Towards Implementing a Smaller Context-Window Size in LENS for Visual-Question Answering

Independent Study — Project Report

Advait Deshmukh, Ashwath Ramakrishnan
advait.deshmukh@colorado.edu ,ashwath.ramakrishnan@colorado.edu

May 8, 2024

1 Introduction

The advent of multimodal Large Language Models (LLMs) has significantly impacted the field of artificial intelligence, with state-of-the-art (SOTA) models like GPT-4 [1] and Gemini [7] spearheading advancements in achieving multimodality. Despite these developments, the quest to develop an open-source LLM capable of performing visual tasks efficiently continues. While the prevailing strategy among leading models tends to involve scaling up the volume of data and increasing model parameters, it is imperative to reassess the methodologies employed to integrate visual capabilities into these predominantly text-based systems.

Our investigation began with a comprehensive review of the literature within this domain, focusing initially on pioneering approaches such as ViLBERT and Vision Transformers. These models typically rely on some form of joint training that leverages both text and images, aiming to synchronize these modalities through various objectives. However, recent innovations are shifting the paradigm on how vision is integrated into text-based models. One such novel approach is encapsulated in the LENS framework, which served as the cornerstone for our experiments.

This study aims to delve deeper into the LENS [2] approach, exploring its potential and limitations. Furthermore, we propose an enhancement to this system through a sophisticated filtering mechanism designed to optimize the processing efficiency of the model.

2 Related Work

The exploration of multimodal integration within Large Language Models has led to various innovative approaches. This section reviews significant contributions in this domain, focusing on different strategies employed by ViLBERT, Vision Transformers (ViT), CLIP, BLIP, and the LENS framework.

2.1 ViLBERT: Joint Multimodal Learning

ViLBERT (Vision-and-Language BERT) [5] utilizes a dual-stream architecture involving co-attentional layers that process visual and textual inputs in parallel. This model trains on mixed datasets of image captions and visual question answering tasks, fostering its ability to understand complex multimodal data dynamically. This approach emphasizes deep integration of both modalities through joint attention mechanisms, enhancing the model's ability to perform a variety of visual and textual tasks.

2.2 Vision Transformers (ViT): Applying Transformers to Images

Vision Transformers [3] revolutionize image processing by treating images as sequences of patches, applying a pure transformer architecture without traditional convolutional layers. ViT demonstrates that transformers can effectively learn contextual relationships within images, providing a powerful basis for extending transformer models to multimodal applications where they process both visual and textual data.

2.3 CLIP: Contrastive Language-Image Pre-training

CLIP (Contrastive Language-Image Pre-training) [6] by OpenAI advances multimodal learning by training on a vast range of images and text pairs collected from the internet. Unlike traditional approaches that require specific dataset curation, CLIP learns visual concepts from natural language descriptions, enabling it to understand and categorize a wide array of images that were never seen during training. Its robust performance across diverse visual tasks illustrates the potential of learning from natural language supervision, thus broadening the scope of language models to comprehend visual content directly.

2.4 BLIP: Bootstrapped Language Image Pre-training

BLIP [4] enhances the capabilities of multimodal models by focusing on generating informative image captions and text-based image editing, pushing the boundaries of image-language pre-training. It bootstraps from existing models and introduces novel training strategies that improve both the relevance and the richness of the text generated in response to visual stimuli. BLIP's approach to synthesizing text that accurately reflects complex visual scenes further advances the state-of-the-art in language-driven image understanding.

2.5 LENS: Leveraging Independent Modules for Multimodal Integration

The LENS framework introduces a modular approach to multimodality by employing independent "vision modules" like CLIP and BLIP to first convert visual inputs into comprehensive textual descriptions. These descriptions are then processed by a standard LLM, allowing the model to apply its text-based processing capabilities to visual content. This strategy enables the use of pre-trained text-only LLMs for visual tasks without additional multimodal training, significantly reducing computational overhead and increasing flexibility. The use of CLIP and BLIP within LENS exemplifies how cutting-edge image understanding technologies can be integrated into traditional LLM frameworks to enhance their multimodal capabilities without extensive retraining.

3 Methodology

3.1 Motivation

The primary motivation for optimizing the LENS framework stems from its inherent limitations in handling detailed imagery with rich content. Current strategies may lead to potential information loss due to the constrained context size of smaller LLMs. This limitation becomes particularly evident when the system fails to include captions that are crucial for answering complex questions, thereby necessitating a more effective approach to manage and utilize the limited context capacity efficiently.

3.2 Proposed Methodological Enhancements

To overcome these challenges, we propose a selective filtering mechanism within the LENS framework. This methodological enhancement aims to dynamically evaluate and filter captions based on their relevance to the task at hand, allowing for a more effective use of the LLM's context size. The proposed filtering mechanism involves the development of an additional sophisticated module that assesses each caption's pertinence to the visual question or context. By integrating advanced machine learning techniques, this filter will selectively pass only the most relevant captions to the LLM, ensuring that each piece of text processed contributes meaningfully to the task outcome.

3.3 Expected Benefits

The integration of a selective caption filtering mechanism is expected to enhance the LENS framework by:

- **Improving information retention** through the capability to initially process a larger number of captions without overwhelming the LLM.
- **Optimizing context utilization** by ensuring that only pertinent information is considered, thus maximizing the efficiency of the limited context space available.
- **Increasing accuracy and reliability** of the system's outputs in answering complex visual questions through focused and relevant data processing.

4 Experiments

The central component in optimizing the LENS framework involves the efficient processing of hundreds of captions to identify the most relevant information quickly. Given the natural suitability of the attention mechanism in transformer architectures for tasks that require relevance judgment, we opted to utilize transformers to develop our filtering module. This attention mechanism enables the transformer to focus on the most pertinent parts of the input data, making it an ideal choice for determining the relevance of each caption in relation to a specific visual query. However, fine-tuning a transformer for this binary classification task—deciding whether a caption is relevant or not—required a specialized dataset consisting of caption-question-label rows.

As such a dataset was not readily available, we started off our experiments by attempting to create this dataset, laying the groundwork for subsequent experiments aimed at refining the caption filtering process within the LENS framework.

4.1 Exp. 1: Generating Descriptive Captions Using LENS

In our first experiment, we employed the LENS framework to generate descriptive captions for a selection of 85 diverse images from the VQA 2.0 dataset, resulting in a total of 838 captions. This process utilized intense captioning modules that analyze and interpret visual content to produce textual descriptions encapsulating key elements of each image. The primary aim was to assess the descriptiveness, accuracy, and relevance of these captions to evaluate how effectively the LENS framework can bridge the gap between visual data and textual understanding without the direct visual training of the LLM. The successful generation of a significant number of relevant captions demonstrates the potential of LENS to enhance the multimodal capabilities of LLMs, setting a solid foundation for subsequent experiments focused on refining the caption selection process.

4.2 Exp. 2: Labeling Image-Caption Pairs with GPT-4 and Gemini-Pro

To transform the generated 5000 image-caption pairs into image-caption-answerable tuples, we utilized GPT-4, a state-of-the-art LLM, to automate the labeling process due to the impracticality of manual labeling given the scale of the dataset. We designed a specific prompt: *Caption: {caption}\nQuestion: {question}\nAnswer only if certain, otherwise say 'Cannot be inferred':*. This format was intended to distinguish answerable from not-answerable captions, mapping "Cannot be inferred" to label "1" and any definitive answer to label "0".

Despite the initial setup, we noticed that GPT-4 occasionally mislabeled data. To improve the reliability of our dataset, we incorporated a second LLM for a cross-validation step. This process involved generating labels from the second LLM for the same data and comparing them. We discarded any tuples where the labels disagreed, reducing our initial dataset size from 5000 to 4560. The refined dataset contained 502 positive (answerable) labels and 4058 negative (not-answerable) labels, providing a robust basis for training smaller models on detecting answerable captions.

4.3 Exp. 3: Sequence Classification Using Roberta-base-squad-2 and Canine-s

With a refined dataset of image-caption-answerable tuples, the next step involved fine-tuning transformer models to perform the binary classification task of determining caption relevance. Although this task does not align perfectly with traditional question-answering, its similarity to extractive QA models prompted us to utilize 'roberta-base-squad-2'. This model, a fine-tuned version of RoBERTa base on the SQuAD 2.0 dataset, was adapted for our purposes by modifying its final layer to output binary classification labels rather than spans of text.

The training process involved adjusting the model to effectively discern between 'answerable' and 'not-answerable' captions based on the features extracted from the captions and their contextual alignment with the corresponding questions. In parallel, we experimented with 'canine-s', a model known for its ability to understand and represent multiple languages. This capability suggests that CANINE can extract nuanced features from text data, which are crucial

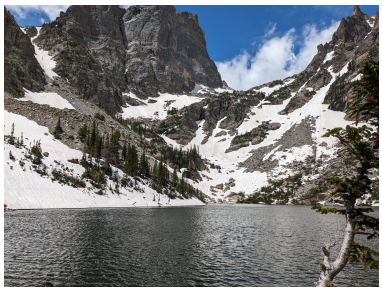
for downstream tasks such as ours. For our experiment, CANINE-s served as a feature extractor, where the embedded representations it produced were used as inputs for a standard classifier tasked with the binary classification.

The choice of these two models allowed us to explore different aspects of feature extraction and classification in the context of LLMs adapted for multimodal tasks. The results from these experiments would not only test the efficacy of the adapted models but also refine our understanding of how transformer models can be optimized for specific tasks in multimodal AI, particularly in scenarios involving complex and varied datasets.

5 Results

We provide results from each of the three experiments, showcasing the output from leveraging existing models, as well as the output from our proposed fourth vision module.

5.1 Generating Descriptive Captions Using LENS



(a) Input image

```
output['intensive_captions']

[['a beautiful view of a very big, mountainous lake landscape',
  'high altitude snow mountains in the distance behind a body of water',
  'the view of lake and mountains is photographed from an viewpoint point',
  'many snow - capped mountains towering over a body of water',
  'sheep grazing in a meadow with snow capped mountains',
  'a view of a body of water and mountains',
  'a view of a valley with large mountains with a few snow capped mountains',
  'a cloudy blue sky covers a lake near a grassy area in the mountains',
  'a mountain range with many mountains all in bright red on one side',
  'a small lake in the middle of rural land']]
```

(b) Intensive captions

```
output['caption']

['a large body of water surrounded by mountains']
```

(c) Short caption

```
output['attributes']

[['lake which has large body of water surrounded by land',
  'destination which has Natural features such as mountains, rivers, or lakes',
  'mountains which has snow-capped peaks',
  'place which has natural features like mountains, rivers, and lakes',
  'scenery which has mountains, valleys, and hills']]
```

(d) Image Attributes

```
output['tags']

[['Mountain range',
  'Glacial lake',
  'mountains',
  'Body of water',
  'Lake district']]
```

(e) Image Tags

Figure 1: LENS output

Figure 1 illustrates the results produced by LENS, which generates a set of outputs including a list of intensive captions, a single short caption, and a list of attributes and tags found in the image. This manner of output was repeated for the entire selection of 85 images from the VQA 2.0 dataset, leading to 838 descriptive captions. In this manner, we were able to create a LENS-specific dataset for our VQA task.

5.2 Labeling Image-Caption Pairs with GPT-4 and Gemini-Pro

Caption	Question	Answer	Generated Answer_gemini	Label_gemini	Generated Answer_gpt4	Label_gpt4
a single zebra standing on a dirt field	how many zebras are in the picture?	1	Answer: 1	0	One	0
a lone zebra grazes in the middle of the ground	is this animal sunbathing?	no	Cannot be inferred	1	Cannot be inferred	1
a lone zebra grazes in the middle of the ground	how many zebras are there?	1	Answer: 1	0	One	0
a lone zebra grazes in the middle of the ground	can shadows be seen on the ground?	yes	Cannot be inferred	1	Cannot be inferred	1
a lone zebra grazes in the middle of the ground	how many baby zebras?	1	Cannot be inferred	1	Cannot be inferred	1
a lone zebra grazes in the middle of the ground	what is the zebra doing in this image?	grazing	Answer: Grazing	0	The zebra is grazing in the middle of the ground.	0
a lone zebra grazes in the middle of the ground	is there weeds here?	yes	Cannot be inferred	1	Cannot be inferred	1
a lone zebra grazes in the middle of the ground	is the zebra alone?	yes	Answer: Yes	0	Yes	0
a lone zebra grazes in the middle of the ground	are these zebras facing the camera?	yes	Cannot be inferred	1	Cannot be inferred	1
a lone zebra grazes in the middle of the ground	how many zebras are in the picture?	1	Answer: 1	0	One	0

Figure 2: Answers generated using GPT-4 and Gemini

Figure 2 provides a small blurb that is representative of the output from GPT-4 and Gemini-Pro APIs. A similar type of output was generated across 5000 rows of captions. Captions for which both GPT-4 and Gemini agreed on the answer were used as positive samples and answers that could not be inferred by both were used as negative samples, while captions which the models disagreed on the answer were discarded.

5.3 Sequence Classification Using Roberta-base-squad-2 and Canine- s

Model	Epoch	Training Loss	Validation Loss	Accuracy	F1 Score
Canine-s	1	0.436600	0.324983	0.90019	0.852949
	2	0.426600	0.324622	0.900219	0.852949
	3	0.442000	0.324596	0.900219	0.852949
RoBERTa-base-squad-2	1	0.186100	0.194724	0.964912	0.963998
	2	0.001300	0.205305	0.964912	0.963177
	3	0.015300	0.146460	0.973684	0.973421

Table 1: Model comparison for Sequence Classification

Table 1 shows our results from training RoBERTa-base-squad-2 and Canine-s on the dataset developed in experiment 2. Each model was trained for three epochs on 80% of the data and validated on the remaining 20%. As observed, RoBERTa-base-squad-2 provides the best results with a high accuracy and F1 score showcasing its ability in sequence classification.

6 Conclusion

In this study, we investigated the LENS framework for visual-question answering (VQA) tasks and proposed enhancements to improve its efficiency and effectiveness. Our exploration highlighted the potential of LENS in integrating independent vision modules with large language

models (LLMs) to process visual data effectively. By leveraging three distinct vision modules, LENS demonstrates a novel approach to multimodal integration, enabling the use of pretrained LLMs for visual tasks without extensive retraining.

We proposed a fourth module, namely, a selective caption filtering mechanism within the LENS framework, to optimize context utilization and enhance information retention. This enhancement aims to dynamically evaluate and filter captions based on their relevance to the task at hand, ensuring that only pertinent information is considered. Through experiments involving sequence classification tasks, we demonstrated the efficacy of our proposed methods.

In this project, our exploration was limited to the scope of determining caption relevance. Promising future directions include building upon the current progress by leveraging the knowledge of caption relevance to enhance the capabilities of the LENS framework; specifically, integrating the selective caption filtering mechanism into the larger context of the VQA pipeline. This would entail incorporating the filtered captions into the processing pipeline of LENS to guide the model's attention towards relevant visual and textual information. To conclude, our work has shown great promise and has provided a starting point to build on top of LENS for improved information retention, optimized context utilization and increased accuracy for answering complex visual questions.

7 Weekly Paper Reading Questions

7.1 Towards Language Models That Can See: Computer Vision Through the LENS of Natural Language [https://arxiv.org/pdf/2306.16410.pdf]

MCQ What is one difference between multimodal LLMs and LENS?
(A) LENs requires additional pre-training, multimodal LLMs do not
(B) Multimodal LLMs extract task-specific visual information from an image, LENS tries to extract all visual information
(C) LENS converts textual information retrieval tasks to vision tasks
(D) LENS is highly task-specific, multimodal LLMs are more generic methods

MCQ Which is true about LENS?
(A) Performance does not depend on the number of generated tokens
(B) During inference, one vision module is chosen depending on the task
(C) Only task-specific information about the image is extracted
(D) Inference time in LENS may be longer than in multimodal LLMs depending on the input

7.2 IMAGEBIND: One Embedding Space To Bind Them All [https://arxiv.org/pdf/2305.05665.pdf]

MCQ Which emergent zero shot task is possible if pre-trained classification models are trained on (image, thermal) embeddings and (image, audio) embeddings from IMAGEBIND?
(A) Segmentation on audio using thermal prompts
(B) Classification on thermal using audio prompts

(C) Classification on image using audio prompts
(D) Object detection on thermal using audio prompts

MCQ What is one application demonstrated using the multimodal embedding space of IMAGEBIND?
(A) Video generation using music audio embeddings
(B) Analyzing the effectiveness of different image encoders
(C) Upgrading text-based object detection models to audio-based
(D) Enhancing virtual reality experiences through IMU embeddings

7.3 SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [https://ieeexplore.ieee.org/abstract/document/7803544]

MCQ What is the primary motivation behind the development of SegNet?
(A) Optimizing deep architectures for category prediction to improve on AutoNet's performance in real-time segmentation
(B) Improving the mapping of low-resolution features to input resolution for pixel-wise classification
(C) To solve self-driving
(D) Enhancing the efficiency of fully connected layers in VGG16

MCQ What aspect of SegNet's architecture makes it suitable for real-time applications like autonomous driving?
(A) Reliance on fully connected layers for feature extraction
(B) Ability to delineate objects based on their shape despite their small size
(C) Emphasis on training with large sets of weakly labeled data for improved performance
(D) Efficient encoder network design and memory-computation trade-offs

7.4 Hierarchical Text-Conditional Image Generation with CLIP Latents [https://arxiv.org/abs/2204.06125]

MCQ To improve the **(robustness | diversity)** of the **(downsamplers | upsamplers)** during training, the two methods used are **(Gaussian blur | dropout regularization)** and **(Markov blur | BSR degradation)**.

MCQ Classifier-free guidance is enabled for both the AR and diffusion **(prior | decoder)** models during **(training | inference)** by **(randomly | selectively)** dropping **(text | image)** conditioning information 10% of the time.

7.5 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [https://arxiv.org/pdf/2103.14030.pdf]

MCQ Swin Transformer addresses the challenge of computational complexity of transformers for computer vision by computing **(self-attention | cross-attention) (locally | globally)** within **(overlapping | non-overlapping)** windows.

MCQ A key advantage of the **(shifted | sliding)** window approach in Swin Transformer is that it enhances **(modeling | generalization)** power and reduces **(latency | storage size)** compared to **(sliding | shifted)** window methods.

7.6 Automatic Cardiac Cine MRI Segmentation and Heart Disease Classification [https://www.sciencedirect.com/science/article/pii/S0895611121000124]

MCQ One limitation addressed by this paper was poor tissue **(contrast|similarity)** in images that made it difficult to delineate the epicardium accurately. **(3D|4D)** signal processing-based ROI extraction reduced the **(computational load|program runtime)** and **(increased|decreased)** the class imbalance.

MCQ **(Cartesian|polar)** to **(cartesian|polar)** transformation was used to characterize the myocardium thickness and circularity variability, which helped in augmenting the **(quality|size)** of feature inputs.

7.7 Mamba: Linear-Time Sequence Modeling with Selective State Spaces [https://arxiv.org/abs/2312.00752]

MCQ Selective SSMs incorporate a selection mechanism that makes model parameters **(time-invariant|input-dependent)**, enabling the model to selectively remember or ignore inputs based on their **(relevance|position)**.

MCQ The selection mechanism facilitates adaptive processing of input sequences, resulting in improved **(model scalability|convergence)** and **(accuracy|throughput)**.

7.8 ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks [https://arxiv.org/pdf/1908.02265]

MCQ The five tasks ViLBERT was scored on were Visual Question Answering (VQA), **(Textual|Visual)** Commonsense Reasoning, Grounding Referring Expressions, 'Zero-shot' Caption-based Image Retrieval, and **(Caption|Image)**-based **(Caption|Image)** Retrieval.

MCQ In Zero-Shot Caption-based Image Retrieval, ViLBERT is used without any **(fine-tuning|pre-training)** to demonstrate its ability to **(overfit|generalize)** to visual and linguistic variation with-

out additional task-specific knowledge.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*, 2023.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[5] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[7] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.