International Journal of Computer Applications (0975 - 8887)
Volume 186 - No.61, January 2025

X-VL: Injecting External Knowledge into Vision-Language Models for Better Answering

Shyam Agarwal
Department of Computer Science
University of California, Davis
One Shields Ave, Davis, CA 95616

Amey Bharat Gohil
Department of Computer Science
University of California, Davis
One Shields Ave, Davis, CA 95616

Smit Nautambhai Modi
Department of Computer Science
University of California, Davis
One Shields Ave, Davis, CA 95616

ABSTRACT

In recent years, there has been significant growth in the vision and language community, especially with the advent of large models. Visual Question Answering (VQA) is a task in computer vision and natural language processing that is both unique as well as difficult in its framing because it demands a holistic understanding of images and language for accurate responses and requires the model to integrate multiple modalities of data. The conventional VQA approach processes the entire image to answer a posed question, often missing nuanced contextual information. This research work aims to improve VQA systems by incorporating external knowledge into the system and analyzing the performance. This study utilizes MMLFT (MultiModal Late Fusion Transformer), used with three pre-trained models for textual embeddings: BERT, RoBERTa, and ALBERT, and three pre-trained models for image encoding: ViT, DeiT, and BEiT. Experiments are conducted across various possible combinations of these text and image encoders to assess the impact of incorporating external knowledge into the system. Captions from a pre-trained image captioning model, BLIP, are utilized as a form of external knowledge to the model, and the investigation focuses on whether this addition enhances the model's evaluation metrics. Although much work has been done in improving VQA models by adding external knowledge to them, this study is believed to be the first to approach the topic from a data-specific point of view, closely analyzing the data entries and attempting to justify why the results improve or not. A simple but novel way to cheaply generate inferences about an image is also presented, showcasing its potential to assist with future VQA tasks. The conclusion drawn is that adding external information contributes to better results, but the mode of knowledge addition needs to be well-constructed.

General Terms

Machine Learning, Deep Learning, Aritficial Intelligence

Keywords

Visual Question Answering (VQA), External knowledge injection, Multimodal Fusion models, Common Sense Transformer

1. INTRODUCTION

There has been a huge growth in the vision and language community recently, especially with the advent of large models. The task of Visual Question Answering (VQA) [1] is a task in computer vision and natural language processing that requires the model to carefully understand both the image as well as the question and how they both interact. It is important to note that building such vision-language models goes beyond specialized algorithms like object detection or image segmentation and it also requires the model to integrate multiple modalities of data.

This research work aims to improve VQA systems by trying to incorporate external knowledge and use it to better solve the problem at hand. The general idea is to learn some context about the image and use this context with some knowledge from outside sources to better inform the Vision-Language model about what to do. This is also useful for building common sense reasoning within these systems - a cognitive task effortless for humans but challenging for machines. This would help in not only improving the current baseline from popular benchmark models but also help in finding ways that are different from merely increasing the number of parameters in the existing model. This would be crucial when considering computation time, storage costs, and the harms to the environment that the large number of parameters create. In fact, many applied research works have also shown consideration for saving carbon emissions in their methodologies recently. By means of utilizing external knowledge, i.e., knowledge not coming from the models, we can hope to better solve the problems by not increasing the number of parameters in the models and saving money as well as time - obviously, given that knowledge injection is less computationally expensive but at least we can claim that this would mean that instead of training the model multiple times on any new information that is available, we can probably find an alternative method to leverage it (via injecting external knowledge).

Apart from these reasons, in recent times, both Natural Language Processing and Computer Vision have shown an increase in demand, especially because of the successful applications that they have found in day-to-day human lives. Moreover, the amount of textual and visual information available online has become a lot more accessible. These factors further motivate us to work on the problem - precisely, the accessibility of resources as well as the expansiveness of possibilities in the same.

There are a bunch of possible applications for VQA systems, ranging from the healthcare industry assisting doctors by generating detailed responses focused on specific regions or anomalies within medical images and enhancing diagnostic processes, to increasing accessibility by assisting people with visual impairment by providing them answers to the questions that they ask about the visual object (or real-time video). There are various other commercial applications like autonomous vehicles, social media content moderation, robots helping in households, virtual assistants, and so on.

Historically speaking, Visual Question Answering arose from image captioning [1] which was a task aimed at examining the capabilities of machines to understand images by combining computer vision and natural language processing. However, VQA tasks are significantly more complex than basic image captioning, as more often than not, it requires some understanding or reasoning that is not present in the image otherwise. This reasoning is easy for humans to understand, but significantly difficult for machines to build. In this context, VQA tasks are truly AI-complete tasks - tasks that might be called "real intelligence" by some because they deal with multimodal data that goes beyond a single domain or field. This is one of our primary reasons for working on this problem.

Prior works have approached this problem with an emphasis on different ways to add the information, but have not provided a good justification for why their method fails in particular instances. However, we take a data-first approach, trying to explain the reasoning of why the model might interpret or misinterpret the scenario, given the original information or the added information, which is in the form of image captions [10] that we provide to it. Moreover, we present a novel method to generate inferences for an image-text pair which is inspired by work done by [14] but suggests an alternative and simpler route that is cheap and easy to implement.

The task of VQA is in itself quite challenging. However, access to limited GPU and other hardware resources made it even more challenging for us to perform experiments and compare them to standard baselines. Thus, we generate our own baseline by running all the experiments with only 5 epochs so we can compare them across different evaluation scenarios. We hope that this work will motivate future research in this direction and benefit the vision-language community.

2. RELATED WORK

Vogel and Schiele [16] used local image regions and combined them into a global representation to generate semantic classes. Hodosh et al. [7] proposed a framework to generate sentence-based image captions and rank them. Similarly, [6], [9], studied the problem as a task to retrieve information by associating the image and its description in equivalent latent space. Malinowski et al. [12] introduced a way that mixes the Bayesian method with segmentation and semantic parsing to create training set samples using the nearest neighbor search. Geman et al. [5] proposed a method of generating binary questions automatically and using them.

All the above methods are limited by specific query types. However, with the advent of deep learning, architectures using CNN and LSTM, either directly or indirectly, have been popular. H. Gao et al. [4] and M. Malinowski et al. [13] encoded question sentence using CNN and LSTM, Malinowski et al. [13] combined CNN and LSTM into an end-to-end learning, whereas, Gao et al. extracted question representation using one LSTM and stored the linguistic context using another.

In multiple papers [11] [3], they suggest utilizing knowledge graphs to create a context for the image and pass it as a fact to the model while predicting information. However, this method is not very scalable as knowledge graphs are hard to generate and store. In [8], they suggest utilizing COMET to find inferences for an image and use them along with the question for better efficiency. However, the way they suggest choosing the best inference is quite complex. In this study, we observed that we could make the process simpler by utilizing just a simple BERT classifier to decide the type of information needed which is not something that has been done before to the best of our knowledge but more about this is discussed later in Section IV.

Fusion Model Number	Image Encoder	Text Encoder
0	albert-base-v2	facebook/deit-base-distilled-patch16-224
1	albert-base-v2	google/vit-base- patch16-224in21k
2	albert-base-v2	microsoft/beit-base-patch16-224-pt22k-ft22k
3	bert-base-uncased	facebook/deit-base-distilled-patch16-224
4	bert-base-uncased	google/vit-base- patch16-224-in21k
5	bert-base-uncased	microsoft/beit-base-patch16-224-pt22k-ft22k
6	roberta-base	facebook/deit-base-distilled-patch16-224
7	roberta-base	google/vit-base- patch16-224-in21k
8	roberta-base	microsoft/beit-base-patch16-224-pt22k-ft22k

Fig. 1: Fusion Model Mappings

3. DATASET

We used a smaller version of DAQUAR dataset, proposed in [12] where we use 3825 training examples and 834 evaluation examples for our example. The dataset was created to have images that have high demands on "understanding" the visual input and those that check for a whole chain of perception, language understanding, and deduction. The dataset has one unique correct answer for each of the image-question pairs as well. The choice of this dataset was guided by the fact that the images are of reasonable quality - not too bad to hinder the progress and not too high quality to limit our ability to use multiple of them. Since we created our own baselines by running the model on a fixed number of epochs for both cases - with and without a caption, we did not have to worry much about comparing the results with different papers.

4. METHODOLOGY

In this section, we discuss our approach which is rooted in three major steps - building our baseline model, adding captions, and evaluating performances, and finally, our novel approach to generate inferences which is inspired by [8].

4.1 Baseline model

As mentioned beforehand, we use multimodal fusion models which are highly useful in capturing information from different modalities (image and text here) but also preserving any cross-modal interactions. The process can be summarized in three basic steps:

4.1.1 Image and Text Featurization. Image and text featurization involves converting raw input data—such as images and text—into structured and meaningful representations that can be effectively processed by machine learning models. For images, this process involves extracting visual features that capture important patterns, textures, and structures. There are various methods for generating image embeddings, ranging from traditional techniques like SIFT descriptors and Histogram of Oriented Gradients (HOG) to modern approaches using Convolutional Neural Networks (CNNs). In this study, pre-trained transformer-based image models, namely ViT

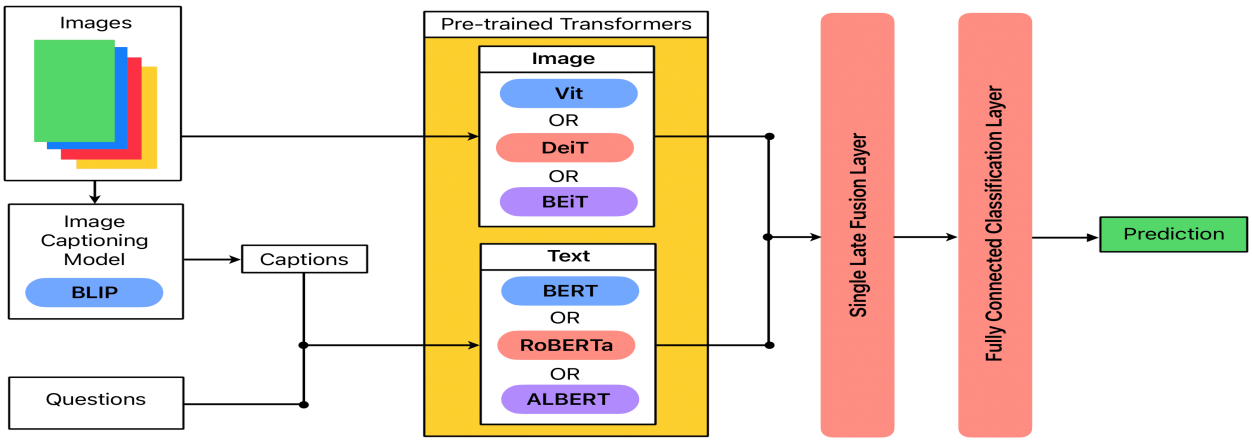


Fig. 2: MMLFT Architecture, with the injection of captions as additional knowledge

(Vision Transformer), DeiT (Data-efficient Image Transformer), and BEiT (Bidirectional Encoder representation from Image Transformers), are utilized for image encoding. These models have been trained on large-scale datasets and are designed to capture global and local contextual features, leveraging their self-attention mechanisms to produce robust and expressive embeddings.

For text, featurization involves representing textual data in numerical form while retaining the semantic and syntactic meaning of the input. Traditional approaches, such as word embeddings (Word2Vec, GloVe) or sequential models like Long Short-Term Memory Networks (LSTMs), have been widely used. However, for this study, transformer-based pre-trained language models, namely BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (A Robustly Optimized BERT Pretraining Approach), and ALBERT (A Lite BERT), are employed. These models are pre-trained on extensive corpora and can leverage external knowledge to produce rich and contextual embeddings for textual data.

The use of these pre-trained models for both image and text processing is motivated by several factors. First, these models have been trained on diverse and extensive datasets, enabling them to incorporate external knowledge beyond the dataset used in this study. This external knowledge provides a broader understanding of real-world concepts and relationships, which is beneficial for tasks like Visual Question Answering (VQA) that require reasoning across modalities. Second, using pre-trained models reduces the computational overhead associated with training large-scale models from scratch, making the process more efficient and resource-friendly.

The adoption of a late fusion strategy further supports the decision to utilize pre-trained embeddings. In late fusion, features from different modalities are processed independently and combined at a later stage, allowing for the integration of highly specialized embeddings. Pre-trained models are particularly well-suited for this strategy, as they provide rich and task-agnostic feature representations that can be effectively combined to achieve superior performance. This approach ensures that the featurization process captures the essential attributes of both images and text, enabling the model to perform well in downstream tasks.

We create 9 baseline models - each combination of image and text encoder to ensure that our results are generalized. Fig. 1 defines each possible combination and gives a unique name that we will use to refer to that combination in the paper moving forward.

4.1.2 Fusion of Features. The fusion of features is a critical step in Visual Question Answering (VQA) systems, as the task inherently requires a model to integrate information from two distinct modalities: textual data from the question and visual data from the image. The output of this fusion process directly influences the model's ability to understand the semantic relationship between the question and the image content, which is vital for generating accurate and contextually appropriate answers.

In this study, the fusion process involves combining the two feature vectors generated during the image and text featurization stages into a single, joint representation. Each feature vector represents rich embeddings derived from the respective modality. The image embeddings encapsulate the visual semantics and structural details of the image, while the text embeddings capture the linguistic and semantic nuances of the question. Effectively merging these modalities ensures that the model can reason across them to infer the correct answer.

The fusion process begins by concatenating the two feature vectors along their respective dimensions. Concatenation is a reasonably straightforward yet effective approach to preserve all information from both modalities, enabling the model to leverage the richness of the pre-trained embeddings without discarding any critical details. This concatenated vector serves as the initial step in constructing a unified representation.

To refine this representation and facilitate meaningful interactions between the modalities, the concatenated vector is passed through a linear layer. The linear layer applies a learned weight transformation to the joint representation, projecting it into a new embedding space where the relationship between the text and image features can be more effectively modeled. This transformation allows the model to identify and emphasize relevant cross-modal patterns while filtering out less pertinent information. Additionally, the linear layer reduces dimensionality, optimizing the joint representation for subsequent processing and improving computational efficiency.

4.1.3 Answer Generation. Since the dataset consists exclusively of single-word answers or short phrases, the task does not require generating free-form responses. Instead, the problem can be effectively formulated as a classification task. In this setup, the model predicts the correct answer from a fixed set of possible outputs, corresponding to the vocabulary space of the dataset.
To achieve this, a fully connected layer is used as the final layer of the model. The size of this output layer is set to match the size of the vocabulary, where each node represents a potential answer. The model processes the joint representation generated during the feature fusion stage and outputs a probability distribution over the vocabulary space. This distribution indicates the likelihood of each possible answer.
The cross-entropy loss function is employed as the objective function to train the model. Cross-entropy loss is well-suited for classification tasks as it quantifies the difference between the predicted probability distribution and the true target label. By minimizing this loss, the model learns to accurately associate the input features with the correct answers, ensuring effective performance on the Visual Question Answering task. The loss equation is given by:

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{1}$$

- M represents the number of classes or categories.
- $y_{o,c}$ denotes the ground truth label or target probability associated with the correct class c for a specific observation o.
- $p_{o,c}$ signifies the predicted probability of class c for the observation o as inferred by the model.
This equation calculates the cross-entropy loss between the predicted probabilities $p_{o,c}$ and the ground truth labels $y_{o,c}$ for a given observation in a classification task.

4.2 Adding captions via BLIP Image Captioning

To enhance the Visual Question Answering (VQA) system, captions are generated for each image using the pre-trained Bottom-Up and Top-Down with Lightweight Image Processing (BLIP) model [10]. Additionally, Optical Character Recognition (OCR) is applied to extract textual information embedded in the images, such as labels, signs, or numbers. The nine combinations of text and image models described earlier are then re-evaluated with these captions incorporated as additional input. The architecture employed for this approach is depicted in Fig. 2. The process of adding captions is summarized as follows:

4.2.1 Overview of BLIP. The BLIP model generates detailed image captions by employing a combination of bottom-up and top-down attention mechanisms. The bottom-up attention focuses on specific regions of interest within the image, extracting localized features that capture fine-grained details. Concurrently, the top-down attention synthesizes these features into coherent captions by leveraging global contextual information, ensuring that the description is both accurate and meaningful. BLIP is also capable of performing OCR to extract any text present within the image, providing an additional layer of information that is particularly useful for VQA tasks where textual content plays a role.

4.2.2 Concatenating captions and questions using structured formatting. To incorporate the generated captions into the VQA pipeline, the caption is concatenated with the corresponding question. The caption is placed before the question, separated by a single space, to maintain a structured input format. This modification allows the model to access a descriptive summary of the image alongside the question, enhancing its ability to reason about the content. By integrating the caption, the model gains additional contextual knowledge that can clarify ambiguities in the question and provide relevant details about the image, leading to improved performance on the VQA task.

4.3 Novel approach of inference generation

Apart from generating captions and adding that to our images, we also present a simple, cheap, and novel technique that is inspired from [14]. The architecture can be found in Fig. 3. We use a 4 step process that is summarized in the five bullets below:

4.3.1 Question to Declaration Conversion. The first step in the process involves transforming the given question into a declarative form, which provides a more structured and coherent input for subsequent reasoning tasks. For example, as illustrated in the figure, the question *"What is a likely liquid to find in these glasses?"* is reformulated as the declarative statement *"A likely liquid to find in these glasses is likely a"*. The use of declarative statements will be justified later in the discussion on COMET.
To achieve this transformation, GPT-3.5 is utilized, along with carefully designed prompt tuning. The prompts are crafted to explicitly instruct the model to generate declarative statements while retaining the semantic meaning of the original question. After conducting multiple trials with various prompt designs, it was observed that the most effective results were obtained using simple prompts that directly instructed the model to perform the conversion without additional complexity or instructions. This straightforward approach yielded high-quality declarative statements consistently.

4.3.2 Object Detection Using YOLO v7. After converting the question into a declarative statement, the next step involves extracting objects from the image using the You Only Look Once (YOLO v7) object detection model [17]. YOLO v7 is a highly efficient and accurate object detection framework that identifies and localizes objects within an image in real time. The detected objects are then concatenated with the declarative statement generated earlier, enriching the input with relevant visual information from the image.
The intuition behind this approach is that generating commonsense inferences about the image requires not only the question's textual context but also specific information about the image content. As can be seen in the example, the object that we get is "wine glass" - which is obviously quite helpful in helping the model understand that the liquid is more likely to be wine (since it is in a wine glass). By integrating object detection outputs with the declarative statement, the system benefits from an additional layer of contextual knowledge. The detected objects act as visual anchors, helping the model make more accurate inferences. This integration ensures that the reasoning process incorporates both textual and visual cues, leading to more informed and precise outputs.

4.3.3 Overview of Common Sense Transformer. COMET [2], short for the Common Sense Transformer, is a sophisticated language model specifically designed to integrate commonsense knowledge into tasks requiring natural language understanding and generation. By leveraging external knowledge bases such as ConceptNet [15] and ATOMIC [8], COMET enhances an AI system's capacity to reason, infer, and generate responses that are contextually relevant and enriched with implicit commonsense reasoning. This makes it particularly effective for applications such as question answering, dialogue systems, and other tasks where deeper contextual understanding is required.

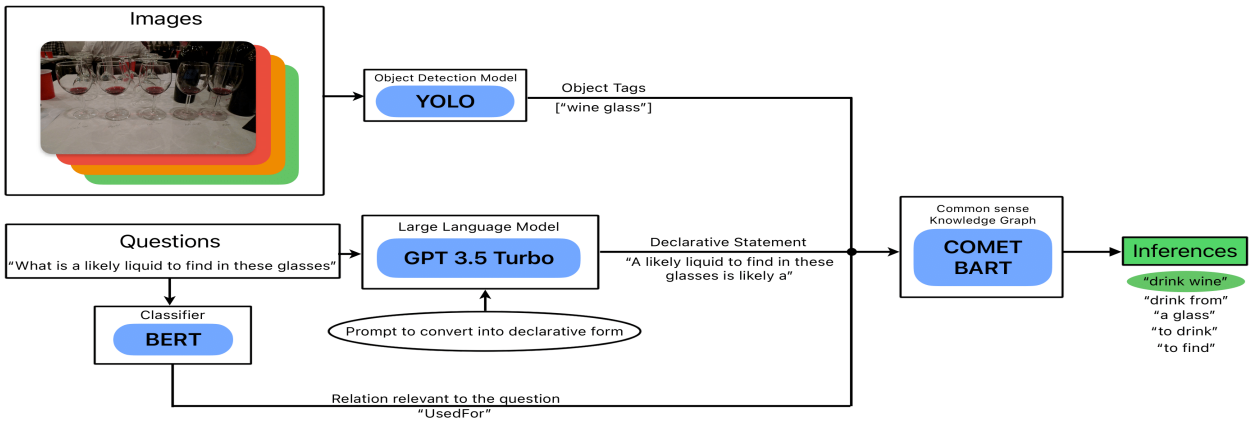


Fig. 3: Architecture for Inference Generation

The architecture of COMET allows it to go beyond the explicit textual information present in the input by drawing on the rich, structured knowledge stored in these external databases. For instance, ConceptNet provides a graph-based representation of commonsense relationships between concepts, while ATOMIC [8] includes more specific event-based knowledge about causes, effects, and preconditions. COMET utilizes this information to bridge the gap between what is explicitly stated and the implicit commonsense knowledge necessary for human-like reasoning.
Since COMET is trained on declarative sentences, we converted our questions to declaratives above.

4.3.4 BERT for Question Classification. A significant deviation from the methodology in [14] lies in the handling of inferences. Unlike [14], which involves searching for and ranking potential inferences, this approach avoids such a strategy due to its lack of scalability as the size of the knowledge graph grows. Instead, the proposed method employs BERT to classify the question into one of the predefined relation categories outlined in ATOMIC [8]. These relations, such as UsedFor, Causes, or Effects, are limited in number compared to the vast set of all possible inferences, making the classification process computationally efficient.
The process involves feeding the declarative form of the question into a fine-tuned BERT model trained on ATOMIC relation categories. The model predicts the most relevant relation category for the input, such as UsedFor. By focusing on this specific relation, the scope of inferences is significantly reduced, allowing the system to retrieve and evaluate only the relevant inferences associated with the predicted relation. Among these, the most suitable inference is selected based on predefined criteria. This approach ensures efficiency and scalability while maintaining high-quality results.

4.3.5 Utilizing COMET to get relevant inferences. Having acquired the declarative form of the question from GPT-3.5, the relevant concept relation/map from BERT ("UsedFor" in this case), and the object tags from YOLOv7, we combine these elements to craft a statement that approximates natural language, albeit not perfect. A dictionary is introduced to map relation categories to corresponding natural language phrases, enabling the construction of a statement that approximates natural language. By employing this dictionary, along with the aforementioned components, we generate a statement with sufficient natural language nuances. This statement is then fed into the COMET model, resulting in five inferences. From these, the best inference is selected based on its contextual alignment with the input. In the example case, the selected inference is *"drink wine"*, which aligns well with the context of the question and the visual content. These inferences are another instance of external knowledge and should serve as valuable inputs for enhancing the Visual Question Answering (VQA) task.

text_model	image_model	num_parameters	eval_accuracy & No_caption	eval_accuracy & Yes_caption	eval_f1 score & No_caption	eval_f1 score & Yes_caption
albert-base-v2	facebook/deit-base-distilled-patch16-224	98531078	0.22012831	0.22253408	0.02003806	0.02080793
albert-base-v2	google/vit-base-patch16-224-in21k	98531078	0.13993585	0.20449078	0.00776740	0.01390788
albert-base-v2	microsoft/beit-base-patch16-224-pt22k-ft22k	98531078	0.14314354	0.14755413	0.01153312	0.01233645
bert-base-uncased	facebook/deit-base-distilled-patch16-224	196958534	0.20609463	0.22614274	0.01750686	0.01946350
bert-base-uncased	google/vit-base-patch16-224-in21k	196958534	0.21090617	0.21732157	0.01515973	0.01673650
bert-base-uncased	microsoft/beit-base-patch16-224-pt22k-ft22k	196958534	0.21331195	0.21732157	0.02041540	0.02086263
roberta-base	facebook/deit-base-distilled-patch16-224	212120390	0.24178027	0.22534082	0.02412841	0.02051992
roberta-base	google/vit-base-patch16-224-in21k	212120390	0.22534082	0.22654370	0.01701874	0.01914525
roberta-base	microsoft/beit-base-patch16-224-pt22k-ft22k	212120390	0.24659182	0.22774659	0.02488570	0.02227226

Fig. 4: Accuracy and F1 Score values for different combinations of fusion models

5. RESULTS

We provide the results of all the experiments that we discussed throughout the paper here:

(1) First of all, we observed that using captions reduced the width of our box plot for accuracy, recall, and f1 score, and also shifted it slightly above the older results across all 9 fusion models as can be seen in Fig. 7 and Fig. 8. This suggests that adding captions ensures that evaluations are closer to average, i.e., not very spread apart, and also improves the evaluation metrics in most cases. The precision was seen to be more spread though which we can attribute to the specific characteristics of our data - at least we cannot see a link between adding captions and more spread precision which might be a good problem to consider.

(2) We also plot bar plots in Fig. 9 and Fig. 10 for each of the 4 evaluation metrics, namely accuracy, recall, precision, and f1 score so that the change in the metric can be easily observed. The exact values for accuracy and f1 score are in Fig. 4. As shared earlier in the paper, the mapping from plot labels to the combination of text and image encoders can be found in Fig. 1.

(3) We already described the results of our inference generation technique while explaining it in the methodology section.

(4) We also provide samples of some random images that did not work well without a caption, but do work well when used with a caption which can be found in Table 1 with the corresponding questions in Fig. 6. Let's consider one such example as mentioned below where the model was asked, *"what is beneath the picture?"* and it initially predicted the answer to be *"table"*, but after providing the caption (*"there is a living room with a couch, coffee table, and a clock what is beneath the picture"*) to the model, its prediction was accurately *"sofa"*.

Fig. 5: Random Image Sampled form our Dataset

(5) We noted that the addition of captions resulted in minimal improvement. In some cases, there was a slight increase in accuracy, but, for the most part, it was negligible. This can be attributed to the constraint of running only 5 epochs due to time and resource limitations.

(6) Furthermore, we found that precision does not exhibit significant improvement in most cases. It appears that when the question directly pertains to the image content, the similarity between the image caption and the image itself becomes so pronounced that the addition of caption as external knowledge becomes somewhat redundant. In some cases, the captions were totally indifferent to what the question was - the question asked about the color but the image description did not refer to it at all.

6. CONCLUSION

In summary, our late fusion model showed some improvement by adding image captions, though it's crucial to keep expectations realistic. The changes in accuracy, recall, and f1 score were noticeable but not game-changing. We should interpret these results with caution, considering the brief 5-epoch evaluation due to time and resource constraints. The reduction in variability across fusion models suggests a positive impact from image captions, but it doesn't signal a major performance boost. It's essential to acknowledge these incremental gains, understanding that more substantial improvements may emerge with longer training periods and increased resources in future studies.
Also, the fact that the descriptions did not correspond well to the question that was being asked suggests that the knowledge injected needs to be well structured. Our inference generation technique offers potential for the future.

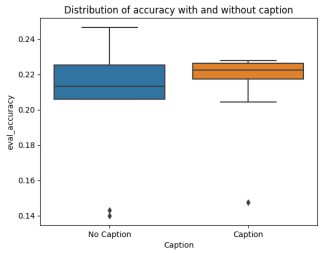
Table 1. : Samples that showed improvement after adding captions

Q. what is the object on the stool	Q. what is the blue object that is in front of the counter and to the right of the photocopying machine?	Q. What is the object in front of the window?
Q. what is on the left side of the table?	Q. what is in the bottom left corner?	Q. what is found on the right side?
Q. what is the largest object?	Q. what is the object on the floor in front of the wall	Q. what is to the left of the lamp?

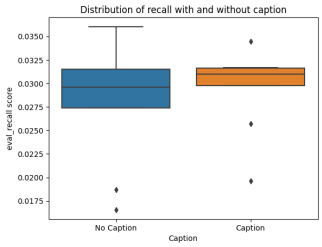
Fig. 6: Questions

7. DISCUSSION OF FUTURE WORK AND LIMITATIONS/CHALLENGES

Looking ahead, several areas for improvement and expansion have been identified to enhance the performance and applicability of the proposed model.
One key direction involves refining how insights are gathered from COMET, the commonsense knowledge source utilized in this work. While the current approach leverages COMET to generate inferences, there is potential to explore more sophisticated methods of integrating these inferences as external knowledge. Specifically, incorporating the generated inferences as an additional input stream to the model could enrich its understanding and reasoning capabilities. This would allow the system to draw on external knowledge in a more structured and impactful manner, improving its performance on Visual Question Answering (VQA) tasks.
Another significant avenue for future work involves extending the training process by providing the model with additional computa-

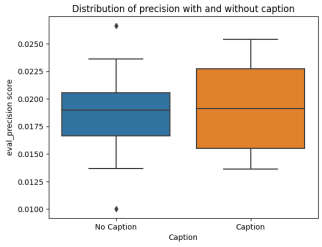


(a) Box Plot for Accuracy

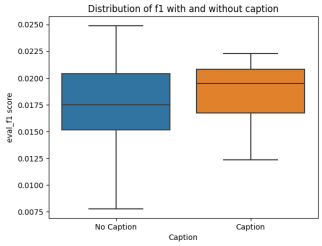


(b) Box Plot for Recall

Fig. 7: Box Plots for Accuracy and Recall



(a) Box Plot for Precision

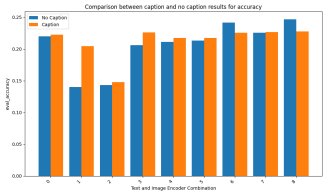


(b) Box Plot for F1

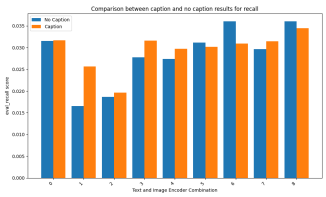
Fig. 8: Box Plots for Precision and F1

tional time and resources. The current implementation operates under certain constraints that limit the duration and intensity of training. By relaxing these constraints, it would be possible to allow the model to explore its parameter space more thoroughly, leading to improved generalization and robustness. Such adjustments could result in better performance, especially on complex VQA tasks requiring deep reasoning.

The task of classifying relation categories using a BERT classifier also presents notable challenges. The limited availability of labeled data for training the BERT model on ATOMIC relation categories restricts its effectiveness. The scarcity of resources for this specific classification task highlights a gap in the current research land-

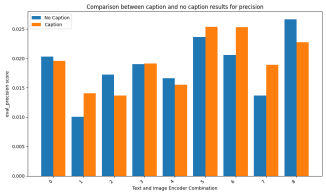


(a) Bar Plot for Accuracy

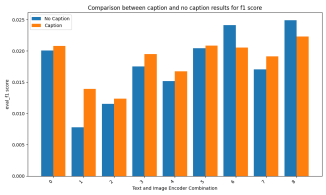


(b) Bar Plot for Recall

Fig. 9: Bar Plots for Accuracy and Recall



(a) Bar Plot for Precision



(b) Bar Plot for F1

Fig. 10: Bar Plots for Precision and F1

scape. Despite these challenges, this work aims to pave the way for further exploration in this direction. Addressing these limitations could involve developing more extensive and diverse datasets or employing semi-supervised or unsupervised learning techniques to enhance the classifier's performance.

In summary, future efforts will focus on improving the integration of COMET-generated inferences, extending training capabilities, and addressing challenges in relation classification. These enhancements hold the potential to make the model more accurate, robust, and scalable, thereby contributing to advancements in the field of Visual Question Answering and related domains.

8. REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Com-

monsense transformers for automatic knowledge graph construction. *arXiv*, 1906.05317, 2019. Available: https://arxiv.org/abs/1906.05317.

[3] Zhexue Chen et al. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. *arXiv*, 2022. Available: https://arxiv.org/abs/2207.12888.

[4] Hongyang Gao, Jingjing Mao, Jian Zhou, Zhicheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2296–2304, 2015.

[5] Donald Geman, Stuart Geman, Neil Hallonquist, and Larry Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, pages 3618–3623, 2015.

[6] Yunchao Gong, Linjie Wang, Margaret Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision (ECCV)*, pages 529–545, 2014.

[7] Margaret Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[8] Jena D. Hwang et al. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. 2020.

[9] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Learning cross-modality similarity for multinomial data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2407–2414, 2011.

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv*, 2201.12086, 2022. Available: https://arxiv.org/abs/2201.12086.

[11] Haitian Lu. Open-ended generative commonsense question answering with knowledge graph-enhanced language models. *Semantic Scholar*, 2021. Available: https://www.semanticscholar.org/paper/Open-Ended-Generative-Commonsense-Question-with-Lu/a201c722c7de07b7354dda9cdabf9baf7e6e2ec0.

[12] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1682–1690, 2014.

[13] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015.

[14] Siddharth Ravi, Abhishek Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. Vlc-bert: Visual question answering with contextualized commonsense knowledge. *arXiv*, 2022. Available: https://arxiv.org/abs/2210.13626.

[15] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv*, 2018. Available: https://arxiv.org/abs/1612.03975.

[16] Joachim Vogel and Bernt Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*, 72(2):133–157, 2007.

[17] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv*, 2207.02696, 2022. Available: https://arxiv.org/abs/2207.02696.