Contrastive Recognition Through Extraction

Seth Grief-Albert   Aidan Kinnear   Dharsan Ravindran   Vivek Chokkalingam   Arya Farivar

Queen's University, **QMIND**
{seth.griefalbert, 21adk1, 20vdc1, 19dr26, 22sw59}@queensu.ca

***Abstract*—The world as we perceive it is multimodal: we learn about and interact with our environments through our perceptions. Large Language Models (LLMs) have emerged in artificial intelligence as natural language processing engines that form textual generations based on their inputs. We observe that LLMs form relationships between objects and their environments during pre-training. We propose Contrastive Recognition Through Extraction (CoRTEx), a novel framework that expands the information content of image-captioning models while grounding generations in observed visual features through a discriminator. We leverage the reasoning capabilities of a base model and encourage it to create vivid descriptions through fine-tuning on a descriptive instruction-tuning dataset. We release code on GitHub and model adapter on Hugging Face: https://github.com/SethGA/CoRTEx, https://huggingface.co/SethGA/neocortex-grounded.**

I. Introduction

Multimodal LLMs have shown increasing prevalence throughout the field of artificial intelligence. These LLMs can be leveraged for their pre-training knowledge to form linguistic shortcuts that aid in the task of downstream question-and-response on visual content [Ko et al., 2023]. These models, however, can often over rely on questions and become ungrounded to observed visual features, leading to linguistic bias. A research area of interest in artificial intelligence is Computer Vision in the Wild, which aims to develop generalizable computer vision that can be deployed in novel contexts [Li et al., 2022]. Linguistic bias is a key unsolved problem in this area. Recently, BLIP-2 has emerged as a State-of-the-Art image captioning model [Li et al., 2023]. Particularly compelling is its ability to produce zero-shot captions that are well-grounded in image features, but lack in detail and complexity. Yet, if the hallucination problem is re-framed as an inference problem, then by controlling the level of hallucination, more comprehensive descriptions can be generated. We leverage this existing capability of LLMs to form a cohesive understanding of environments captured in videos.

Our Contributions: We explore the grounding of zero-shot image captioning generations for video understanding, and use a discriminator model to construct a scene memory that can be used for downstream tasks including Visual Question Answering. We expand on the work of LLaMA-VQA [Ko et al., 2023] and utilize the latent knowledge of LLMs learned during pre-training, fine-tuning LLaMA-2 to create rich descriptions from image captions. The discriminator computes the similarity between these descriptions and the grounded visual feature memory to supervise the hallucination rate of our model.

II. Related Works

Multimodal Instruction Tuning:
Although end-to-end fine-tuning of LLMs is computationally expensive, parameter-efficient fine-tuning techniques including Low-Rank Adaptation have demonstrated comparable results through the introduction of a smaller weight matrix concatenated with the language model [Xu et al., 2023] [Hu et al., 2021]. LLaMA-Adapter proposed a zero-initialized attention mechanism to adapt a relatively small number of learnable parameters to a frozen base model such as LLaMA, which lent itself well to image inputs using a visual encoder like CLIP [Zhang et al., 2023] [Radford et al., 2021]. During pre-training, LLMs learn the relationships between objects. If language models can be made to "see" through the introduction of visual encoders, existing understanding of these relationships can be effectively leveraged.

Visual Question Answering:
Several recent advances in developing general-purpose assistants have included visual question answering (VQA) as a key focus, including Flamingo and Vicuna [Alayrac et al., 2022] [Chiang et al., 2023]. By fine-tuning an LLM on the output of a visual encoder, a model can answer natural-language questions about both visual and video context [Liu et al., 2023]. However, many such fine-tuning implementations have utilized multiple choice question-and-answer instruction tuning, which can lead to hallucinations from over-reliance on question content [Ko et al., 2023]. Other VQA approaches have used detailed prompting to assist the model in generating descriptions of visual content, like in VideoChat [Li et al., 2024]. Our approach moves towards zero-shot video descriptions through fine-tuning. We build upon the work of LLaMA-VQA to capitalize on *linguistic shortcuts* for semantic relationship modelling while reducing hallucinations. Our model architecture includes a *discriminator* that compares descriptive generations to a grounded *visual feature memory*.

World Model:
A world model for a LLM is an internal state representation of an observed environment. Language models have been shown to have more rigorous reasoning capabilities in physical contexts when exposed to embodied experience data [Xiang et al., 2023]. Reasoning via Planning is a recent framework to adapt world-modelling and enhanced reasoning capabilities to LLMs [Hao et al., 2023]. Our work builds upon the knowledge that an LLM has learned during pre-training to construct a small semantic world model we call *world memory*.

III. Visual Feature Memory

LLMs are engines of language. Efforts to extend their capabilities to the real world have proved challenging, especially due to their lack of visual perception. A key advantage that LLMs, and transformer architectures in general, have over traditional machine learning approaches is their scalability. It seems worthwhile to explore how textual capabilities can be applied to real-world tasks, but a key problem arises: How can we effectively interface LLMs with the world if they cannot process their environments? To build semantic context from a video environment, we first propose a visual feature memory F_m to ground the language model in real observations. We utilize BLIP-2 inference on each relevant video frame to form our grounding captions. A caption usually consists of a short, 5-10 word description of one image frame. The Python Natural Language Toolkit is used to extract the object features present in each frame, capturing their frequency and storing them in a permanent feature list. This list acts as a query database, which can be indexed and processed to aid in reducing hallucinations downstream.

IV. CoRTEx

A. Architecture

We aim to build a semantic world memory M_W through a video input V. Each video frame $\langle v_1\rangle, \langle v_2\rangle, \dots, \langle v_n\rangle$ captured at time $t_1, t_2 \dots, t_n$ is input into BLIP-2 to create a list of captions $\vec{c} = [c_1, c_2, \dots, c_n]$. The generator model G and visual feature memory F_m take $\vec{c}$ to generate a detailed scene description and feature list respectively 1 2.

$$G(\vec{c}) = d \tag{1}$$

$$F_m(\vec{c}) = [f_1, f_2, \dots, f_k] \tag{2}$$

Where d is the generated description and f is an image feature.

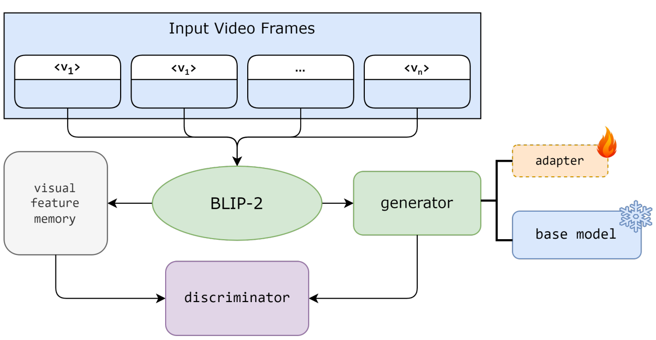

Fig. 1: CoRTEx Architecture Diagram

B. Training

We construct the neocortex_grounded_23k dataset in alpaca instruction tuning format from the detailed description dataset proposed in Visual Instruction Tuning [Liu et al., 2023]. Both datasets are based upon the train_2014 COCO dataset [Lin et al., 2015]. In addition to the generations within detail_23k, we add the human annotations as input to the model, corresponding to a similar BLIP-2 caption.

Video: $\langle v_1\rangle, \langle v_2\rangle, \dots, \langle v_n\rangle$
Descriptive question: q
Human Annotation: $\langle c_{H_1}\rangle, \langle c_{H_2}\rangle, \dots, \langle c_{H_5}\rangle$
Synthetic Annotation: $\langle c_S\rangle$

Table 1: neocortex-grounded-23k dataset construction

We fine-tune our model using 2×A40 GPUs, on LLaMA-2-7b-hf from Meta and NousResearch. We fine-tune on the neocortex_grounded_23k dataset for 3 epochs, with a learning rate of 2e-4.

C. Few-shot Scene Discrimination

The goal of the discriminator is to compare how similar the generated scene description is to the visual feature memory by measuring the hallucination occurrence through cosine similarity.

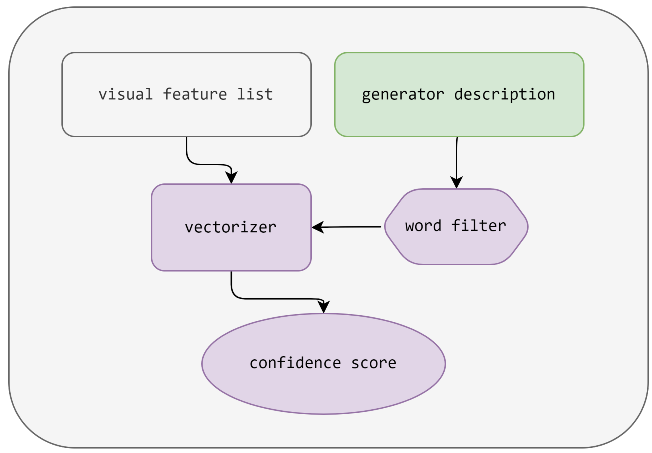

Fig. 2: CoRTEx Discriminator Diagram

The discriminator retrieves a list of features from the visual feature memory, and the corresponding generated description of the input video. First, the generated description is filtered to remove insignificant words and punctuation, resulting in a string of words each separated by a space, to match the format of the feature list. This creates a pool of words, both obtained by the visual feature memory and the generator. We create a vector for both the feature list and the generated description. The vectors contain the frequency of the words used in each string, respectively. The size of the vector is determined by the number of words in the union of the feature list and generated description.

We compare the vectors on word commonality with respect to the corpus, using Term Frequency-Inverse Document Frequency (TF-IDF) to return a float value between 0-1 for each word in the vector 3.

$$w_{i,j} = tf_{i,j} \cdot \log\left(\frac{N}{d_i}\right) \tag{3}$$

Where $tf_{i,j}$ is the frequency of i in j divided by the total number of words in j, N is the total number of documents, and d_i is the number of documents containing i. A cosine similarity function is used to determine a confidence score that the generated description is grounded in the visual feature memory 4. Cosine similarity is especially useful for computing the similarity of sparse data within text documents.

$$S_C(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

Where A and B are vectors of the same length. The cosine similarity of the two vectors (S_C) is appended to the generated description to form the world memory $M_W = d, S_C$ for use in downstream tasks.

V. Experiments

We test our model on multiple video environments with varying scene complexity. The captions generated from BLIP-2 give grounded depictions of each frame, allowing for our fine-tuned model to build upon these captions. Our model builds upon the image captions and the pre-training knowledge of LLaMA-2 to construct a human-like description of the scene. In 3, our model uses its existing knowledge of beaches to essentially ”dream” about what it is looking at. This description includes the experience of the beach-goers in regards to their sentiment and the weather, which the BLIP-2 captions provide no information on; the generator infers that the presence of many people is an implication that the weather is pleasant. These insights provide a more detailed understanding of the big picture of the scene and shows the capabilities of our model for grounded textual expansion. The corresponding confidence score given by the discriminator shows reasonable similarity from the generated description and feature list.

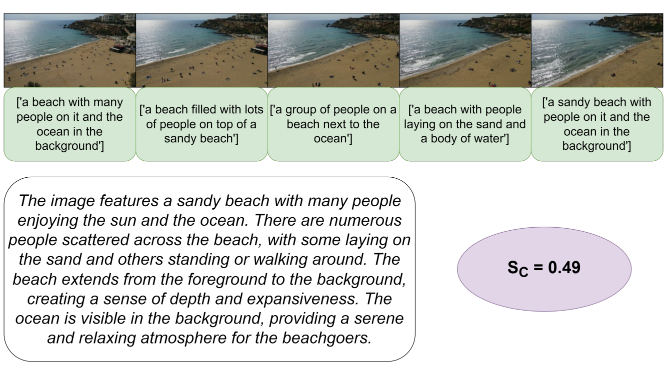

Fig. 3: Beach Scene Experiment

Longer temporal context can be absorbed by our model to improve its world memory over time. In 4, the camera is moved in a panorama around a busy intersection. The model pieces together the various frame captions into a cohesive understanding of the environment being captured.

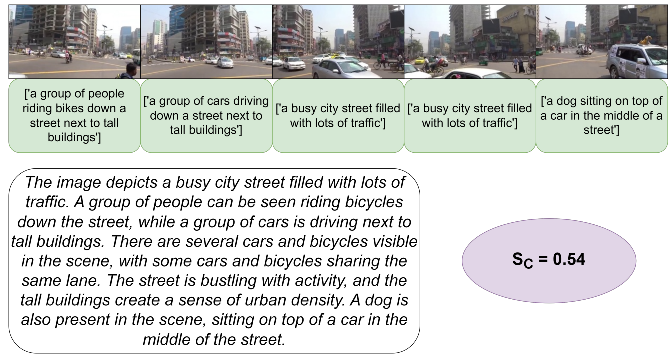

Fig. 4: Traffic Scene Experiment

A. Limitations

The CoRTEx architecture creates a mental model of a video scene we call world memory. One limitation to this approach is that specific events within the video can be lost, as our generator model produces a description of the entire scene as if it was a static image. In the future, we can explore forming multiple and continuous world memories to be used in downstream visual question answering. The discriminator is implemented to gauge the hallucination rate of the generator. Establishing a threshold for an acceptable hallucination rate such that the model could effectively use its pre-training knowledge could open a path to training the generator and discriminator with a traditional GAN architecture [Goodfellow et al., 2014]. Although reliable data accessibility remains a challenge, novel approaches towards synthetic data creation could prove lucrative. Additionally, we rely on BLIP-2 as a ”grounded” image captioning model. Future work could implement CLIP as a visual encoder, in which image features to the word embedding space through a linear layer, as proposed in Visual Instruction Tuning [Liu et al., 2023].

VI. Conclusion

We propose CoRTEx, an architecture leveraging LLaMA-2, fine-tuned on synthetic video annotations, and BLIP-2 captions to build a rich understanding of environments portrayed in videos. We create a discriminator model to understand the extent of which a generated description is grounded vs. hallucinated. We form a small-scale *world memory*, as a precursor to a generalized world model for LLMs.

References

[Alayrac et al., 2022] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning.

[Chiang et al., 2023] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

[Goodfellow et al., 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

[Hao et al., 2023] Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. (2023). Reasoning with language model is planning with world model.

[Hu et al., 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

[Ko et al., 2023] Ko, D., Lee, J. S., Kang, W., Roh, B., and Kim, H. J. (2023). Large language models are temporal and causal reasoners for video question answering.

[Li et al., 2022] Li, C., Liu, H., Li, L. H., Zhang, P., Aneja, J., Yang, J., Jin, P., Hu, H., Liu, Z., Lee, Y. J., and Gao, J. (2022). Elevater: A benchmark and toolkit for evaluating language-augmented visual models.

[Li et al., 2023] Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

[Li et al., 2024] Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. (2024). Videochat: Chat-centric video understanding.

[Lin et al., 2015] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

[Liu et al., 2023] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

[Xiang et al., 2023] Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., and Hu, Z. (2023). Language models meet world models: Embodied experiences enhance language models.

[Xu et al., 2023] Xu, L., Xie, H., Qin, S.-Z. J., Tao, X., and Wang, F. L. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment.

[Zhang et al., 2023] Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., and Qiao, Y. (2023). Llama-adapter: Efficient fine-tuning of language models with zero-init attention.