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

Anonymous ACL submission

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 complementary 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 - **K**nowledge **A**ugmented **T**ransformer (KAT) - which achieves a strong state-of-the-art result (+6% absolute) on the open-domain multimodal task of OK-VQA. Our approach integrates implicit and explicit knowledge in an encoder-decoder architecture, while still jointly reasoning over both knowledge sources during answer generation. Additionally, explicit knowledge integration improves interpretability of model predictions in our analysis.

1 Introduction

There has been a revival of interest in knowledge-intensive tasks which require an external knowledge source for humans to perform. Many applications in real-world scenarios, such as autonomous AI agents, need to seamlessly integrate implicit (*i.e.*, commonsense) and explicit knowledge (*e.g.*, Wikidata) to answer questions. In this work, we investigate how to effectively integrate implicit and explicit knowledge for reasoning. Tasks like Outside Knowledge Visual Question Answering (OK-VQA) (Marino et al., 2019) require that models use knowledge not present in the input to answer questions, making it an ideal test bed for investigating this implicit-explicit knowledge trade-off.

Consider the examples from OK-VQA shown in Figure 1. To answer the question in the left example, the system needs to both ground *organism* to bird through explicit knowledge and then apply the implicit knowledge *birds evolved from reptiles* to answer the question. Similarly for the question in the right example, the system needs to recognize boats and harbor and requires the implicit knowledge *anchors are used to stop boats from moving*. A 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., 2022; Garderes et al., 2020). Existing methods first retrieve external knowledge from 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.

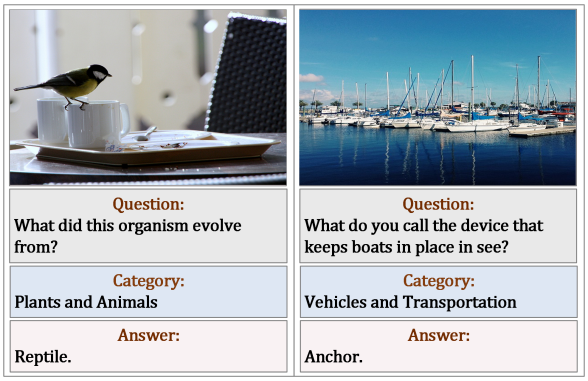

Figure 1: Examples of knowledge-based VQA that requires external knowledge. Success on this task requires not only visual recognition, but also logical reasoning to incorporate external knowledge about the world.

However, most existing approaches have several drawbacks. First, explicit knowledge retrieved using keywords from questions or image tags may be too generic, which leads noise or irrelevant knowledge during knowledge reasoning. Second, existing work mainly focuses on explicit knowledge which is often in the form of encyclopedia articles or knowledge graphs. While this type of knowl-

edge can be useful, it is insufficient to answer many knowledge-based questions. As shown in Figure 1, questions require the system to jointly reason over explicit and implicit knowledge, which is analogous to the way humans do. To address these challenges, we propose an approach, **KAT**, to effectively integrate implicit and explicit knowledge during reasoning. 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 in an encoder-decoder transformer. We design a novel reasoning module in *KAT* to perform jointly 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 ta 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., 2022; 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. Some Knowledge-based visual language tasks 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 a 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., 2022; 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., 2022; Luo et al., 2021), Microsoft (Yang et al., 2022), and OCR (Luo et al., 2021; Wu et al., 2022) are used to enrich the associated knowledge. Finally, pre-trained transformer-based language models (Yang et al., 2022) or multimodal models (Wu et al., 2022; Luo et al., 2021; Wu et al., 2022; Garderes et al., 2020; Marino et al., 2021) are leveraged as implicit knowledge bases for answer predictions.

Different from previous approaches, Our work aims to develop a single, unified architecture, by jointly reasoning over explicit and implicit knowledge to augment generative language models. While part of our approach is similar to PICa (Yang et al., 2022) which considers GPT-3 as implicit knowledge base, our model takes one step further

by showing that how explicit and implicit knowledge can be integrated during knowledge reasoning. Another similar work Vis-DPR (Luo et al., 2021) collects a knowledge corpus from training set by Google Search which is specific to a certain dataset. Our proposed model is more generic by collecting entities from Wikidata and not limited to the training set.

Open-Domain Question Answering (ODQA). ODQA is the NLP task of answering general domain questions, in which the evidence is not given as input to the system. Several approaches (Chen et al., 2017; Karpukhin et al., 2020) propose to predict the answers by first retrieving support document from Wikipedia, before extracting answers from the retrieved document. Recent works (Izacard and Grave, 2020; Lewis et al., 2020b) combine text retrieval models with language generative models which achieve state-of-the-art performance on knowledge-intensive natural language processing tasks. Similar to these works as part of our method, we extend this framework to VQA domain and show the effectiveness of aggregating explicit and implicit knowledge for knowledge-based VQA.

3 Method

3.1 Overview

When humans reason about the world, they process multiple modalities and combine external and internal knowledge related to these inputs. Inspired by this idea, we introduce a new *KAT* approach. The overview of the proposed *KAT* model is shown in Figure 2. We define the knowledge from explicit knowledge bases as the explicit knowledge, and the knowledge stored in large-scale language models as the implicit knowledge (*i.e.*, implicit commonsense knowledge). We describe the retrieval method of our explicit knowledge (§3.2) and the retrieval method of our implicit knowledge (§3.3). Next, we introduce the details of our knowledge reasoning module which jointly reasons over both explicit and implicit knowledge (§3.4).

Problem Formulation. We apply our *KAT* on OK-VQA task in this paper. Formally, given a training dataset $\mathbb{D} = \{(v_i, q_i, a_i)\}_{i=1}^{s}$, where v_i denotes the ith training image; s is the total number of the training images; q_i and a_i represent the ith question and its corresponding answer, respectively. We use a sequence-to-sequence model that is composed of an encoder and a decoder, which is a comparison method of T5 (Raffel et al., 2020) or BART (Lewis et al., 2020a). Let θ 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 auto-regressive manner.

3.2 Explicit Knowledge Retrieval

Given an image v_i and corresponding question q_i, it is important to ground image regions with fine-grained descriptions, which is conducive to understanding both the image content and the question with the referred items. Existing approaches (Radford et al., 2021; Jia et al., 2021) on OK-VQA apply object detectors to generate image tags which are used for explicit knowledge retrieval. Such image tags can be generic and have a limited vocabulary size, leading noise or irrelevant knowledge. Motivated by the recent progress of visual-semantic matching approaches (Radford et al., 2021; Jia et al., 2021), we leverage a contrastive-learning-based model to associate image regions with external knowledge bases.

Similar to the previous work (Marino et al., 2021; Luo et al., 2021) which uses a subset of external knowledge, we construct an explicit knowledge base that covers the 8 categories of 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 $\mathcal{K}$. Each knowledge entry e from $\mathcal{K}$ is a concatenation of the entity and its corresponding description.

The goal of our explicit knowledge retriever is to index all knowledge entries in d_r-dimensional dense representations by a dense encoder $E_{ent}(\cdot)$, such that it can efficiently retrieve the top m knowledge entries 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_r-dimensional dense representation, and retrieves k knowledge entries from $\mathcal{K}$ whose representations are closest to the patch-level representation. 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:

$$sim(v_i^j, e) = E_{ent}(e)^T E_{img}(v_i^j). \quad (1)$$

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

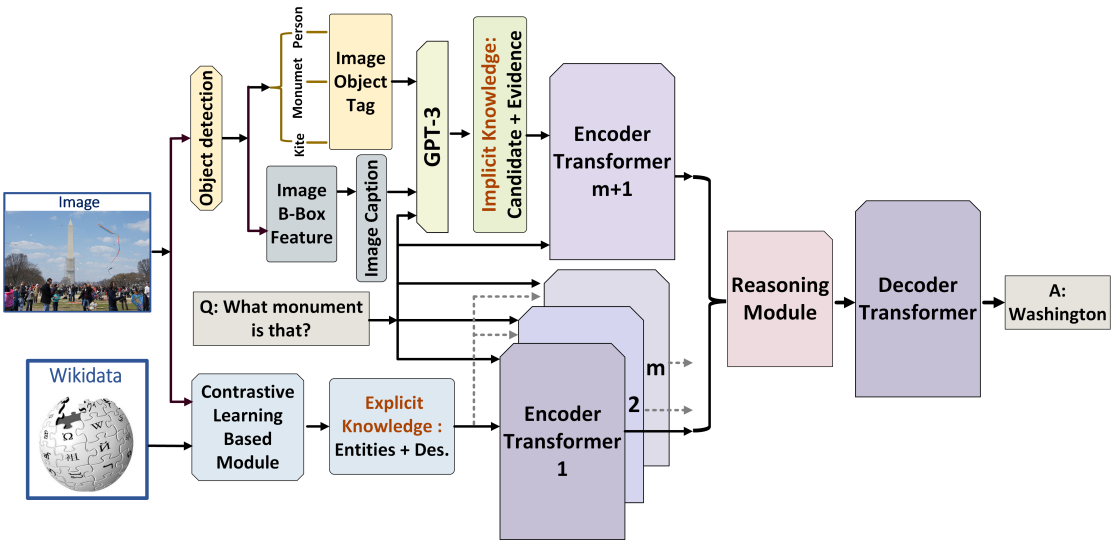

Figure 2: Our *KAT* model uses a contrastive-learning-based module to retrieve knowledge entries from an explicit knowledge base, and uses GPT-3 to retrieve implicit knowledge with supporting evidence. The integration of knowledge is processed by the respective encoder transformer, and jointly with reasoning module and the decoder transformer as an end-to-end training with the answer generation.

In principle, the image and knowledge entry 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. We pre-extract representations of the knowledge entries in the knowledge base $\mathcal{K}$ using the entity encoder E_{ent} and index them using FAISS (Johnson et al., 2019). The qualitative example for the extracting explicit knowledge model is presented in Appendix A.

3.3 Implicit Knowledge Retrieval

While our explicit knowledge retriever focuses on semantic matching between image regions and knowledge entries, it lacks implicit commonsense knowledge (*e.g.*, *Lemons are sour*) which is usually stored in large-scale language models. In this section, we retrieve implicit knowledge with supporting evidence by prompting from a large-scale pre-trained language model.

We design our implicit knowledge retriever with inspirations from the previous work (Yang et al., 2022). 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 a re semantically most similar to the current image-question pair (see Figure 7 in Appendix B for a concrete example). 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.

To gain deeper insights from the implicit knowledge coming out of GPT-3 and its rationale, we design another prompt to query GPT-3 for supporting evidence behind the tentative answer candidate that it generates. 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_i)? (answer a). This is because" to query GPT-3 for supporting evidence (see Figure 6 in Appendix B for a concrete example). We finally compile both the tentative answers and the corresponding supporting evidence from GPT-3 as implicit knowledge source x^{imp}.

3.4 KAT Model

As showed in the Figure 2, the explicit knowledge entries are from an image, which are concerned with semantic matching of the image regions. These knowledge entries could be noisy or irrelevant to its corresponding question. Moreover, some of the supporting evidence prompted from GPT-3 is generic or not related to image content. Simple concatenation of different knowledge may introduce noise during model training. We design a knowledge reasoning module with inspirations from the previous work (Karpukhin et al., 2020).

Our knowledge reasoning module encodes each question and knowledge pair separately, and jointly reason over both explicit and implicit knowledge when generating an answer.

Encoder. We concatenate question q_i with each knowledge as a question-knowledge pair. Firstly, we add sentinel tokens `question:`, `entity:` and `description:` before the question, the retrieved entity, and its description separately. Similarly, we add sentinel tokens `question:`, `candidate:` and `evidence:` before the question, the tentative answer, and its evidence. Secondly, we use an embedding layer followed by a sequence of encoder layers to encode the question-knowledge pairs separately. 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, m and p are the embedding dimension, the number of explicit knowledge x^{exp}, and the number of implicit knowledge x^{imp}, respectively.

Reasoning Module. To jointly reason over implicit and explicit knowledge, we concatenate the embeddings of explicit and implicit knowledge form a global representation $X \in \mathbb{R}^{(m+p) \times d}$. The cross-attention module takes the global representation X of the encoder as the input. 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:

$$Q_v = softmax(\frac{QK^T}{\sqrt{d}})V, \quad (2)$$

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, and implies that our model performs a joint reasoning over explicit and implicit knowledge when generating answers.

Decoder. We feed the embeddings of explicit and implicit knowledge to a sequence of decoder layers for answer generation. We train our model with a cross-entropy loss:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{n} \log p_\theta(y_t | y_{<t}, x^{exp}; x^{imp}), \quad (3)$$

where y_t is predicted autoregressively.

4 Experiment

We first describe the knowledge base construction (§4.1) our experimental setup (§4.2-§4.3). In §4.4, we compare our model to existing approaches.

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 explicit knowledge base. Above are 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 explicit knowledge source base which contains $95,870,584$ entities. Each data item is stored in a structured format constituted of 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 all entities whose string labels or corresponding descriptions are empty or non-English. This results in a total of $423,520$ 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 in order to be answered correctly. The dataset contains $14,031$ images and $14,055$ questions covering a variety of knowledge categories. We follow the standard evaluation metric recommended by the VQA challenge (Antol et al., 2015).

4.3 Implementation Details

For the knowledge reasoning module, we initialize our model with the pre-trained T5 model (Raffel et al., 2020). We compare two model sizes, base

	Method	Knowledge Resources	Acc (%)
No knowledge	Q only (Marino et al., 2019)	-	14.93
	Vanilla T5	-	18.56
	MLP (Marino et al., 2019)	-	20.67
	BAN (Marino et al., 2019)	-	25.1
	MUTAN (Marino et al., 2019)	-	26.41
With knowledge	BAN+AN (Marino et al., 2019)	Wikipedia	25.61
	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
	Vis-DPR (Luo et al., 2021)	Google Search	39.2
	MAVEx (Wu et al., 2022)	Wikipedia+ConceptNet+Google Images	39.4
GPT-3	PICa-Base (Yang et al., 2022)	Frozen GPT-3 (175B)	43.3
	PICa-Full (Yang et al., 2022)	Frozen GPT-3 (175B)	48.0
	KAT-explicit (w/ reasoning)	Wikidata	44.25
	KAT-implicit (w/ reasoning)	Frozen GPT-3 (175B)	49.72
	KAT (w/o reasoning)	Wikidata+Frozen GPT-3 (175B)	51.97
	KAT	Wikidata+Frozen GPT-3 (175B)	**54.41**

Table 2: Results of OK-VQA comparing to standard baselines show that our KAT (large size) model achieves state-of-the-art performance on OK-VQA full testing set. It is important (see table sections) to compare methods based on their access to increasingly large implicit sources of knowledge and utilization of explicit knowledge sources. Our four KAT models variants make the relative importance of these decisions explicit.

and large, each containing $220M$ and $770M$ parameters respectively. We fine-tune the models on OK-VQA dataset, using AdamW (Loshchilov and Hutter, 2019). We use a learning rate of $3e-5$ to warm up for $2K$ iterations and train for $10K$ iterations. Limited by the computational resources, we set the number of retrieved entities to 40. The model is trained with a batch size of 32, using 16 V100 GPUs with 32Gb of memory each. Unless otherwise specified, all results reported in this paper as KAT use this model which we found to perform best. We evaluate our predictions 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 on the OK-VQA dataset 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, our model achieves 44.25% which is 4.85% and 5.9% higher than MAVEx and KRISP, respectively. Our model uses contrastive-learning-based model to extract knowledge, leaving headroom by incorporating supervised pre-trained models, such as pre-trained object detectors. It should be noted that our proposed model is working on a more challenging problem. As the generated answer could contain an arbitrary number of words from the entire vocabulary. 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 the explicit knowledge base. (2) If we take the implicit knowledge from GPT-3 as the additional input, our model outperforms PICa-Full by 6.41% which indicates it is important to integrate knowledge of different types when generating answers. The detailed comparison can be found in Table 3.

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 (%)
Base	Large	Explicit	Implicit	
√				18.56
√		√		40.93
	√	√		44.25
√			√	47.60
	√		√	49.72
√		√	√	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.

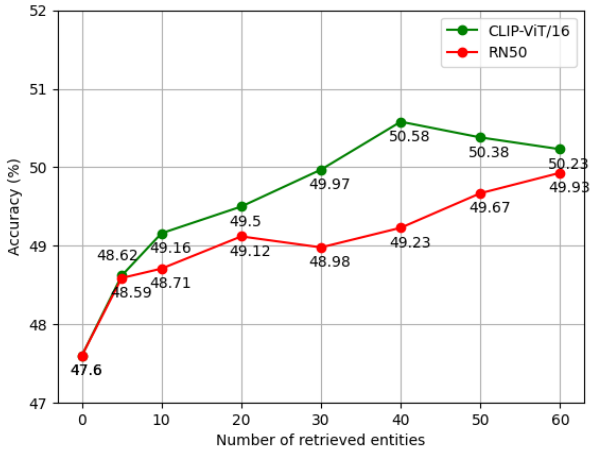

Figure 3: Our model achieves consistent improvement when aggregating more knowledge entries from an explicit knowledge base. However, as CLIP-ViT/16 and RN50 are very different explicit knowledge retrieval backbones we see the choice of backbone and number of sources to include are intimately related.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

Specifically, as shown in Table 3, our KAT-large shows a consistent improvement over using KAT-base. This larger model has more capacity for implicit knowledge reasoning. The integration of explicit and implicit knowledge achieves a performance gain of $\sim 4\%$, supporting the intuition that these two types of knowledge provide complementary pieces of knowledge.

5.1 Effectiveness of Knowledge Reasoning

Method	Accuracy (%)
KAT (w/o reasoning)	51.97
KAT	**54.41**

Table 4: Comparison with KAT (w/o reasoning) which uses the concatenated knowledge as inputs without the knowledge reasoning module.

To verify the effectiveness of our knowledge reasoning module, we use a KAT without the knowledge reasoning module which is denoted as KAT (w/o reasoning). This model concatenates explicit and implicit knowledge as a sentence and adopts a maximum length of 256 tokens. We train this variant with the same parameter settings. As shown in Table 4, simply concatenating knowledge sources is 2.43% lower than our proposed model. It indicates that KAT (w/o reasoning) may introduce noise to relevant knowledge during encoding. Our model adaptively attend different knowledge sources for answer generation that can reduce the influence of irrelevant knowledge.

5.2 Extracting Explicit Knowledge

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 more explicit knowledge. While a more extensive knowledge set may include more distracting knowledge, retrieved knowledge entries can share either visually or semantically 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.

5.3 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.

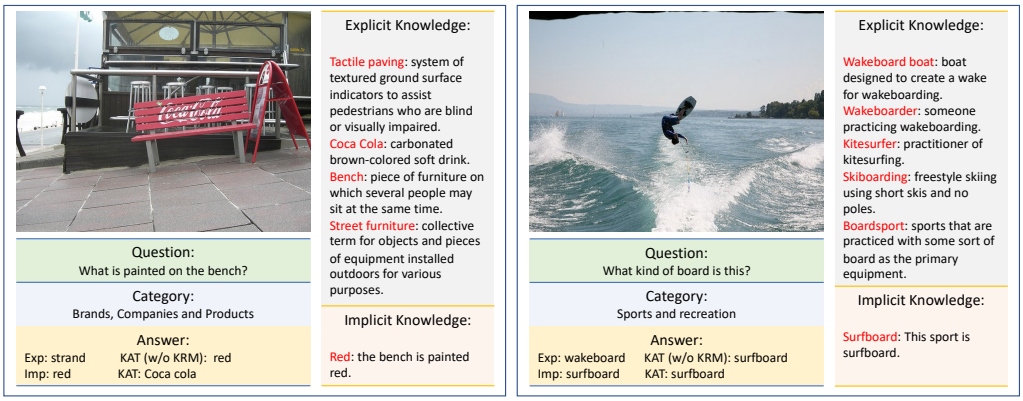

Figure 4: Two examples from OK-VQA dataset that our model generates correct answers by jointly reasoning over both implicit and explicit knowledge. (exp: predictions by using explicit knowledge only and imp: predictions by using implicit knowledge only). More examples and analysis can be found in Appendix C.

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 supporting 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 descriptions in the images is important for these categories to answer corresponding questions.

Question Type	Exp	Imp	**Ours**	Δ
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)

5.4 Qualitative Analysis

Analyzed in previous sections, jointly reasoning over both knowledge sources during answer generation improves the explicit-only and implicit-only models by large margins. Figure 4 shows two examples comparing 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 knowledge entries for 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, failing to recognize its logo. 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 knowledge entries, 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

This paper takes a step towards understanding the complementary role of implicit knowledge gained from continuing to scale models and explicit knowledge from structured knowledge bases. Importantly, it appears that there is headroom in both directions (i.e. improving retrieval and reasoning). Our conceptually simple yet effective approach for knowledge-based VQA makes these relationships explicit while still achieving a significant improvement against state-of-the-art results. Additional challenges remain, for example how best to align image regions with meaningful external semantics deserves and how to efficiently and accurately integrate multiple knowledge bases.

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 Ziaeefard, Baptiste Abeloos, 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, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

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. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

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

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, Lijuan 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. 2019. Decoupled weight decay regularization. In *ICLR*.

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. In *EMNLP*.

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. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Denny Vrandecic and Markus Krotzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017a. Fvqa: Fact-based visual question answering. *TPAMI*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2017b. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *AAAI*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*.

Appendix

A Figure of Explicit Knowledge

In this section, we show one example Figure 5 to extract explicit knowledge from an image, which use the CLIP model to conduct the explicit knowledge retrieval with the image and a wiki knowledge base.

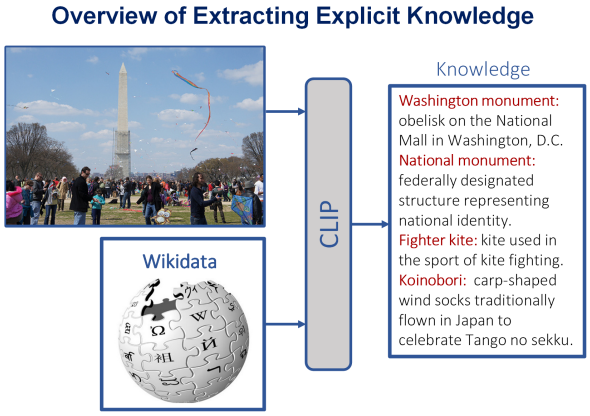

Figure 5: Overview of the explicit knowledge extraction. We use a sliding window to crop image regions and retrieve knowledge entries from an explicit knowledge base by CLIP.

B Examples of Prompts of Implicit Knowledge

In this Section B of the Appendix, we show two concrete examples (Figure 6 and Figure 7) for the prompts that constructed to query GPT-3 for implicit knowledge in our experiments:

```
Query Prompt:
What is the shape? circle. This is because

Response from GPT-3:
'the circle is the only shape that has no corners.'
```

Figure 6: An example of the evidence of rationale that we obtain from GPT-3 by using a combination of question and answer candidate to query it.

C Analysis on More Examples

In this section, we showcase more predictions from variants of our model. As shown in Figure 8, we analyze the predictions based on different type of knowledge from several aspects:

Effectiveness of explicit knowledge retriever. Our explicit knowledge retriever can retrieve fine-grained knowledge entries from the explicit knowledge base, such as *golden retriever* (a fine-grained breed of dogs), *cucumber sandwich* (a specific type of sandwich) and *Macbook Pro* (a specific model

```
Please answer the question according to the above context and the provided knowledge if given.
===
Context: a keyboard sitting in front of a computer monitor
===
Q: Is this a laptop or desktop?
A: desktop
===
......
===
Context: This picture shows one man on ski's and one youngster on a snowboard.
===
Q: Who is a famous participant in this sport?
A: bode miller
===
Context: a black motorcycle parked in a parking lot.. motorcycle, parked, parking, ground, outdoor, wheel, fender, brake, vehicle, exhaust system, land vehicle, tire, muffler, motorcycling, rim, suspension, automotive exhaust, auto part, motorcycle racing, motorcycle fairing, automotive tire, fuel tank, vehicle brake, disc brake, cruiser, motorbike, bike

Knowledge:
A Benelli Adiva is a motor scooter model.
A Yamaha TMAX is a scooter.
A mofa is a moped with a maximum speed of 20 to 25 km/h.
A motorcycle based vehicle is a motor vehicle with two or more wheels, based on cycle technology.
A sport bike is a motorcycle designed for sporty riding regardless of other characteristics.
A motorcycle development rider is a person who for their profession participates in tests to develop a motorcycle.
A Bravo is a Piaggio moped.
A Suzuki X-90 is a motor vehicle.
A Polaris Industries is a Designs, engineers and manufactures powersports vehicles.
A Monark is a Bicycle, moped and motorcycle manufacturer.
===
Q: What sport can you use this for?
A:
```

Figure 7: An example of the prompts that we use to query GPT-3 in our knowledge-aumented GPT-3 query system.

of Apple products). These fine-grained entities are hardly obtained from existing object detection models, which can constraint the search space of our model and are beneficial to our answer generation process.

Effectiveness of implicit knowledge retriever. Our implicit knowledge retriever can retrieve supporting evidence from GPT-3, such as *Thomas: the train is named after the man who designed it.* and *Refrigerator: the refrigerator is used to keep food cold*. These kinds of knowledge are highly related to commonsense knowledge which needs further inference based on entities and provide complementary explanation to explicit knowledge.

Answer generation & classification. As most previous work on OK-VQA task, such as KRISP or MAVEx method, implement OK-VQA as a classification task. The prediction vocabulary is dataset-specific and assumes the training and test set are sharing a similar vocabulary. The limitation of these methods is the generalization ability. Our proposed KAT model initialize and perform the OK-VQA as a kind of open-end generation task. From these examples we found, our model can generate answers like *Iphone* or *Hercules* that are visually and semantically reasonable. Our proposed

Figure 8: More examples from OK-VQA dataset that our model generates answers by jointly reasoning over both implicit and explicit knowledge.

novel KAT model using the explicit and implicit knowledge is designed to enhance semantic alignment and generate representations with stronger knowledge-awareness.