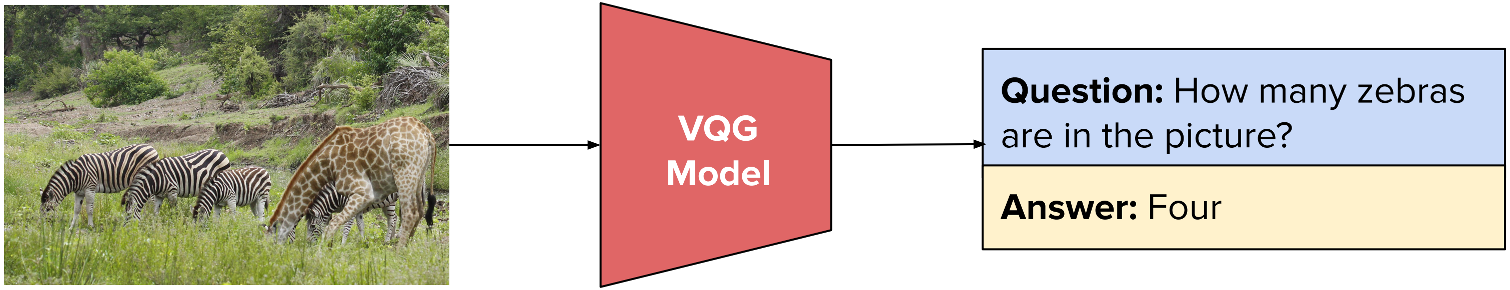
Applying the Answer-Clue-Style approach to VQG

Chrys Ngouma
Kuruba Vijaya Lakshmi

What we will cover

- Visual Question Generation/Answering (VQG/VQA)
- System architecture
- Datasets
- Evaluation methods and results
- Challenges

The task of Visual Question Generation (VQG)

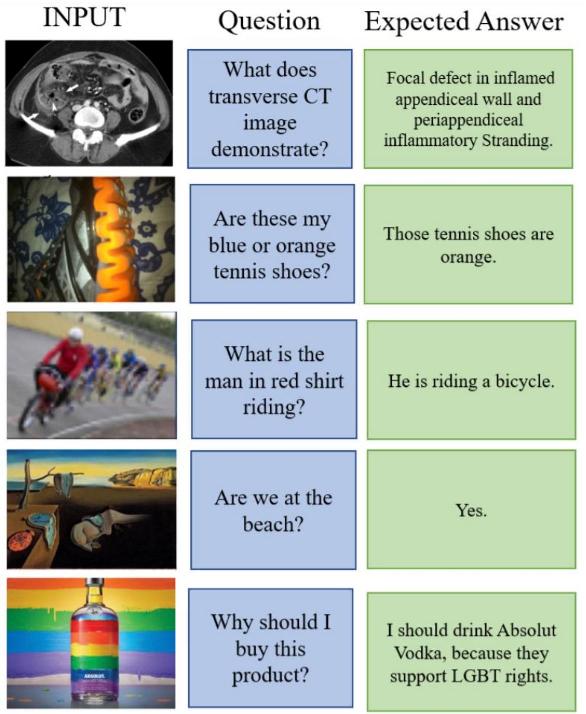


- Interdisciplinary task involving computer vision and NLP
- VQG consists of generating meaningful questions based on an input image

Applications of VQA

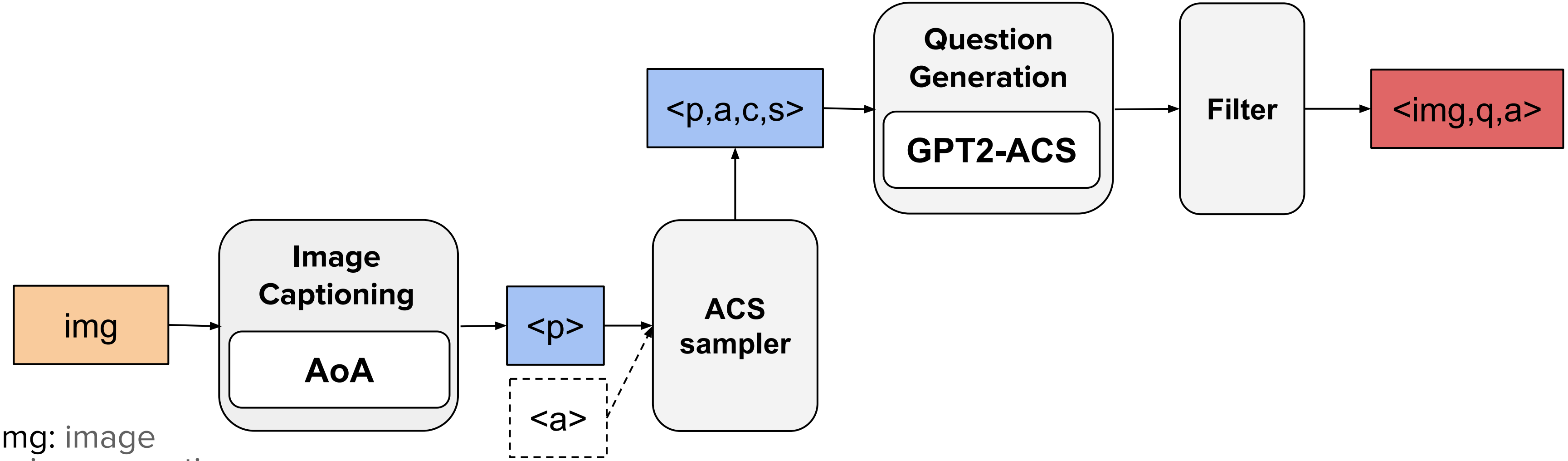
- Medical VQA (Abacha et al., 2019)
- VQA for visually impaired people (Gurari et al., 2018)
- Video Surveillance (Toor et al., 2019)
- Education and cultural heritage (Bongini et al., 2020)
- Advertising (Husain et al., 2017)
- And more

Our goal is generate visual QA pairs from unlabelled images



Ref: arXiv:2103.02937

System Architecture Overview



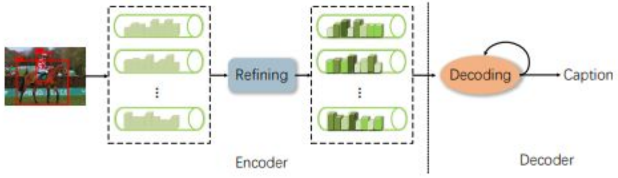
img: image
p: image caption
q: question
a: answer
c: clue
s: question style

[Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus (umontreal.ca)]

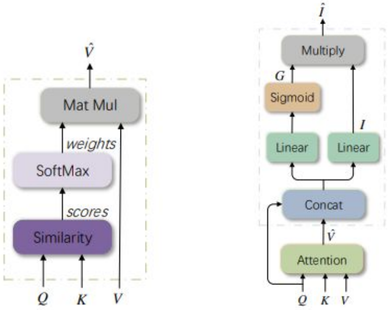
Datasets

- GPT2-ACS was trained on the SQuAD1.1 dataset
- AoA is a model pretrained on MSCOCO dataset
- We evaluate our approach on the VQA validation set

Attention on Attention for Image Captioning (AoA)



AOA Model Architecture

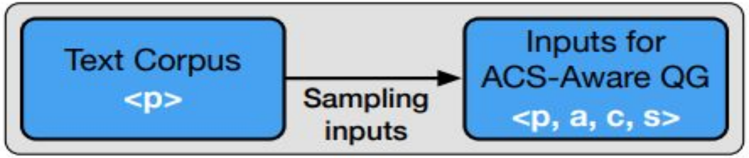


(a) Attention (b) Attention on Attention

- Attention on Attention (AoA) module, an extension to the conventional attention mechanism, to determine the relevance of attention results.
- Apply AoA to both the encoder and decoder to constitute AoANet: in the encoder, AoA helps to better model relationships among different objects in the image; in the decoder, AoA filters out irrelative attention results and keeps only the useful ones.
- When this paper was released, this method achieved a new state-of-the-art performance on MS COCO dataset with 129.6 CIDEr-D (C40) score on the official online testing server.
- We have used pretrained model from the authors published code in Github to get captions on images.

Ref :1908.06954.pdf (arxiv.org)

ASC-aware Question Generation



$$P(a, c, s|p) = P(a|p)P(s|a, p)P(c|s, a, p)$$

$$\begin{aligned} P(a|p) &= P(a|POS(a), NER(a), length(a)), \\ P(s|a, p) &= P(s|POS(a), NER(a)), \\ P(c|s, a, p) &= P(c|POS(c), NER(c), DepDist(c, a)). \end{aligned}$$

- **Sequential sampling**
- **Learn from existing dataset**

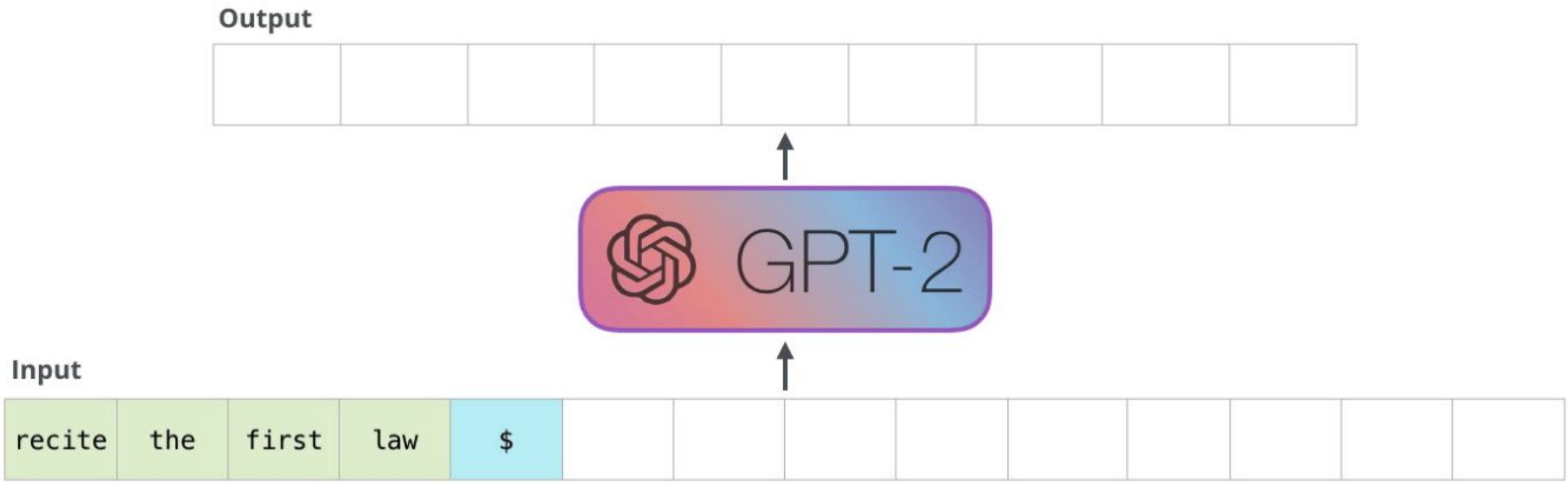
Answer-Clue-Style Sampler

Problem: the caption most likely do not contain the exact answer

Our approach:

- We first look for a candidate chunk that is the most similar (at least 70%) to the answer e.g. doughnut ➜ donut, elephant ➜ animal, or young woman ➜ blonde person
- If no match, we sample a random answer the way the ACS sampler intended

GPT2 Pretrained Model



- **Finetune pretrained language model**

Results

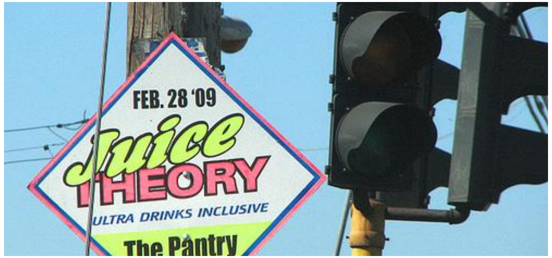
- 47% of the questions are well formatted while 83% are relevant to the images
- Most of the answers are not correct or irrelevant

q: “What is in the back of the truck?”

a: “elephant”

q: “What is a common transportation option in Portugal?”

a: “cars”



q: “What is on a pole and on what?”

a: “a sign on a pole”

Human Evaluation

We use three volunteers to vote on whether the generated questions and answers are well formatted, meaningful, and relevant to the images.

Experiment		***Ours***
Question is Relevant	Yes	83%
	No	17%
Question is Well-formed	Yes	47%
	Partially	36%
	No	17%
Answer is Correct	Yes	38%
	Partially	12%
	No	50%

Performance comparison

This approach underperforms the baselines on most metrics

Models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	ROUGE-L	METEOR	CIDEr	Relevance
IA2Q	32.43	15.49	9.24	6.23		11.21	36.22	90.00
V-IA2Q	36.91	17.79	10.21	6.25		12.39	36.39	92.20
Krishna et al. (2019)	**47.40**	**28.95**	**19.93**	**14.49**	**49.10**	**18.35**	**85.99**	**97.20**
Ours	19.45	7.50	3.31	1.38	23.94	15.32	21.24	83.00

Challenges and Observations

- The AoA model fails to capture small details in the picture leading to loss of information
- The generated questions sometimes contain unrelated and/or repeated information
- Model struggles with “yes/no” question types

caption: “A sandwich on a plate on a table”

original q: “What kind of bread was used to make the sandwich?”

original a: “weat”

q: “What was on a table on the day of the assassination?”

caption: “A group of people riding on the back of an elephant”

original q: “Are these people in the jungle?”

original a: “yes”

q: “Is riding on the back of an elephant a group of people or a group of people?”

Challenges and Observations

Writing the caption in more active voice improves the relevance of the questions

c: "A sandwich on a plate on a table"

q: "What was on a table on the day of the assassination?"

→

c: "A sandwich **is** on a plate on a table"

q: "What is on a plate?"

c: "A clock on the wall above a table with a clock"

q: "What type of clock is above a table in the Notre Dame library?"

→

c: "A clock on the wall **is** above a table with a clock"

q: "What is above a table on the wall?"

Possible Improvements

Visual information extraction

- Using a region captioning model such as Multi-level Scene Description Network (MSDN)
- End-to-end training on the VQA dataset

Question generation

- Extend the question styles list (e.g. “how many”, “what kind of”, “what type of”, “which of”, etc.
- Exclude words with certain POS tags when sampling answers (e.g. DT, CC, EX, etc.)
- To date the GPT2-ACS reports 53.5 % of correct answers and 74.5% of good questions on Wikipedia datasets. Further work on the GPT2-ACS could improve our results.

Questions?