LLaVA(Visual Instruction Tuning)

ソース

- https://arxiv.org/pdf/2304.08485
- https://github.com/haotian-liu/LLaVA

概要

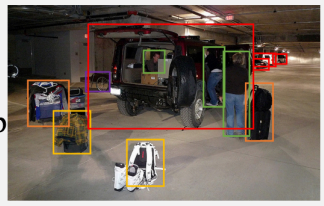
- LLaVAを提案
  - Large Language and Vision Assistant
- 概要は4点
- Multimodal instruction-following data
  - vision-language instruction-following dataが不足している
  - 上記の課題に対してGPT-4を使って画像とテキストのペアをinstruction-following formatに変換するデータ生成パイプラインを提案
- Large multimodal models
  - 大規模マルチモーダルモデル(LMM)の開発
  - CLIPのvisual encoderとVicunaのlanguage decoderを接続
  - 生成したinstructional vision-language dataを使ってend-to-endでfine-tuning
- Multimodal instruction-following benchmark
  - 多様な画像、指示、詳細な注釈を含むLLaVA-Benchという2つのチャレンジングなベンチマークを発表
- Open-source
  - 生成したmultimodal instruction data、コードベース、モデルチェックポイント、ビジュアルチャットデモを公開

データ生成

- 画像から2つのsymbolic representationを得る
  - Captions
  - Bounding boxes
- レスポンスとして以下3つ用意
  - conversation

    - detailed description
    - complex reasoning
- 論文のデータ例がわかりやすい

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



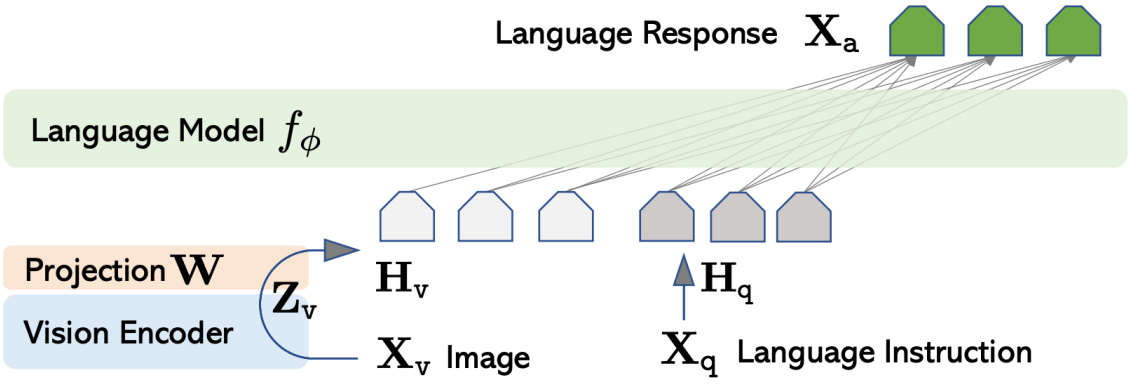
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

(論文より引用)

Architecture

- LLaVAのnetwork architectureは下図



(論文より引用)

- X_Vが画像, X_qがLanguage Instruction, X_aがLanguage Response
- X_qからsequence of tokens H_qを得る

- CLIPのVision Encoder(ViT-L)gを使用してZ_Vを得る
- 得られたZ_Vの次元をword embedding spaceと一致させるためにProjectionする

$$
\begin{aligned}
Z_V &= g\left(X_V\right) \\
H_V &= W \cdot Z_V
\end{aligned}
\tag{1}
$$

- 得られたH_V, H_qからLLM $f_\phi\left(\cdot\right)$を通してX_aを得る
- f_θはVicunaのdecoder

Training

- X_Vからmulti-turn conversation data $\left(X_q^1, X_a^1, \cdots, X_q^T, X_a^T\right)$を生成
  - Tはturnの合計数
- t番目のturn X_{instruct}^tを以下で定義

$$
X_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [X_q^1, X_V] \text{ or } [X_V, X_q^1], & \text{the first turn } t = 1 \\ X_q^t, & \text{the remaining turns} \quad t > 1 \end{cases}
$$

- このとき以下でX_aが計算される

$$
p\left(X_a \mid X_V, X_{\text{instruct}}\right) = \prod_{i=1}^{L} p_\theta\left(x_i \mid X_V, X_{\text{instruct},<i}, X_{a,<i}\right)
$$

- x_iは以下の画像の緑部分

$\mathbf{X}_{\text{system-message}}$ <STOP>
Human : $\mathbf{X}_{\text{instruct}}^1$ <STOP> Assistant: $\mathbf{X}_{\text{a}}^1$ <STOP>
Human : $\mathbf{X}_{\text{instruct}}^2$ <STOP> Assistant: $\mathbf{X}_{\text{a}}^2$ <STOP> $\cdots$

(論文より引用)

- 学習は以下の2stageから構成される
  - Stage 1: Pre-training for Feature Alignment
  - Stage 2: Fine-tuning End-to-End

Stage 1: Pre-training for Feature Alignment

- CLIPのencoderとLLMのdecoderをfreezeする

- つまり、学習可能なパラメータを$\theta = W$とする
- これはvisual tokenizerの学習とみなせる

Stage 2: Fine-tuning End-to-End

- 次にCLIPのencoderをfreezeしてdecoderと行列を学習する
- つまり、学習可能なパラメータを$\theta = \{W, \phi\}$とする