Look and Answer the Question: On the Role of Vision in Embodied Question Answering

Nikolai Ilinykh and Yasmeen Emampoor and Simon Dobnik

Centre for Linguistic Theory and Studies in Probability,

Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden

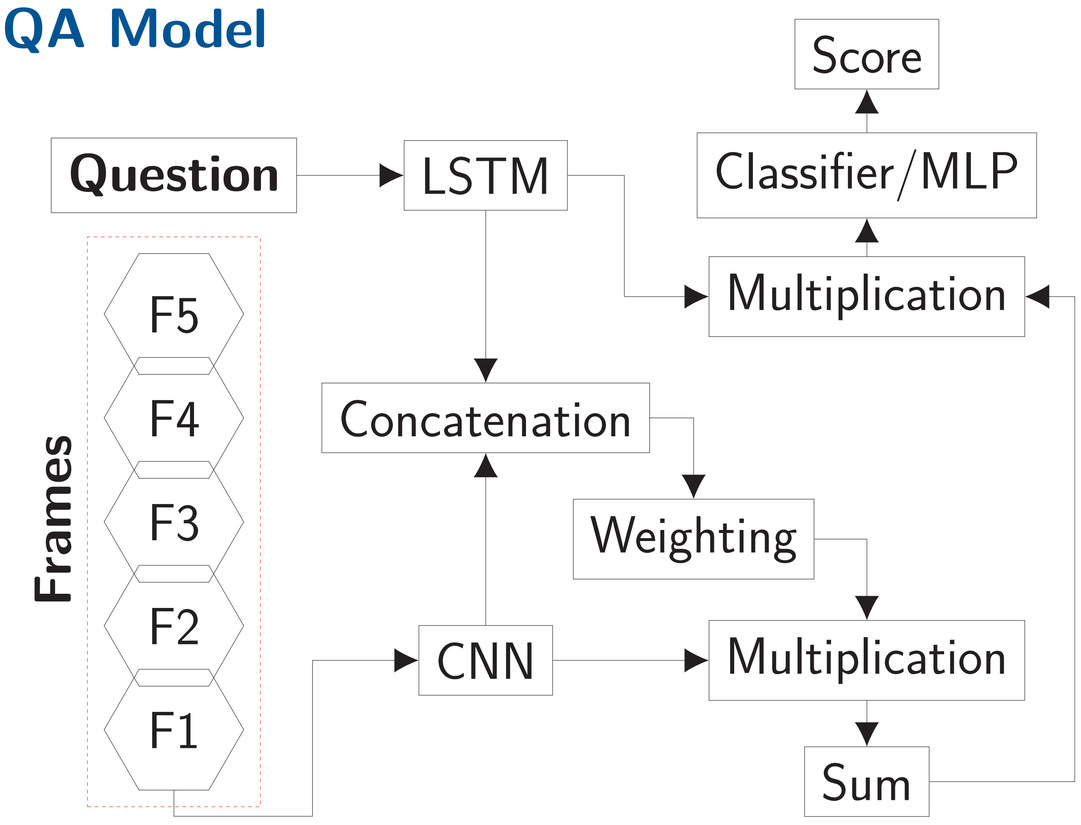
nikolai.ilinykh@gu.se, gusemampya@student.gu.se, simon.dobnik@gu.se

We show that the EQA model captures low-level visual information, not high-level visual features which are required to answer questions correctly and overcome dataset biases and reliance on language.

EQA task

Given the last fives images from the end of the navigation, answer the question: *What colour is the fireplace?*

QA Model



Experiment I: is there learning from vision?

Train and evaluate the model on original images (Vis-L), black images (B) or language only (Ø-L)

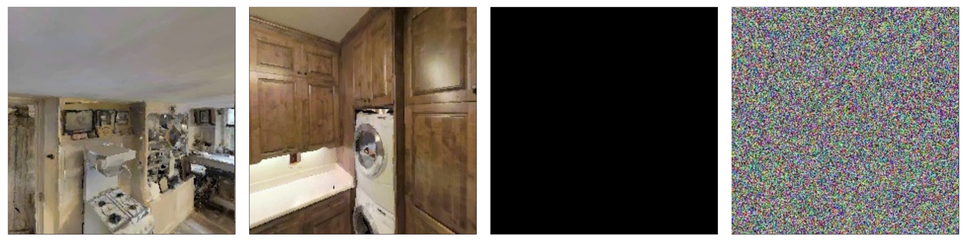
Metric	Vis-L	B	Ø-L
↓ Overall Mean Rank (MR)	4.352	4.454	**3.685**
MR, Color Room Questions	3.611	**3.157**	3.247
MR, Color Questions	2.693	**2.261**	2.304
MR, Location Questions	10.137	13.667	**7.611**
↑ Overall Accuracy (A)	**0.38**	0.323	0.362
A, Color Room Questions	**0.374**	0.348	0.337
A, Color Questions	**0.528**	0.478	0.522
A, Location Questions	0.222	0	**0.278**
Kappa Score	-0.005	0.014	**0.024**

- Vision artificially inflates accuracy for a wrong reason
- Vision is not used for every question type - location questions can be answered from language alone
- Model's approximations are better without vision

Experiment II: "how much" vision?

Evaluate the **Vis-L** model on different perturbations and see *how much it can learn from original visual representations?*

Q: What colour is the stove in the kitchen?



- **Vis-L**, structure +, content +, context +
- **S, Shuffled**, structure +, content +, context -
- **B, Black**, structure +, content -, context -
- **R, Random**, structure -, content -, context -

Metric	Vis-L	S	B	R
↓ Overall Mean Rank (MR)	4.352	5.145	5.508	6.899
MR, Color Room Questions	3.611	4.157	4.562	5.512
MR, Color Questions	2.693	3.035	3.087	3.319
MR, Location Questions	10.137	12.722	13.278	18.33
↑ Overall Accuracy (A)	0.38	0.266	0.246	0.211
A, Color Room Questions	0.374	0.264	0.258	0.258
A, Color Questions	0.528	0.307	0.217	0.194
A, Location Questions	0.222	0.222	0.222	0
Kappa Score	-0.005	0.013	0.004	-0.005

- Patterns, structure and some low-level general knowledge are learned (original > shuffled > black > random)

What did we learn about QA in EQA?

- The model does not have a deeper vision understanding
- There are dataset and modelling biases which suggest directions for future research:
  - Improve vision by implementing cognitive attention,
  - Split question answering into several subtasks,
  - Use pre-trained multi-modal transformer

CLASP centre for linguistic theory and studies in probability