Zero-shot Translation of Attention Patterns in VQA Models to Natural Language

Leonard Salewski[1], A. Sophia Koepke[1], Hendrik P. A. Lensch[1], Zeynep Akata[1,2]

[1]University of Tübingen, [2]MPI for Intelligent Systems

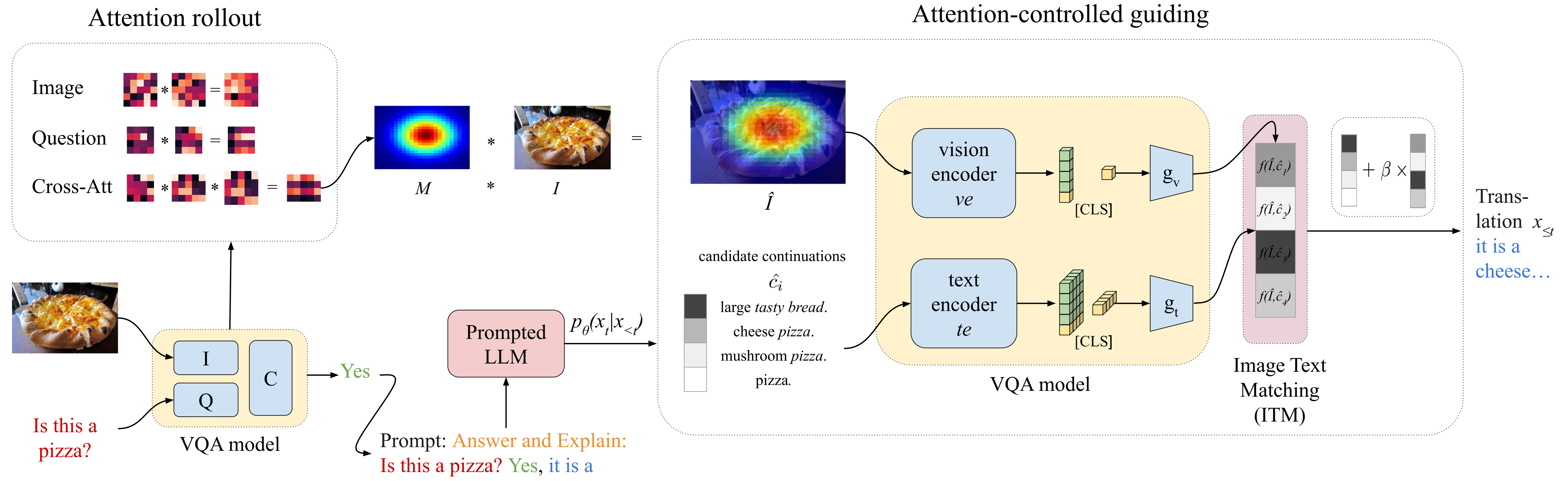
Motivation

- Deep learning algorithms need to be transparent and accessible.
- Natural language can be more intuitive than salient visual regions.
- Translating model internals into text should work in a zero-shot fashion without training.

The ZS-A2T Framework

- Zero-Shot Attention-to-Text
- Attention-controlled guiding based on
  - Attention Rollout of the VQA model
  - Image Text Matching (ITM)

$$f(\hat{I}, \hat{c}_i) = \frac{e^{\kappa \cdot \mathrm{CosSim}(g_v(ve(\hat{I})), g_t(te(\hat{c}_i)))}}{\sum_{j \in 1,\dots,k} e^{\kappa \cdot \mathrm{CosSim}(g_v(ve(\hat{I})), g_t(te(\hat{c}_j)))}}$$



Experiments

- Language model: *OPT*
- VQA model: *ALBEF*
- Datasets: *GQA-REX* and *VQA-X*
- Metrics: *Bleu-4*, *Meteor*, *Rouge-L*, *Cider*, *Spice*
- Inference speed: 7.2s / sample

Quantitative Results

Setting	Framework↓	GQA-REX [14,8] B4	M	RL	C	S	VQA-X [15] B4	M	RL	C	S
Zero-shot	ZeroCap $_{\text{GPT-2}}$ [41]*	1.4	4.6	12.3	16.9	5.3	0.7	4.7	14.0	5.8	2.0
	EPT $_{\text{GPT-2}}$ [40]*	0.0	3.3	3.2	2.6	2.8	0.9	6.5	14.9	6.7	2.9
	MAGIC $_{\text{GPT-2}}$ [37]*	2.3	10.8	18.8	41.1	18.8	1.0	8.8	19.3	10.6	7.1
	MAGIC $_{\text{OPT 6.7B}}$ [37]*	3.3	11.6	22.2	48.8	21.4	1.9	9.5	20.5	14.7	8.9
	Socratic Models $_{\text{OPT 6.7B}}$ [47]*	3.3	14.1	22.8	40.5	19.3	3.6	12.8	25.7	19.9	10.1
	ZS-A2T $_{\text{OPT 6.7B}}$ (ours)	**10.2**	**18.2**	**35.0**	**113.5**	**31.4**	**8.5**	**13.8**	**34.2**	**38.1**	**10.5**
Supervised	NLX-GPT [31] $_{\text{GPT-2}}$	-	-	-	-	-	23.8	20.3	47.2	89.2	18.3
	VisualBert-REX [8] $_{\text{LSTM}}$	54.6	39.2	78.6	464.2	46.8	-	-	-	-	-

Ablations

- Influence of attention masking
- Influence of text continuations
- Different language models
- Visual explanation methods
- Few-shot prompt ablations

Guiding Input	B4	M	RL	C	S
Full Image	8.1	13.7	34.1	37.6	**10.8**
No Continuation	6.2	12.5	31.1	28.2	9.4
ZS-A2T (Rel. Masking + Cont.)	**8.5**	**13.8**	**34.2**	**38.1**	10.5

LM (#Params)	B4	M	RL	C	S
GPT-2 (125M)	3.6	11.0	26.6	19.8	7.7
OPT (125M)	3.4	10.7	26.5	18.5	7.1
OPT (350M)	3.9	11.6	27.9	20.2	7.9
OPT (1.3B)	7.1	13.0	32.8	28.9	9.2
OPT (2.7B)	7.1	13.3	32.5	31.0	10.1
OPT (6.7B)	**8.5**	**13.8**	**34.2**	**38.1**	**10.5**

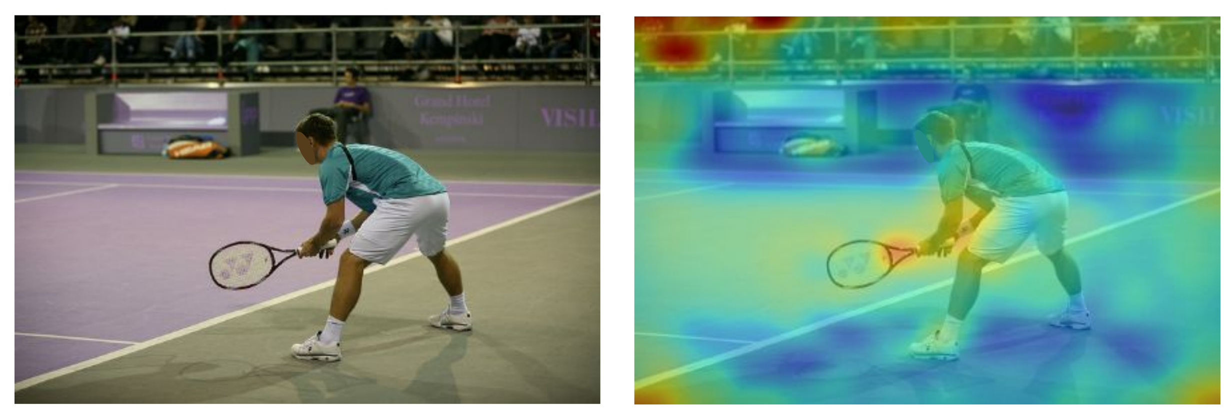
Model	n	B4	M	RL	C	S
ZS-A2T	0	8.5	13.8	34.2	38.1	10.5
ZS-A2T	1	9.8	14.5	34.8	42.7	11.7
ZS-A2T	5	11.9	15.3	37.5	49.6	12.4

Qualitative Results

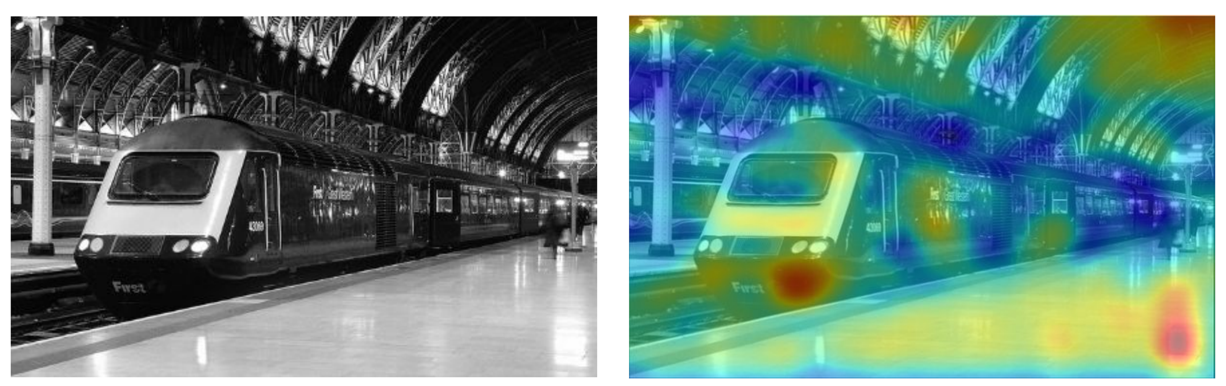
- Attention rollout selects relevant image regions that correspond to the question.
- The translations are fluent and refer to visual elements.
- The translations contain visual information from the attention patterns.

What kind of vehicle is in the front of the photo? The answer is truck because it's a pickup.

What is this man doing? The answer is playing tennis because he is a tennis player.



What kind of place is this? The answer is train station because it's the only place the train goes.



Does the bus like the kids? The answer is yes because the bus is a school bus.

Conclusion

- ZS-A2T translates aggregated the attention of a VQA model into natural language.
- Our method does not need training and can be used with any language model or visual explanation method.

The authors thank IMPRS-IS for supporting Leonard Salewski. This work was partially funded by the Max Planck Society, the BMBF Tübingen AI Center (FKZ: 01IS18039A), DFG (EXC number 2064/1 – Project number 390727645), ERC (853489-DEXIM), and DFG-CRC 1233 (Project number 276693517).