデルにどれくらい画像の認識能力があるのかの検証を行いました。視覚言語モデルを用いて画像特徴量を得て、テキストに関連した画像を検索するような予備的なタスクの設計と、画像の時間情報を考慮する画像質問応答に関する新しいアーキテクチャの検討を行いました。画像質問応答に関しては、大規模言語モデルを用いて自動的に重要度を付与したデータセットを作成し、重要度を自動的に推定するアーキテクチャを考案し、実験したところ、提案手法はベースラインの手法と比較し、画像の時間情報を考慮した質問応答システム性能の向上に寄与することが示唆されました。

背景と目的

画像やテキストを統合的に利活用するマルチモーダルな大規模多言語言語モデル（視覚言語モデル）に関する研究が盛んに行われています。視覚言語モデルがどれくらい言語と視覚のインタラクションを扱うことができるかは未知数であるため、視覚言語モデルの認識能力がどれくらいあるのかを検証しました。また、画像質問応答において、関連性が低い画像を用いて質問応答を行う可能性があるため、重要度の高い画像を推定しつつ質問応答ができるアーキテクチャの考案を行いました。

概要

画像やテキストを統合的に利活用するマルチモーダルな大規模多言語言語モデルに関する研究を行いました。ViCLIP を用いて画像に対する画像特徴量を得てテキストに関連した画像を検索するような予備的なタスクの設計と、それを用いた画像質問応答のアーキテクチャに関する検討を行いました。データセットは大規模言語モデル（GPT-4o, Gemini）によって評価および説明の生成を行い、質の高いデータセットとなるようにフィルタリングを行いました。また、画像質問応答については画像に付与された前後の時間情報を考慮したアーキテクチャにすることで、画像の質問に対する関連度（重要性）を適切に推定することを狙いました。

結果および考察

画像とテキストを CLIP によって統合的に扱うことが可能である、ということが検証できました。画像質問応答に関しては検索された画像の重要度を推論するアーキテクチャを考案し、重要度の判断を自動付与したデータセットを作成してファインチューニングすることで、重要度を自動的に推定できることも示しました。直接的評価、間接的評価の両方でベースラインと比較して性能の向上を達成することができました。一方、前後の時間情報を両方考慮するようなアーキテクチャには必ずしも大きな性能の向上が得られないことも分かりました。

まとめ、今後の課題

マルチモーダル大規模言語モデルの利活用に関する研究に取り組みました。具体的には画像質問応答において、画像とテキストのインタラクションを用いて検索する新しいタスクを提案し、時間情報を考慮しつつ関連度の高い画像を用いて質問応答する新しいアーキテクチャを考案し、実験的に有効性を検証しました。

ダウンストリームのタスクでハイパーパラメータの探索をすることに時間がかかることが分かり、どのように TSUBAME のプラットフォームを使いこなしていくのかについて研究グループ内で知見を溜めていく必要があることが分かりました。