Curriculum Learning for Compositional Visual Reasoning

Wafa Aissa, Marin Ferecatu, Michel Crucianu

Cedric laboratory, CNAM Paris

Visual Reasoning

- Reasoning about the visual world: manipulate previously acquired knowledge to **understand** an image and reason about the different objects, environment, actions ...
- Evaluate the reasoning skills with Visual Question Answering (VQA) tasks.

Problem

- Visual scene understanding goes beyond visual recognition and object detection/segmentation.
- Visual reasoning about **complex scenes** is extremely hard.
- **Multi-modal** reasoning requires good image and text representations.
- Machines are still far from **human-like learning and reasoning**.

Curriculum learning for VQA

This work proposes a compositional reasoning framework trained by a CL strategy on real world images.

Related work

Fusion methods

- CNN/Transformer extract image features.
- LSTM/Transformer extract question embeddings.
- Multi-modal attention.

+ Performance gains due to the power of DNNs.

– Lack interpretability.

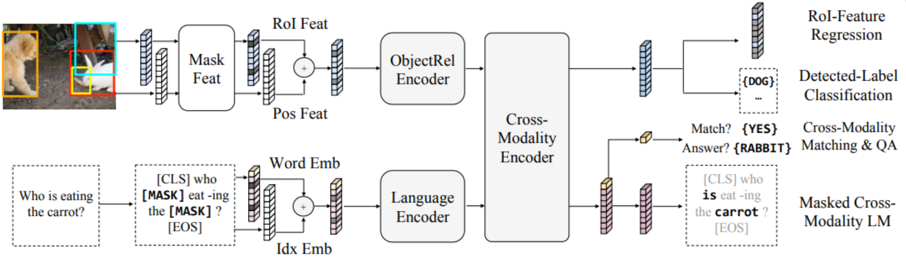


Figure 1: LXMERT: Learning Cross-Modality Encoder Representations from Transformers [1].

- Visual reasoning is inherently **compositional**.
- Break down the question into **modular** sub-problems.
- **Reasoning skills**: transitive and logical relations, counting and comparisons...
- Sets a road towards more explainable and human logic inference.

Figure 2: What color is the fruit on the right side, red or green?
Is there any milk in the bowl to the left of the apple?

- A modular, composable and jointly trained NNs framework for VQA.
- Language parser Andreas et al. [2] or recently LSTM to generate the layouts.
- **Transparency** and **explainability**.
- Each module operates to accomplish a different **sub-task**.

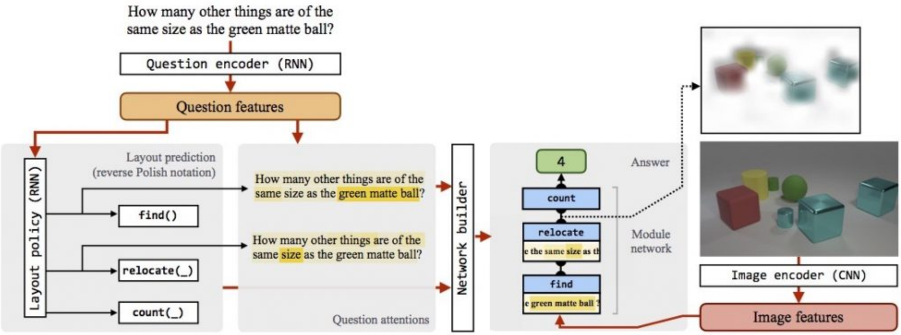


Figure 3: Learning to reason model overview [3].

- Real world reasoning augmented with supervised **knowledge guidance**.
  - **Bboxes** of relevant visual regions for attention modules.
  - KL divergence between predicted attention maps and knowledge guidance.
- Model layout generation is similar to Hu et al. [3].
- **Expected scores** over candidate answers for the other modules.

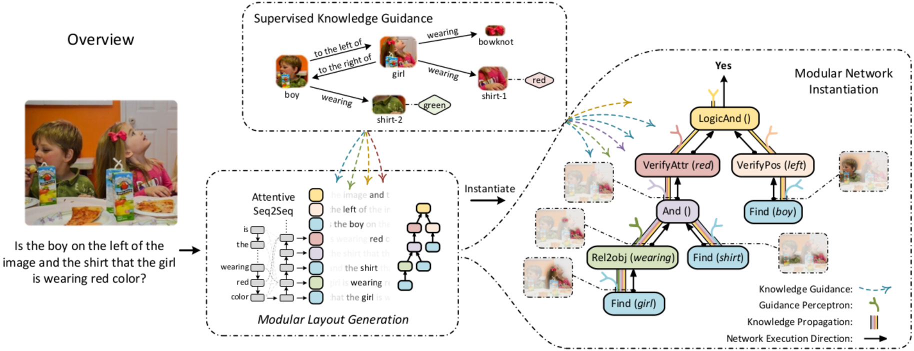


Figure 4: Perceptual Visual Reasoning overview Li, Wang, and Zhu [4].

Curriculum Learning CL

- CL is a **start small** training strategy similar to human learning [5].
- Start by easy training examples then gradually increase the complexity of the examples.
- Successfully applied to ML tasks and recently to textual QA and synthetic images VQA.
- Speed up convergence and use less training examples [5].

How to define the complexity ?

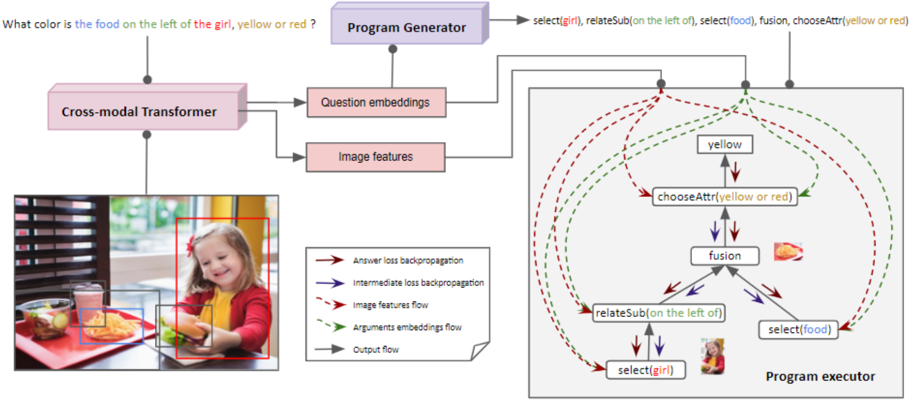
- A term frequency selector and a grammar selector [6].
- Answer loss as the hardness measure [7] inspired by Self-Paced Learning.
- Heuristics: program length, answer hierarchy and loss-based hardness [8].

CL for VQA

This work proposes a compositional reasoning framework trained by a CL strategy on real world images.

Modular VQA framework

- Extract **aligned cross-modal embeddings** for words and objects, freeze LXMERT encoder [1].
- Generator: Transformer decoder to decompose the reasoning task into a modules program.
- **Executor**: instantiate and run the program over the image and answer the question.
- **Intermediate modules losses** to supervise the modules.
- **Textual argument** to indicate the desired module's facet.



- Modules perform **reasoning sub-tasks**: object detection, filter attribute, logic ...
- **Dependencies** to get information from the previous module.
- Three module groups: **attention, boolean, answer**.
- Modules are basic algorithmic operations such as dot products and MLPs.

Name	Dependencies	Output	Definition
Select	–	attention	$x = r(W\,t), Y = r(WV)$ $o = S(W(Y^T x))$
RelateSub	[a]	attention	$x = r(W\,t), Y = r(WV), z = S(W(Y^T x))$ $o = S(W(x \odot y \odot z))$
VerifyAttr	[a]	boolean	$x = r(W\,t), y = r(W(V\,a)$ $o = \sigma(W(x \odot y))$
And	[b_1,b_2]	boolean	$o = b_1 \times b_2$
ChooseAttr	[a]	answer	$x = r(W\,t), y = r(W(V\,a)$ $o = S(W(x \odot y))$
QueryName	[a]	answer	$y = r(W(V\,a))$ $o = S(W\,y)$

Table 1: Sample module definitions. S: softmax, σ: sigmoid, r: RELU, W_i: weight matrix, a: attention vector (36×1), V: visual features (768×36), t: text features (768×1), $\odot$: Hadamard product.

Curriculum Learning (CL) for Visual Question Answering (VQA)

CL training protocol

- **Number of objects** in the question is the a priori **difficulty criterion** for CL.
- CL **difficulty refinement** based on **program length**.
- **CL scheduler** to update the curriculum (1M examples).
- Weight the examples using a **sampling function**.
- Balance the occurrence probabilities of the different types of answer modules.
- Avoid catastrophic forgetting by retaining few previous examples (20%).

Question: Which color do you think the train car is?'

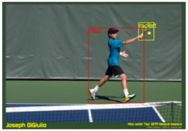
Program: select(train car), queryAttr(color).

Difficulty level: 1

Question: Is the racket to the right or to the left of the person in the middle of the picture?'

Program: select(person), filterPos(middle), select(racket), choosePos(to the left or to the right).

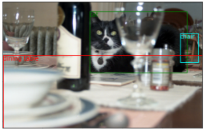
Difficulty level: 2

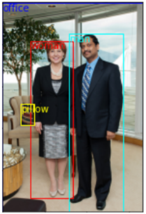


Question: What do you think is the piece of furniture to the right of the white animal that is lying on the dining table?'

Program: select(dining table), relateSub(lying on), select(animal), fusion, filterAttr(white), relateSub(to the right of), select(furniture), fusion, queryName.

Difficulty level: 3





Question: Do both the man in the office and the woman to the right of the pillow look happy?

Program: select(office), relateSub(in), select(man), fusion, verifyAttr(happy), select(pillow), relateSub(to the right of), select(women), fusion, verifyAttr(happy), and, answerLogic.

Difficulty level: 4

Figure 5: *Dataset samples with different difficulty levels.*

Experimental details & analysis

Dataset & implementation details

- GQA dataset: Balanced $\subset$ Unbalanced.
- Train on the unbalanced split to have more examples.
- Experiment on pre-processed GQA programs and investigate CL on **Program Executor**.
- Cross entropy loss for intermediate attention modules and BCE for boolean modules.
- **Weight sharing** between compatible modules.
- CL uses **sampling with replacement**.

Split	Train	Val	Testdev	Test	Challenge
Balanced	943.000	132.062	12.578	95.336	50.726
Unbalanced	14.305.356	2.011.853	172.174	1.340.048	713.449

Table 2: GQA dataset partitioning

Evaluated methods

- **Unbalanced**: Unbalanced GQA with random batch training (14M per epoch).
- **Balanced**: Balanced GQA split with random batch training (1M per epoch).
- **Random**: 1M random examples from unbalanced GQA at every iteration.
- **CL**: CL trainining strategy with increasing difficulty. 1M programs meticulously sampled from unbalanced GQA every iteration.
  - **Length (L)**: Filter by length (short, medium, or long) withing each difficulty level.
  - **Weights (W)**: Sampling weights: 'uniform', 'answer module'(W.a), 'modules loss'(W.b).
  - **Pretrain (P)**: Parameters initialisation from the 2nd iteration of Random variant.
  - **Repeat (R)**: Repeat the same CL-iteration twice.

Comparison of CL methods

Model	CL configuration			Iterations	Number of examples ($\leq$)	Accuracy
	weighting	pretraining	iterations/level			
CL+W.a	answer	–	1	4	4 M	0.642
CL+W.b	losses	–	1	4	4 M	0.635
CL+L	uniform	–	1	11	11 M	0.650
CL+L+W.a	answer	–	1	12	12 M	0.655
CL+W.a+P	answer	2 iterations	1	[2] + 3	5 M	0.670
CL+W.a+P+R	answer	2 iterations	2	[2] + 5	7 M	**0.681**

Table 3: Results on `testdev-all` for several CL strategies.

- 'answer' weighting **W.a** is the most effective weighting function.
- **CL+L** refinement improves the results over CL but the experiments are expensive.
- **P** "warms up" the model to the modular aspect of VQA framework.
- **R** helps to better learn the task without augmenting the data size.
- **CL+W.a+P+R** model is the **best modular VQA model** scoring 68.1% accuracy after 7 training iterations using less than 7M examples, *i.e.* less than half of the training data.

Impact of CL

Model	Comp. cost	# examples	Accuracy
Unbalanced	9×14 M	14 M	**0.702**
Balanced	50×1.4 M	1.4 M	0.678
Random	12×1 M	≤ 12 M	0.694
CL+W.a+P+R	7×1 M	< 7 M	0.681

Table 4: Comparaison of our CL model (CL+W.a+P+R) with no-CL models (Unbalanced, Balanced, and Random) on the `testdev-all` set. Computation cost is the number of seen examples per iteration times the number of iterations.

- Unbalanced model (14M) has the highest accuracy (70.2%) but has the highest cost.
- Balanced model achieves lower results than the Unbalanced mode.
- CL+W.a+P+R provides very significant gains in terms of computational cost.
- CL+W.a+P+R the best trade-off between performance and training cost.

Conclusion

Conclusion

- Modular neural network for visual reasoning in a **real world** VQA context.
- Decompose the reasoning task to a series of easier and more general sub-tasks.
- Benefit from **cross-modal representations** [1] for Compositional VQA.
- NMN trained with CL.
- **Number of questioned objects** is an adequate CL difficulty criterion.
- Reduce experimental cost by half.
- Find a **trade-off between cost and performance.**

References

[1] Hao Tan and Mohit Bansal. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019.

[2] Jacob Andreas et al. "Neural Module Networks". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 39–48.

[3] Ronghang Hu et al. "Learning to Reason: End-to-End Module Networks for Visual Question Answering". In: *CoRR* abs/1704.05526 (2017). arXiv: 1704.05526.

[4] Guohao Li, Xin Wang, and Wenwu Zhu. "Perceptual Visual Reasoning with Knowledge Propagation". In: MM '19. Nice, France: Association for Computing Machinery, 2019, pp. 530–538. ISBN: 9781450368896.

[5] Yoshua Bengio et al. "Curriculum Learning". In: ICML '09. Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 41–48. ISBN: 9781605585161.

[6] Cao Liu et al. "Curriculum Learning for Natural Answer Generation". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden, 2018.

[7] Mrinmaya Sachan and Eric Xing. "Easy Questions First? A Case Study on Curriculum Learning for Question Answering". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, Aug. 2016, pp. 453–463.

[8] Narjes Askarian et al. "Curriculum learning effectively improves low data VQA". English. In: *ALTA 2021*. Association for Computational Linguistics (ACL), 2021, pp. 22–33.

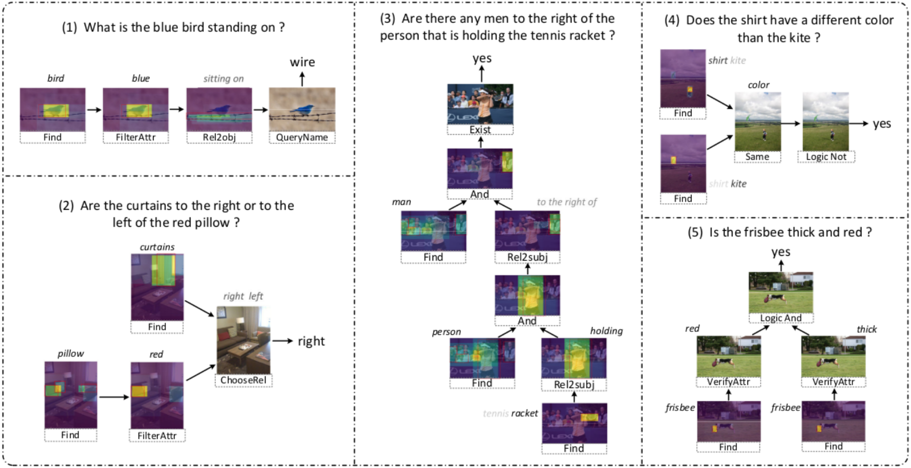


Figure 6: Examples visualizing the reasoning process [4]