Expert Systems With Applications 212 (2023) 118669

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Image captioning for effective use of language models in knowledge-based visual question answering

Ander Salaberria *, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, Eneko Agirre

HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country (UPV/EHU), M. Lardizabal 1, Donostia 20018, Basque Country, Spain

ARTICLE INFO

Code link: https://github.com/salanueva/CBM

Keywords:
Visual question answering
Image captioning
Language models
Deep learning

ABSTRACT

Integrating outside knowledge for reasoning in visio-linguistic tasks such as visual question answering (VQA) is an open problem. Given that pretrained language models have been shown to include world knowledge, we propose to use a unimodal (text-only) train and inference procedure based on automatic off-the-shelf captioning of images and pretrained language models. More specifically, we verbalize the image contents and allow language models to better leverage their implicit knowledge to solve knowledge-intensive tasks. Focusing on a visual question answering task which requires external knowledge (OK-VQA), our contributions are: (i) a text-only model that outperforms pretrained multimodal (image-text) models of comparable number of parameters; (ii) confirmation that our text-only method is specially effective for tasks requiring external knowledge, as it is less effective in standard a VQA task (VQA 2.0); and (iii) our method attains results in the state-of-the-art when increasing the size of the language model. We also significantly outperform current multimodal systems, even though augmented with external knowledge. Our qualitative analysis on OK-VQA reveals that automatic captions often fail to capture relevant information in the images, which seems to be balanced by the better inference ability of the text-only language models. Our work opens up possibilities to further improve inference in visio-linguistic tasks.

1. Introduction

Most visio-linguistic tasks are framed in such a way that all the necessary information to solve them is in the images and texts provided in the dataset. That is the case of visual question-answering (VQA) (Antol et al., 2015) or visual entailment (Xie, Lai, Doran, & Kadav, 2019). In addition, some tasks require access to external knowledge in order to solve them. In this work we dive in *Outside Knowledge VQA* (OK-VQA) (Marino, Rastegari, Farhadi, & Mottaghi, 2019), where the image content is not sufficient to answer the questions. Contrary to self-contained VQA tasks, which can be solved grounding images and text alone, these tasks require methods that leverage external knowledge resources and are able to do inference on that knowledge.

External knowledge useful for OK-VQA can be broadly classified into two categories, according to Marino, Chen, Parikh, Gupta, and Rohrbach (2021): (i) symbolic knowledge, which can be represented using graphs, for example ConceptNet (Speer, Chin, & Havasi, 2017), and (ii) implicit knowledge, which is encoded in the weights of neural networks trained in different datasets. Supporting the later case, transformer-based language models (LM) pretrained in large corpora like BERT (Devlin, Chang, Lee, & Toutanova, 2019) have been successfully used as implicit knowledge bases (Petroni et al., 2019).

In this paper we focus on the use of implicit knowledge in the form of pretrained LMs. While using LMs is relatively common in OK-VQA, they are usually integrated into multimodal transformers by diverse means, so as to integrate the visual and textual inputs of the task. Given that LMs were originally designed to process textual input and are extensively trained in textual corpora, we hypothesized that a system that relies exclusively on text will allow LMs to better leverage their implicit knowledge. Because OK-VQA is a visio-linguistic task, we propose to use automatic image captioning as a way to verbalize the information in the image, where the captions are descriptions of the images which are used as input to the LMs. Once the captions are generated, all the inference in our method is done using text-only models. We are aware that captions do not contain all the information in an image, and want to check whether the text-only models can compensate for that initial loss of information. The approach proposed in this paper, named Caption-based Model or CBM, can be seen in Fig. 1.

* Corresponding author.
E-mail addresses: ander.salaberria@ehu.eus (A. Salaberria), gorka.azcune@ehu.eus (G. Azkune), oier.lopezdelacalle@ehu.eus (O. Lopez de Lacalle), a.soroa@ehu.eus (A. Soroa), e.agirre@ehu.eus (E. Agirre).

https://doi.org/10.1016/j.eswa.2022.118669
Received 1 April 2022; Received in revised form 12 July 2022; Accepted 21 August 2022
Available online 28 August 2022

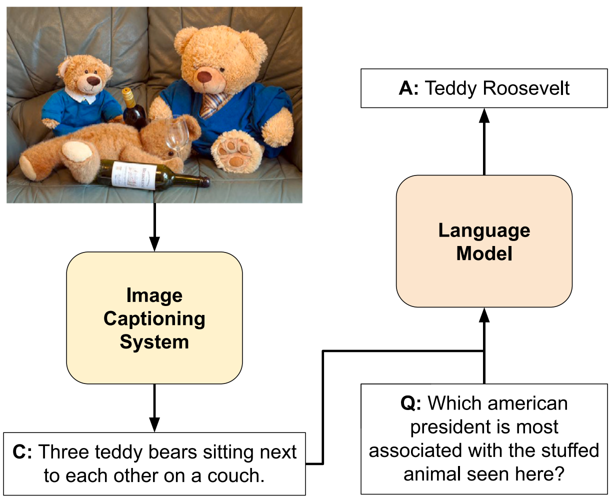


Fig. 1. Given a question and image, we verbalize the contents of the image and apply a pretrained language model for inference. We show that current text-only models are better in generalization and inference than multimodal models for knowledge-based VQA.

To validate our hypothesis, we present an extensive experimentation on the OK-VQA dataset. We compare our proposed caption-based model with the *de facto* standard of visio-linguistic tasks, i.e. multimodal transformers, which are widely used in VQA tasks to process the questions (text) and images. We also focus on language models of different sizes, to see the impact of model capacity on OK-VQA.

The contributions of this research are as follows:

- Captions are more effective than images for OK-VQA when models of similar size are used as is, and achieve similar results when both are fine-tuned on additional VQA datasets.
- Increasing the size and the capacity of language models allows to reach state-of-the-art results, outperforming by a large margin current multimodal transformers. Furthermore, we observe a trend of improvement that has not yet stabilized.
- The complex use of in-context-learning as in PICa (Yang et al., 2022) does not beat fine-tuning our smaller model, that is, our system based on T5 (Raffel et al., 2020) obtains results comparable to an ensemble of five GPT-3 runs which are 15-times larger in parameters.
- The larger contribution of captions on OK-VQA with respect to results on a regular VQA dataset (Goyal, Khot, Summers-Stay, Batra, & Parikh, 2017) show that text-only systems are specially effective when external knowledge is needed.

Our VQA system can be adapted for real life applications that range from aiding visually-impaired or blind people (Gurari et al., 2018) to improving current virtual assistants (Tulshan & Dhage, 2018). Our model is specially beneficial for questions that require world knowledge. Therefore, it could also be used for educational and recreational purposes.

Our code is available at https://github.com/salanueva/CBM.

2. Related work

We now present a brief introduction of VQA datasets and multimodal transformers before discussing different approaches to tackle the OK-VQA task and the use of generated captions in VQA tasks found in the literature.

2.1. Visual question-answering datasets

There are many VQA datasets in the literature (Antol et al., 2015; Goyal et al., 2017; Johnson et al., 2017), where, given an image and a question about the contents of that image, a system has to provide a textual answer. Some VQA datasets also demand leveraging external knowledge to infer the answer and, thus, they are known as knowledge-based VQA tasks. Good examples are KB-VQA (Wang, Wu, Shen, Dick, & van den Hengel, 2017a), KVQA (Shah, Mishra, Yadati, & Talukdar, 2019), FVQA (Wang, Wu, Shen, Dick, & Van Den Hengel, 2017b) and OK-VQA (Marino et al., 2019). KVQA requires knowledge about named entities (e.g. Barack Obama, White House, United Nations) and that knowledge is already provided as a graph. FVQA annotates questions by selecting a fact from a fixed knowledge base but its size is relatively small. KB-VQA is even smaller, presenting template-based questions whose answers can be obtained reasoning over commonsense resources or Wikipedia. In contrast, OK-VQA requires knowledge from unspecified external resources and, although smaller than KVQA in terms of the number of images and question-answer pairs, it is considerably bigger than the other knowledge-based VQA datasets. Therefore, we have chosen OK-VQA for our experiments.

2.2. Multimodal transformers

Currently, these transformers are the most successful systems for VQA and can be broadly classified into two types: single-stream and double-stream transformers. A good example of the former is VisualBERT (Li, Yatskar, Yin, Hsieh, & Chang, 2019), where the BERT architecture (Devlin et al., 2019) is used, adding visual features obtained by an object detector as input and using visio-linguistic pretraining tasks, such as image-text matching. OSCAR (Li et al., 2020) also follows a very similar philosophy, adding object tags to the input and proposing different pretraining strategies. Among double-stream transformers, VilBERT (Lu, Batra, Parikh, & Lee, 2019) and LXMERT (Tan & Bansal, 2019) use a dedicated transformer for each modality (text and image) to fuse them with a cross-modal transformer. Their differences lie mainly on some architectural choices and pretraining task selection (Bugliarello, Cotterell, Okazaki, & Elliott, 2020).

2.3. OK-VQA systems

Multimodal transformers have also been used to provide implicit knowledge from pretraining tasks. For example, VilBERT uses a pretrained BERT to encode the questions, so it uses the implicit knowledge that BERT acquired during its pretraining. Additionally, VilBERT is further trained on Conceptual Captions (Sharma, Ding, Goodman, & Soricut, 2018), a large image-caption dataset from where additional knowledge can be acquired. Those multimodal transformers are the backbone of most models used for OK-VQA, which also use symbolic knowledge to bring some extra performance. This information is summarized in Table 1.

ConceptBert (Gardères, Ziaeefard, Abeloos, & Lecue, 2020) was the first system to use multimodal transformers and symbolic knowledge for OK-VQA. It is based on a combination of a pretrained BERT to encode questions, a graph convolutional neural network to encode triples extracted from the ConceptNet knowledge graph (Speer et al., 2017) and a multimodal transformer (VilBERT) to jointly represent and reason over image features and encoded question tokens.

A similar approach was followed by KRISP (Marino et al., 2021), combining again a multimodal transformer with symbolic knowledge. In this case, the multimodal transformer, MM_{BERT}, is based on VisualBert (Li et al., 2019) and initialized with the weights of a pretrained BERT. Additionally, authors built a knowledge graph fusing DBPedia (Auer et al., 2007), ConceptNet (Speer et al., 2017), VisualGenome (Krishna et al., 2017) and hasPart KB (Bhakthavatsalam, Richardson, Tandon, & Clark, 2020). They used different image feature

Table 1
Summary of OK-VQA systems. See text for references.

System	Implicit Knowledge	Symbolic Knowledge	Is multimodal?
ConceptBERT	VilBERT	ConceptNet	Yes
KRISP	BERT	ConceptNet, DBPediahasPart KB, Visual Genome	Yes
MAVEx	VilBERT	ConceptNet, Google ImagesWikipedia	Yes
RVL	LXMERT	ConceptNet, Wikidata	Yes
PICa	GPT-3	None	No
CBM (ours)	BERT, T5	None	No

encoders and the question tokens to obtain a subset of the full graph relevant to the target question and image. Finally, using a graph convolutional neural network, they combined the symbolic and implicit knowledge to predict the final answer.

Some recent approaches, named MAVEx (Wu, Lu, Sabharwal, & Mottaghi, 2022) and RVL (Shevchenko, Teney, Dick, & van den Hengel, 2021) showed different ways to combine implicit and symbolic knowledge. MAVEx used a pretrained VilBERT to generate various candidate answers which were later reranked using answer-specific knowledge retrieval. Authors used both textual and visual knowledge resources, including images searched using Google, sentences from Wikipedia articles, and concepts from ConceptNet. On the other hand, RVL trained the two-stream multimodal transformer LXMERT (Tan & Bansal, 2019) with an auxiliary objective that aligned its representations with knowledge graph embeddings retrieved from ConceptNet and Wikidata.

These models (Gardères et al., 2020; Marino et al., 2021; Shevchenko et al., 2021; Wu et al., 2022) make use of different symbolic knowledge sources and apply different methods to use them. Nevertheless, we have noticed that the improvement obtained by adding symbolic knowledge is minor in these models. The only one that obtains a significant improvement is MAVEx (Wu et al., 2022). However, due to its design, the model is limited to give an answer from a set of answer candidates generated by only accessing to implicit knowledge. This shows the dependency of current systems on the encoded knowledge found in multimodal transformers. So, in this work we focus on the use of implicit knowledge (as opposed to explicitly encoded knowledge) which we exploit by first verbalizing images and then feeding these captions to a pretrained LM.

2.4. Captions for VQA

Integrating annotated captions, or other types of text related to the image, in multimodal systems benefit several multimodal challenges. Examples range from fake news detection (Kumari & Ekbal, 2021) to image classification tasks such as flower (Bae, Park, Lee, Lee, & Lim, 2020) and crisis (Ahmad, Jindal, N.S., Ekbal, & Bhattachharyya, 2022) classification. However, regarding the use of automatically generated captions for VQA, to the best of our knowledge, Mucko (Zhu et al., 2020) and PICa (Yang et al., 2022) are the only systems that explore this idea.

Mucko uses dense captions (Johnson, Karpathy, & Fei-Fei, 2016) to query a knowledge graph to extract relevant information to answer the question. The reported results on OK-VQA are well below the state-of-the-art. Dense captions describe different regions of an image using short sentences. Our method differs in the use of a single caption which is the input to the LM, and does not require neither knowledge graphs nor the use of OCR systems that have recently been integrated in some other VQA models (Sharma & Singh Jalal, 2022; Singh et al., 2019).

On the other hand, PICa takes advantage of the implicit knowledge found in GPT-3 (Brown et al., 2020) via prompt-engineering. Instead of supervised fine-tuning, PICa adapts to the task with a few in-context examples during inference time using both captions and object tags to describe the image, defining the current state-of-the-art with an ensemble of GPT-3's and clever selection of those examples. However, GPT-3 is only accessible via OpenAI's paid API and it has limited functionalities.

Summary. Out of several VQA datasets that demand external knowledge, we select OK-VQA for our experiments. An analysis on current OK-VQA systems shows a common trend: symbolic knowledge does not contribute too much to the performance of the systems, which mainly rely on the performance of the backbone transformer, i.e. the implicit knowledge. Therefore, we define a text-only train and inference method in order to take advantage of the implicit knowledge of LMs. For that purpose, we use textual captions generated by an image captioning system combined with the question as the input for these pretrained LMs.

3. Implemented models

In this section we describe the implemented models. We use Pytorch (Paszke et al., 2019), Pytorch Lightning and the Transformers library (Wolf et al., 2020) for all the implementation work.

3.1. Caption-based model (CBM)

Our caption-based model, denoted by CBM, is divided in two steps: (i) a caption generation system that generates a short description of a given image and (ii) a language model that takes this caption and a question in order to answer it.

We use **OSCAR** (Li et al., 2020) to generate captions from images, a transformer encoder that produces state-of-the-art results on several multimodal tasks including image captioning. As it is common in multimodal transformers, OSCAR uses a pretrained object detector called FasterRCNN (Ren, He, Girshick, & Sun, 2015) to obtain region features from images and their respective labels. Both features and labels alongside manually annotated captions are then fed to the transformer during pretraining, following the work of Anderson et al. (2018). The performance on image-captioning of both base and large models is similar, so we use OSCAR-base as our image-captioning system for all of our experiments.

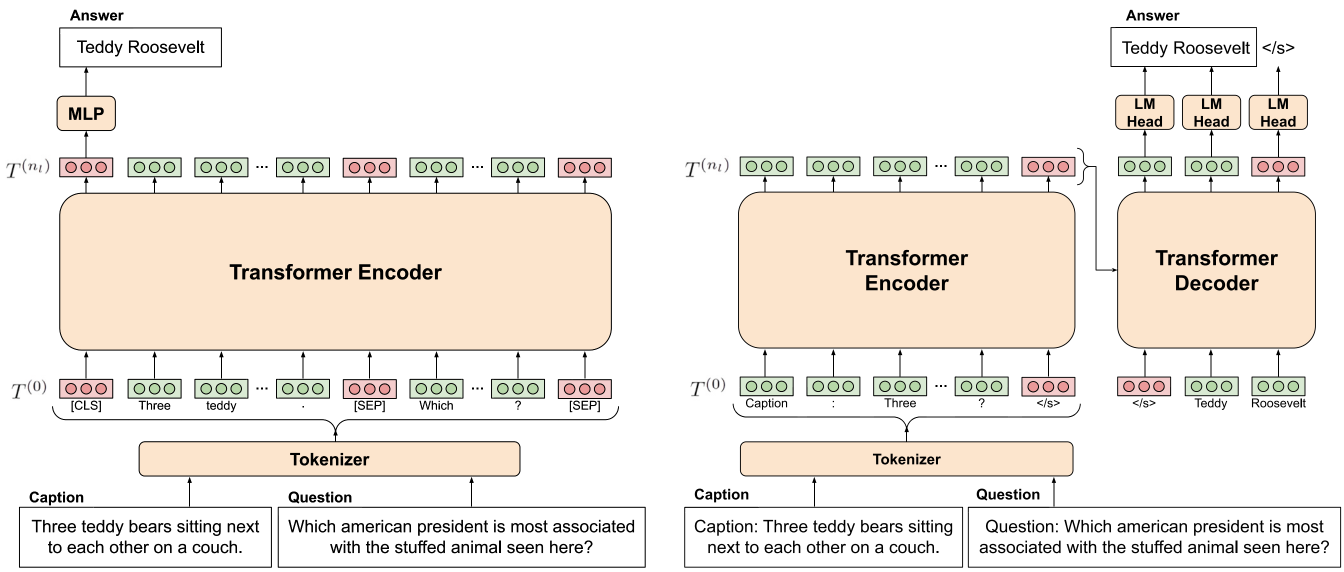
During OSCAR's fine-tuning step on image captioning, some of OK-VQA's test split images and gold captions are used. In order to ensure fairness and avoid any contamination in our experiments, we fine-tune a pretrained OSCAR model on image-captioning removing these instances from its training process.

For the second step, we explore two different language models: BERT, to perform comparative experiments with current multimodal transformers, and the T5 family, to explore the performance of LMs of increasing size.

3.1.1. CBM_{BERT}

In this first approach, we use a pretrained **BERT-base** transformer encoder (Devlin et al., 2019) as our language model. We feed sequences of tokenized captions and questions $T^{(0)} = \{\mathbf{t}_i^{(0)} | i = 1, \ldots, n_t\}$ to BERT, and take the output of the $[CLS]$ or first token of the sequence $\mathbf{t}_1^{(n_l)}$, where n_t is the number of tokens in the sequence and n_l is the number of transformer layers (see Fig. 2(a)).

In order to fine-tune the language model for VQA tasks, we add a **classification head** to the $[CLS]$ embedding. Although VQA (Antol et al., 2015; Goyal et al., 2017) and OK-VQA (Marino et al., 2019) were



(a) CBM_{BERT} (b) CBM_{T5}

Fig. 2. Detailed view of our proposed CBM models.

defined with open-ended answers, recent models (Marino et al., 2021; Zhang et al., 2021) cast these tasks as classification problems, building a fixed vocabulary of answers from the training dataset. Following this trend, our classification head is a multilayer perceptron (MLP) with one hidden layer after $\mathbf{t}_1^{(n_l)}$. We define our MLP in Eq. (1).

$$\begin{aligned} \mathbf{h} &= \text{LayerNorm}(\text{GELU}(\mathbf{W}_h \mathbf{t}_1^{(n_l)} + \mathbf{b}_h)) \\ \hat{\mathbf{y}} &= \text{Softmax}(\mathbf{W}_{\hat{y}} \mathbf{h} + \mathbf{b}_{\hat{y}}) \end{aligned} \tag{1}$$

We use a GELU activation function as well as layer normalization (Ba, Kiros, & Hinton, 2016). The trainable parameters are $\mathbf{W}_h \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_h \in \mathbb{R}^{d_h}$, $\mathbf{W}_{\hat{y}} \in \mathbb{R}^{d_h \times n_{label}}$ and $\mathbf{b}_{\hat{y}} \in \mathbb{R}^{n_{label}}$, where n_{label} equals to the number of labels on a given classification task and d_h equals to 768.

3.1.2. CBM_{T5}

In our second approach, we use pretrained **T5** encoder–decoder transformers (Raffel et al., 2020), as they are the state-of-the-art models for text-only question-answering tasks and are available in different sizes, ranging from 60M parameters to 11B. Following CBM_{BERT}, we also feed sequences of tokenized captions and questions $T^{(0)} = \{\mathbf{t}_i^{(0)} | i = 1, \dots, n_t\}$ to the T5 model. Nevertheless, in this case we add text prefixes before each sentence, such as *'caption:'* and *'question:'*. This is mainly done to mimic the input prompts used during the pretraining process of the T5 model, helping the language model to better leverage what it has learnt before. Differently from BERT, T5 is a generative LM, so instead of classifying an answer, T5 produces it in an open-ended text generation manner. Thus we do not use any classifier head for this approach (see Fig. 2(b)).

3.2. Multimodal transformer (MM_{BERT})

We compare our CBM_{BERT} model with the multimodal transformer-based MM_{BERT} (Marino et al., 2021), a variant of BERT that uses the question text and image region features as input. While BERT is designed to only process textual inputs, MM_{BERT} adapts its embedding layer in order to be able to process features from images.

We use a FasterRCNN with a ResNeXt-152 (Xie, Girshick, Dollár, Tu, & He, 2017) as its backbone to extract a total of n_v region features $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_{n_v}\}$ per image. Each of these $\mathbf{v}_i \in \mathbb{R}^{d_v}$ features represents an object that appears in the image, where d_v equals to 2048. $\mathbf{V}$ lacks the positional information between objects, which can be solved concatenating the corresponding bounding box coordinates to each feature. Upon some initial experiments, we concluded that this extra information does not improve performance in any of VQA 2.0 and OK-VQA. We use MMF Multimodal Framework (Singh et al., 2020) to extract the image region features that are fed into MM_{BERT}.

In order to allow for easier comparison between our CBM and MM_{BERT} we use the output representation for $[CLS]$ to feed into the classification multilayer perceptron (see). Note that this is slightly different from the original MM_{BERT} (Marino et al., 2021), which uses the average of all token representations in the last transformer layer.

3.3. Question-only baseline (Q_{BERT})

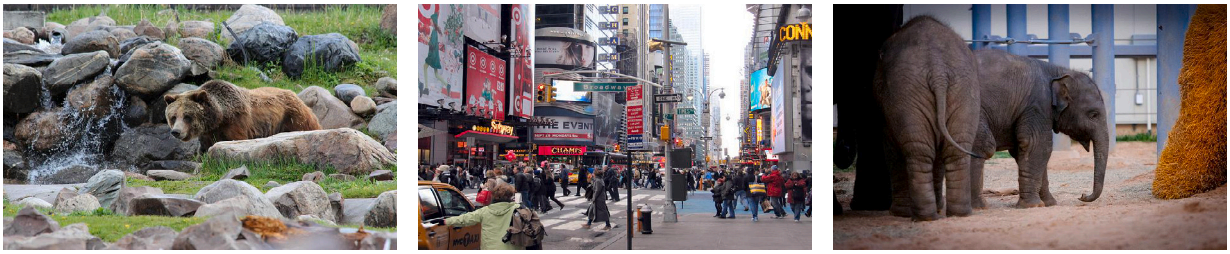
In order to assess the contribution of captions, we also trained a model which only had the question in the input, without any information about the image or caption, denoted as Q_{BERT}. This model can be seen as an ablation of CBM_{BERT}.

3.4. Loss function

Contrary to previous works in VQA, we do not use binary cross-entropy loss for our classification models, as initial experiments showed that cross-entropy loss with soft labels (SCE) converges faster with similar results. SCE loss is defined in Eq. (2), where $\mathbf{y}$ is the ground truth vector with probabilities proportional to the VQA evaluation metric (Eq. (3)) assigned to each class.

$$\mathcal{L}_{SCE}(\mathbf{y}, \hat{\mathbf{y}}) = -\mathbf{y} \cdot \log \hat{\mathbf{y}} \tag{2}$$

Regarding CBM_{T5}, we fine-tune this generative model via teacher forcing using cross-entropy loss. Therefore, the model learns to map each input sequence with its respective target sequence. However, training the model using the teacher forcing paradigm causes a discrepancy with the human annotations, as each question in OK-VQA has multiple valid answers. We fix this by randomly choosing a target sequence on each epoch. Initial experiments also showed that randomly choosing an answer from all annotated answers is slightly detrimental, as some answers are not spelled correctly, are empty strings or do not make sense as an answer. Therefore, during training we exclude answers that do not obtain a full score in the VQA score, that is, we



VQA: What color is the bear? brown
OK-VQA: What species of bear is this? grizzly

VQA: What is the weather like? cloudy
OK-VQA: Why would one suspect that this is not chicago? sign

VQA: Are the animals in captivity? yes
OK-VQA: Which valuable material grows on this animal's face? ivory

Fig. 3. Some examples of VQA 2.0 and OK-VQA datasets for the same images. VQA questions are about image contents, while OK-VQA questions require outside knowledge.

choose answers that are annotated by at least two annotators on a given question.[1]

4. Datasets

This section describes the two datasets that we experiment on, providing several details and examples on each one.

The main dataset for our experiments is OK-VQA (Marino et al., 2019), since it allows us evaluating the usage of the implicit knowledge of LMs in a multimodal task. But we also run experiments on the VQA 2.0 dataset (Goyal et al., 2017) with a double motivation: (i) to use it as additional pretraining before applying the model to OK-VQA; (ii) to analyze the performance differences among models on a knowledge-based VQA dataset and a standard VQA dataset. Examples of both datasets can be found in Fig. 3.

4.1. VQA 2.0

This dataset contains open-ended questions about images where questions focus mainly on identifying objects in the image and their attributes, detecting relations between them, as well as counting those objects. The dataset is composed of 204 K images taken from the COCO dataset (Lin et al., 2014) and 1.1M questions, each question having 10 (possibly repeated) annotations as accepted answers. Following the classification setting of VQA tasks, which is currently the dominant paradigm, VQA 2.0 has 3129 different possible answers, extracted from the most frequent answers of the training split.

VQA 2.0 is divided in three splits named train, dev and test. Some of the images from the development split of VQA 2.0 are reused in OK-VQA's test split. So, in order to avoid any contamination, we do not use the VQA 2.0 dev set for any training or hyper-parameter tuning.

Antol et al. (2015) proposed a standard evaluation metric for VQA tasks where a system answer is considered totally correct if it appears at least three times in the ten ground-truth annotations. Considering that a given answer appears x times in a question's annotations, this accuracy metric is defined in Eq. (3).

$$\mathrm{acc} = \min\left(\frac{x}{3}, 1\right) \tag{3}$$

4.2. OK-VQA

The OK-VQA dataset is built upon 14,031 images from the COCO dataset and 14,055 crowd-sourced questions. Each question has ten annotated answers (possibly repeated), and the evaluation metric is the same as in VQA 2.0 (Eq. (3)). As a knowledge-based VQA dataset, OK-VQA requires outside knowledge to answer the questions. However, this outside knowledge is neither provided nor identified, i.e. there is not a list of available knowledge sources for this task, making the task more challenging.

There are two versions of this dataset, depending on how the stemming of the answers provided by the crowd-sourcers is handled. The stemming used in OK-VQA v1.0 results in some "non-word" answers (such as "poni tail" instead of "pony tail"). OK-VQA v1.1 applied a different stemming algorithm, resulting in a more coherent answer vocabulary. We use OK-VQA v1.1 through our experiments.

5. Experiments and results

This section provides details about our experimental settings, shows results of the models defined in Section 3 and compares them with the state-of-the-art.

5.1. Experimental settings

We use the same hyperparameters as Marino et al. (2021) for fine-tuning $\mathrm{CBM_{BERT}}$, $\mathrm{MM_{BERT}}$ and $\mathrm{Q_{BERT}}$ models both in VQA 2.0 and OK-VQA tasks. We train our models for 88 K steps using AdamW optimizer (Loshchilov & Hutter, 2017). Our batch size is of 56 with a maximum learning rate of $5 \cdot 10^{-5}$ following a cosine schedule with a linear warmup of 2K steps.

Regarding $\mathrm{CBM_{T5}}$, there are 5 different T5 models that vary on size. They range from 60M to 11B parameters and we show the performance of all five models on OK-VQA. To do so, we have chosen to keep the same hyperparameters as before with the following changes:

- As models of different sizes need different amounts of training steps in order to converge, we propose the following methodology. We use 20% of the training instances to define a validation split, train the models using the remaining 80% for 20K steps. Then, we decide the final number of steps by taking the step with the best VQA score in the validation split. This process is done three times using the same validation split. After that, we compute the average number of steps of all three runs as our final number of training steps.
- As the number of training steps varies among different model sizes, we have decided to use a fixed learning rate of $5 \cdot 10^{-5}$ during the fine-tuning process. Hence, we do not use any learning-rate scheduler that depends on warmup steps or total number of training steps.

All experiments regarding classification models have been run in a single GPU with 12 GB of vRAM and their runtimes are at most of 12 h. Regarding the much larger $\mathrm{CBM_{T5}}$ we used up to 4 NVIDIA A100 GPUs (each with 80 GB of vRAM), changing both hardware and hyper-parameters to keep the same effective batch size across model sizes, and used DeepSpeed's ZeRO Stage 2 optimization algorithm with CPU offload (Rajbhandari, Rasley, Ruwase, & He, 2020) when fine-tuning the biggest model. However, their runtimes are at most of 4 h,

[1] If a question does not have answers that fulfill this rule, the question is discarded from training, which amounts to a total of 112 instances in the OK-VQA's training split.

Table 2
Performance on OK-VQA for three classification models (respectively, question only, image-based and caption-based) without and with additional pretraining on VQA 2.0. Mean VQA score and standard deviation across 3 different runs.

Model	Score	+ VQA pretraining	Parameters
Q_{BERT}	21.2 ±0.2	23.0 ±0.2	112M
MM_{BERT}	29.6 ±0.6	35.7 ±0.3	114M
CBM_{BERT} (ours)	**32.5** ±0.4	**36.0** ±0.4	112M

Table 3
Performance on OK-VQA of our generative CBM_{T5} models.

Model	Score	Parameters
$CBM_{T5\text{-}Small}$	29.2 ±0.2	60M
$CBM_{T5\text{-}Base}$	36.1 ±0.5	220M
$CBM_{T5\text{-}Large}$	40.8 ±0.4	770M
$CBM_{T5\text{-}3B}$	44.0 ±0.7	3B
$CBM_{T5\text{-}11B}$	**47.9** ±0.2	11B

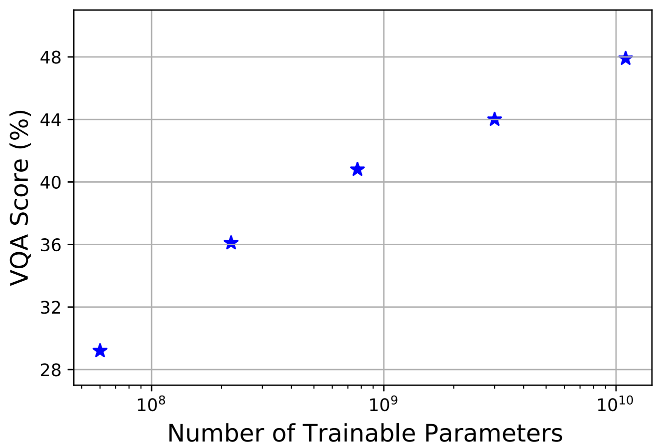


Fig. 4. Correlation between the size of CBM_{T5} models and their performance. The horizontal axis is in logarithmic scale.

as less training steps are needed for CBM_{T5}, compared to the rest of our models.

In order to get consistent results we make each experiment three times and provide the mean VQA score and standard deviation in all of our results.

5.2. Results for images vs. captions

Table 2 shows the results for the three models presented in Section 3, which share the same architecture, size and initial parameters. We show the results for the models fine-tuned on OK-VQA, as well as the same models which have been previously fine-tuned on VQA 2.0.

We observe that the sole use of questions (Q_{BERT}) offers poor performance compared to the other two systems, achieving up to 13 points less accuracy. This shows that having any representation of the image (captions or image region features) is key to answer questions correctly. This is further justified comparing the improvement that VQA pretraining entails, as Q_{BERT} improves less than 2 points, whereas the other two improve their accuracy between 4–6 points.

Contribution of captions. When we compare the performance of CBM_{BERT} and MM_{BERT}, we see that, when there is no visio-linguistic pretraining involved, CBM_{BERT} performs better in OK-VQA. However, when we pretrain these models in a similar multimodal task like VQA 2.0, their accuracy increases by 4–6 points and both obtain similar performance. As OK-VQA's training is comparatively smaller (9K instances vs. VQA's 410K instances), we hypothesize that training MM_{BERT} on OK-VQA is not enough to adapt the model to the new input modality. However, as CBM_{BERT} uses only text, the fine-tuning with such small training is more effective.

5.3. T5 and larger models

As T5 has been pre-trained on several question answering tasks, we directly fine-tune it on OK-VQA alone.

In Table 3 we show the results of five differently sized CBM_{T5} models on OK-VQA. Note that our T5-Base model obtains results comparable to our BERT-base model pre-trained on VQA 2.0. This was expected, as both models have been pre-trained with question answering datasets and both have comparable model sizes, T5-base being composed of two BERT-base encoder and decoder.

The results in Table 3 are plotted in Fig. 4, showing that the size of our models is logarithmically proportional to its score, which follows the scaling laws mentioned in Kaplan et al. (2020). This trend is followed even by our biggest model and does not seem to slow down yet. These results show the importance of the model's capacity in the results. All models have been pretrained with the same corpus and downstream tasks, but the difference in size helps bigger models to better leverage the information learnt from that corpus in order to incorporate the external knowledge needed to solve OK-VQA. Our largest model performs much better than the multimodal model.

In fact, it is not clear whether larger multimodal models could match our largest text-only caption based model. We cannot currently test this hypothesis, as, to the best of our knowledge, there are no publicly available multimodal transformers with comparable numbers or parameters. Still, we hypothesize that in the case of knowledge-intensive datasets such as OK-VQA, current multimodal transformers (Li et al., 2019; Lu et al., 2019; Tan & Bansal, 2019) will underperform our system, as the only textual data fed to these models during their pretraining is mostly composed by captions or small descriptions attached to images. This means that these models only see a limited vocabulary from a limited corpus, compared to the rich, diverse and much larger corpora used to build models such as T5.

5.4. Comparison with the state of the art

In Table 4, we show the results of various state-of-the-art models in three groups: (i) classification models based on multimodal transformers, which additionally include the usage of symbolic knowledge; (ii) GPT-3 based generative models that use in-context learning; (iii) our caption-based models.

The state-of-the-art classification models like KRISP (Marino et al., 2021), MAVEx (Wu et al., 2022) and RVL (Shevchenko et al., 2021) show similar results on the implicit-only versions of their models, even though they are based on different multimodal transformers and pretraining tasks. Note that RVL has a contamination issue as images from OK-VQA's test split were used to train their multimodal transformer. We also observe that using symbolic knowledge improves the results around 2 points, the exception being MAVEx, which combines knowledge found in ConceptNet (Speer et al., 2017), Wikipedia and Google Images.[2]

PICa (Yang et al., 2022) takes advantage of GPT-3 (Brown et al., 2020) to define a new state-of-the-art in a generative manner using in-context learning. Its base results (PICa-Base) already surpass the ones seen before without any need of symbolic knowledge. An ensemble of 5 GPT-3 models and a clever selection of annotated examples from the training data to build the input prompt further improves its results (PICa-Full). Table 4 reports two results for each PICa model: the results using automatically generated captions alone (like us), and the results

[2] This result is obtained with an ensemble of 3 MAVEx models that share the same multimodal transformer. A unique MAVEx model achieves an accuracy of 40.3%.

Table 4
Comparison to the state-of-the-art on OK-VQA. +sym. stands for systems additionally using symbolic knowledge, and +tags for the additional use of object tags. Results of models marked with * are in OK-VQA v1.0 and † specifies contaminated results (see main text).

Model	Score		Parameters
ConceptBERT (Gardères et al., 2020) *	31.4	(+sym. 33.7)	348M
MAVEx (Wu et al., 2022)	35.2	(+sym. 41.4)	353M
KRISP (Marino et al., 2021)	37.1	(+sym. 38.9)	116M
RVL (Shevchenko et al., 2021) *†	37.3	(+sym. 39.0)	208M
PICa-Base (Yang et al., 2022)	42.0	(+tags 43.3)	175B
PICa-Full (Yang et al., 2022) (Ensemble)	46.9	(+tags **48.0**)	175B
CBM_{BERT} (ours)	36.0		112M
$CBM_{T5\text{-}11B}$ (ours)	**47.9**		11B

Table 5
Performance on the dev split of VQA 2.0 of the multimodal model MM_{BERT} and two text-only models: PICa-Full and CBM_{BERT}.

Model	Score
MM_{BERT}	**65.8**
PICa-Full	56.1
CBM_{BERT} (ours)	59.6

Table 6
Performance on OK-VQA for early and late fusion models.

Model	Score	+ VQA pretraining
Early Fusion	32.5 ±0.4	38.2 ±0.8
Late Fusion	34.0 ±0.4	**38.6** ±0.2

when also using object tags automatically obtained from the image, which slightly improve the results.

Our CBM_{BERT} system performs on par with the multimodal transformers, which is remarkable, since we do not use directly any visual features in our models and only use the caption. Note that all those systems have models of comparable size. When scaling up our generative models, we see that $CBM_{T5\text{-}11B}$ outperforms current multimodal models by a large margin and is on par with the results obtained by PICa-Full. Indeed, $CBM_{T5\text{-}11B}$ achieves slightly better results than the PICa version which uses captions alone, even if our model is 15 times smaller.

6. Analysis

In this section we perform additional experiments. We first contrast the results on OK-VQA with those obtained in VQA 2.0, discussing the reasons for the different performance. We then combine our text only model with its counterpart multimodal model to analyze if they are complementary. Afterwards, we compare the performance of CBM_{BERT} with manually annotated captions or the ones generated by OSCAR (Li et al., 2020). Finally, we present some qualitative analysis.

6.1. Results on VQA 2.0

Even though both unimodal and multimodal methods perform similarly in OK-VQA, we observed a different trend in VQA 2.0. Table 5 shows that CBM_{BERT} obtains 59.6, while MM_{BERT} achieves 6 points more. We think this is due to the information loss when converting an image into a caption, as relevant information that is needed to answer the question can be lost. This is specially important for VQA 2.0, where the questions refer directly to image contents, spatial relations and object attributes (see Fig. 3). A similar behavior can be observed for PICa (Yang et al., 2022). Interestingly, PICa also uses object tags to minimize the information loss when verbalizing the image, but it does not perform as well as our system. Even with 1000 times less parameters, our CBM_{BERT} outperforms PICa, showing the importance of fine-tuning in contrast to in-context-learning, specially when large training data is available, as in VQA 2.0.

The difference between VQA and OK-VQA performances suggests that captions contain enough information to effectively use the implicit knowledge of language models for knowledge-intensive multimodal tasks like OK-VQA. However, it seems that in datasets where the answer can be found in the image, multimodal models are preferable.

6.2. Combining visual information and captions

Given the different nature of the inputs, we wanted to check whether CBM_{BERT} and MM_{BERT} are complementary. Our hypothesis is that the former can take advantage of the implicit knowledge acquired by the language model, whereas the latter has access to more fine-grained information found in image regions. Therefore, we define two different approaches to check how they complement each other.

Early fusion. For each question we feed both caption and image features alongside the question to the language model. This system can be seen as a MM_{BERT} which processes a multimodal input composed by a question (text), a caption (text) and image region features. We initialize the weights of this model with the weights of the base language model (BERT-base) and fine-tune it on the target train data.

Late fusion. We train CBM_{BERT} () and MM_{BERT} (Section 3.2) separately, each of them with their corresponding inputs, and combine their outputs in inference time to obtain the final answer. The combination is done by multiplying output probabilities of both models for each class and taking the answer with the highest value. We show their performance in Table 6.

These fusion models improve the performance of both CBM_{BERT} and MM_{BERT} by 2–3 points in almost all cases. The only case where there is no improvement comparing to CBM_{BERT} is in the early fusion without VQA pretraining. This may be caused again by the small training split of OK-VQA, causing difficulties to learn how to ground textual and visual modalities. However, this is solved when VQA pretraining is added to the model, increasing vastly the amount of data seen by the models and showing similar performance on both early and late fusion models.

Additionally, we also observed the complementarity of both modalities in the VQA dataset. Early fusion obtains 67.8% and late fusion 67.7% in the dev split of VQA 2.0, improving the performance of MM_{BERT} by 2 points. The results validate our hypothesis, showing that image region features and captions are complementary in this setting.

6.3. Ground truth captions

In order to measure the effects of the image captioning system to our proposed CBM model, we check the gap of performance between the use of generated captions and gold captions. As OK-VQA is built upon images from the COCO dataset (Lin et al., 2014), each image has five different annotated captions. We use these captions and fine-tune CBM_{BERT} on OK-VQA without VQA pretraining following the same experimental settings. We repeat this experiment three times, as it is done through the entire work. On each run we select a different set of captions, that is, for each image we just choose one gold caption

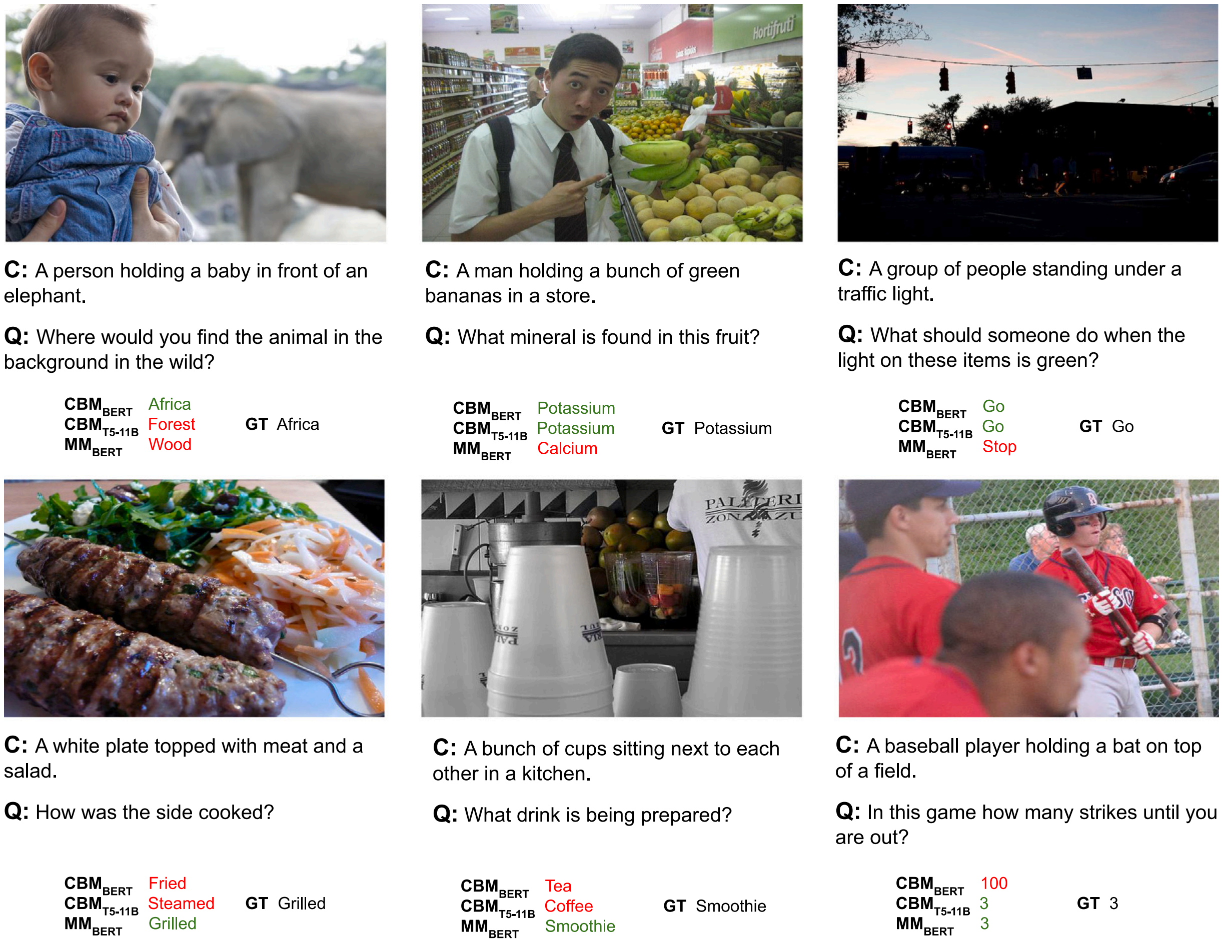


Fig. 5. Examples of OK-VQA questions where only one between CBM_{BERT} and MM_{BERT} answers correctly according to the ground truth (GT). We also show answers given by $\text{CBM}_{\text{T5-11B}}$ for further comparisons. C refers to captions generated by OSCAR. Correct answer in green, incorrect in red.

randomly and use it during the entire training process. As we also have several captions in all of OK-VQA's test split, we test each fine-tuned model three times following the same caption selection process.

Table 2 already shows that we achieve an accuracy and standard deviation of 32.5 ± 0.4 using generated captions on OK-VQA's test split. However, when we use gold captions we get an average accuracy of 32.3 ± 0.3 in all of our runs. In both cases we obtain similar results, showing that captions generated by OSCAR contain enough information for CBM_{BERT} to perform comparably on this specific task.

6.4. Qualitative analysis on OK-VQA

Both CBM_{BERT} and its multimodal counterpart perform similarly (see Table 2), but in 38.7% of the test examples their output differs and only one of them is correct. Fig. 5 shows some OK-VQA test examples together where the outputs of CBM_{BERT} and MM_{BERT} with VQA pretraining differ. We also add answers from $\text{CBM}_{\text{T5-11B}}$ for further comparisons.

Starting with the top-left example, CBM_{BERT} can infer that elephants are native to Africa whereas MM_{BERT} does not. In fact, the generated caption includes the information that the animal found in the image is an elephant, performing the first step needed to answer the question. This way, the LM can focus on using its implicit knowledge in order to answer correctly. CBM_{T5} generates 'forest' as an answer. Although the answer may be considered as valid to us, the answer is not within the list of ground truth answers, making it incorrect. The other two examples in the top row behave similarly. The caption facilitates the grounding between the question and the image. Whenever a question refers to the image ("this fruit" and "these items"), if the caption already mentions these objects ("bananas" and "traffic light", respectively), the LM seems to better leverage its implicit knowledge and reasoning capabilities to answer the question. The top-right example is interesting in this regard. While the image shows red traffic lights, the question asks about the effects of green lights. This may trick MM_{BERT} into answering the effect that red lights produce, not the green ones.

The bottom row of Fig. 5 shows two examples where the caption does not give enough information to infer the answer for CBMs. In the first case CBMs cannot decide whether the meat is steamed, fried or grilled by only examining the caption. Nevertheless, MM_{BERT} does have access to visual cues of the image, where we can see that the meat is grilled. This also happens in the second example, as the caption does not specify any ingredient of the beverage while we can see fruits in the image. The rightmost example illustrates an example where the caption does support the inference, but where our BERT based CBM gets it wrong. With the given caption, "this game" refers to baseball. However, CBM_{BERT} is unable to infer that three strikes are enough for a strikeout whereas both $\text{CBM}_{\text{T5-11B}}$ and MM_{BERT} manage to give the correct answer.

All in all, these examples support our hypothesis that visual features and captions are complementary. They also show that our system has some advantages regarding the interpretability of the system, specially in the cases our method is wrong. In some cases like the two leftmost examples in the bottom row, the object or feature needed to answer the question is missing from the caption. In other cases, the required information is in the caption, but the inference is erroneous.

7. Conclusions

In this paper we present a VQA system which describes images with a caption to then work only with textual data. We show that such a system performs surprisingly well in OK-VQA, where the questions cannot be answered with the image alone, requiring access to external knowledge. Our analysis indicates that the loss of information when summarizing the image into a caption is compensated by the better inference ability of text-only pretrained language models. We also show the importance of a language model's capacity when leveraging the implicit knowledge found in it, achieving state-of-the-art results, outperforming current multimodal models by a large margin and matching a 15-times larger ensemble model. Compared to multimodal models, orders of magnitude bigger text-only LMs are available, which we show to be an advantage for knowledge-intensive tasks. In the future we would like to explore whether richer descriptions of images might improve results further, and whether large text-only language models can still benefit from incorporating symbolic knowledge graphs.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Our code will be available at https://github.com/salanueva/CBM.

Acknowledgements

Ander is funded by a PhD grant from the Basque Government (PRE_2021_2_0143). This work is based upon work partially supported by the Ministry of Science and Innovation of the Spanish Government (DeepKnowledge project PID2021-127777OB-C21), and the Basque Government (IXA excellence research group IT1570-22).

References

Ahmad, Z., Jindal, R., N.S., M., Ekbal, A., & Bhattachharyya, P. (2022). Multimodality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification. *Expert Systems with Applications*, *195*, Article 116626. http://dx.doi.org/10.1016/j.eswa.2022.116626.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 6077–6086). IEEE Computer Society, http://dx.doi.org/10.1109/cvpr.2018.00636.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., et al. (2015). VQA: Visual question answering. In *2015 IEEE international conference on computer vision* (pp. 2425–2433). http://dx.doi.org/10.1109/access.2020.3041503.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web* (pp. 722–735). Berlin, Heidelberg: Springer, http://dx.doi.org/10.1007/978-3-540-76298-0_52.

Ba, L. J., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. http://dx.doi.org/10.48550/arXiv.1607.06450, arXiv preprint arXiv:1607.06450.

Bae, K. I., Park, J., Lee, J., Lee, Y., & Lim, C. (2020). Flower classification with modified multimodal convolutional neural networks. *Expert Systems with Applications*, *159*, Article 113455. http://dx.doi.org/10.1016/j.eswa.2020.113455.

Bhakthavatsalam, S., Richardson, K., Tandon, N., & Clark, P. (2020). Do dogs have whiskers? a new knowledge base of haspart relations. http://dx.doi.org/10.48550/arXiv.2006.07510, arXiv preprint arXiv:2006.07510.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), *33*, *Advances in neural information processing systems* (pp. 1877–1901). URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html,

Bugliarello, E., Cotterell, R., Okazaki, N., & Elliott, D. (2020). Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. http://dx.doi.org/10.48550/arXiv.2011.15124, arXiv preprint arXiv:2011.15124.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423.

Gardères, F., Ziaeefard, M., Abeloos, B., & Lecue, F. (2020). Conceptbert: Concept-aware representation for visual question answering. In *Findings of the association for computational linguistics* (pp. 489–498). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.44.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 6325–6334). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2017.670.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., et al. (2018). VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. http://dx.doi.org/10.1109/cvpr.2018.00380.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2901–2910). http://dx.doi.org/10.1109/cvpr.2017.215.

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. In *2016 IEEE conference on computer vision and pattern recognition*. IEEE, http://dx.doi.org/10.1109/cvpr.2016.494.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. http://dx.doi.org/10.48550/arXiv.2001.08361, arXiv preprint arXiv:2001.08361.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73. http://dx.doi.org/10.1007/s11263-016-0981-7.

Kumari, R., & Ekbal, A. (2021). AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*, *184*, Article 115412. http://dx.doi.org/10.1016/j.eswa.2021.115412.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. http://dx.doi.org/10.48550/arXiv.1908.03557, arXiv preprint arXiv:1908.03557.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision* (pp. 121–137). Springer, http://dx.doi.org/10.1007/978-3-030-58577-8_8.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), *Computer vision* (pp. 740–755). Cham: Springer International Publishing.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. http://dx.doi.org/10.48550/arXiv.1711.05101, arXiv preprint arXiv:1711.05101.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *32*, (pp. 13–23). URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html,

Marino, K., Chen, X., Parikh, D., Gupta, A., & Rohrbach, M. (2021). KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14111–14121). http://dx.doi.org/10.1109/CVPR46437.2021.01389.

Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). OK-VQA: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF conference on computer vision and pattern recognition* (pp. 3190–3199). IEEE, http://dx.doi.org/10.1109/cvpr.2019.00331.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, R. Garnett (Eds.), *32*, *Advances in neural information processing systems*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., et al. (2019). Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 2463–2473). http://dx.doi.org/10.18653/v1/d19-1250.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67, URL http://jmlr.org/papers/v21/20-074.html.

Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis* (pp. 1–16). IEEE, http://dx.doi.org/10.1109/sc41405.2020.00024.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster r-CNN: towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28: annual conference on neural information processing systems 2015* (pp. 91–99). URL https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

Shah, S., Mishra, A., Yadati, N., & Talukdar, P. P. (2019). Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01* (pp. 8876–8884). http://dx.doi.org/10.1609/aaai.v33i01.33018876.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2556–2565). http://dx.doi.org/10.18653/v1/p18-1238.

Sharma, H., & Singh Jalal, A. (2022). A framework for visual question answering with the integration of scene-text using PHOCs and fisher vectors. *Expert Systems with Applications*, *190*, Article 116159. http://dx.doi.org/10.1016/j.eswa.2021.116159.

Shevchenko, V., Teney, D., Dick, A., & van den Hengel, A. (2021). Reasoning over vision and language: Exploring the benefits of supplemental knowledge. In *Proceedings of the third workshop on beyond vision and language: integrating real-world knowledge* (pp. 1–18). Kyiv, Ukraine: Association for Computational Linguistics, URL https://aclanthology.org/2021.lantern-1.1.

Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., et al. (2020). MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., et al. (2019). Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8317–8326). http://dx.doi.org/10.1109/cvpr.2019.00851.

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence, vol. 31, no. 1* (pp. 4444–4451). http://dx.doi.org/10.1609/aaai.v31i1.11164.

Tan, H., & Bansal, M. (2019). LXMERT: learning cross-modality encoder representations from transformers. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 5099–5110). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1514.

Tulshan, A. S., & Dhage, S. N. (2018). Survey on virtual assistant: Google assistant, siri, cortana, alexa. In *International symposium on signal processing and intelligent recognition systems* (pp. 190–201). Springer, http://dx.doi.org/10.1007/978-981-13-5758-9_17.

Wang, P., Wu, Q., Shen, C., Dick, A., & van den Hengel, A. (2017a). Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 1290–1296). http://dx.doi.org/10.24963/ijcai.2017/179.

Wang, P., Wu, Q., Shen, C., Dick, A., & Van Den Hengel, A. (2017b). Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(10), 2413–2427. http://dx.doi.org/10.1109/tpami.2017.2754246.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6.

Wu, J., Lu, J., Sabharwal, A., & Mottaghi, R. (2022). Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence, vol. 36, no. 3* (pp. 2712–2721). http://dx.doi.org/10.1609/aaai.v36i3.20174.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500). http://dx.doi.org/10.1109/cvpr.2017.634.

Xie, N., Lai, F., Doran, D., & Kadav, A. (2019). Visual entailment: A novel task for fine-grained image understanding. http://dx.doi.org/10.48550/arXiv.1901.06706, arXiv preprint arXiv:1901.06706.

Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., et al. (2022). An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence, vol. 36, no. 3* (pp. 3081–3089). http://dx.doi.org/10.1609/aaai.v36i3.20215.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., et al. (2021). VinVL: Making visual representations matter in vision-language models. http://dx.doi.org/10.48550/arXiv.2101.00529, CoRR abs/2101.00529, arXiv:2101.00529.

Zhu, Z., Yu, J., Wang, Y., Sun, Y., Hu, Y., & Wu, Q. (2020). Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *IJCAI* (pp. 1097–1103). http://dx.doi.org/10.24963/ijcai.2020/153.