Visual Question Answering with Contextualized Commonsense Knowledge

by

Aditya Aravind Chinchure

B.Sc. (Honours) in Computer Science, University of British Columbia, 2021

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2024

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Visual Question Answering with Contextualized Commonsense Knowledge

submitted by **Aditya Aravind Chinchure** in partial fulfillment of the requirements for the degree of **Master of Science** in **Computer Science**.

Examining Committee:

Leonid Sigal, Professor, Department of Computer Science, UBC
Supervisor

Renjie Liao, Assistant Professor, Department of Electrical and Computer Engineering, UBC
Co-Supervisor

Vered Shwartz, Assistant Professor, Department of Computer Science, UBC
Supervisory Committee Member

Abstract

There has been a growing interest in solving Visual Question Answering (VQA) tasks that require the model to reason beyond the content present in the image. In this work, we focus on questions that require commonsense reasoning. In contrast to previous methods which inject knowledge from static knowledge bases, we investigate the incorporation of contextualized knowledge using Commonsense Transformer (COMET), an existing knowledge model trained on human-curated knowledge bases. We propose a method to generate, select, and encode external commonsense knowledge alongside visual and textual cues in a new pre-trained Vision-Language-Commonsense transformer model, VLC-BERT. Through our evaluation on the knowledge-intensive OK-VQA and A-OKVQA datasets, we show that VLC-BERT is capable of outperforming existing models that utilize static knowledge bases. Furthermore, through a detailed analysis, we explain which questions benefit, and which don't, from contextualized commonsense knowledge from COMET.

Lay Summary

Visual Question Answering (VQA) is the task of answering a question given an image. In our work, we focus on the challenging problem of commonsense knowledge based VQA, where external knowledge about the world is necessary for a model to answer the question. We propose VLC-BERT, a model that can incorporate external commonsense knowledge by using a knowledge generation language model, COMET, to obtain knowledge in a contextual manner, specific to the question and the image. This method yields a model that outperforms models of its size on two datasets, OK-VQA and A-OKVQA. Our investigation reveals the possibility of building smaller language models while incorporating commonsense knowledge into them.

Preface

This thesis is an original work of the author, Aditya Aravind Chinchure, under the supervision of Professor Leonid Sigal in UBC's Department of Computer Science and Professor Renjie Liao in UBC's Department of Electrical and Computer Engineering. The outcomes of this work are accepted to the IEEE/CVF Winter Conference on Applications of Computer Vision 2023 (WACV 2023).

The research work presented in this thesis was done in collaboration with Sahithya Ravi, and her advisor Dr. Vered Shwartz, also at UBC's Department of Computer Science. Sahithya primarily focused on implementing the knowledge generation and selection steps and the error analysis, whereas I focused on developing the core model and running subsequent experiments and ablation studies. We contributed equally to doing the qualitative analysis and writing the conference paper, and share first-authorship on the publication.

The publication associated with the work done in this thesis is:

1. **VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge**
   Sahithya Ravi, Aditya Chinchure*, Leonid Sigal, Renjie Liao, Vered Shwartz (*equal first authors)*
   Accepted at IEEE/CVF Winter Conference on Applications of Computer Vision 2023 (WACV 2023)

Table of Contents

Appendix

List of Tables

List of Figures

Acknowledgments

I am incredibly grateful to my supervisors, Dr. Leonid Sigal and Dr. Renjie Liao for their continued guidance and encouragement throughout my Masters. Their expertise in the field and their passion for research has allowed me to become a better researcher. I am thankful to Dr. Vered Shwartz for her mentorship during my project, and welcoming me to be a part of her group. I also appreciate Dr. Giuseppe Carenini for his valuable feedback and discussions during the early stages of this project, Dr. Fredrick Tung for his mentorship during my internship at Borealis AI, and Dr. Matthew Turk and Dr. Kartik Hosanagar for their guidance on my ongoing research projects.

My Masters would not have been as enjoyable or fruitful without the support of my peers. I would especially like to thank Sahithya Ravi, Felipe González-Pizarro, Gaurav Bhatt, and Pushkar Shukla, who have not only served as invaluable mentors, but also great friends who I share my achievements with. I am also appreciative of the support of my friends and colleagues from the Vision, DSL and NLP groups.

I am indebted to my family and friends for their continuous encouragement and belief in my abilities. A heartfelt thank you to my mom and dad for their selfless motivation to pursue my passions. My life in Vancouver would be incomplete without my friends, who have continually inspired me to keep learning.

Finally, my work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chair, NSERC CRC, NSERC DG and Accelerator Grants, and a research gift from AI2. Hardware resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute[1]. Additional hardware support was provided by John R. Evans Leaders Fund CFI grant and Compute Canada under the Resource Allocation Competition award.

[1] www.vectorinstitute.ai/partners

Chapter 1

Introduction

Recent progress in multimodal vision-language learning has been fueled by large-scale annotated datasets for Visual Question Answering (VQA) [1, 6, 12, 38, 50], in which models are presented with questions about an image. To answer questions correctly, models are required to perform scene understanding and learn meaningful connections between the two modalities. In recent years, transformer-based vision and language (VL) models [8, 21, 45], pre-trained on large-scale multimodal corpora, have reached impressive accuracies on standard VQA datasets.

VQA often necessitates not only visual comprehension of the scene depicted by the image (e.g., "A plate with meat, potatoes and bread") but also making inferences about plausible stories behind the image (e.g., "The plate is likely found at a restaurant"). Humans make such inferences based on prior experience and commonsense knowledge (e.g., "This is likely a lunch or dinner at a restaurant, people may be enjoying themselves..."). Most existing methods rely on world knowledge implicitly encoded by language models, which often lacks in both accuracy and coverage [33]. This is primarily due to the fact that commonsense knowledge is extremely broad, and frequently assumed. Commonsense knowledge learned from text suffers from reporting bias [11]: over-representation of exceptional facts (e.g., "people die in accidents") in text corpora, at the expense of rarely discussed trivial facts known to everyone (e.g., "people eat").

Several visual question answering benchmarks were proposed, in which the questions require either factual [28, 46] or commonsense knowledge [37, 50] beyond the visual scene comprehension. This prompted the development of neurosymbolic methods combining transformer-based representations with knowledge bases (KBs) [9, 29, 48]. However, retrieving relevant facts directly from a KB is challenging due to lack of coverage, and because KB facts are only appropriate in certain contexts.

In this work, we propose VLC-BERT (Vision-Language-Commonsense BERT), a model designed to incorporate contextualized commonsense knowledge into a Vision-Language transformer built on VL-BERT [42]. As an alternative to the retrieval paradigm often used in knowledge-based VQA,

Figure 1.1: A question from the OK-VQA [28] dataset: *Where might one buy this?* In order to answer this question, a model requires commonsense knowledge about the contents of the image (the food on the plate), and where such food may be obtained (restaurant), that humans often infer from past experiences and their world knowledge.

our model generates contextualized commonsense inferences on the question phrase combined with image object tags using COMET [2, 15], a language model trained on commonsense knowledge graphs. We augment sentence transformers [32] to rank, filter and embed the commonsense inferences. We incorporate the filtered inferences into VLC-BERT using an attention-driven fusion mechanism that learns to focus on the most important inferences for each question. Commonsense knowledge may not be necessary for answering every question, as some questions are either purely visual, factual, or straight-forward. To eliminate injecting noisy knowledge in such cases, we employ weak supervision to help us discriminate between situations when commonsense knowledge may or may not be valuable.

Our evaluations on the challenging OK-VQA [28] and A-OKVQA [37] datasets confirm that leveraging commonsense is consistently useful for knowledge intensive visual question answering tasks. We analyze the successful predictions and show how the commonsense inferences help answering difficult questions. Ultimately, VLC-BERT performs favourably compared to other similarly sized models on both datasets.

Chapter 2

Related Work

2.1 Vision-Language Transformer Models

Pre-trained Vision-Language models based on BERT [8] have shown impressive performances on downstream multimodal tasks such as Visual Question Answering. ViLBERT [25] and LXMERT [43] use a two-stream architecture to first encode language and vision modalities independently, and then apply a cross-modality encoder to align textual and visual tokens. VL-BERT [42], OSCAR [22] and OSCAR+ [51] use a single-stream architecture to directly learn inter-modality interactions. Large-scale pre-training is commonly done using the Conceptual Captions [39] dataset, with objectives that are designed to encourage interaction between modalities, such as predicting masked tokens or image regions [22, 25, 42, 43], and using contrastive loss between modalities [22]. As a result, such models inherently capture some commonsense knowledge through their pre-training regime. While these models perform impressively on downstream tasks such as VQA [1], they typically perform worse on questions requiring reasoning about knowledge beyond the image content or involving multiple reasoning hops.

In more recent years, the emergent capabilities of large-scale vision-language models (VLLMs) such as OpenAI's GPT-4V [30] have shown the potential of implicit commonsense knowledge obtained through an extensive training regime. However, these models are often monetarily expensive to train and use.

In our work, we introduce VLC-BERT, a multimodal transformer model based on VL-BERT that explicitly incorporates external knowledge to alleviate the knowledge gap in pre-trained VL-models, while being significantly smaller and less expensive to train and use compared to recent VLLMs.

2.2 Knowledge-based Visual Question Answering

In recent years, several VQA datasets were designed specifically to require reasoning about external knowledge beyond the image, whether using fac-

tual and web information (FVQA [46], WebQA [5], a provided text passage (VLQA [35]), commonsense-driven reasoning (VCR [50]), or external commonsense knowledge (OK-VQA [28], A-OKVQA[37]). This motivated a line of work on knowledge-enhanced VL transformer models. External knowledge is typically retrieved from a structured knowledge base like ConceptNet [41], in the form of a subgraph, and integrated into the VL transformer as an additional input [9, 20, 29, 48]. Alternative sources of knowledge include image captions [34], Google Search results [26], and textual and visual knowledge from Wikipedia, and Google Images [48]. In contrast to most of the preceding work, PICa [49] and Knowledge Augmented Transformer (KAT) [13] attempt to use GPT-3 [3] in a few-shot setting on the VQA task, by building prompts containing the caption and object tags generated using the image, followed by the question statement, asking the model to produce an answer.

In our proposed model, we focus on a specific subset of the knowledge-intensive datasets that require commonsense knowledge. Our approach, that uses COMET [15] for generating relevant commonsense knowledge, is distinctly different, far simpler, and more cost-effective than other alternatives described above.

2.3 Knowledge incorporation in NLP

Structured knowledge bases, or KBs, like ConceptNet [41] and ATOMIC [36] are widely used in NLP tasks to provide additional commonsense knowledge to models. ConceptNet contains 3.4M assertions focusing on concept and entity relations (such as RelatedTo, Synonym, IsA, MadeOf). ATOMIC contains 1.33M triplets focusing on event-centric social commonsense about causes, effects, mental states of the event participants. Several approaches were proposed for incorporating symbolic knowledge from these KBs into downstream NLP tasks such as encoding subgraphs of relevant knowledge [9, 23] and pre-training on commonsense knowledge bases or tasks [52].

Despite the performance improvements, incorporating knowledge directly from KBs suffers from two limitations: lack of coverage and lack of consideration for context. Commonsense Transformer, COMET [15], attempts to alleviate these issues by fine-tuning pre-trained language models on KBs. COMET can generate inferences for the various KB relations dynamically for new inputs. It has been successfully used for generating knowledge in language tasks [4, 27, 40, 44]. Inspired by the success of these models, we chose to use COMET [15] to generate relevant contextual expansions rather

than directly retrieving knowledge from KBs. To the best of our knowledge, we are the first to incorporate commonsense knowledge using COMET in VQA tasks.

Newer COMET variants [31, 47] are less applicable to OK-VQA and A-OKVQA as they focus more on event commonsense than entities. While obtaining implicitly learned commonsense from LLMs is a more recent alternative to using KBs [13, 49], this method is prohibitively expensive, both monetarily and in terms of compute resources.

Chapter 3

VLC-BERT

We briefly outline the overall architecture of our model and then delve deeper into its individual components. Figure 3.1 illustrates the VLC-BERT pipeline. Given an image with corresponding image regions I precomputed using Fast RCNN [10] and a question Q related to the image, we generate commonsense inferences C on the events and entities in the question phrase and two object tags O, and select the set of commonsense inferences which is the most useful for answering the question, $C = \{C_1, C_2, ..., C_k\}$ (§3.1). Finally, we embed Q, I and C, as input to VLC-BERT and train it to predict an answer A to Q (§3.2).

3.1 Structured knowledge generation and selection

3.1.1 Knowledge Generation

To generate commonsense knowledge, we employ the most recent version of COMET [15] initialized using BART [19] in a zero-shot setting. COMET is trained to complete 50 relation types from both ConceptNet [41] (such as AtLocation, Madeof) and ATOMIC [36] (such as xNeed, xWants), thus capturing concept as well as event oriented knowledge. We generate inferences based on 30 relation types most relevant to our work and supported by COMET.[2]Consider the example shown in Figure 3.2. For the given question, *"What is the purpose of the umbrella?"* we first process each question using AllenNLP's constituency parser [17] and convert it into a declarative sentence, since COMET was mainly trained on declarative sentences. In the example shown, *"What is the purpose of the umbrella?"* is rephrased as *"The purpose of the umbrellas is"*. We then adopt a state-of-the-art object detection model, YOLOv5 [16], to translate the corresponding image into object tags that COMET can understand. We select the top two most confident object tags and combine it with the question phrase to obtain

[2]We include the full list of relation types in the supplementary material.

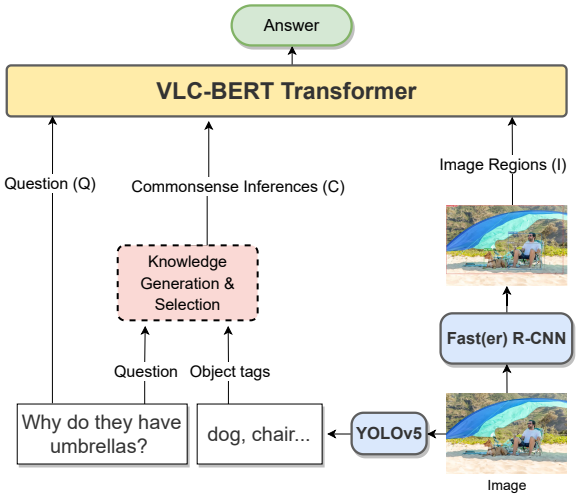

Figure 3.1: **Architecture of VLC-BERT**: Given an image, VLC-BERT generates commonsense inferences for the question-object phrase using COMET. These inferences are relevance ranked, and top ones (C) are selected and fed along with image regions (I) and the question (Q) into a VL-Transformer in order to produce an answer.

a question-object (QO) phrase, *"The purpose of the umbrella is, with dog and chair"*. We restrict the number of the object tags used in COMET's input to two because the addition of multiple tags make the inferences more conflated and noisy. In this manner, we can obtain inferences that can provide additional knowledge about both the visual and language inputs to VLC-BERT.

We use beam search to decode the top 5 inferences for each relation type, ranked according to the model's confidence. Overall, we get $30 \times 5 = 150$ inferences for each input phrase. Finally, we convert each inference to a sentence in natural language using relation-specific templates as defined in [7]. In the shown example, the assertion $< umbrella, Located\ At, store >$ is expressed as "You are likely to find umbrella at store". In order to remove redundant sentences of the same relation type, we measure the lexical overlap by measuring the percentage of common words between two given sentences. We exclude the sentences which have more than 70% overlap with previously constructed sentences of the same relation.

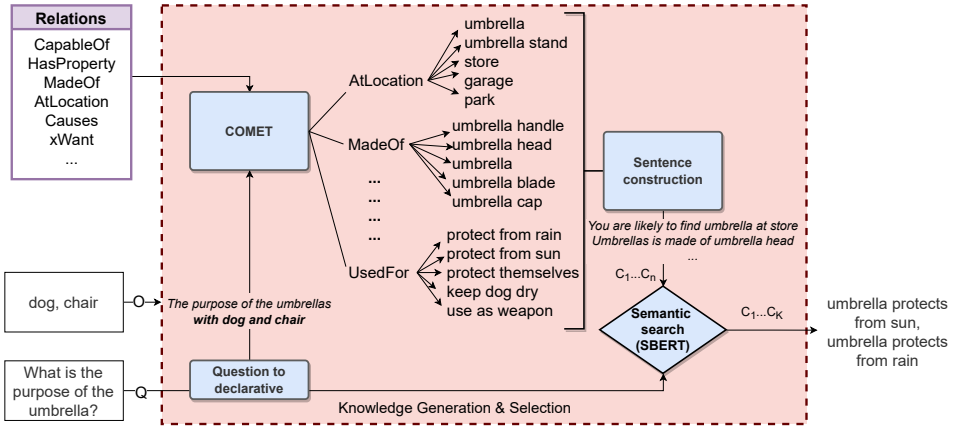

Figure 3.2: **Knowledge generation and selection**: We generate knowledge using COMET for fixed set of relations, using the object tags (O) and the question (Q). Semantic search is used to rank the most relevant knowledge associated with the question, to obtain a list of commonsense inferences (C).

3.1.2 Knowledge Selection

Due to the high cost of computation, and the noise associated with feeding such a large number of text tokens, feeding up to 150 COMET inferences into the VL Transformer model is impractical. In order to rank and select the inferences, we employ semantic search based on sentence transformers (SBERT) [32], which are pre-trained on tasks that retrieve candidate answers to a search query. In this method, the question and the inferences are embedded into the same vector space using SBERT [32] and cosine similarity between the question and the inference embeddings is used to rank the inferences. We prune the set of inference sentences C by picking $K = 5$ inferences which are expected to be the most useful for answering the question Q.

Augmented-SBERT We augment the SBERT used for semantic search by starting with a pre-trained SBERT model and continuing to train it for 2 epochs on question-inference instances from the *training set* of our datasets. To achieve this, we label the inferences for each question with similarity scores based on the proportion of overlap with the human-annotated answers. Since SBERT is trained on corpora that are distinct from our task, the augmentation ensures that the model understands the nature of query-

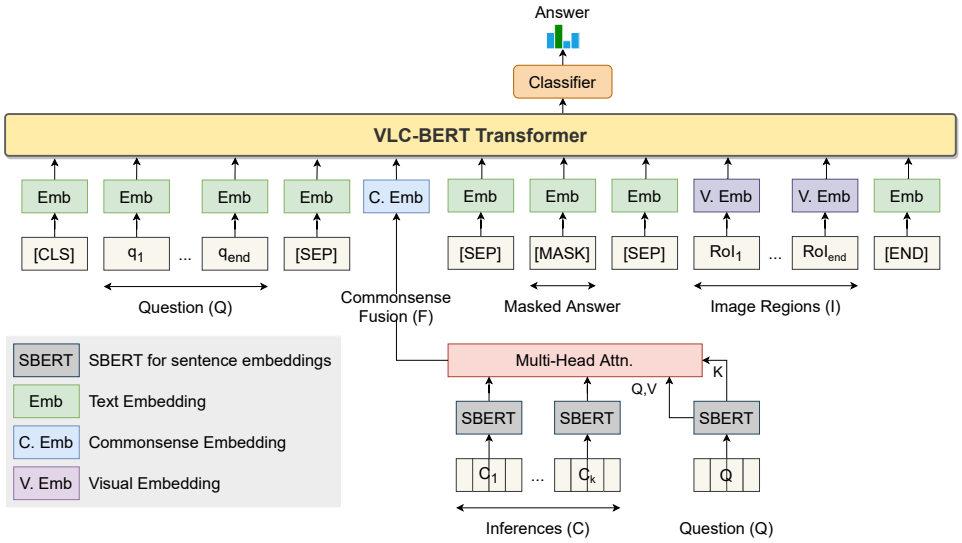

Figure 3.3: **VLC-BERT Transformer** is a single-stream Transformer that can attend across language, vision, and commonsense representations. We use the MHA block to fuse commonsense inferences into a useful commonsense representation.

inference pairings in our tasks. The augmented SBERT especially helps with narrowing down the right relations to the question. For instance, the question in shown in Figure 3.2 benefits most from the relations that talk about what the umbrella (*UsedFor*) is used for or capable of (*CapableOf.*)

3.2 VLC-BERT

We use a single-stream multimodal transformer encoder, VL-BERT [42], as the basis of VLC-BERT. VL-BERT is pre-trained on large-scale vision-language and language-only datasets with a goal of aligning the visual and linguistic features and building robust multimodal representations for downstream tasks. It is trained on the vision-language Conceptual Captions dataset [39], to predict regions-of-interests (RoIs) from language cues, and on the language-only BookCorpus [53] and English Wikipedia corpora, with a masked language modeling objective.

Figure 3.3 shows the VLC-BERT Transformer architecture. In the following paragraphs, we share how the input sequence is constructed and how the predicted answer is selected.

3.2.1 Inputs

Like VL-BERT, VLC-BERT accepts word token embeddings for language inputs and RoI token embeddings from the image for vision inputs. The architecture of VLC-BERT Transformer is shown in Figure 3.3. We use the [CLS] in the beginning of the sequence, [END] to mark the end of the sequence, and the separator token [SEP] between different inputs. We feed the question Q as a sequence of word tokens and the image regions I as sequences of RoIs. A [MASK] token is used to represent the unknown answer. In addition, we introduce a commonsense fusion token, F, to the input sequence, to incorporate our commonsense inferences.

A straightforward way to leverage the commonsense inferences $C = \{C_1, C_2, ..., C_k\}$ is to embed each word token in every inference sentence as an input token. However, this would lead to a very long input sequence, where the majority of inputs consist of inferences, thus potentially drawing the model's attention away from the other inputs. To overcome the challenge, we summarize the information contained in each inference sentence C_i into a single token representation $\vec{C_i}$, by embedding the inference using SBERT [32]:

$$\vec{C_i} = \text{SBERT}(C_i) \tag{3.1}$$

Next, in order to obtain a fused representation of the k commonsense inferences, we attend to the corresponding SBERT embeddings, $[\vec{C_i}...\vec{C_k}]$ against the SBERT embedding of the question, $\vec{Q} = \text{SBERT}(Q)$. The intuition behind this approach is that the model learns to assign a higher score to the most important inference to the question. The key (K_A), query (Q_A) and value (V_A) are assigned as shown below,

$$K_A = \vec{Q} \tag{3.2}$$

$$Q_A, V_A = \text{append}([\vec{C_i}...\vec{C_k}], \vec{Q}) \tag{3.3}$$

$$\vec{F} = \text{MHA}(K_A, Q_A, V_A) \tag{3.4}$$

where MHA is the standard multi-head attention [45], that delivers a *single vector* incorporating all relevant commonsense knowledge required to answer the question. Note that we append the question embedding $\vec{Q}$ to list of commonsense inference embeddings for Q and V because there may be cases where none of the inferences are useful to answer the question. In such a case, the model may choose to ignore the inferences by attending to the question embedding $\vec{Q}$ instead.

Weak Supervision In order to train the MHA block effectively, we employ weak supervision on the attention weights. For a small subset of the questions in the training set, we obtain label attention weights by following these steps: (1) we initialize a vector $\hat{A}$ of length $k+1$ where all values are 0.05, (2) for each C_i, if C_i contains a word in the ground-truth answer list, then we set the $\hat{A}_i$ to 0.8, (3) if none of the C inferences contain answer words, we assign a weight of 0.8 to $\hat{A}_{k+1}$ so that the question has the largest weight, and (4) we normalize $\hat{A}$ so that its values sum up to 1. We then apply cross-entropy loss between the predicted attention weights from MHA and our label attention weights $\hat{A}$, and sum this with the answer prediction loss.

Finally, a positional encoding is added to all input tokens following the method described in VL-BERT. In addition, a different segment type encoding is applied to the four segments in the input sequence: the question segment, the commonsense segment, the masked answer segment, and the image region segment.

3.2.2 Answer Selection

We use the encoded [MASK] token to represent the answer, thereby making VQA a masked language modelling task with visual cues. To predict the final answer, we apply a classifier over the entire answer vocabulary, as done in VL-BERT. During training, we follow VL-BERT and use a cross-entropy loss over picking the correct answer from an answer vocabulary.

Chapter 4

Datasets

We perform experiments on the OK-VQA [28] and A-OKVQA [37] datasets. In order to utilize the existing VL-BERT model effectively, we pre-train VLC-BERT on the larger VQA 2.0 [12].

4.1 Dataset Descriptions

OK-VQA In the Outside-Knowledge VQA dataset [28], questions require external knowledge in addition to the information in the images. The dataset is composed of 14,031 images and 14,055 questions, and the crowsourced questions are divided into ten knowledge categories: Vehicles and Transportation; Brands, Companies and Products; Objects, Materials and Clothing; Sports and Recreation; Cooking and Food; Geography, History, Language and Culture; People and Everyday Life, Plants and Animals; Science and Technology; and Weather and Climate. OK-VQA only contains open-ended questions with five human-provided answers. Since OK-VQA does not have a validation set, we dedicate 1,000 of the 9,009 training questions for validation.

A-OKVQA A-OKVQA [37] is the augmented successor to OK-VQA and consists of 25K questions that require a combination of commonsense, visual, and physical knowledge. In contrast to other knowledge-based visual question answering datasets, the questions in A-OKVQA are conceptually diverse, involving knowledge that is not contained in the image, and cannot be resolved by a simple knowledge base query. A-OKVQA is split into training, validation, and test sets based on images used from the COCO 2017 [24] dataset. Moreover, all questions in the dataset have human annotated direct answers as well as multiple-choice options, but we focus on the direct answers. The A-OKVQA test set is blind, requiring us to submit to the leaderboard to obtain a test accuracy.

VQA 2.0 The Visual Question Answering (v2.0) dataset contains 1.1 million crowdsourced questions about 204,721 images from the COCO dataset [24]. Each question is annotated with 10 ground truth answers obtained using Amazon Mechanical Turk. A majority of the questions in this dataset do not require external commonsense knowledge.

4.2 Evaluation Metric

Both datasets use the same accuracy-based evaluation metric. Each question has a set of 10 ground truth answers provided by different annotators. Accuracy is calculated as the percentage of predicted answers that were proposed by at least 3 human annotators: $\text{acc} = \min(\frac{\#\text{ humans gave the answer}}{3}, 1)$.[3]

[3]Following the same evaluation, each of the 5 answers in OK-VQA is used twice

Chapter 5

Experiment Setup

The implementation of our model builds on VL-BERT [42]. To that end, we follow the fine-tuning steps provided in the official codebase of the VL-BERT model for VQA 2.0, and modify it to support the OK-VQA and A-OKVQA datasets. We maintain the recommended hyperparameter values, and train the $BERT_{BASE}$ size of the model, with a hidden feature dimension of 768. The model is trained for 20 epochs on the OK-VQA and A-OKVQA datasets. For all models, we use a batch size of 16 and gradient accumulation step size of 4. We train the models presented in the main result thrice and report the average test accuracy on the OK-VQA dataset, and the best (leaderboard) test accuracy on the A-OKVQA dataset.

Answer Vocabulary Due to the large number of unique answers to questions in visual question answering datasets, it is infeasible to use all answers in the answer vocabulary. For the OK-VQA dataset, following KRISP [29], we build an answer vocabulary of 2,249 answers by selecting all answers in the training set that appear at least 10 times. This answer vocabulary ignores the empty space answer, and includes an <UNK> answer token. During training, if a ground truth answer is not present in the answer vocabulary, we assign it to the (<UNK>) token. For the A-OKVQA dataset, we use the answer dictionary that is already provided in the dataset [37].

VQA Pre-Training (VQA P.T) Following the idea that pre-training is beneficial for Transformer models, we initialize VLC-BERT with weights obtained after fine-tuning VL-BERT on the VQA 2.0 dataset for 5 epochs. Note that KRISP [29] benefits from pre-training on the VQA 2.0 dataset, and PICa [49] and KAT [14] utilize GPT-3, a large-scale pre-trained model, for external commonsense. Furthermore, because OK-VQA and A-OKVQA are significantly smaller than VQA 2.0, this initialization favourably benefits the training process and gives us a stronger baseline to work with.

Chapter 6

Results

In this chapter, we focus on evaluating VLC-BERT on the OK-VQA and A-OKVQA datasets and comparing against existing state-of-the-art models for VQA with external commonsense knowledge. Table 6.1 highlights our performance improvements on the test set for OK-VQA and A-OKVQA against other models. Later in this chapter, we ablate on the components of our model.

6.1 Main Results

Table 6.1 specifies which knowledge sources each model leverages. In the top section, we consider models that utilize knowledge bases such as ConceptNet and Wikipedia, as well as models that utilize web search APIs to obtain external knowledge. VLC-BERT incorporates COMET, which is trained on ConceptNet and ATOMIC, and we compare favourably against these models. Notably, VLC-BERT achieves an accuracy of 43.14 on OK-VQA, outperforming KRISP (Wikipedia + ConceptNet + VQA P.T.) by over 4 points, and MAVEx (Wikipedia + ConceptNet + Google Images) by about 2 points. While our model clearly outperforms previous methods that use knowledge bases, it does not outperform models with large-scale pre-training and large number of parameters such as GPT-3 [3] and GPV2 [18], which incorporate implicit commmonsense knowledge and require extensive resources to train. However, on OK-VQA, we achieve very similar results to PICa-Base [49], despite not having access to GPT-3. We expect that the use of a large pre-trained model like GPT-3 can further boost the performance of VLC-BERT.

6.2 Ablation Tests

We perform comprehensive ablations on the validation set of the A-OKVQA dataset, as represented in Table 6.2.[4]

[4]We present additional ablations in supplementary material Sec 2.3

Table 6.1: Accuracy of our model against other models for OK-VQA and A-OKVQA datasets. Our model improves upon existing knowledge base based models due to the contextualized commonsense inferences from COMET, which is trained on ConceptNet and ATOMIC. We compare favourably against the highlighted models that utilize external knowledge bases. Note: P.T. stands for Pre-Training, W stands for Wikipedia, and CN stands for ConceptNet.

Method	Knowledge Sources	OK-VQA	A-OKVQA	Appx. Params
ViLBERT [37]	-	-	25.85	116M
LXMERT [37]	-	-	25.89	-
BAN + AN [28]	W	25.61	-	-
BAN + KG-AUG [20]	W + CN	26.71	-	-
MUTAN + AN [28]	W	27.84	-	-
ConceptBert [9]	CN	33.66	-	118M
KRISP [29]	W + CN	32.31	27.1	116M
KRISP [29]	W + CN + VQA P.T.	38.9	-	116M
Visual Retriever-Reader [26]	Google Search	39.2	-	-
MAVEx [48]	W + CN + Google Images	41.37	-	-
GPV2 [18, 37]	Web10k + COCO P.T.	-	40.7	220M
PICa-Base [49]	GPT-3	43.3	-	175B
PICa-Full [49]	GPT-3	48.0	-	175B
KAT [14]	Wikidata + GPT-3	54.41	-	175B
VLC-BERT (Ours)	VQA P.T. + COMET	43.14	38.05	118M

VQA P.T We begin by training A-OKVQA on the baseline VL-BERT model without VQA pre-training. This gives us a score of 36.24. Next, obtain a new baseline for our model with VQA pre-training, where we then initialize VLC-BERT with pre-trained weights on the VQA 2.0 dataset, and further train it on the A-OKVQA dataset. This results in a score of 43.46, over 7 points better, highlighting the impact of pre-training with a large-scale dataset. This model is a strong baseline for our VQA tasks.

Comm. Inference Representation In the full model, we use SBERT to summarize each commonsense inference into a single vector, and use the multi-head attention block to capture useful information from the list of inference vectors. To test the effectiveness of our commonsense inference representation method, we first ablate SBERT, i.e., we incorporate all inferences as an additional text input for VLC-BERT, feeding them token-by-token. This results in an accuracy score of 43.44, which is slightly lower than our baseline with VQA pre-training. Next, we use SBERT to summarize inferences, and feed the SBERT embeddings directly into VLC-BERT

Table 6.2: Ablation of various components in VLC-BERT, evaluated on the A-OKVQA validation set. We observe that all the components of our model play a critical role in empirical performance.

VQA P.T.	Aug. SBERT	SBERT	Attn.	Val
VQA Pre-training				
–	–	–	–	36.24
✓	–	–	–	43.46
Comm. Inference Representation				
✓	✓	–	–	43.44
✓	✓	✓	–	43.64
✓	✓	✓	✓	**44.95**
Augmentation of SBERT				
✓	–	✓	✓	44.10
✓	✓	✓	✓	**44.95**

with only a linear projection layer rather than the MHA block. This variant performs worse than the model with the MHA block by 1.25 points.

Augmented SBERT In order to familiarize SBERT with our question-inference pairs, we fine-tune SBERT on the training set of A-OKVQA and OK-VQA (Sec 3.1.2). We perform an ablation by evaluating our model on SBERT that has never been exposed to the question-inference-pairs. This results in a drop of 0.85 points in accuracy, which shows that our augmentation of SBERT is effective.

Chapter 7

Analysis

7.1 Commonsense subsets

Questions in OK-VQA and A-OKVQA datasets are diverse and require commonsense reasoning, visual understanding, as well as factual knowledge. While COMET can generate contextualized commonsense knowledge, it does not help with questions that require scene understanding (*e.g.*, "What is to the left of the computer?"), factual knowledge (*e.g.*, "Where was this food invented?"), or text/symbol recognition (*e.g.*, "What does this sign say?"). Moreover, averaging results on the entirety of OK-VQA and A-OKVQA obfuscates the improvements brought about to a subset of questions that truly require commonsense knowledge. We propose subsets to assess the performance of our model on questions that are more likely to require external commonsense knowledge. We obtain the subsets by eliminating questions that are mostly factual or visual, and hence do not require commonsense, following these conditions: (1) *factual*: The question or answer contains named entities (*e.g.*, "USA"); (2) *numerical*: The answers contain numbers or number words (*e.g.*, "twenty") or the question has date or time words (*e.g.*, "century"); (3) *visual*: The question contains directional words (*e.g.*, "left of") and words referring to symbols (*e.g.*, "mascot").

Table 7.1: Evaluation on the subsets of OK-VQA test (OK_s) and A-OKVQA validation (A-OK_s) sets, where factual, numerical and visual questions are pruned. The performance gain observed on the subsets shows a better picture of where external commonsense is effective.

Method	OK	OK_s	A-OK	A-OK_s
Base	42.29	47.4	43.46	46.52
w/ COMET	**43.14**	**48.21**	**44.95**	**49.53**

In Table 7.1, we show that VLC-BERT with COMET performs 3 points better on the A-OKVQA subset, and maintains an 0.8 point improvement

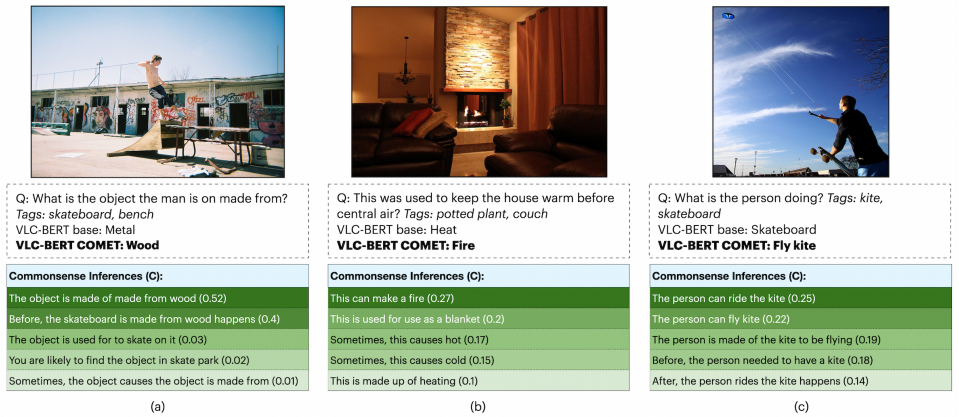

Figure 7.1: **Attention analysis:** (a) is from A-OKVQA, and (b) and (c) are from OK-VQA. We observe that the weakly supervised attention layer in VLC-BERT accurately picks useful commonsense inferences. In (c), we observe how object tags are useful to guide COMET to produce contextualized knowledge.

on the OK-VQA subset. This substantiates our claim that utilizing our COMET pipeline substantially increases VLC-BERT's ability to answer questions that require external knowledge.

7.2 Attention Analysis

In this section, we show qualitative examples to demonstrate questions where VLC-BERT benefits from contextualized commonsense knowledge from COMET. We also show the corresponding attention weights, to show the effectiveness of the proposed weakly-supervised attention mechanism. Fig 7.1a shows an example from A-OKVQA, where COMET's inferences on the question and the object tags, weighted by the attention score, results in the correct answer. Fig 7.1b shows an example from OK-VQA where VLC-BERT COMET exhibits higher attention towards the fire despite the object tags missing the fireplace. This is an example where deriving inferences from the question phrase is equally important as doing so with the object tags. Fig 7.1c shows that inferences on the object tag *kite* drove the model to answer correctly. The supplementary material includes additional examples of improvements and failures.

Chapter 8

Discussion

VLC-BERT is more capable than other models of its size in its ability to obtain relevant commonsense knowledge, and effectively use it while answering questions. However, our analysis of VLC-BERT highlighted a few limitations of our model and the datasets we evaluate on. We highlight some of these limitations in the following paragraphs:

Limitations of object tags: Some questions require a deeper understanding and linking of multiple entities and events in the image, that object tags lack, for deriving relevant commonsense inferences. In future work, it would be valuable to experiment with more complex image descriptions in the form of captions or a larger set of object tags, or even scene graphs generated from the image.

Semantic compression of inferences: Condensing the commonsense inferences using SBERT and MHA leads to a compressed representation will likely cause the model to lose some information. In some cases, this information loss can be detrimental to the model.

Limitations of COMET: Our model is limited by COMET, and the knowledge bases it is trained on, as we observe that large-scale models like GPT-3 outperform it. The increasing availability of LLMs and VLMs may reduce the necessity of a commonsense LM such as COMET. In future work, with the democratization of large scale models, we may attempt to incorporate open-source LLMs and VLMs to build and summarize contextual commonsense inferences, in line with recent works.

Chapter 9

Conclusion

We presented Vision-Language-Commonsense BERT (VLC-BERT) for external knowledge-driven VQA tasks. VLC-BERT outperforms previous models based on knowledge bases on the OK-VQA and A-OKVQA datasets by incorporating contextualized commonsense knowledge from COMET and combining it with visual and linguistic inputs. Through our evaluation, we show the effectiveness of our knowledge generation, selection, and incorporation strategies, and the positive impact of VQA pre-training.

We view our work as a first step in analyzing the potential of *generative commonsense incorporation*, and exploring approaches to *decide when commonsense is needed*. We plan to investigate the potential of multi-hop reasoning with COMET to bridge the question and image-based expansions closer, in addition to experimenting with novel large-scare models.

Bibliography

[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: Commonsense transformers for automatic knowledge graph construction. In *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.

[4] Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. It's not rocket science : Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics (TACL)*, 2022.

[5] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[6] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Joe Davison, Joshua Feldman, and Alexander Rush. Commonsense knowledge mining from pretrained models. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lécué. Conceptbert: Concept-aware representation for visual question answering. In *FINDINGS*, 2020.

[10] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[11] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Workshop on Automated Knowledge Base Construction*, page 25–30, 2013.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language, 2021.

[14] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, July 2022.

[15] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.

[16] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, February 2022.

[17] V. Joshi, Matthew E. Peters, and Mark Hopkins. Extending a parser to distant domains using a few dozen partially annotated examples. In *ACL*, 2018.

[18] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models, 2022.

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[20] Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. *28th ACM International Conference on Multimedia*, 2020.

[21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[22] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *European Conference on Computer Vision (ECCV)*, 2020.

[23] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*, 2019.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[25] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[26] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *EMNLP*, 2021.

[27] Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206. Association for Computational Linguistics, November 2020.

[28] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14111–14121, 2021.

[30] OpenAI. Gpt-4 technical report, 2024.

[31] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*, 2020.

[32] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

[33] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[34] Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. Image captioning for effective use of language models in knowledge-based visual question answering. *CoRR*, 2021.

[35] Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. Visuo-linguistic question answering (VLQA) challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, November 2020.

[36] Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019.

[37] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022.

[38] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6642–6651, 2019.

[39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[40] Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629. Association for Computational Linguistics, November 2020.

[41] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017.

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[43] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

[44] Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. HypoGen: Hyperbole generation with commonsense and counterfactual knowledge. In *Association for Computational Linguistics: EMNLP 2021*, pages 1583–1593, 2021.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

[46] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427, oct 2018. ISSN 0162-8828.

[47] Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *NAACL*, 2022.

[48] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-Modal Answer Validation for Knowledge-based VQA. In *AAAI*, 2022.

[49] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa, 2021.

[50] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[51] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021.

[52] Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Improving question answering by commonsense-based pre-training. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 16–28. Springer, 2019.

[53] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, 2015.

Appendix A

Supporting Material

A.1 Implementation Details

In this section, we provide additional information about the implementation of each of the components of VLC-BERT.

A.1.1 Object Tags with YOLO

As described in Sec 3 of our paper, we utilize object tags to incorporate image context for generating commonsense inferences. In order to obtain the object tags, we use an off-the-shelf YOLO model for PyTorch, YOLOv5 *by Ultralytics* [16]. We use the pretrained *yolov5l* model to obtain object bounding boxes and the associated class name for each bounding box, on COCO 2014 and 2017 images for OK-VQA and A-OKVQA datasets respectively. We then use a confidence threshold of 0.5 to prune out objects that are unlikely to be useful. In addition, we prune out the *person* object name as well as the objects already present in the question phrase, to avoid unnecessary tags or repetitions. Finally, the two object names associated with the highest confidence bounding boxes are picked as the object tags for our model, O.

A.1.2 Knowledge Generation

As described in Sec 3.1.1 of our paper, we generate commonsense inferences from COMET [15] by inputting the question followed by "with" and two object tags into it. If S is the sentence consisting of the question and object tags and R is the relation type we want to generate from COMET, we provide it to COMET in the form $S^{(i)}$ $R^{(i)}$ `[GEN]` and let comet generate the commonsense inferences. Though COMET can support 50 relation types, we cherry-pick 30 relation types by removing duplicate relations and relations that are irrelevant to our work (*e.g.*HasPainIntensity). Table A.1 provides the list of the 30 relations we used to generate commonsense expansions from COMET and the corresponding templates we used to convert COMET's

output to natural language sentences. {0} usually indicates the subject in the input sentence to COMET, and {1} indicates the generated expansion.

Table A.1: Relations used for generating expansions from COMET and their corresponding sentence templates

#	Relation	Sentence template
1	AtLocation	You are likely to find {0} in {1}
2	CapableOf	{0} can {1}
3	Causes	Sometimes {0} causes {1}
4	CreatedBy	{1} is created by {0}
5	Desires	{0} wants {1}
6	HasA	{0} has {1}
7	HasFirstSubevent	The first thing you do when you {0} is {1}
8	HasProperty	{0} is {1}
9	HinderedBy	{0} is hindered by {1}
10	IsA	{0} is {1}
11	isAfter	{0} happens before {1}
12	isBefore	{1} happens before {0}
13	LocatedNear	{0} is located near {1}
14	MadeOf	{0} is made of {1}
15	MadeUpOf	{0} is made up of {1}
16	NotCapableOf	{0} is not capable of {1}
17	NotHasProperty	{0} does not have the property of {1}
18	NotIsA	{0} is not {1}
19	NotMadeOf	{0} is not made of {1}
20	ObjectUse	{0} is used for {1}
21	PartOf	{1} has {0}
22	SymbolOf	{0} is a symbol of {1}
23	UsedFor	{0} is used for {1}
24	xAttr	{0} is seen as {1}
25	xEffect	{0} then {1}
26	xIntent	Because {0} wanted {1}
27	xNeed	Before {0} needed {1}
28	xReact	As a result {0} feels {1}
29	xReason	{0} reasons {1}
30	xWant	As a result {0} wants {1}

Table A.2: Hyperparameters of our model

Hyperparameter	VQA P.T.	OK-VQA	A-OKVQA
Batch Size	16	16	16
Gradient Accumulation	4	4	4
Epochs	5	20	20
Learning Rate	6.25e-7	6.25e-7	6.25e-7
Visual Size	768	768	768
Hidden Size	768	768	768
Warmup Method	linear	linear	linear
Warmup Steps	1000	1000	1000
MHA Heads	–	3	3
MHA Dropout	–	0.1	0.1

A.1.3 Knowledge Selection

As described in Sec 3.1.2 of our paper, we augment S-BERT to perform semantic search and filter and rank the relevance of commonsense inferences. In order to perform semantic search, we utilize the sentence-transformers package for SBERT[5] [32]. We initialize our SBERT model from the pre-trained *msmarco-roberta-base-ance-firstp* model and train this model for 2 epochs on the training set of the corresponding task. To create the labels for this augmentation, we measure the overlap of the expansions to human annotated answers and assign a similarity score of *0.8* for overlapping expansions and a score of *0.2* for non-overlapping expansions. This augmented S-BERT model is then used to encode the question and commonsense sentences, before computing the sentence similarity between every commonsense sentence and the question, and picking the top k ($K = 5$) sentences to use in the input sequence of the VLC-BERT transformer.

A.1.4 VLC-BERT Transformer

Our implementation of the VLC-BERT transformer encoder is based on the publicly available implementation of VL-BERT[6] [42]. The hyperparameters we use for training VLC-BERT on the VQA 2.0 (only for pre-training), OK-VQA and A-OKVQA datasets are given in Table A.2.

For generating sentence embeddings for commonsense inferences that are fed into the MHA block, we use the *all-mpnet-base-v2* pre-trained model from SBERT.

[5]https://www.sbert.net
[6]https://github.com/jackroos/VL-BERT

A.1.5 Implementation of Commonsense Subsets

In Sec 7.1, we describe the need for commonsense-specific subsets of OK-VQA and A-OKVQA, to show that our model improves on the baseline significantly. The lack of any annotations for the type of reasoning required to answer the question led us to develop our own method to obtain the subsets. Below, we have the exact details required to re-create the subsets:

Named Entities. We use spaCy's entity recognizer[7]. If any word in the question or list of answers is recognized as an entity, we prune the question.

Numerical. We first attempt to check if a string is a number using Python's built-in function, `isdigit()`. If it is not a digit, then we use the `word2number` package[8] to attempt to convert words (*e.g.*"twenty") into numbers. If it is successful in doing so, we deem the word to be a number. If any word in the question or list of answers is recognized as a number, we prune the question.

Directional. We list commonly used directional words: `right, left, top, bottom, behind, under, inside, over, front, back, near, next`. If any word in the question is recognized as a directional word, we prune the question.

Symbol. We list commonly used symbol words: `logo, symbol, name, company, mascot, word, brand`. If any word in the question is recognized as a symbol word, we prune the question.

Color. We list commonly used color words: `blue, green, red, black, white, grey, purple, pink, yellow, orange`. If any word in the question or list of answers is recognized as a color word, we prune the question.

Time. Finally, we list commonly used time words: `century, year, time, month, day`. If any word in the question is recognized as a time word, we prune the question.

As the task of recognizing the type of a question is challenging in itself, we tried to simplify it to a basic, reproducible method, in order to better evaluate on commonsense reasoning specific questions on the OK-VQA test set and the A-OKVQA validation set.

[7] https://spacy.io/api/entityrecognizer
[8] https://pypi.org/project/word2number/

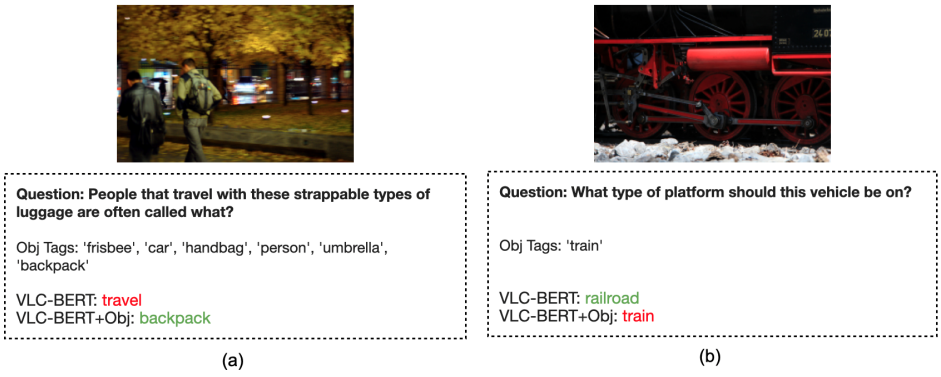

Figure A.1: **Qualitative examples with Obj Tags:** (a) Object Tags may include contents in the image that may have been missed in the model, giving us the right answer. (b) In some cases, object tags may lead to the model making erroneous predictions.

A.2 Additional Results

A.2.1 Main Evaluation

The standard deviation on our scores for the OK-VQA test set is 0.20 and for the A-OKVQA validation is 0.47.

A.2.2 Including the Object Tags in the VL Model

We extend VLC-BERT to incorporate the generated object tags by YOLOv5 [16] (obtained in our Knowledge Generation step). Object tags serve as additional natural language information to the VLC-BERT model, that could be useful for answering certain questions where the information in the caption is limited. In this setting, a comma separated list of all object tags obtained using YOLOv5 are fed into the VLC-BERT Transformer as text tokens before the [MASK] token (Figure 3.3).

We test this method only on the OK-VQA test set. As we run these experiments separately from our main results, we have re-run the baseline model. Our baseline VLC-BERT model obtains a test set score of $\boxed{44.86}$, while the model trained with the object tags achieves $\boxed{45.43}$, indicating a significant improvement in performance. We show an example of where object tags can be useful, and where they may fail, in Figure A.1

Table A.3: Performance of our model on OK-VQA question categories.

Category	Base	w/ COMET
Vehicles and Transportation	40.1	**41.16**
Plants and Animals	**42.58**	41.65
People and Everyday Life	**40.09**	39.95
Sports and Recreation	51.53	**52.31**
Cooking and Food	42.36	**45.04**
Objects, Material and Clothing	39.86	**39.95**
Science and Technology	37.38	**38.57**
Weather and Climate	**50.7**	48.99
Brands, Companies and Products	33.6	**35.81**
Geog, Hist, Language and Culture	40.14	**43.4**
Other	41.23	**42.68**

A.2.3 Evaluation on OK-VQA Question Categories

The results provided in our paper only show the overall scores of our models on the OK-VQA [28] dataset. In Table A.3, we share the results for each question category in the OK-VQA dataset. The OK-VQA dataset has questions divided into 11 different categories [28]. The results show that our model with external knowledge from COMET improves upon the baseline in all but three categories. Across all the models, we see that the 'Brands, Companies and Products' is the most challenging category, with low accuracy for both the baseline and the VLC-BERT with COMET models. This is expected, because the questions in this category often require the model to read text or symbols in the image, or identify company names and logos, which are challenging tasks outside the domain of our model.

A.2.4 Ablations

In this section, we present additional ablations to show the impact of different components of the VLC-BERT pipeline.

Ablation on number of sentences. In order to test the impact of the number of commonsense inferences K, we report the performance with different K values. We ran our latest model with $K = 10$ and $K = 15$ sentences. On the A-OKVQA validation set, we obtain the following results: $\boxed{K = 5 : 44.95;\ K = 10 : 44.57;\ K = 15 : 43.93}$. We thus feed $K = 5$ com-

monsense inference sentences into VLC-BERT transformer, because we had observed that adding too many commonsense inferences also adds unnecessary noise in the model, which hurts performance.

Use of Object tags. In order to assess the importance of the number of object tags used in deriving commonsense inferences, we ran experiments with no (0) object tags, as well as all (>2) tags. For zero tags, we get $\boxed{44.42}$, and for all tags, we get $\boxed{44.62}$. These are slightly worse than the two tags version ($\boxed{44.95}$). This is in line with what we expected, since COMET is not designed to deal with complex sentences containing multiple entities, and 2 object tags stands as a good trade-off. Furthermore, in our qualitative results, we show examples of where object tags are useful in providing image context.

Impact of weak attn. supervision. Disabling weak attn. supervision, we obtain a result of $\boxed{44.89}$ which is slightly worse compared to $\boxed{44.95}$ with supervision. However, qualitative analysis shows that our model with supervision produces stronger attention weights for useful inferences compared to the model without.

Appendix B

Error Analysis and Examples

In this chapter, as mentioned in Section 7.2 of our paper we provide additional qualitative examples along with their attention scores, of where VLC-BERT improved as well as failed.

B.1 Error examples

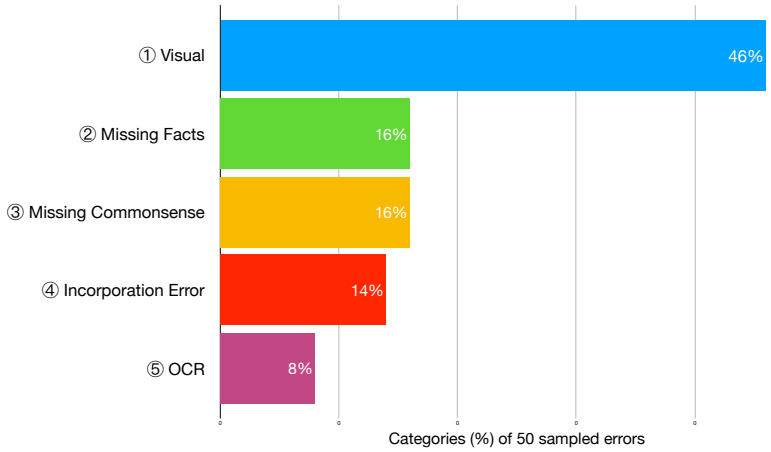

Figure B.1: **Error analysis:** Percentage of error categories from AOKVQA

We analyze the errors from the best version of VLC-BERT model. We randomly sample 50 erroneous examples from the validation set of A-OKVQA, analyze the errors, and classify them into five categories as shown in Figure B.1. We provide an example of each category in Figure B.2.

1 **Visual**: The model is lacking deep scene understanding that either required to answer the question, or to generate relevant commonsense inferences. This includes cases where the object tags are insufficient for describing the scene. A majority of the errors we see in VLC-BERT fall

in this category, in line with our conclusions and motivations for future work on commonsense models that involve deep scene understanding.

2 **Missing Facts**: The model failed due to missing factual knowledge about named entities, types of entities and well-known facts.

3 **Missing Commonsense**: The final commonsense inferences provided to VLC-BERT are missing the commonsense knowledge required to answer the question, either due to COMET not capturing this knowledge or semantic search not picking the right inferences.

4 **Incorporation Error**: Though the answer is provided in the commonsense inferences, and we attended highly to these inferences, it is still ignored by VLC-BERT, probably because the visual representation took priority. The commonsense inferences being much more condensed compared to other inputs of VLC-BERT could be one of the reasons for this.

5 **OCR**: The question involves reading text in the images and requires the VLC-BERT to support Optical Character Recognition (OCR).

B.2 Improvement examples

In Figure B.3, we provide additional qualitative examples where commonsense from COMET helped in driving the model to make the right prediction.

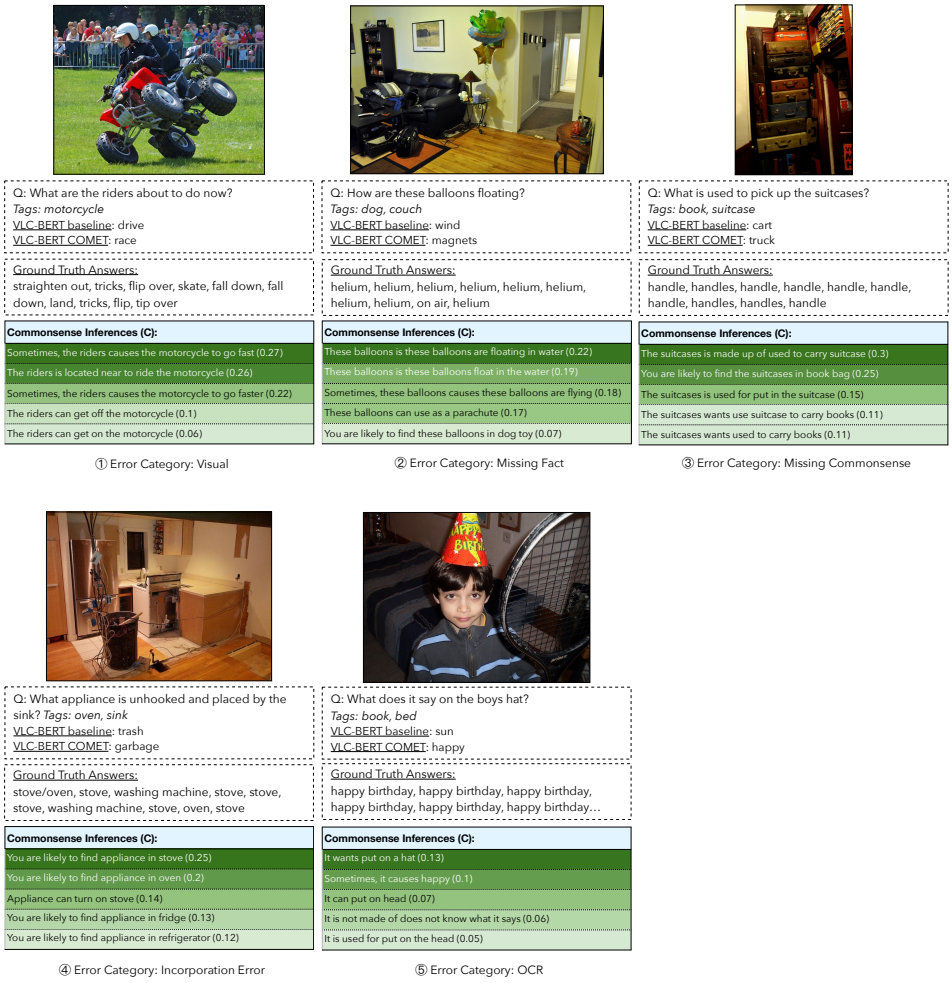

Figure B.2: **Error analysis:** We sample 50 erroneous examples from the A-OKVQA validation set, and categorize it into five categories.

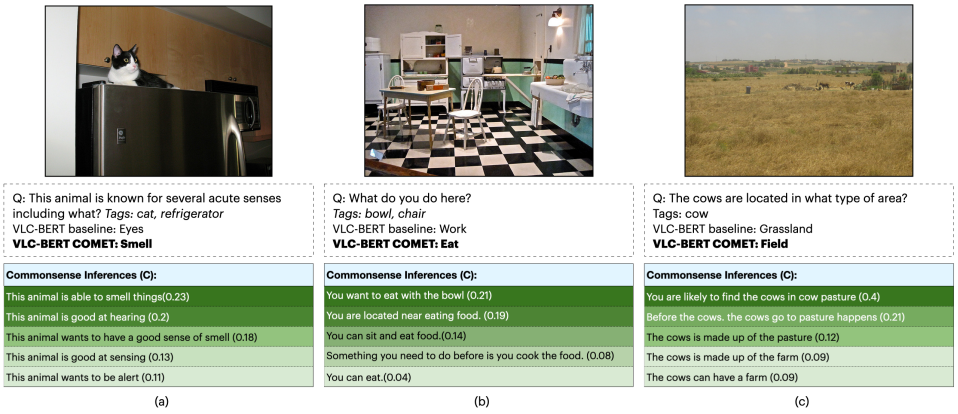

Figure B.3: **Qualitative examples:** (a) is from A-OKVQA, and (b) and (c) are from OK-VQA.

Appendix C

Additional Works

In addition to this work, the author has contributed to other research activities during the Master of Science degree. This includes:

1. **PD-EST: Process-disentangling Event Sequence Transformer**
   Aditya Chinchure, Fredrick Tung, Leonid Sigal
   Research conducted as a part of an internship at Borealis AI. Not included in this thesis.