KAT: A Knowledge Augmented Transformer for Vision-and-Language

Liangke Gui†‡
Alex Hauptmann§
Borui Wang†‡
Yonatan Bisk§‡
Qiuyuan Huang‡
Jianfeng Gao‡
§Carnegie Mellon University  †Yale University  ‡Microsoft Research
{liangkeg, alex, ybisk}@cs.cmu.edu
borui.wang@yale.edu, {qihua, jfgao}@microsoft.com

Abstract

The primary focus of recent work with large-scale transformers has been on optimizing the amount of information packed into the model’s parameters. In this work, we ask a different question: Can multimodal transformers leverage explicit knowledge in their reasoning? Existing, primarily unimodal, methods have explored approaches under the paradigm of knowledge retrieval followed by answer prediction, but leave open questions about the quality and relevance of the retrieved knowledge used, and how the reasoning processes over implicit and explicit knowledge should be integrated. To address these challenges, we propose a novel model - Knowledge Augmented Transformer (KAT) - which achieves a strong state-of-the-art result (+6 points absolute) on the open-domain multimodal task of OK-VQA. Our approach integrates implicit and explicit knowledge in an end to end encoder-decoder architecture, while still jointly reasoning over both knowledge sources during answer generation. An additional benefit of explicit knowledge integration is seen in improved interpretability of model predictions in our analysis.

1 Introduction

Recently, there has been a revival of interest in knowledge-intensive tasks which require an external knowledge source for humans to perform. These tasks address many applications in real-world scenarios that require precision and domain specific knowledge not covered by commonsense. Virtual assistants and autonomous AI agents, which participate in our everyday life, need to seamlessly integrate implicit (i.e., commonsense) and external knowledge when answering questions. In this work, we investigate how to effectively combine implicit and explicit knowledge for knowledge reasoning.

Work done when Liangke and Borui interned at Microsoft.

Figure 1: Examples of knowledge-based VQA that requires external knowledge. Success on this task requires not only visual recognition, but also logical reasoning to incorporate external knowledge about the world.

We use the Outside-Knowledge Visual Question Answering (OK-VQA) dataset as a case study.

Consider the examples from OK-VQA shown in Figure 5. To answer the first question “What did this organism evolve from?”, the system needs to both ground organism to bird and then apply the external knowledge “birds evolved from reptiles” to answer the question. The key challenge here is to accurately link image content to abstract external knowledge. There have been a number of recent developments demonstrating the feasibility of incorporating external knowledge into Question Answering models (Wang et al., 2017b; Li et al., 2020b; Marino et al., 2021; Wu et al., 2021; Garderes et al., 2020). Existing methods first retrieve external knowledge from multiple external knowledge resources, such as DBPedia (Auer et al., 2007) and ConceptNet (Liu and Singh, 2004) before jointly reasoning over the retrieved knowledge and image content to predict an answer.

Most existing retrieval-based approaches have several drawbacks: 1. Knowledge retrieved using keywords from questions or image tags may be too generic, which leads noise or irrelevant knowledge during knowledge reasoning. 2. Existing
work mainly focuses on explicit knowledge which is often in the form of encyclopedia articles or knowledge graphs. While this type of knowledge can be useful, it is insufficient to answer many knowledge-based questions. For example, to answer the question “What type of event might the couple be preparing for?”, people need to recognize the dress and relationships of people from the corresponding image. Such inference relies on commonsense knowledge to make default assumptions about the unknown which is analogous to the way people do. To address these challenges, we propose an approach, KAT, to effectively aggregate implicit and explicit knowledge during reasoning. Motivated by how humans understand the world – images and questions are centered around a few key concepts which require jointly reasoning over implicit and explicit knowledge. The main contributions of our work are as follows:

i) Knowledge extraction. We adopt two novel methods for knowledge extraction that significantly improve the quality and relevance of extracted knowledge: for implicit knowledge, we design new prompts to extract both tentative answers and supporting evidence from a frozen GPT-3 model; for explicit knowledge, we design a contrastive-learning-based explicit knowledge retriever using the CLIP model, where all the retrieved knowledge are centered around visually-aligned entities.

ii) Reasoning for e2e encoder-decoder transformer. We design a novel reasoning module in KAT to perform joint reasoning over explicit and implicit knowledge during answer generation, which is trained by using an end-to-end encoder-decoder transformer architecture.

iii) OK-VQA performance. KAT sets a new state-of-the-art on the challenging OK-VQA (Marino et al., 2019) benchmark, and significantly outperforms existing approaches.

2 Related Work

Vision-Language Transformer. Multimodal transformers have made significant progress over the past few years, by pre-trained on large-scale image and text pairs, then finetuned on downstream tasks. VisualBERT (Li et al., 2019), Unicoder-VL (Li et al., 2020a) and VL-BERT (Su et al., 2020) propose the single-stream architecture to work on both images and text. ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) propose a two-stream architecture to process images and text independently and fused by a third transformer in a later stage. While these models have shown to store in-depth cross-modal knowledge and achieved impressive performance on knowledge-based VQA (Marino et al., 2021; Wu et al., 2021; Luo et al., 2021), this type of implicitly learned knowledge is not sufficient to answer many knowledge-based questions (Marino et al., 2021). Another line of work for multimodal transformers, such as CLIP (Radford et al., 2021) or ALIGN (Jia et al., 2021), aligns visual and language representations by contrastive learning. These models achieve state-of-the-art performance on image-text retrieval tasks. Different from existing work that uses multimodal transformers as implicit knowledge bases, we focus primarily on how to associate images with external knowledge. Importantly, our model only relies on multimodal transformers learned by contrastive learning which do not require any labeled images. This makes our model more flexible in real-world scenarios.

Knowledge-based VQA. Knowledge-based VQA requires external knowledge beyond the image to answer a question. Early exploration, such as FVQA (Wang et al., 2017a), creates a fact-based VQA dataset by selecting a fact (e.g., <Cat, CapableOf, ClimbingTrees>) from a fixed knowledge base. A recent Outside knowledge VQA (OK-VQA) dataset is a more challenging dataset, covering a wide range of knowledge categories. In our work, we focus on OK-VQA due to its large-scale knowledge-based questions as well as its open-ended nature.

Recent approaches have shown great potential to incorporate external knowledge for knowledge-based VQA. Several methods explore aggregating the external knowledge either in the form of structured knowledge graphs (Garderes et al., 2020; Narasimhan et al., 2018; Li et al., 2020b; Wang et al., 2017a,b) or unstructured knowledge bases (Marino et al., 2021; Wu et al., 2021; Luo et al., 2021). In these methods, object detectors (Ren et al., 2015) and scene classifiers (He et al., 2016) are used to associate images with external knowledge. Further, external APIs, such as Google (Wu et al., 2021; Luo et al., 2021), Microsoft (Yang et al., 2021), and OCR (Luo et al., 2021; Wu et al., 2021) are used to enrich the associated knowledge. Finally, pre-trained transformer-based language models (Yang et al., 2021) or multimodal models (Wu et al., 2021; Luo et al., 2021;
While part of our approach is similar to PICa (Yang et al., 2021; Garderes et al., 2020; Marino et al., 2021) which integrates implicit knowledge for the instinctive responses and explicit knowledge for thoughtful complex processing. We define the knowledge from external knowledge bases as explicit knowledge, and the knowledge stored in large-scale language models as implicit knowledge (i.e., implicit commonsense knowledge). We describe our explicit knowledge retrieval in Section 3.2 and implicit knowledge retrieval in Section 3.3. Next, we describe the design of our knowledge reasoning module which is inspired by FiD (Karpukhin et al., 2020) to jointly reason over implicit and explicit knowledge with the e2e encoder-decoder transformer.

**Problem Formulation.** We focus on the knowledge-based VQA task in this paper. Formally, given a training dataset \( \mathcal{D} = \{(v_i, q_i, a_i)\}_{i=1}^s \), where \( v_i \) denotes the \( i^{th} \) training image, \( s \) is the total number of training images, and \( q_i \) and \( a_i \) represent the question and its corresponding answer, respectively. We use a sequence-to-sequence model that is composed of an encoder and decoder, such as T5 (Raffel et al., 2020) or BART (Lewis et al., 2019). Let \( \theta \) be the parameters of the model \( p \) that needs to be trained. Our goal is to take \( v_i \) and \( q_i \) as inputs and generate the answer \( a_i \) in an autoregressive manner.

### 3.2 Explicit Knowledge Retrieval

Given an image \( v_i \) and its question \( q_i \), grounding objects with entities from external knowledge bases is important for our model to understand both the question with referred items and the image content. Similar to previous work (Marino et al., 2021; Luo et al., 2021) which uses a subset of external knowledge, we construct a knowledge base that covers animals, vehicles, and other common objects from Wikidata (Vrandecic and Krotzsch, 2014). The details can be found in Section 4.1. We denote the constructed knowledge base as \( K \). Each entity description \( e \) from \( K \) is a concatenation of the label and its corresponding description of one entity.

The goal of our explicit knowledge retriever is to index all entity descriptions in \( d_e \)-dimensional dense representations by a dense encoder \( E_{ent}(\cdot) \), such that it can retrieve efficiently the top \( m \) entity descriptions relevant to each input image. Given an image \( v_i \), we use a sliding window with a stride to generate \( N \) image patches \( \{v_i^1, ..., v_i^N\} \). Then an image encoder \( E_{img}(\cdot) \) is applied to map each patch to a \( d_e \)-dimensional dense representation, and retrieves \( k \) entity descriptions from \( K \) of which representations are closest to the patch-level represen-
Figure 2: Our KAT model uses a contrastive-learning-based module to retrieve entities and descriptions from a knowledge resource as explicit knowledge, and uses GPT-3 to retrieve implicit knowledge with its evidence. The integration-knowledge are processed by the respective encoder transformer, and jointly with reasoning module and the decoder transformer as end-to-end training with the answer generation.

To define the similarity score between the image patch $v^i_j$ and the entity $e$, we use the inner product of their normalized representations:

$$sim(v^i_j, e) = E_{ent}(e)^T E_{img}(v^i_j).$$

(1)

In total, we retrieve the top $N \times k$ entity descriptions relevant for image $v_i$. We keep top-$m$ entity descriptions ranked by similarity scores as explicit knowledge source $x^{exp}$.

The image and entity encoders can be implemented by any multimodal transformer. We use the CLIP model (ViT-B/16 variant) (Radford et al., 2021) in our work and take the [CLS] as representations. This leads to a representation of dimension $d_e = 512$ for both image and entity encoders.

We pre-extract representations of the entity descriptions in the knowledge base $\mathcal{K}$ using the entity encoder $E_{ent}$ and index them using FAISS (Johnson et al., 2019). FAISS is a library for efficient similarity search and clustering of dense vectors.

### 3.3 Implicit Knowledge Retrieval

While our explicit knowledge retriever focuses on semantic matching between image regions and entities from external knowledge bases, it lacks a global description to infer the relations among objects. In this section, our implicit knowledge retriever aims to learn logical reasoning from images by prompting from a large-scale pre-trained language model.

PICa (Yang et al., 2021) shows that GPT-3 can act as an implicit knowledge base with a well-constructed prompt and a few examples. Inspired by PICa, we leverage GPT-3 as an implicit language knowledge base and treat VQA as an open-ended text generation task. For each image-question pair, we first convert the image $v_i$ into a textual description $C$ via a state-of-the-art image captioning model (Li et al., 2020c), and then construct a carefully designed text prompt consisting of a general instruction sentence, the textual description $C$, the question, and a set of context-question-answer triplets taken from the training dataset that are semantically most similar to the current image-question pair. We then input this text prompt to the GPT-3 model in its frozen version and obtain the output from GPT-3 as the tentative answer candidate to the current image-question pair.

In order to gain deeper insights from the implicit knowledge coming out of GPT-3 and its rationale, we also design another prompt to query GPT-3 for supporting evidence behind the tentative answer candidate that it generate. In our experiment, we observe that GPT-3 is not only good at generating tentative answers to knowledge-based visual questions, but can also often generate quite meaningful rationale and explanation text for question-answer pairs. More specifically, for each image-question pair $(v_i, q_i)$, and for a tentative answer $a$ generated by GPT-3, we construct the prompt in the form of: “(question $q_j$)? (answer $a$). This is because” to query GPT-3 for supporting evidence. We fi-
nally compile both the tentative answers and the corresponding supporting evidence from GPT-3 as implicit knowledge source $x_{imp}$.

### 3.4 Knowledge Reasoning Module

As entity descriptions from explicit knowledge are image-specific that are concerned with semantic matching of image regions, some of the entity descriptions are irrelevant to corresponding questions. Furthermore, questions from knowledge-based VQA are by definition under-specified and require more language understanding from both explicit and implicit knowledge. Inspired by FiD (Karpukhin et al., 2020), our knowledge reasoning module encode each question and knowledge pair separately, but still jointly reasons over both knowledge sources during answer generation. Our model is trained in an end-to-end fashion.

**Encoder.** We concatenate question $q_i$ with each piece of knowledge as a question-knowledge pair. We add sentinel tokens `question:`, `entity:`, and `description:` before the question, retrieved entity, and its description. Similarly, we add sentinel tokens `question:`, `candidate:`, and `evidence:` before the question, tentative answer, and its evidence. Suppose $m$ and $p$ are the number of explicit knowledge $x_{exp}$ and implicit knowledge $x_{imp}$, respectively. We use an embedding layer followed by a sequence of encoder layers to encode question-knowledge pairs separately. Then we average the token embeddings of each question-knowledge pair from the last encoder layer, which results in an embedding matrix of explicit knowledge $X_{exp} \in \mathbb{R}^{m \times d}$ and implicit knowledge $X_{imp} \in \mathbb{R}^{p \times d}$, where $d$ is the embedding dimension.

**Reasoning Module.** To jointly reason over explicit and implicit knowledge, we concatenate the embeddings of explicit and implicit knowledge to from a global representation $X \in \mathbb{R}^{(m+p) \times d}$. The cross-attention module explicitly takes as input the global representation $X$ of the encoder. Let $H \in \mathbb{R}^d$ be the output of the previous self-attention layer of the decoder. By definition (Vaswani et al., 2017), the scaled dot-product attention can be expressed as

$$Q_v = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} V \right),$$  \hspace{1cm} (2)

where queries $Q$, keys $K$, and values $V$ are computed by applying linear transformations: $Q = W_Q H, K = W_K X, V = W_V X$. The attended representation $Q_V$ is a weighted sum of the values which implies that our model performs a joint reasoning over explicit and implicit knowledge when generating answers.

**Decoder.** We pass the embeddings of explicit and implicit knowledge to a sequence of decoder layers for answer generation. We train our model through a cross entropy loss function:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{n} \log p(y_t|y_{<t}, x_{exp}; x_{imp}),$$  \hspace{1cm} (3)

where $y_t$ is predicted in an autoregressive generation manner.

### 4 Experiment

In this section, we first describe our experimental setup (§4.2-§4.3). In §4.4, we compare our model to existing state-of-the-art methods. We then conduct ablation studies to validate the effectiveness of proposed model in §5. With the best models at hand, we investigate the gains of individual models in §5.4, where we showcase that jointly reasoning over explicit and implicit knowledge serves as the key facilitator. In §5.5, we further show some failure and success cases from our model and existing state-of-the-art methods.

| Subclass                  | Number  |
|---------------------------|---------|
| Role                      | (Q214339) | 162,027 |
| Point of interest         | (Q960648) | 85,900  |
| Tool                      | (Q39546)  | 78,621  |
| Vehicle                   | (Q42889)  | 44,274  |
| Animal                    | (Q729)    | 18,581  |
| Clothing                  | (Q11460)  | 17,711  |
| Company                   | (Q891723) | 12,173  |
| Sport                     | (Q349)    | 4,233   |
| Total                     |          | 423,520 |

Table 1: We collect a subset of Wikidata that covers common objects in real-life scenarios as our external knowledge base. Here we show the statistics of these subclasses.

### 4.1 Knowledge Base Construction

We use the English Wikidata (Vrandecic and Krotzsch, 2014) dump from Sep. 20, 2021 as the source knowledge base which contains 95,870,584 entities. Each data item is stored in
Table 2: Results of OK-VQA comparing to standard baselines show that our KAT model achieves state-of-the-art performance on OK-VQA full testing set comparing to knowledge VLP baselines.

| Method                        | Knowledge Resources               | Accuracy (%) |
|-------------------------------|-----------------------------------|--------------|
| Q only (Marino et al., 2019) | -                                 | 14.93        |
| MLP (Marino et al., 2019)    | -                                 | 20.67        |
| BAN (Marino et al., 2019)    | -                                 | 25.1         |
| BAN+AN (Marino et al., 2019) | Wikipedia                         | 25.61        |
| MUTAN (Marino et al., 2019)  | -                                 | 26.41        |
| BAN+KG-AUG (Li et al., 2020b)| Wikipedia+ConceptNet              | 26.71        |
| MUTAN+AN (Marino et al., 2019)| Wikipedia                      | 27.84        |
| ConceptBERT (Garderes et al., 2020)| ConceptNet  | 33.66        |
| KRISP (Marino et al., 2021)  | Wikipedia+ConceptNet              | 38.35        |
| Visual Retriever-Reader (Luo et al., 2021)| Google Search   | 39.2         |
| MAVEx (Wu et al., 2021)      | Wikipedia+ConceptNet+Google Images| 39.4         |
| PICa-Base (Yang et al., 2021) | GPT-3 (175B)                     | 43.3         |
| PICa-Full (Yang et al., 2021) | GPT-3 (175B)                     | 48.0         |
| Knowledge-T5 (ours)          | Wikidata+GPT-3 (175B)            | 51.97        |
| KAT (ours)                   | Wikidata+GPT-3 (175B)            | 54.41        |

Table 2: Results of OK-VQA comparing to standard baselines show that our KAT model achieves state-of-the-art performance on OK-VQA full testing set comparing to knowledge VLP baselines.

a structured format through property-value pairs. Properties are objects and have their own Wikidata pages with labels, aliases, and descriptions. We extract a subset that covers common objects in real-world scenarios. We remove the entities if string labels or corresponding descriptions are empty or non-English. This results in 423,520 entity triplets in the end (e.g., <Q2813, Coca-Cola, carbonated brown colored soft drink>) (See Table 1).

4.2 Dataset

OK-VQA (Marino et al., 2019) is currently the largest knowledge-based VQA dataset. The questions are crowdsourced from Amazon Mechanical Turkers and require outside knowledge beyond the images. The dataset contains 14,031 images and 14,055 questions covering a variety of knowledge categories. We follow the standard evaluation metric recommended in the VQA challenge (Antol et al., 2015):

$$\text{Acc}(\text{ans}) = \min\left(\frac{\#\{\text{humans said ans}\}}{3}, 1\right) \quad (4)$$

4.3 Implementation Details

We use the pre-trained CLIP model (ViT-B/16 variant) (Radford et al., 2021) for similarity calculation. For the knowledge reasoning module, we initialize our model with the pre-trained T5 models (Raffel et al., 2020). We compare two model sizes, base and large, containing 220M and 770M parameters respectively. We fine-tune the models on OK-VQA dataset, using AdamW (Loshchilov and Hutter, 2017). We use a learning rate of $2 \times 10^{-5}$ and warm up over $2K$ iterations and train for $10K$ iterations. Limited to computation resources, we set the number of retrieved entities as 40. The model is trained with a batch size of 32, using 16 V100 32Gb. Unless otherwise specified, all results reported in this paper as KAT use this model which we found to perform best. We evaluate our generated answers with ground-truth after normalization. The normalization step consists of lowercasing, and removing articles, punctuation and duplicated whitespace (Chen et al., 2017; Lee et al., 2019).

4.4 Comparison with Existing Approaches

We compare our model against existing approaches for knowledge-based VQA and the results are summarized in Table 2. Our model outperforms state-of-the-art methods by significant margins. We compare our model with existing approaches from two aspects. (1) If we only consider using explicit knowledge (i.e., knowledge bases), our model achieves 44.25%, 4.85% higher than previous state-of-the-art methods (e.g., KRISP or MAVEx) which involve several object detection/classification modules. Our model is slightly better than PICa-Base which is a plain version of PICa-Full without example engineering. It implies that our single, unified architecture can effectively associate images with
external knowledge. (2) If we take the implicit knowledge from GPT-3 as the input, our model outperforms PICa-Full by 6.41% which indicates it is important to aggregate knowledge of different types when generating answers. The detailed comparison can be found in Table 5.

5 Ablation Study

To unpack the performance gain and understand the impact of different components, we ablate and compare different model architectures, types of knowledge and the number of explicit knowledge.

| Model architecture | Knowledge | Accuracy (%) |
|--------------------|-----------|--------------|
| T5-base            | Explicit  | 40.93        |
| T5-large           | Implicit  | 44.25        |
|                    |           | 47.60        |
|                    |           | 50.58        |
|                    |           | 54.41        |

Table 3: Ablation study on model architectures and types of knowledge. Our experiments show that larger model has more capacity for implicit knowledge reasoning and jointly reasoning over both knowledge sources has a consistent improvement with baselines.

5.1 Model Architecture & Types of Knowledge

As shown in Table 5, our model using T5-large as the backbone shows a consistently improvement than that using T5-base. This demonstrates larger model has more capacity for implicit knowledge reasoning. The combination of explicit and implicit knowledge achieves a performance gain of around 4%, supporting the intuition that these two types of knowledge with different focuses together provide complementary pieces of knowledge.

5.2 Effectiveness of Knowledge Reasoning Module

| Method       | Accuracy (%) |
|--------------|--------------|
| Knowledge-T5 | 51.97        |
| KAT          | 54.41        |

Table 4: Comparison between our model with knowledge-T5 which uses the concatenated knowledge as inputs without the knowledge reasoning module.

To verify the effectiveness of our knowledge reasoning module, we use a vanilla T5 module as our baseline which contains a standard encoder-encoder model (Raffel et al., 2019). We concatenate implicit and explicit knowledge as a sentence and adopt a maximum length of 256. The model is denoted as knowledge-T5. We use the same hyper-parameters to train our baseline model. The comparison is shown in Table 4. Our results show that simply concatenating both knowledge sources may introduce noise to relevant knowledge. Our model adaptively attend different knowledge sources for answer generation that can reduce the influence of irrelevant knowledge.

5.3 The Role of the Extractor and Scale of Explicit Knowledge

Figure 3: Our model achieves consistent improvement when aggregating more entity descriptions from an external knowledge base. However, as CLIP-ViT/16 and RN50 are very different explicit knowledge retrieval backbones we see the choice of backbone and number of sources to include are intimately related. Here we use T5-base as the backbone for demonstration.

From Figure 3 we can see, the performance of our model is directly affected by the size of retrieved explicit knowledge.

When only considering the implicit knowledge (i.e., the number of retrieved entities is 0), our model achieves 47.6% which is slightly worse than PICa-Full baseline. It indicates that solely increasing model complexity cannot improve the performance. This also demonstrates the importance of explicit knowledge.

Our model shows a consistent improvement by incorporating explicit knowledge (i.e., 3% improvement when considering 40 entity descriptions). While a more extensive knowledge set may include more distracting knowledge, retrieved entity descriptions can share either visually or se-
mantically similar knowledge as the relevant ones. Thus this can massively reduce the search space and/or reduce spurious ambiguity.

We compare different explicit knowledge retrieval module. Though ViT/16 has a large classification improvement over ResNet-50 (e.g., 6.9% on ImageNet) (Radford et al., 2021), there is a less gap between these two backbones. As the number of retrieved entities increases, our knowledge reasoning module can further migrate this gap by adaptively attending to different explicit knowledge.

5.4 Category Results on OK-VQA

Here we present quantitative analyses to illustrate how explicit and implicit knowledge influence the final predictions. Based on the types of knowledge required, questions in OK-VQA are categorized into 11 categories and the accuracy results of each category are reported in Table 5. We re-train our model under the same settings with only either explicit or implicit knowledge, denoted as “exp” and “imp” respectively.

For most categories, the model using only explicit knowledge performs worse than that using only implicit knowledge. As implicit knowledge comes from the results of state-of-the-art object detection, image captioning models and logical evidence by prompting GPT-3. While explicit knowledge is retrieved based on semantic matching between images and entities from knowledge bases, it contains richer but more distracting knowledge. Note that using explicit knowledge performs better for category “Brands, Companies, and Products” and “Weather and Climate”. It indicates that accurately recognizing objects with fine-grained attributes in the images is important for these categories to answer corresponding questions.

In all categories, our model outperforms either explicit-only or implicit-only model by a large margin. Specifically, on categories “Geo, History, Lang, and Culture”, “Sports and Recreation” and “Brands, Companies, and Products”, our model provides the most significant improvements. This observation illustrates that jointly reasoning over explicit and implicit knowledge serves as the key facilitator to our model.

5.5 Qualitative Analysis

As analyzed in previous sections, jointly reasoning over both knowledge sources during answer generation can improve the explicit-only and implicit-only model by large margins. Figure 5 shows two examples comparing the answers generated by different models along with retrieved knowledge. The left example shows that while explicit knowledge retrieved from the knowledge base contains the necessary entity descriptions for knowledge reasoning, it fails to generate the answer which requires the relation between bench and Coca Cola logos. On the other side, implicit knowledge retrieved from GPT-3 can only infer the bench is painted red without recognizing the logo on it. By jointly considering both knowledge sources, our model can associate the color of Coca Cola logo with the painted color of the bench which derives the correct answer. The right example shows that though explicit knowledge does not contain the right entity descriptions, it provides visually similar descriptions of this sport which further constrains the search space of our model and verifies the correctness of the implicit knowledge.

6 Conclusion

In this paper, we demonstrate a conceptually simple yet effective approach for knowledge-based VQA. Our model shows superior performance compared with previous work. Nonetheless, several challenging and promising directions can be considered in the future. First, most existing knowledge-based VQA models use several object detection or classification modules to incorporate external knowledge which makes the models complicated and error-prone. How to align image regions with meaningful external semantics deserves further investigation. Second, recent advances on large-scale vision-language models have shown promising results on knowledge reasoning tasks. How to aggregate such

| Question Type                     | Exp | Imp | Ours | ∆    |
|-----------------------------------|-----|-----|------|------|
| Plants and Animals                | 42.2| 51.5| 54.7 | +3.2 |
| Science and Technology            | 44.4| 43.3| 52.8 | +8.3 |
| Sports and Recreation             | 49.7| 53.8| 60.4 | +6.7 |
| Geo, History, Lang, and Culture   | 45.6| 45.4| 55.8 | +10.2|
| Brands, Companies, and Products   | 41.7| 38.2| 48.5 | +6.8 |
| Vehicles and Transportation       | 41.5| 42.9| 51.3 | +8.4 |
| Cooking and Food                  | 47.9| 47.7| 52.7 | +4.8 |
| Weather and Climate               | 51.7| 46.3| 54.8 | +3.1 |
| People and Everyday               | 43.1| 44.4| 51.5 | +7.1 |
| Objects, Material and Clothing    | 42.9| 45.4| 49.3 | +3.9 |
| Other                             | 41.5| 50.2| 51.2 | +1.0 |

Table 5: Accuracy of question types in OK-VQA full testing set. Our models outperforms exp and imp models by a large margin on all categories. (exp: explicit-only model and imp: implicit-only model)
Recent advances with external knowledge bases is one future direction.

Acknowledgement

We are especially grateful to Jianwei Yang, Daniel McDuff, Dragomir Radev, Harkirat Behl, Hao Chen, Chunyun Li, Baolin Peng for their suggestion, comments, and their pointers to the modeling generation and literature. We thank Zhe Gan, Zhengyuan Yang, Lijuan Wang from Cognition Service team of Microsoft for their work and their generous helps and feedback for the project. We are grateful to Subhojit Som from Turing team of Microsoft for his enormous support and encouragement. We appreciate Kenneth Marino from DeepMind, and Roozbeh Mottaghi from the AllenAI for their supporting and helps of the work.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In CVPR.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In The semantic web, pages 722–735. Springer.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.

François Garderes, Maryam Ziaeeafard, Baptiste Abeeloo, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In EMNLP.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In CVPR.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. In EACL.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yuhhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. IEEE Transactions on Big Data.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In EMNLP.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. arXiv preprint arXiv:1906.00300.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation
for knowledge-intensive nlp tasks. arXiv preprint arXiv:2005.11401.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. Proceedings of AAAI.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020b. Boosting visual question answering with context-aware knowledge aggregation. In ACM MM.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Liujian Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In ECCV.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211–226.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In NeurIPS.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. arXiv preprint arXiv:2109.04014.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In CVPR.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In CVPR.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. NeurIPS.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.
A. Appendix

Figure 5: Examples from OK-VQA dataset that our model generates correct answers by jointly reasoning over both implicit and explicit knowledge. (exp: predictions by using explicit knowledge only and imp: predictions by using implicit knowledge only)