ConceptBert: Concept-Aware Representation for Visual Question Answering

François Gardères[*]
Ecole Poyltechnique
Paris, France

Maryam Ziaeefard[†]
McGill University
Montreal, Canada

Baptiste Abeloos
Thales
Montreal, Canada

Freddy Lecue
Inria, France
Thales, Canada

Abstract

Visual Question Answering (VQA) is a challenging task that has received increasing attention from both the computer vision and the natural language processing communities. Current works in VQA focus on questions which are answerable by direct analysis of the question and image alone. We present a concept-aware algorithm, ConceptBert, for questions which require common sense, or basic factual knowledge from external structured content. Given an image and a question in natural language, ConceptBert requires visual elements of the image and a Knowledge Graph (KG) to infer the correct answer. We introduce a multi-modal representation which learns a joint Concept-Vision-Language embedding. We exploit ConceptNet KG for encoding the common sense knowledge and evaluate our methodology on the Outside Knowledge-VQA (OK-VQA) and VQA datasets. Our code is available at https://github.com/ZiaMaryam/ConceptBERT

1 Introduction

Visual Question Answering (VQA) was firstly introduced to bridge the gap between natural language processing and image understanding applications in the joint space of vision and language (Malinowski and Fritz, 2014).

Most VQA benchmarks compute a question representation using word embedding techniques and Recurrent Neural Networks (RNNs), and a set of object descriptors comprising bounding box coordinates and image features vectors. Word and image representations are then fused and fed to a network to train a VQA model. However, these approaches are practical when no knowledge beyond the visual content is required.

Incorporating the external knowledge introduces several advantages. External knowledge and supporting facts can improve the relational representation between the objects detected in the image, or between entities in the question and objects in the image. It also provides information on how the answer can be derived from the question. Therefore, the complexity of the questions can be increased based on the supporting knowledge base.

Organizing the world's facts and storing them in a structured database, large scale Knowledge Bases (KB), have become important resources for representing the external knowledge. A typical KB consists of a collection of subject-predicate-object triplets also known as a fact. A KB in this form is often called a Knowledge Graph (KG) (Bollacker et al.) due to its graphical representation. The entities are nodes and the relations are the directed edges that link the nodes. The triples specify that two entities are connected by a particular relation, e.g., `(Shakespeare, writerOf, Hamlet)`.

A VQA system that exploits KGs is an emerging research topic, and is not well-studied. Recent research has started integrating knowledge-based methods into VQA models (Wang et al., 2017, 2016; Narasimhan et al., 2018; Narasimhan and Schwing, 2018; Zhu et al., 2015; Marino et al., 2019). These methods incorporate the external knowledge through two approaches: i) they exploit a set of associated facts for each question provided in VQA datasets (Narasimhan et al., 2018; Narasimhan and Schwing, 2018), or ii) they collect possible search queries for each question-image pair and use a search API to retrieve the answers (Wang et al., 2017, 2016; Zhu et al., 2015; Marino et al., 2019). However, we go one step further and implement an end-to-end VQA model that is fully trainable. Our model does not require knowledge annotations in VQA datasets or search queries.

[*] This work was done when the author was an intern at Thales. Contact email: francois.garderes@student.ecp.fr

[†] This work was done when the author worked at Thales.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 489–498
November 16 - 20, 2020.

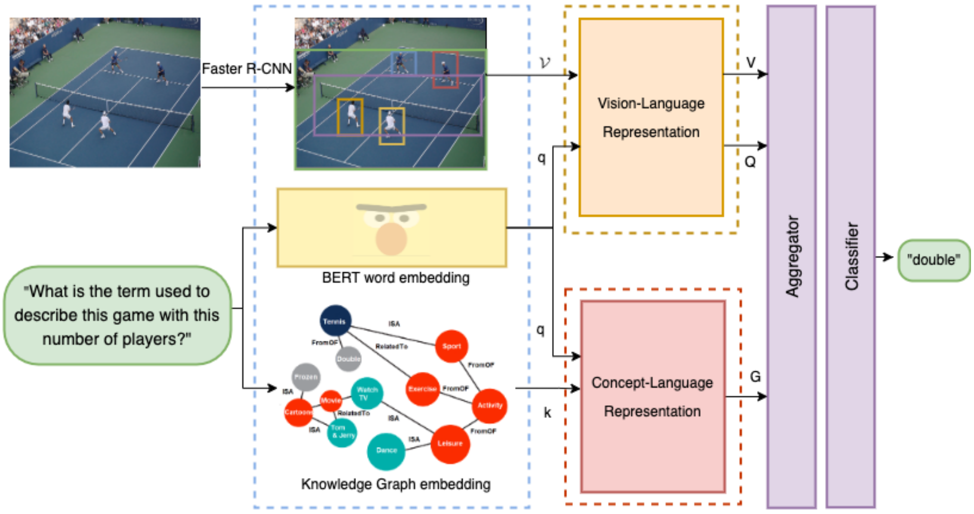


Figure 1: Model architecture of the proposed ConceptBert.

Most of the recent works are still based on the idea of context-free word embeddings rather than the pre-trained language representation (LR) model. While the pre-trained LR model such as BERT (Devlin et al., 2018) is an emerging direction, there is little work on its fusion with KG and image representation in VQA tasks. Liu et al. propose a knowledge-based language representation and use BERT as the token embedding method. However, this model is also a query-based method. It collects entity names involved in questions and queries their corresponding triples from the KG. Then, it injects queried entities into questions.

In this paper, we introduce a model which jointly learns from visual, language, and KG embeddings and captures image-question-knowledge specific interactions. The pipeline of our approach is shown in Figure 1. We compute a set of object, question, and KG embeddings. The embedded inputs are then passed through two main modules: i) the vision-language representation, and ii) the concept-language representation. The vision-language representation module jointly enhances both the image and question embeddings, each improving its context representation with the other one. The concept-language representation uses a KG embedding to incorporate relevant external information in the question embedding. The outputs of these two modules are then aggregated to represent concept-vision-language embeddings and then fed to a classifier to predict the answer.

Our model is different from the previous methods since we use pre-trained image and language features and fuse them with KG embeddings to incorporate the external knowledge into the VQA task. Therefore, our model does not need additional knowledge annotations or search queries and reduces computational costs. Furthermore, our work represents an end-to-end pipeline that is fully trainable.

In summary, the main contributions of our work are:

1. Novel methodology to incorporate common sense knowledge to VQA models (Figure 1)
2. Concept-aware representation to use knowledge graph embeddings in VQA models (Figure 2-b)
3. Novel multimodal Concept-Visual-Language embeddings (Section 3.4)

2 Problem formulation

Given a question $q \in \mathcal{Q}$ grounded in an image $I \in \mathcal{I}$ and a knowledge graph $\mathcal{G}$, the goal is to predict a meaningful answer $a \in \mathcal{A}$. Let Θ be the parameters of the model p that needs to be trained. Therefore, the predicted answer $\hat{a}$ of our model is:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} p_{\Theta}(a|I, q, \mathcal{G}) \tag{1}$$

In order to retrieve the correct answer, we aim to learn a joint representation $z \in R^{d_z}$ of q, I, and $\mathcal{G}$ such that:

$$a^* = \hat{a} = \arg\max_{a \in \mathcal{A}} p_{\Theta}(a|z) \tag{2}$$

where a^* is the ground-truth answer. d_z is a hyperparameter that represents the dimension of the joint space z. d_z is selected based on a trade-off between the capability of the representation and the computational cost.

3 Our approach

3.1 Input representations

The input to our model, ConceptBert, consists of an image representation, a question representation, and a knowledge graph representation module (cf. the blue-dashed box in Figure 1) which are discussed in detail below.

Image representation: We use pre-trained Faster R-CNN features (Anderson et al., 2017) to extract a set of objects $\boldsymbol{\mathcal{V}} = \{v_i \mid i = 1, ..., n_v\}$ per image, where each object v_i is associated with a visual feature vector $\boldsymbol{v_i} \in \mathbb{R}^{d_v}$ and bounding-box coordinates $\boldsymbol{b_i} \in \mathbb{R}^{d_b}$.

Question representation: Given a question consisting of n_T tokens, we use BERT embeddings (Devlin et al., 2018) to generate question representation $\boldsymbol{q} \in \mathbb{R}^{n_T \times d_q}$. BERT operates over sequences of discrete tokens consisting of vocabulary words and a small set of special tokens, i.e., SEP, CLS, and MASK. The representation of each token is a sum of a token-specific learned embedding and encodings for position and segment. Position refers to the token's index in the sequence and segment shows the index of the token's sentence if multiple sentences exist.

Knowledge graph representation: We use ConceptNet (Speer et al., 2016) as the source of common sense knowledge. ConceptNet is a multilingual knowledge base, representing words and phrases that people use and the common sense relationships between them. ConceptNet is a knowledge graph built from several different sources (mostly from Wiktionary, Open Mind Common Sense (Singh et al., 2002) and Games with a purpose such as Ahn et al.). It contains over 21 million edges and over 8 million nodes. In this work, we focus on the English vocabulary which contains approximately 1.5 million nodes. To avoid the step of the query construction and take full advantage of the large scale KG, we exploit ConceptNet embedding proposed in (Malaviya et al., 2020) and generate the KG representation $\boldsymbol{k} \in \mathbb{R}^{n_T \times d_k}$.

This method uses Graph Convolutional Networks (Kipf and Welling, 2016) to incorporate information from the local neighborhood of a node in the graph. It includes an encoder and a decoder. A graph convolutional encoder takes a graph as input, and encodes each node. The encoder operates by sending messages from a node to its neighbors, weighted by the relation type defined by the edge. This operation occurs in multiple layers, incorporating information multiple hops away from a node. The last layer's representation is used as the graph embedding of the node.

3.2 Vision-Language representation

To learn joint representations of language $\boldsymbol{q}$ and visual content $\boldsymbol{\mathcal{V}}$, we generate vision-attended language features V and language-attended visual features Q (cf. the orange box in Figure 1) inspired by VilBERT model (Lu et al., 2019).

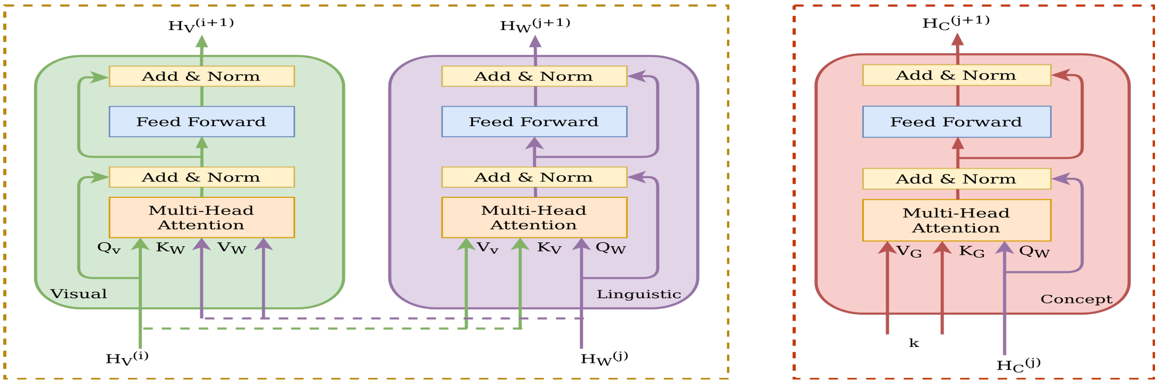
Our vision-language module is mainly based on two parallel BERT-style streams, which operate over image regions and text segments (cf. Figure 2-a). Each stream is a succession of transformer blocks and co-attentional transformer layers to enable information exchange between image and text modalities. These exchanges are restricted between specific layers and the text features go through more processing than visual features. The final set of image features represent high-level information of language features, and final text features include high-level vision features.

3.3 Concept-Language representation

The vision-language module represents the interactions between the image and the question. However, this module alone is not able to answer questions that require insights that are neither in the image, nor in the question. To this end, we propose the concept-language representation to produce language features conditioned on knowledge graph embeddings (cf. the red box in Figure 1). It performs knowledge-conditioned language attention in the concept stream (Figure 2-b). With this system, the model is able to incorporate common sense knowledge to the question, and enhance the question comprehension with the information found in the knowledge graph.

The entities in the knowledge graph have both contextual and relational information that we desire to integrate in the question embedding. To this purpose, we use an attentional transformer layer which is a multi-layer bidirectional Transformer using the encoder part of the original Transformer (Vaswani et al., 2017).

The concept-language module is a series of



a) Vision-Language representation

b) Concept-Language representation

Figure 2: Attention-based representation modules

Transformer blocks that attends to question tokens based on KG embeddings. Given input question tokens $\{w_0, ..., w_T\}$ represented as $\boldsymbol{q}$ and their KG embeddings represented as $\boldsymbol{k}$, our model outputs a final representation G.

The input consists of ”queries” from question embeddings and ”keys” and ”values” of KG embeddings. We use Multi-Head Attention with scaled dot-product. Therefore, we pack a set of $\boldsymbol{q}$ into a matrix Q_w, and $\boldsymbol{k}$ into a matrix K_G and V_G.

$$Att(Q_w, K_G, V_G) = softmax\left(\frac{Q_w \cdot K_G^{\intercal}}{\sqrt{d_k}}\right) \cdot V_G \tag{3}$$

The output of the final Transformer block, G, is a new representation of the question, enhanced with common sense knowledge extracted from the knowledge graph. Figure 2-b shows an intermediate representation H_C.

3.4 Concept-Vision-Language embedding module

We aggregate the outputs of the three streams to create a joint concept-vision-language representation. The aggregator needs to detect high-level interactions between the three streams to provide a meaningful answer, without erasing the lower-level interactions extracted in the previous steps.

We design the aggregator by applying the Compact Trilinear Interaction (CTI) (Do et al., 2019) to question, answer, and image features and generate a vector to jointly represent the three features.

Given $V \in \mathbb{R}^{n_v \times d_v}$, $Q \in \mathbb{R}^{n_T \times d_q}$, and $G \in \mathbb{R}^{n_T \times d_k}$, we generate a joint representation $z \in \mathbb{R}^{d_z}$ of the three embeddings. The joint representation z is computed by applying CTI to each (V, Q, G) :

$$z = \sum_{i=1}^{n_v}\sum_{j=1}^{n_T}\sum_{k=1}^{n_T} \mathcal{M}_{ijk}\left(V_i W_{z_v} \circ Q_j W_{z_q} \circ G_k W_{z_g}\right) \tag{4}$$

where $\mathcal{M}$ is an attention map $\mathcal{M} \in \mathbb{R}^{n_v \times n_T \times n_T}$:

$$\mathcal{M} = \sum_{r=1}^{R} [\![\mathcal{G}_r; VW_{v_r}, QW_{q_r}, GW_{g_r}]\!] \tag{5}$$

where $W_{z_v}, W_{z_q}, W_{z_g}, W_{v_r}, W_{q_r}, W_{g_r}$ are learnable factor matrices, and $\circ$ is the Hadamard product. R is a slicing parameter, establishing a trade-off between the decomposition rate and the performance, and $\mathcal{G}_r \in \mathbb{R}^{d_{q_r} \times d_{v_r} \times d_{g_r}}$ is a learnable Tucker tensor.

The joint embedding computes more efficient and more compact representations than simply concatenating the embeddings. It creates a joint representation in a single space of the three different embedding spaces. In addition, we overcome the issue of dimensionality faced with concatenating large matrices.

The outputs of the aggregator is a joint concept-vision-language representation which is then fed to a classifier to predict the answer.

4 Experiments

We evaluate the performance of our proposed model using the standard evaluation metric recommended in the VQA challenge (Agrawal et al., 2017):

$$Acc(ans) = min\left(1, \frac{\#\{\text{humans provided ans}\}}{3}\right) \tag{6}$$

4.1 Datasets

All experiments have been performed on VQA 2.0 (Goyal et al., 2016) and Outside Knowledge-VQA (OK-VQA) (Marino et al., 2019) datasets.

VQA 2.0 is a public dataset containing about 1.1 million questions and 204,721 images extracted from the 265,016 images of the COCO dataset. At least 3 questions (5.4 questions on average) are provided per image, and each question is associated with 10 different answers obtained by crowd sourcing. Since VQA 2.0 is a large dataset, we only consider questions whose set of answers has at least 9 identical ones. With this common practice, we can cast aside questions which have lukewarm answers. The questions are divided in three categories: Yes/No, Number, and Other. We are especially interested in the ”Other” category, which can require external knowledge to find the correct answer.

OK-VQA: To evaluate the performance of our proposed model, we require questions which are not answerable by direct analysis of the objects detected in the image or the entities in the question. Most of knowledge-based VQA datasets impose hard constraints on their questions, such as being generated by templates (KB-VQA (Wang et al., 2015)) or directly obtained from existing knowledge bases (FVQA (Wang et al., 2016)). We select OK-VQA which is the only VQA dataset that requires handling unstructured knowledge to answer natural questions about images.

The OK-VQA dataset is composed of 14,031 images and 14,055 questions. For each question, we select the unanimous answer as the ground-truth answer. OK-VQA is divided into eleven categories: vehicles and transportation (VT); brands, companies and products (BCP); objects, materials and clothing (OMC); Sports and Recreation (SR); Cooking and Food (CF); Geography, History, Language and Culture (GHLC); People and Everyday Life (PEL); plants and animals (PA); science and technology (ST); weather and climate (WC). If a question was classified as belonging to different categories by different people, it was categorized as ”Other”.

4.2 Implementation details

In this section, we provide the implementation details of our proposed model in different building blocks.

Image embedding: Each image has a total of 36 image region features ($n_v = 36$), each represented by a bounding box and an embedding vector computed by pre-trained Faster R-CNN features where $d_v = 2048$. Each bounding box includes a 5-dimensional spatial coordinate ($d_b = 5$) corresponding to the coordinates of the top-left point of the bounding box, the coordinates of the bottom-right point of the bounding box, and the covered fraction of the image area.

Question embedding: The input questions are embedded using BERT's BASE model. Therefore, each word is represented by a 768-D word embedding ($d_q = 768$). Each question is divided into 16-token blocks ($n_T = 16$), starting with a `[CLS]` token and ending with a `[SEP]` token. The answers are transformed to one-hot encoding vectors.

Knowledge graph embedding: During our experiments, we explored different node embeddings for ConceptNet (e.g. GloVe (Pennington et al., 2014), NumberBatch (Speer et al., 2016), and (Malaviya et al., 2020)). We found that the embedding generated by (Malaviya et al., 2020) works best in our model.

Vision-Language representation: We initialize our vision-language representation with pretrained ViLBERT features. The ViLBERT model is built on the Conceptual Captions dataset (Sharma et al., 2018), which is a collection of 3.3 million image-caption pairs, to capture the diversity of visual content and learn some interactions between images and text. Our vision-language module includes 6 layers of Transformer blocks with 8 and 12 attention heads in the visual stream and linguistic streams, respectively.

Concept-Language representation: We train the concept stream of our ConceptBert from scratch. The module includes 6 layers of Transformer blocks with 12 attention heads.

Concept-Vision-Language embedding: We have tested our concept-vision-language representation with $d_z = 512$ and $d_z = 1024$. The best results were reached using $d_z = 1024$. Our hypothesis is that we can improve the capability of the module by increasing d_z. However, it leads to an increase in the computational cost. We set $R = 32$ in Equation 5, the same value as in the CTI (Do et al., 2019) for the slicing parameter.

Classifier: We use a binary cross-entropy loss with a batch size of 1024 over a maximum of 20 epochs on 8 Tesla GPUs. We use the BertAdam

Dataset	L	VL	CL	CVL
VQA 2.0	26.68	67.9	38.24	**69.95**
OK-VQA	14.93	31.35	22.12	**33.66**

Table 1: Evaluation results on **VQA 2.0** and **OK-VQA** validation sets for ablation study

Model	Overall	Yes/No	Number	Other
Up-Down	59.6	80.3	42.8	55.8
XNM Net	64.7	-	-	-
ReGAT	67.18	-	-	-
ViLBERT	67.9	82.56	54.27	67.15
SIMPLE	67.9	82.70	54.37	67.21
CONCAT	68.1	82.96	54.57	68.00
ConceptBert	**69.95**	**83.99**	**55.29**	**70.59**

Table 2: Evaluation results of our model compared with existing algorithms on **VQA 2.0** validation set.

optimizer with an initial learning rate of 4e-5. A linear decay learning rate schedule with warm up is used to train the model.

4.3 Experimental results

This sub-section provides experimental results on the VQA 2.0 and OK-VQA datasets.

Ablation Study: In Table 1, we compare two ablated instances of ConceptBert with its complete form. Specifically, we validate the importance of incorporating the external knowledge into VQA pipelines on top of the vision and language embeddings. Table 1 reports the overall accuracy on the VQA 2.0 and OK-VQA validation sets in the following setting:

- *L*: Only questions features $\boldsymbol{q}$ are fed to the classifier.
- *VL*: Only the outputs of the Vision-Language representation module $[V; Q]$ are concatenated and fed to the classifier.
- *CL*: Only the output of the Concept-Language representation module G is fed to the classifier.
- *CVL*: ConceptBert complete form; the outputs of both Vision-Language and Concept-Language modules are fused (cf. Section 3.4) and fed to the classifier.

Comparison between *L* and *CL* instances shows the importance of incorporating the external knowledge to accurately predict answers. Adding the KG embeddings to the model leads to a gain of 11.56% and 7.19% in VQA and OK-VQA datasets, respectively.

We also note that the *VL* model outperforms the *CL* model. The reason is that most of the questions in both VQA 2.0 and OK-VQA datasets are related to objects found in the images. Therefore, the accuracy drops without providing the detected object features. Compared to the *VL* and *CL*, the *CVL* model gives the highest accuracy which indicates the effectiveness of the joint concept-vision-language representation.

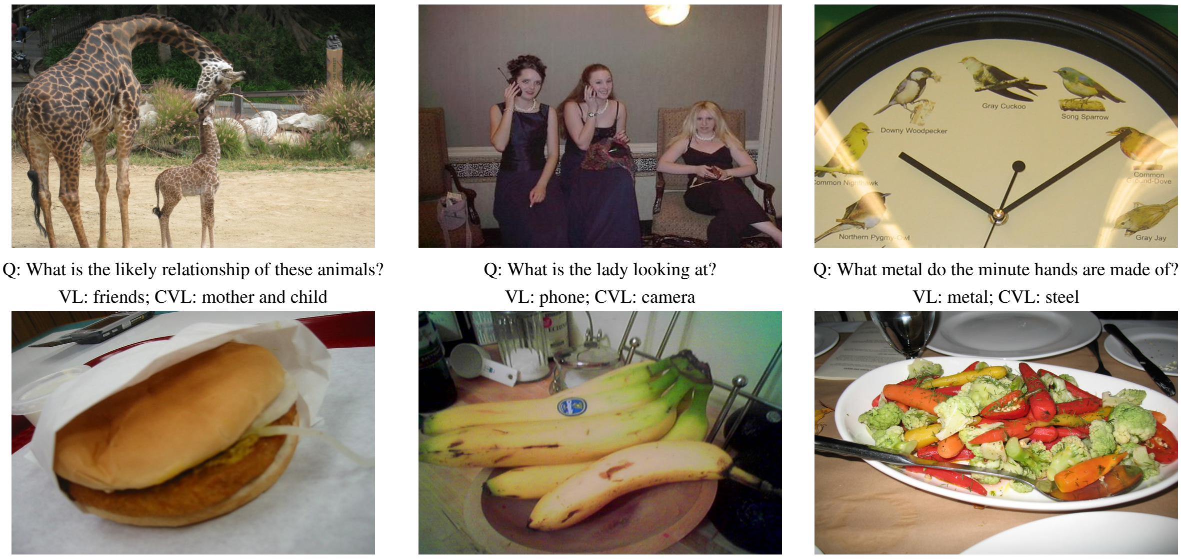
Results on VQA 2.0 dataset: The performance of our complete model on VQA 2.0 validation set is compared with the existing models in Table 2. Up-Down model (Anderson et al., 2017) combines the bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects. XNM Net (Shi et al., 2018) and ReGAT (Li et al., 2019) are designed to answer semantically-complicated questions. In addition to the existing approaches we elaborated two other baselines: (i) SIMPLE: First, we create the embedding G, which is the output of the concept-language module. Then, we use G and the image embedding, feed them to the vision-language module, and send its output to a classifier and check the answer. (ii) CONCAT: we concatenate the embeddings from the question and ConceptNet to form a mixed embedding Q_{KB}. Then, we send Q_{KB} and the image embedding to the vision-language module, and feed its output to a classifier and check the answer. It is worthy to note that SIMPLE and CONCAT do not have CTI involved. The results show that our model outperforms the existing models. Since we report our results on the validation set, we removed the validation set from the training phase, so that the model only relies on the training set.

Results on OK-VQA dataset: Table 3 shows the performance of our complete model on OK-VQA validation set. Since there exists only one work on OK-VQA dataset in the literature, we apply a few state-of-the-art models on OK-VQA and report their performance. We also performed SIMPLE and CONCAT baselines on OK-VQA dataset. In the OK-VQA study (Marino et al., 2019), the best results are obtained by fusing MUTAN and ArticleNet (MUTAN + AN) as a knowledge-based baseline. AN retrieves some articles from Wikipedia for each question-image pair and then train a network to predict whether and where the ground-truth answers appear in the article and in each sentence.

From the table, we observe that our model surpasses the baselines and SOTA models in almost every category which indicates the usefulness of

Model	Overall	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
XNM Net	25.61	26.84	21.86	18.22	33.02	23.93	23.83	20.79	24.81	21.43	42.64	24.39
MUTAN+AN	27.58	25.56	23.95	26.87	33.44	29.94	20.71	25.05	29.70	24.76	39.84	23.62
ViLBERT	31.35	27.92	26.74	29.72	35.24	31.93	**34.04**	26.54	30.49	27.38	46.2	28.72
SIMPLE	31.37	28.12	26.84	29.77	35.77	31.99	29.09	26.99	31.09	27.66	46.28	28.81
CONCAT	31.95	28.66	27.01	29.81	35.88	32.89	31.04	26.94	31.99	28.01	46.33	29.01
ConceptBert	**33.66**	**30.38**	**28.02**	**30.65**	**37.85**	**35.08**	32.91	**28.55**	**35.88**	**32.38**	**47.13**	**31.47**

Table 3: Evaluation results of our model compared with the SOTA algorithms on **OK-VQA** validation set.



Q: What is the likely relationship of these animals?
VL: friends; CVL: mother and child

Q: What is the lady looking at?
VL: phone; CVL: camera

Q: What metal do the minute hands are made of?
VL: metal; CVL: steel

Q: What condiment is hanging out of the sandwich?
VL: mustard; CVL: onion

Q: What is laying on a banana?
VL: nothing; CVL: sticker

Q: What vegetable is on the lower most plate?
VL: celery; CVL: carrot

Figure 3: VQA examples in the category ”Other”: ConceptBert complete form *CVL* outperforms the *VL* model on the question Q.

external knowledge in predicting answers. ConceptBert performs especially well in the ”Cooking and Food” (CF), ”Plants and Animals” (PA), and ”Science and Technology” (ST) categories with a gain larger than 3%. The answers to these type of questions often are entities out of the main entities in the question and the visual features in the image. Therefore, the information extracted from the knowledge graph plays an important role in determining the answer. ViLBERT performs better in the category ”Geography, History, Language and Culture” (GHLC) compared to ConceptBert, since ”dates” are not entities in ConceptNet.

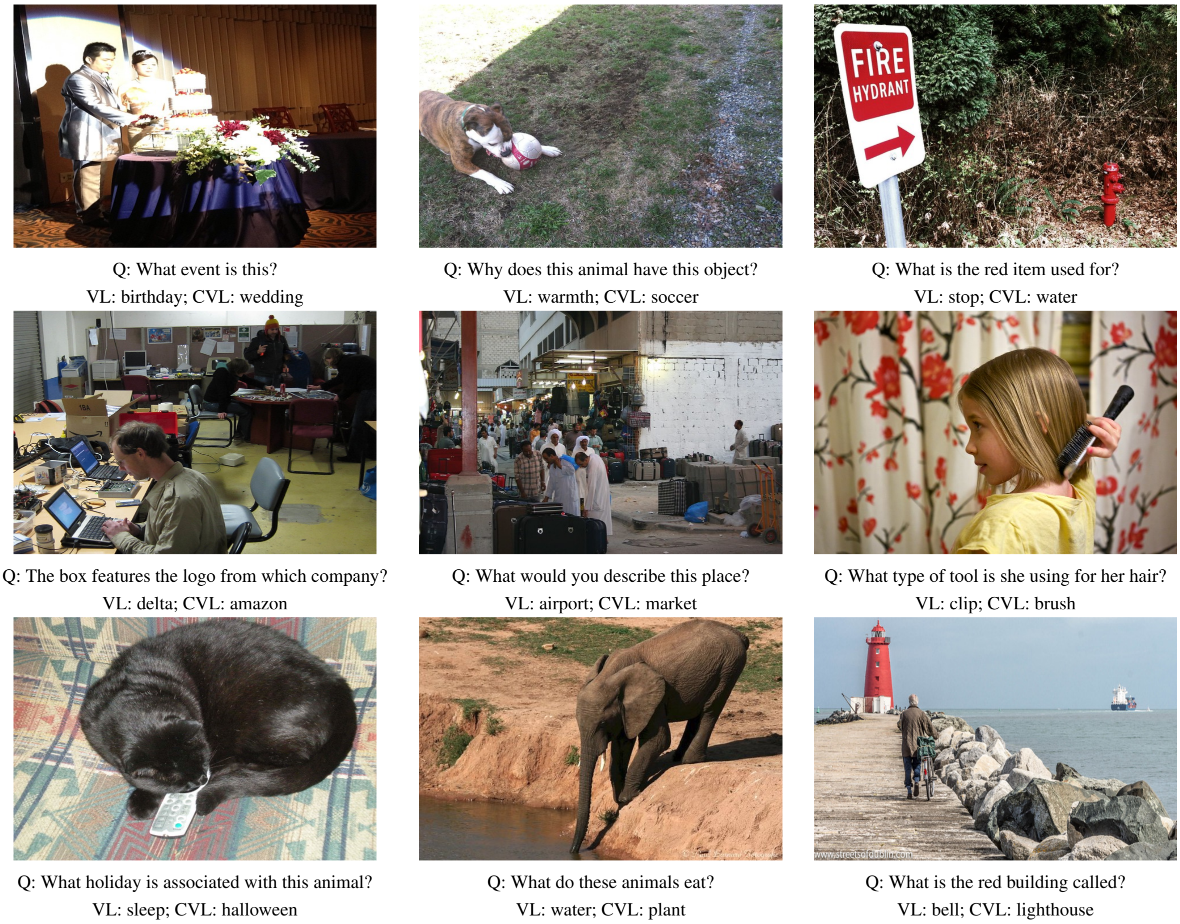
4.4 Qualitative results

We illustrate some qualitative results of ConceptBert complete form *CVL* by comparing it with the *VL* model. In particular, we aim at illustrating the advantage of adding (i) the external knowledge extracted from the ConceptNet knowledge graph, and (ii) concept-vision-language embedding representations.

Figure 3 and Figure 4 illustrate some qualitative results on VQA 2.0 and OK-VQA validation sets, respectively.

From the figures, we observed that the *VL* model is influenced by the objects detected in the picture. However, the *CVL* model is able to identify the correct answer without only focusing on the visual features. For example in the third row in Figure 4, *CVL* model uses the facts that an elephant is herbivorous, and black cat is associated with Halloween to find the correct answers.

It is worthy to note that the *CVL* answers remain consistent from a semantic perspective even in the case of wrong answers. For example, *How big is the distance between the two players?* exposes a distance as opposed to the *VL* model which provides a Yes/No answer (cf. Figure 5). In another example for the question *Sparrows need to hide to avoid being eaten by what?*, the *CVL* model mentions an animal species that can eat sparrows, while the *VL* model returns an object found in the image. From these visualization results, we observe that the knowledge strongly favours the capture of interactions between objects, which contributes



Q: What event is this?
VL: birthday; CVL: wedding

Q: Why does this animal have this object?
VL: warmth; CVL: soccer

Q: What is the red item used for?
VL: stop; CVL: water

Q: The box features the logo from which company?
VL: delta; CVL: amazon

Q: What would you describe this place?
VL: airport; CVL: market

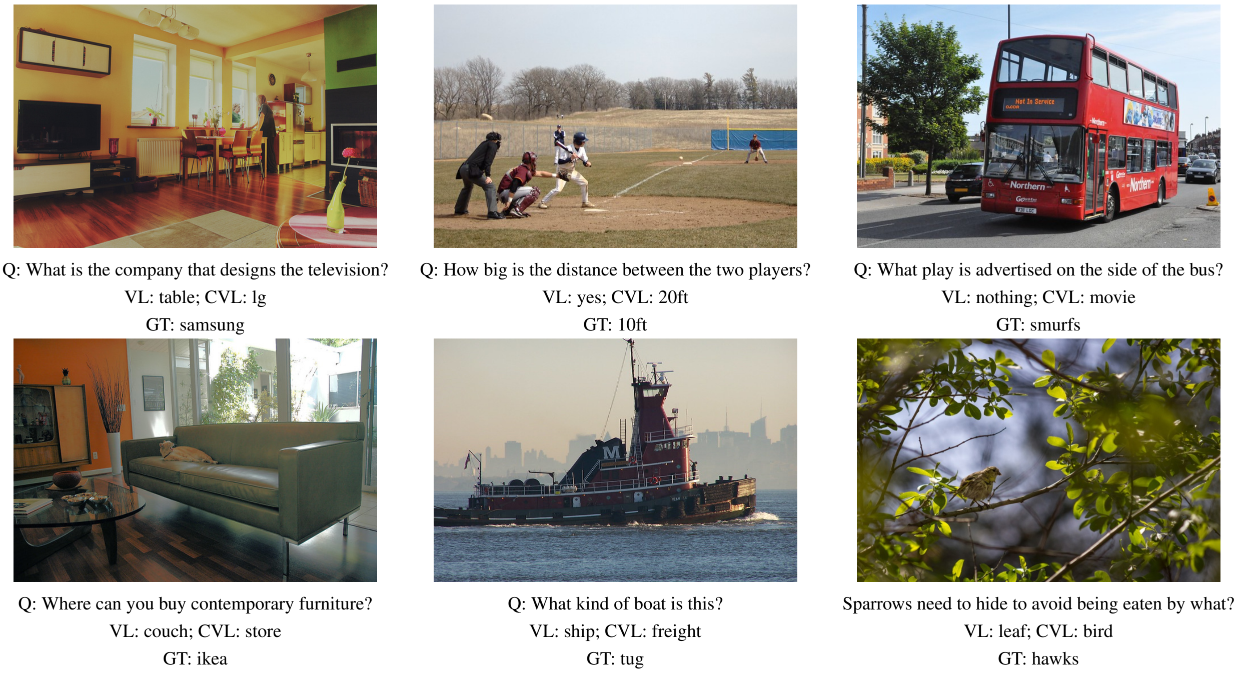
Q: What type of tool is she using for her hair?
VL: clip; CVL: brush

Q: What holiday is associated with this animal?
VL: sleep; CVL: halloween

Q: What do these animals eat?
VL: water; CVL: plant

Q: What is the red building called?
VL: bell; CVL: lighthouse

Figure 4: OK-VQA examples: ConceptBert complete form *CVL* outperforms the *VL* model on the question Q.



Q: What is the company that designs the television?
VL: table; CVL: lg
GT: samsung

Q: How big is the distance between the two players?
VL: yes; CVL: 20ft
GT: 10ft

Q: What play is advertised on the side of the bus?
VL: nothing; CVL: movie
GT: smurfs

Q: Where can you buy contemporary furniture?
VL: couch; CVL: store
GT: ikea

Q: What kind of boat is this?
VL: ship; CVL: freight
GT: tug

Sparrows need to hide to avoid being eaten by what?
VL: leaf; CVL: bird
GT: hawks

Figure 5: ConceptBert complete form *CVL* identifies answers of the same type as the ground-truth answer (GT) compared with the *VL* model on the question Q. VQA and OK-VQA examples are shown in the first and second rows, respectively.

to a better alignment between image regions and questions.

5 Conclusions

In this paper, we present ConceptBert, a concept-aware end-to-end pipeline for questions which require knowledge from external structured content. We introduce a new representation of questions enhanced with the external knowledge exploiting Transformer blocks and knowledge graph embeddings. We then aggregate vision, language, and concept embeddings to learn a joint concept-vision-language embedding. The experimental results have shown the performance of our proposed model on VQA 2.0 and OK-VQA dataset.

For future work, we will investigate how to integrate the explicit relations between entities and objects. We believe that exploiting the provided relations in knowledge graphs and integrating them with relations found between objects in questions/images can improve the predictions.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. Vqa: Visual question answering. *Int. J. Comput. Vision*, 123(1):4–31.

Luis Von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: a game for collecting common-sense facts. In *In Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems, volume 1 of Games*, pages 75–78. ACM Press.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D. Tran. 2019. Compact trilinear interaction for visual question answering.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of AAAI 2020*.

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

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. *CoRR*, abs/1906.00067.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *CoRR*, abs/1811.00538.

Medhini Narasimhan and Alexander G. Schwing. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. *CoRR*, abs/1809.01124.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2018. Explainable and explicit visual reasoning over scene graphs.

Push Singh et al. 2002. The public acquisition of commonsense knowledge.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. 2017. Image captioning and visual question answering based on attributes and external knowledge. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 1290–1296.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2016. Fvqa: Fact-based visual question answering.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *CoRR*, abs/1511.02570.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base for visual question answering. *CoRR*, abs/1507.05670.