CS6101-14

Textualization of Visual Information

Xiao Xu

Abstract

Visual information (image labels, image captions, object labels, etc.) contains the humans' (or visual model's) understanding and analysis of **key features** in the image.

We investigate the performance of LLMs in **directly** accomplishing visual perception/reasoning tasks by transforming visual information in images into **text** via vision experts.

Framework

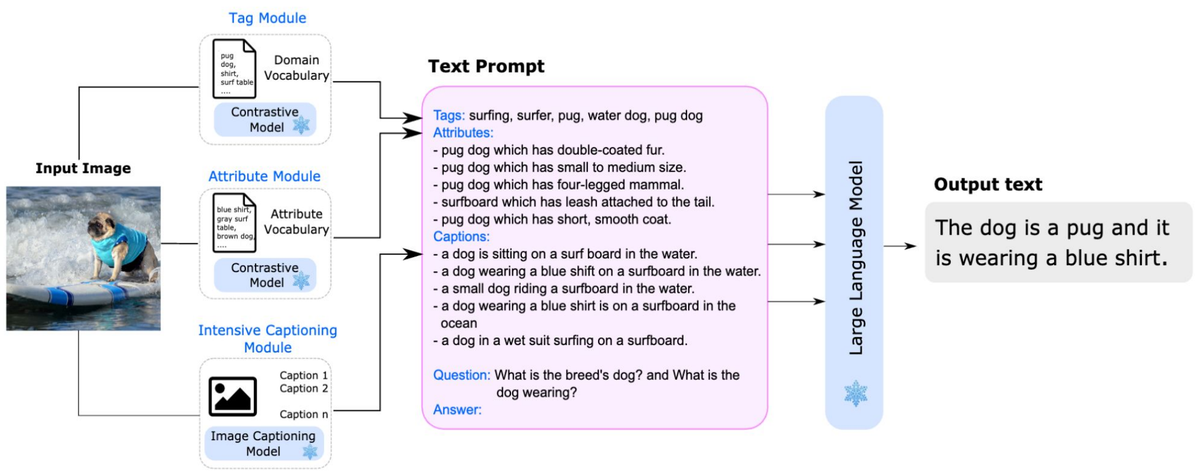

Figure 1: LENS Framework. Figure from LENS paper.

We follow the LENS framework.

1. Construct Tag Vocabulary (~22k) and Attribute Vocabulary (~35k): collect from existing datasets or generate from LLMs.
2. Calculate the matching score between the image and tags or attributes by CLIP.
3. Generate image captions with BLIP.
4. Select Top-N tags, attributes and captions.
5. Design a prompt template and pass them to LLMs.
6. Fill the above, as well as the question, into the designed prompt template. Pass it to the LLM and get the answer.

Evaluation

We evaluate the LENS framework on MMBench.

- **Goal**: Robustly evaluate different abilities of large vision-language models
- **Task**: Image+Question+Options => A/B/C/D
- **Ability**: 20 abilities, ~3000 single-choice question
- **Data source**: Current Benchmarks + Internet + Human Annotators + LLM-Generated
- **Naive evaluation**: input options once in a predefined order
- **Robust evaluation**: to avoid the position bias of options
  - N-option => N times input, options with circular shift
  - Only N answers are all correct => Pass
- **Option extraction**: Rules + ChatGPT-based
- **Experimental setting**
  - LLaMA2-chat-13B as the LLM, CLIP-ViT-H-14 for matching and BLIP-L for captioning.
  - Greedy decoding on ~1.1k Dev samples.
  - We only use rules to extract A/B/C/D without ChatGPT due to 💰, rules fail in 17% runs.

	Parameters	Language Model	Vision Model	Overall
PandaGPT	14B	Vicuna 13B	ImageBind ViT-H/14	45.4
LLaMA-Adapter-v2	7.2B	LLaMA 7B	CLIP ViT-L/14	41
LLaVA-1	7.2B	LLaMA 7B	CLIP ViT-L/14	38.7
Ours	14.5B	LLaMA2-Chat-13B	CLIP-ViT-H-14 BLIP-L	38.4
Qwen-VL	9.6B	Qwen-7B	ViT-G/16	38.2
VisualGLM	8B	ChatGLM 6B	EVA-CLIP	38.1
InstructBLIP	8B	Vicuna 7B	EVA-G	36

Table 1: LENS Framework. Figure from LENS paper.

Analysis and Conclusion

- **Claim**: The comparison is not fair. Except for different Language models and vision models, LENS is not trained on massive multimodal (instruction) data, and extract options only by rules.
- **Performance**: **38.4** overall accuracy (No. 1 is 74.8), which is similar to LLaVA-1 (No. 15).
- **Pros**: When Captions + Tags + Attributes are enough 😋
  - Identity reasoning: What's the profession of the people in this picture?
  - Image scene: What type of environment is depicted in the picture?
  - Function reasoning: What's the function of the demonstrated object?
- **Cons**: When Captions + Tags + Attributes are not enough 😭
  - Image quality, Object localization, Spatial relationship, Future prediction, …
- **Robustness**: Overall accuracy with Naive evaluation is **58 (> 38.4)**.
  - LLaMA2-Chat-13B is less robust against different option orders.
- **Conclusion**
  - Vision experts may misidentify or miss some information. More vision experts, such as object detectors and OCR models, are needed to recognize more information needed in MMBench.
  - Textualization is a strong baseline for MMBench, but it needs a strong LLM to understand and reason on long texts that contain redundant or even contradictory information.

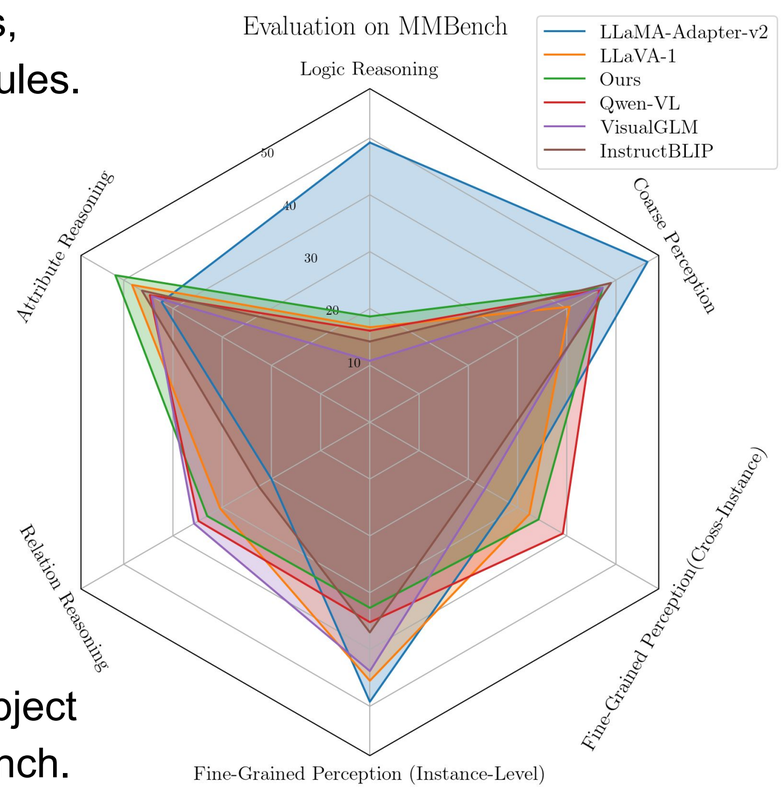

Figure 2: Evaluation on six ability dimensions.