Enhancing Language & Vision with Knowledge - The Case of Visual Question Answering

Freddy Lecue
CortAIx, Thales, Canada
Inria, France
http://www-sop.inria.fr/members/Freddy.Lecue/

Maryam Ziaeefard, François Gardères (as contributors)
CortAIx, Thales, Canada

(Keynote)
2020 International Conference on Advance in Ambient Computing and Intelligence

Introduction

What is Visual Question Answering (aka VQA)?

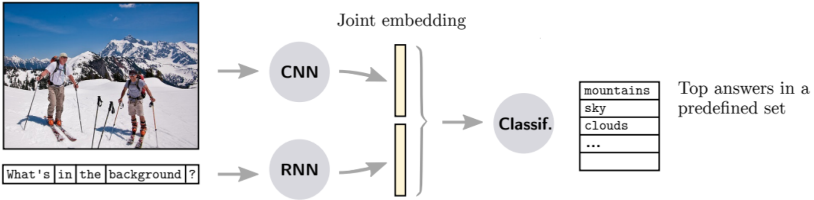
The objective of a **VQA model** combines visual and textual features in order to **answer questions** grounded in an **image**.

What's in the background?

Where is the child sitting?

Classic Approaches to VQA

Most approaches combine **Convolutional Neural Networks** (CNN) with **Recurrent Neural Networks** (RNN) to learn a mapping directly from **input images** (vision) and **questions** to **answers** (language):



Visual Question Answering: A Survey of Methods and Datasets. Wu et al (2016)

Evaluation [1]

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

An answer is deemed 100% accurate if at least 3 workers provided that exact answer.

Example: What sport can you use this for?

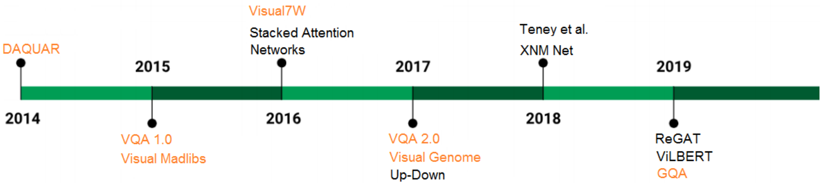
 # {human provided ans}: race (6 times), motocross (2 times), ride (2 times)

Predicted answer: motocross

Acc (motocross): $\min(1, \frac{2}{3}) = 0.66$

VQA Models - State-of-the-Art

Major breakthrough in VQA (models and real-image dataset)



Accuracy Results:

DAQUAR [2] (13.75 %), VQA 1.0 [1] (54.06 %), Visual Madlibs [3] (47.9 %), Visual7W [4] (55.6 %), Stacked Attention Networks [5] (VQA 2.0: 58.9 %, DAQAUR: 46.2 %), VQA 2.0 [6] (62.1 %), Visual Genome [7] (41.1 %), Up-down [8] (VQA 2.0: 63.2 %), Teney et al. (VQA 2.0: 63.15 %), XNM Net [9] (VQA 2.0: 64.7 %), ReGAT [10] (VQA 2.0: 67.18 %), ViLBERT [11] (VQA 2.0: 70.55 %), GQA [12] (54.06 %)

[2] Malinowski et al, [3] Yu et al, [4] Zhu et al, [5] Yang et al., [6] Goyal et al, [7] Krishna et al, [8] Anderson et al, [9] Shi et al, [10] Li et al, [11] Lu et al, [12] Hudson et al

Limitations

- Answers are required to be in the **image**.
- **Knowledge** is limited.
- Some questions cannot be correctly answered as **some levels of (basic) reasoning** is required.

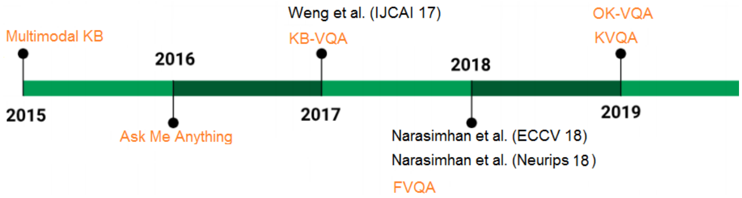
Alternative strategy: Integrating external knowledge such as domain **Knowledge Graphs**.

What sort of vehicle uses this item?

When was the soft drink company shown first created?

Knowledge-based VQA models - State-of-the-Art

- Exploiting **associated facts for each question** in VQA datasets [18], [19];
- **Identifying search queries** for each question-image pair and using a search API to retrieve answers ([20], [21]).



Accuracy Results:

Multimodal KB [17] (NA), Ask me Anything [18] (59.44 %), Weng et al (VQA 2.0: 59.50 %), KB-VQA [19] (71 %), FVQA [20] (56.91 %), Narasimhan et al. (ECCV 2018) (FVQA: 62.2 %) , Narasimhan et al. (Neurips 2018) (FVQA: 69.35 %), OK-VQA [21] (27.84 %), KVQA [22] (59.2 %)

[17] Zhu et al, [18] Wu et al, [19] Wang et al, [20] Wang et al, [21] Marino et al, [22] Shah et al

Our Contribution

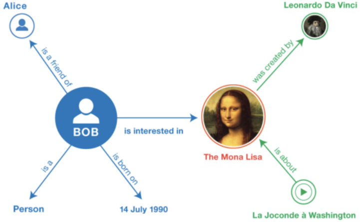
Yet Another Knowledge Base-driven Approach? **No.**

- We go one step further and implement a VQA model that relies on **large-scale** knowledge graphs.
- No dedicated **knowledge annotations** in VQA datasets neither **search queries**.
- Implicit integration of **common sense knowledge** through knowledge graphs.

Knowledge Graphs (1)

- Set of (*subject, predicate, object* – SPO) **triples** - *subject* and *object* are **entities**, and *predicate* is the **relationship** holding between them.
- Each SPO **triple** denotes a **fact**, i.e. the existence of an actual relationship between two entities.

subject	predicate	object
Bob	*is interested in*	*The Mona Lisa*
Bob	*is a friend of*	*Alice*
The Mona Lisa	*was created by*	*Leonardo Da Vinci*
Bob	*is a*	*Person*
La Joconde à W.	*is about*	*The Mona Lisa*
Bob	*is born on*	*14 July 1990*

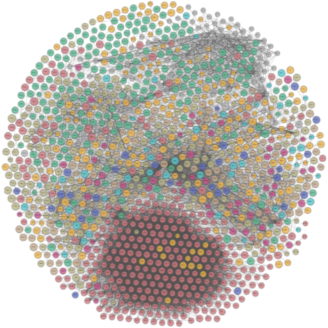


Knowledge Graphs (2)

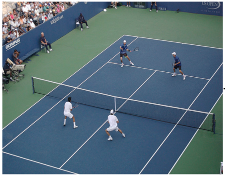
- **Manual Construction** - curated, collaborative
- **Automated Construction** - semi-structured, unstructured

Right: **Linked Open Data cloud** - over 1200 interlinked KGs encoding more than 200M facts about more than 50M entities.
Spans a variety of domains - Geography, Government, Life Sciences, Linguistics, Media, Publications, Cross-domain.

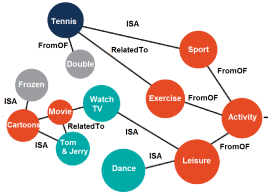
Name	Entities	Relations	Types	Facts
Freebase	40M	35K	26.5K	637M
DBpedia (en)	4.6M	1.4K	735	580M
YAGO3	17M	77	488K	150M
Wikidata	15.6M	1.7K	23.2K	66M
NELL	2M	425	285	433K
Google KG	570M	35K	1.5K	18B
Knowledge Vault	45M	4.5K	1.1K	271M
Yahoo! KG	3.4M	800	250	1.39B



Problem Formulation



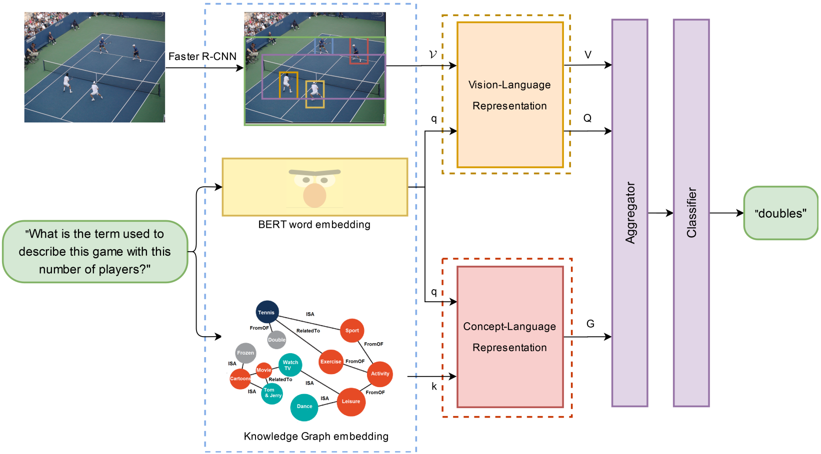
"What is the term used to describe this game with this number of players?"



VQA model

"doubles"

Our Machine Learning Pipeline



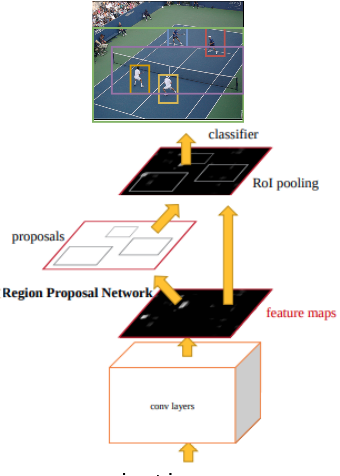
V: Language-attended visual features.
Q: Vision-attended language features.
G: Concept-language representation.

Image Representation - Faster R-CNN

✓ Post-processing CNN with region-specific image features **Faster R-CNN** [24] - Suited for VQA [23].

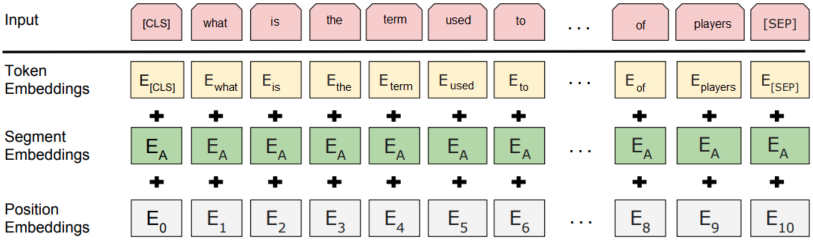
✓ We use pretrained Faster R-CNN to extract 36 objects per images and their bounding box coordinates.

Other region proposal networks could be trained as an alternative approach.



[23] Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. Teney et al. (2017)
[24] Faster R-CNN: towards real-time object detection with region proposal networks. Ren et al. (2015)

Language (Question) Representation - BERT

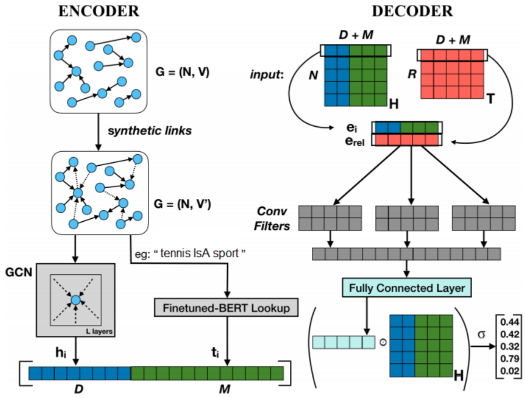


✓ **BERT embedding** [25] for **question representation**. Each question has 16 tokens.

✓ BERT shows the value of transfer learning in NLP and makes use of **Transformer**, an attention mechanism that learns contextual relations between words in a text.

ConceptNet
An open, multilingual knowledge graph

only KG that designed to understand the meanings of word that people use and include common sense knowledge.



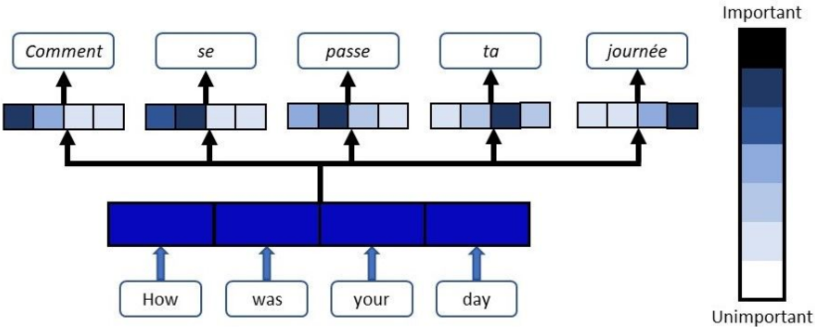
Pre-trained ConceptNet embedding [26] (with dimension = 200).

[26] Commonsense knowledge base completion with structural and semantic context. Malaviya et al. (AAAI 2020)

- Attention learns a context vector, informing about the **most important information** in inputs for given outputs.

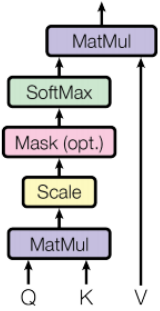
Example

Attention in machine translation (Input: English, Output: French):



Attention Mechanism (More Technical)

Scaled Dot-Product Attention [27].
Query Q: Target / Output embedding.
Keys K, Values V: Source / Input embedding.



Scaled Dot-Product Attention

✓ Machine translation example: Q is an embedding vector from the target sequence. K, V are embedding vectors from the source sequence.

✓ Dot-product similarity between Q and K determines attentional distributions over V vectors.

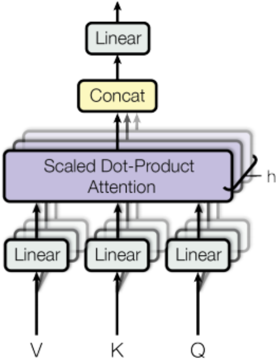
✓ The resulting weight-averaged value vector forms the output of the attention block.

[27] Attention Is All You Need. Vaswani et al. (NeurIPS 2017)

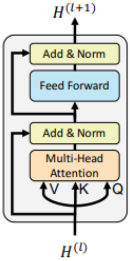
Attention Mechanism - Transformer

Multi-head Attention: Any given word can have multiple meanings $\rightarrow$ more than one query-key-value sets

Encoder-style Transformer Block: A multi-headed attention block followed by a small fully-connected network, both wrapped in a residual connection and a normalization layer.

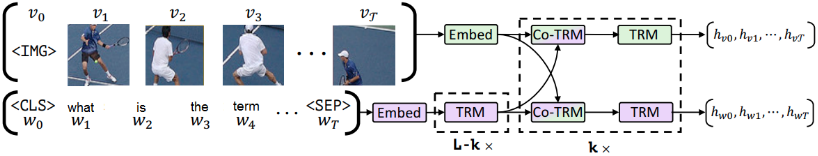


Multi-Head Attention

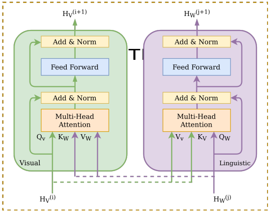


Encoder-style transformer

Vision-Language (Question) Representation



Joint **vision-attended language** features and **language-attended visual** features to learn **joint representations** using VilBERT model [28].



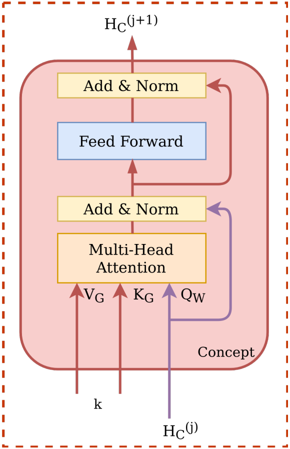
[28] Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Lu et al. (2019)

✓ Questions features are conditioned on knowledge graph embeddings.

✓ The concept-language module is a series of **Transformer** blocks that attends to question tokens based on KG embeddings.

✓ The input consists of **queries** from **question embeddings** and **keys** and **values** of **KG embeddings**.

✓ Concept-Language representation enhances the question comprehension with the information found in the knowledge graph.



Concept-Vision-Language Module

Compact Trilinear Interaction (CTI) [29] applied to each (V, Q, G) to achieve the joint representation of concept, vision, and language.

- V represents language-attended visual features.
- Q shows vision-attended language features.
- G is concept-attended language features.

✓ Trilinear interaction to **learn** the **interaction between V, Q, G**.
✓ By computing the attention map between all possible combinations of V, Q, G. These attention maps are used as weights. Then, the joint representation is computed with a weighted sum over all possible combinations.
(There are $n1 \times n2 \times n3$ possible combinations over the three inputs with dimensions $n1$, $n2$, and $n3$).

[29] Compact trilinear interaction for visual question answering. Do et al. (ICCV 2019)

Implementation Details

- **Vision-Language Module**: 6 layers of Transformer blocks, 8 and 12 attention heads in the visual stream and linguistic streams, respectively.
- **Concept-Language Module**: 6 layers of Transformer blocks, 12 attention heads.
- **Concept-Vision-Language Module**: embedding size = 1024
- **Classifier**: binary cross-entropy loss, batch size = 1024, 20 epochs, BertAdam optimizer, initial learning rate = 4e-5.

- Experiments conducted on NVIDIA 8 TitanX GPUs.

Datasets (1)

VQA 2.0 [30]

- 1.1 million questions. 204,721 images extracted from COCO dataset (265,016 images).
- At least 3 questions (5.4 questions on average) are provided per image.
- Each question: 10 different answers (through crowd sourcing).
- Questions categories: Yes/No, Number, and Other
- Special interest: "Other" category.

[30] Making the v in vqa matter: Elevating the role of image understanding in visual question answering. Goyal et al. (CVPR 2017)

Datasets (2)

Outside Knowledge-VQA (OK-VQA) [31]

- Only VQA dataset that requires external knowledge.
- 14,031 images and 14,055 questions.
- 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), and "Other".

[31] Ok-vqa: A visual question answering benchmark requiring external knowledge. Marino et al (CVPR 2019)

Results and Lessons Learnt (1)

Model	Overall	Yes/No	Number	Other
Up-Down	63.2	80.3	42.8	55.8
XNM Net	64.7	-	-	-
ReGAT	67.18	-	-	-
ViLBERT	68.14	**82.99**	54.27	67.15
ConceptBert	**71.81**	81.56	**61.29**	**72.59**

Table 1: Our Model vs. State-of-the-art Approaches on **VQA 2.0**

- Integrating common sense knowledge improves overall performance (5.3% higher).
- Major improvement in **"Other"** category.
- ViLBERT outperforms on **Yes/No** questions as they are more towards direct analysis of the image.

Results and Lessons Learnt (2)

Model	Overall	VT	BCP	OMC	WC	GHLC
XNM Net	25.24	26.84	21.86	18.22	42.64	23.83
MUTAN+AN	27.84	25.56	23.95	26.87	39.84	20.71
ViLBERT	31.47	26.74	**29.72**	**30.65**	46.20	31.47
ConceptBert	**36.10**	**30.02**	28.92	30.38	**53.13**	**36.91**

Model	CF	PEL	PA	ST	SR	Other
XNM Net	23.93	20.79	24.81	21.43	33.02	24.39
MUTAN+AN	29.94	25.05	29.70	24.76	33.44	23.62
ViLBERT	31.93	26.54	30.49	27.38	35.24	28.72
ConceptBert	**37.04**	**31.55**	**37.88**	**34.38**	**39.85**	**37.08**

Table 2: Our Model vs. State-of-the-art Approaches on **OK-VQA**

- Our model is better in **PA, ST, and CF** categories (14.7% higher).
- ViLBERT outperforms our model on **OMC** and **BCP** categories, respectively. Questions more towards direct analysis of the image.

Qualitative Results (1)

Q: What is the likely relationship ? of these animals?
V: friends
C: mother

Q: What is the lady looking at?
V: phone
C: camera

Figure 1: **VQA 2.0** examples in category "Other": ConceptBert (C) outperforms ViLBERT (V) on Question Q.

Qualitative Results (2)

Q: How big is the distance between the two players?
V: yes
C: 20ft
GT: 10ft

Q: What play is being advertised on the side of the bus?
V: nothing
C: movie
GT: smurfs

Figure 2: **VQA 2.0** examples: ConceptBert (C) identifies answers of the same type as ground-truth GT when compared with ViLBERT (V) on Question Q.

Qualitative Results (3)

Q: What holiday is associated with this animal?
V: sleep
C: halloween

Q: What do these animals eat?
V: water
C: plant

Figure 3: **OK-VQA** examples: ConceptBert (C) outperforms ViLBERT (V) on Question Q.

Qualitative Results (4)

Q: Where can you buy contemporary furniture?

V: couch

C: store

GT: ikea

Q: What kind of boat is this?

V: ship

C: freight

GT: tug

Figure 4: **OK-VQA** examples: ConceptBert (C) identifies answers of the same type as ground-truth GT when compared with ViLBERT (V) on Question Q.

Conclusion and Future Work

- Concept-aware VQA model for questions which require common sense knowledge from external structured content.
- Novel representation of questions enhanced with commonsense knowledge exploiting Transformer blocks and knowledge graph embeddings.
- Aggregation of vision, language, and concept embeddings to learn a joint concept-vision-language embedding for VQA tasks.

Future work

- Integrating explicit relations between entities and objects in knowledge graph.
- Evaluation through a semantic metric.
- Integrating spatial relations or scene graphs to VQA models.