多用途ビジョンエンコーダを使った マルチモーダル大規模言語モデルについて

福岡デザイン＆テクノロジー専門学校

ＡＩクリエーター専攻

ブリゲッタ　ニコル

目次

- 概要
- 特徴
- 事例
- 結果
- まとめ

概要

- マイクロソフト研究所、Georgia Tech（ジョージアテック）とPicsart AI Research（ピックスアートAI研究所）によるVCoder – 多用途ビジョンエンコーダソリューション。
（記事： https://www.marktechpost.com/2023/12/27/researchers-from-microsoft-and-georgia-tech-introduce-vcoder-versatile-vision-encoders-for-multimodal-large-language-models/）
- 現状のマルチモーダル大規模モデル（MLLMs）は視覚言語に関するタスクに才能を魅せてきた。
- しかし、これらのモデルは単純な物体認証タスクで正確に認識と数えることが弱点になっている。
- MLLMsを改善するために施策の特徴は次の通りです。

特徴

- モデルにセグメンテーションや深度マップの追加
  - 高い順位のコンポーネントを認識し、重みの行列を削減することにより、トランスフォーマー内の特定レイヤーに集中できる
- これにより、モデルの物体レベルの認識を向上し、追加訓練とパラメータの必要性がなくなった。

事例

- COCOデータセット
- 出力：COCOセグメンテーションテキスト

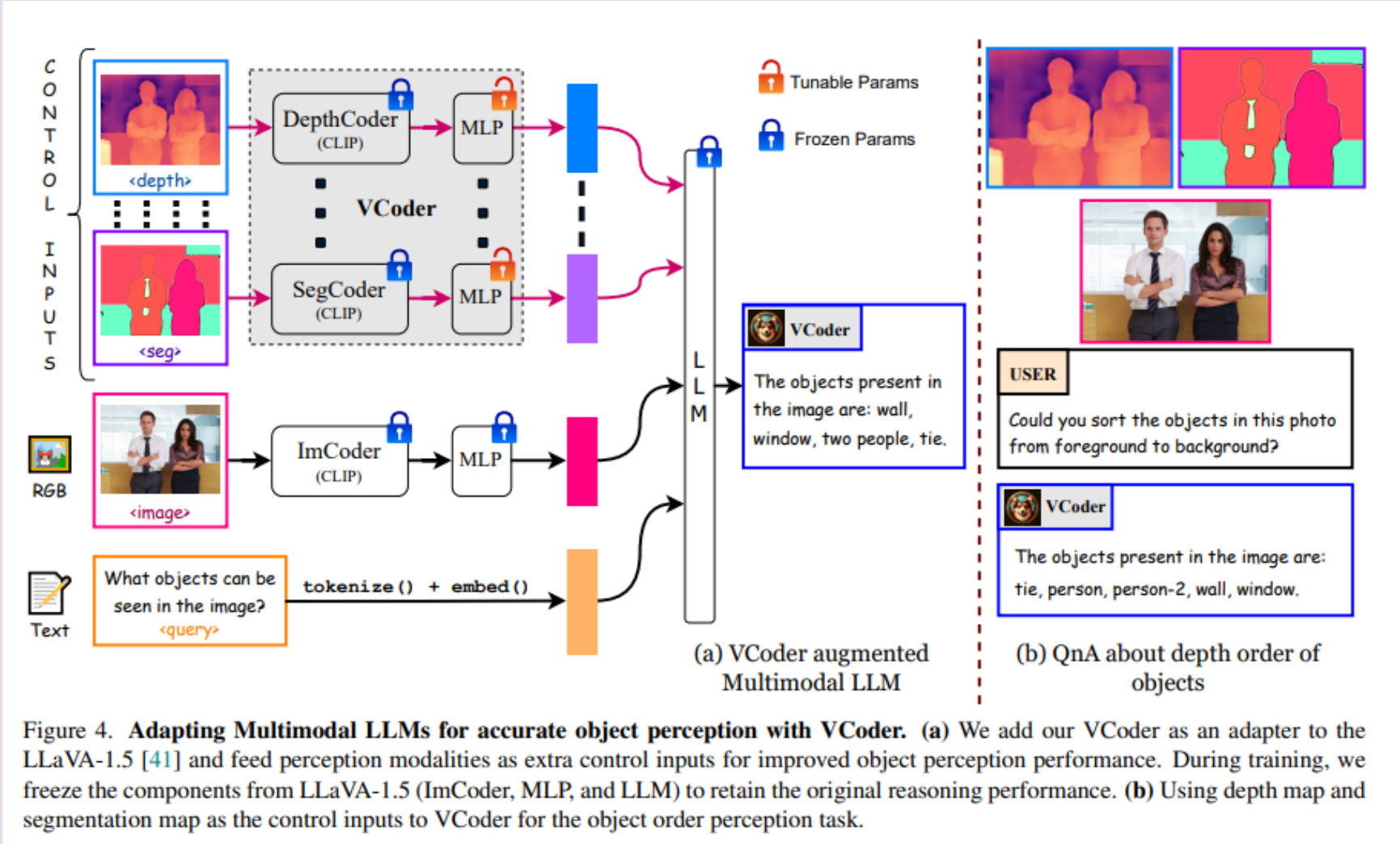

Figure 4. **Adapting Multimodal LLMs for accurate object perception with VCoder.** **(a)** We add our VCoder as an adapter to the LLaVA-1.5 [41] and feed perception modalities as extra control inputs for improved object perception performance. During training, we freeze the components from LLaVA-1.5 (ImCoder, MLP, and LLM) to retain the original reasoning performance. **(b)** Using depth map and segmentation map as the control inputs to VCoder for the object order perception task.

結果

- 効率的に物体認識タスクの精度が上がった
- 正解率の向上、特に訓練データの情報が頻繁に表現がない時
- 複雑な視覚的推論タスクのパフォーマンス向上

まとめ

- VcoderがMLLMsの最適化の進化に貢献した
- パソコンに負担をかけずに、モデルの効率向上に成功した。
- これにより、MLLMsのパフォーマンス向上の他に、複雑な視覚の処理と理解の精度を向上させることができました。
- 今後認識と推論がより効率な言語モデルを開発されるのを期待しています。

ご清聴ありがとうございました