Multimodal Representations for Teacher-Guided 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.
- In the field of VQA, two prominent approaches: monolithic and **compositional**.

Contributions

- Vision and language pre-trained (VLP) representations for Multi-modal compositional VQA.
- Teacher forcing (TF) compositional VQA.

Compositional Visual Reasoning for VQA

- Visual reasoning is inherently **compositional**.
- Break down the question into **modular** sub-problems.
- **Reasoning skills**: object and attribute detection, relation extraction, counting and comparisons...
- Assign each **sub-task** to a different module.
- **Transparency** and **explainability** gains.

Figure 1: 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?

Related work

- Supervised learning task.
- NL and functional programs representing questions with images and answers.
- GQA dataset [1]: real world images.

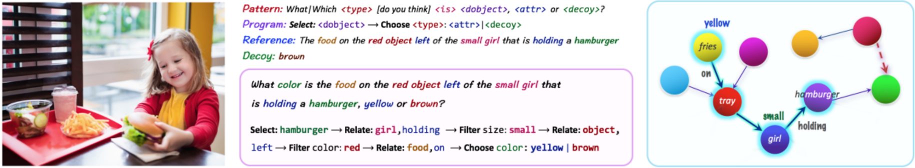

Figure 2: GQA Example: Image on the Left, Functional Program and Question on the middle and image graph on the right.

- Generator: Program generation using LSTM.
- Executor: Executes the program modules.
- NMN augmented with supervised **knowledge guidance**.
  - **Bboxes** of relevant visual regions for attention modules.
  - KL divergence between predicted attention maps and knowledge guidance.

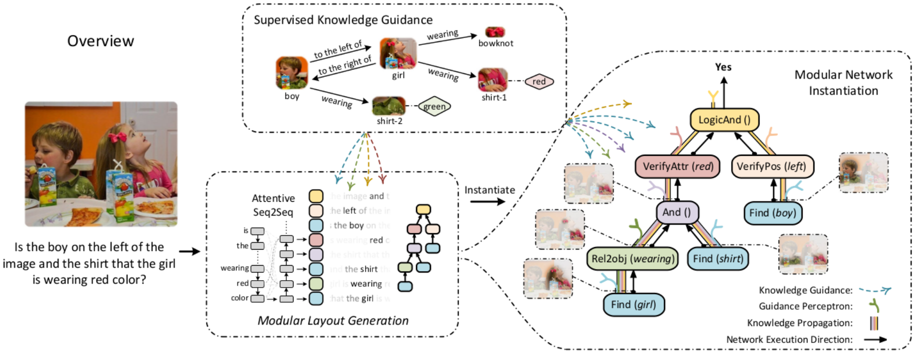

Figure 3: Perceptual Visual Reasoning overview Li, Wang, and Zhu [2].

Teacher forcing (TF)

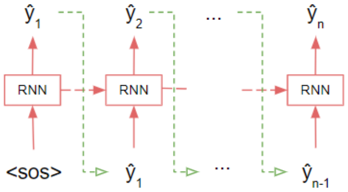

Figure 4: RNN w/o TF

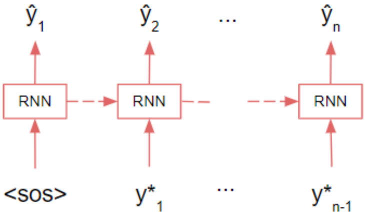

Figure 5: RNN w/ TF

- TF [3] is a widely used training technique in generation tasks.
- Instead of using the model's predicted output, TF uses the true output from the previous step as input.
- Pros of TF: Accelerates learning by providing accurate guidance.
- Cons of TF: Exposure bias. Model isn't exposed to its own errors during training.
- Scheduled sampling (SS) [4]: At each step, randomly choose between using ground truth or model predictions.

Modular VQA framework

Modular VQA framework

- Extract **aligned cross-modal embeddings** for words and objects using VLP model.
- 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.
- **Textual argument** to indicate the desired module's facet.

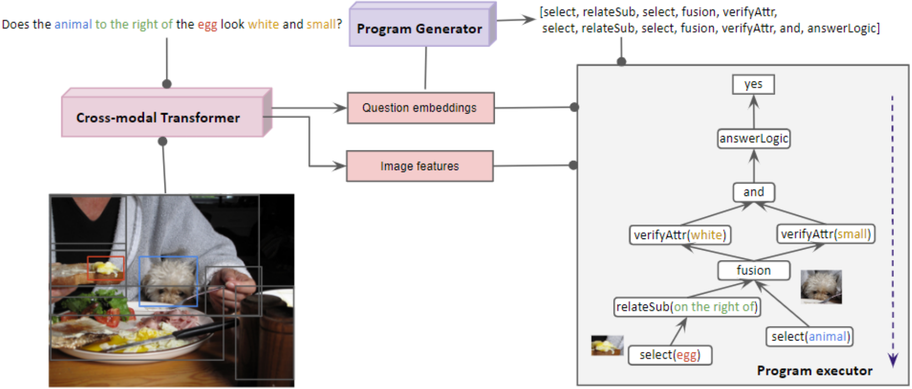

Figure 6: Our modular VQA framework.
Output flow (Plain arrows), MT loss backward flow (dotted arrows).

Neural modules

- 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.

Multimodal Representations for Teacher-Guided Compositional Visual Reasoning

Multimodal Representations from VLP model

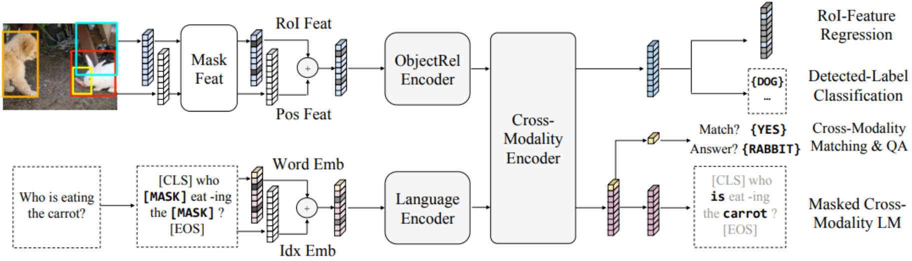

Figure 7: LXMERT model architecture [5].

- We use LXMERT as a feature extractor for question and image.
- LXMERT is trained on massive image-text data.
- Transformer encoder architecture.
- Image is represented by its object regions features [6].
- Cross modality encoder to align image and question features.
- We freeze the weights and discard the classifier.

Teacher guidance for the program execution process

- **Input guidance:** Decaying teacher forcing (TF).
  - Coin flip at each reasoning step (t).
  - Use predicted $\hat{o}_{t-1}$ as input with probability ϵ_e.
  - Use GT o^*_{t-1} as input with probability $1 - \epsilon_e$.
  - Probability ϵ_e decreases as epoch number e increases.
- **Output feedback:** multi-task (MT) loss $L = \alpha L_{att} + \beta L_{bool} + \gamma L_{answer}$

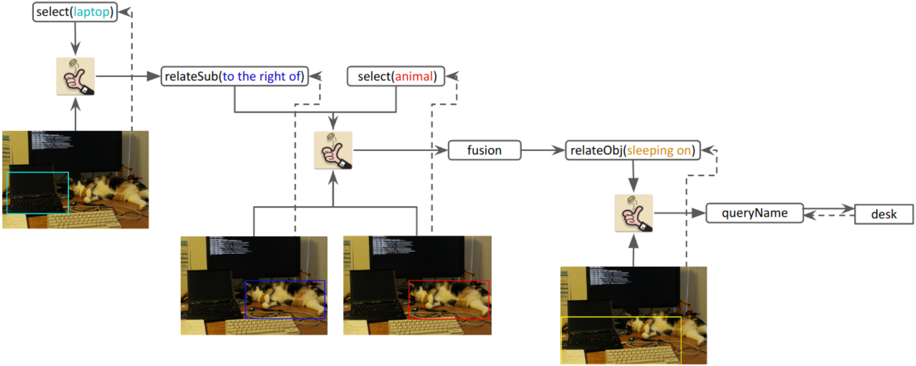

*Figure 8: Teacher guidance: answering 'On what is the animal to the right of the laptop sleeping?'. **input guidance** (Plain arrows) and **output feedback** (dotted arrows).*

Experimental details & analysis

Evaluation protocol

- Investigate teacher guidance on **Program Executor** using pre-processed GQA programs.
- Test the accuracy performance on the `testdev-all` set.

Evaluated methods:

- **LXV**: Cross-modal representations from LXMERT [5].
- **BertV**: Unimodal contextual language using BERT and GQA for bboxes [7, 1]
- **FasttextV**: Unimodal non-contextual fastText embeddings with GQA bounding boxes [8].
- **TF**: Decaying teacher forcing to guide the inputs of the modules.
- **MT**: Multi-task losses to guide the outputs of the modules.
- Matching techniques for aligning ground truth bounding boxes:
  - **Hard**: Hard matching for bboxes: Highest IoU between ground-truth and extractor.
  - **Soft**: Soft matching for bboxes: IoU threshold for multi-label classification.

Results analysis

Table 2: Performance of various training methods on the `testdev-all` set.

Model	accuracy
LXV-TF-hard	0.548
LXV-MT-hard	0.598
LXV-TF-MT-hard	0.630
LXV-TF-soft	0.536
LXV-MT-soft	0.563
LXV-TF-MT-soft	**0.632**
FasttextV-TF-MT-hard	0.495
BertV-TF-MT-hard	0.506
BertV-TF-MT-soft	0.485
FasttextV-TF-MT-soft	0.511

- LXV-TF vs. LXV-MT: MT achieves higher accuracy compared to decaying TF alone.
- Combination of TF and MT achieves highest accuracy: **LXV-TF-MT-soft** at 63.2%.
- Complementary effects: Multi-task loss and decaying teacher forcing enhance training dynamics and performance.
- Cross-modal aligned features (**LXV**) yield accuracy improvements compared to unimodal features (**BertV**, **FasttextV**).

Conclusion

Conclusion

- Neural module 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** [5] for Compositional VQA.
- NMN trained with **Teacher guidance** to enhance model performance.
- Modules learn their reasoning sub-tasks both independently and in collaborative manner.

References

[1] Drew A. Hudson and Christopher D. Manning. "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering". In: (2019).

[2] 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.

[3] Ronald J. Williams and David Zipser. "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks". In: *Neural Computation* 1.2 (June 1989), pp. 270–280.

[4] Samy Bengio et al. "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks". In: *CoRR* (2015). arXiv: 1506.03099.

[5] 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.

[6] Peter Anderson et al. "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *CVPR*. 2018.

[7] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1*. June 2019.

[8] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *Transactions of ACL* 5 (July 2016).

Qualitative results

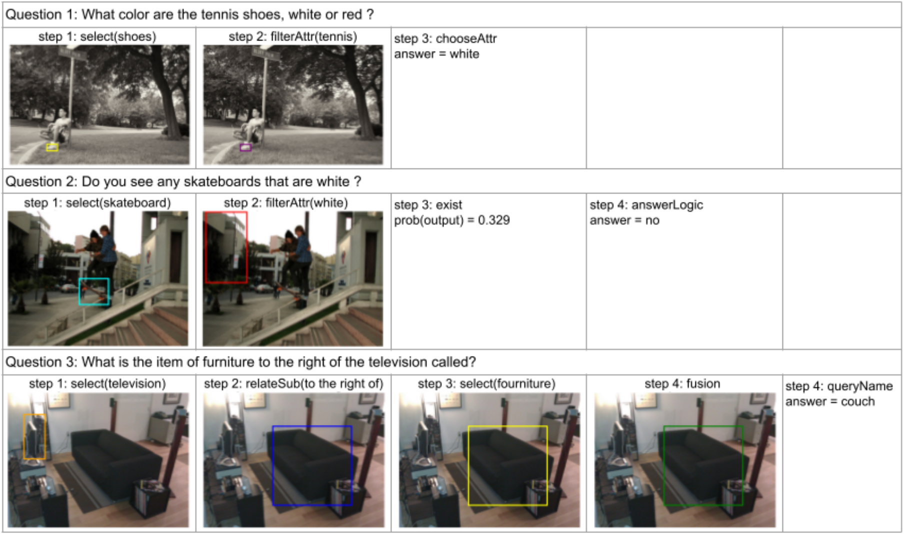

Figure 9: Examples showing the reasoning process.