Generating Natural Questions About an Image

Grant Zhao, Jacob Fernandez

INTRODUCTION

- **Image captioning has focused on literal, surface-level descriptions**
- **This project proposes Visual Question Generation (VQG): automatically generating natural, engaging questions from an image.**

Reason for Study

- **While previous tasks focused on literal descriptions of images, VQG moves beyond that by exploring how questions address abstract events and commonsense inferences that objects in images evoke.**
- **A VQG task is designed to generate questions that are natural sounding, engaging, and prompt deeper thinking about the image.**

METHODOLOGY/MODEL ARCHITECTURE

GRU Cell Dynamics

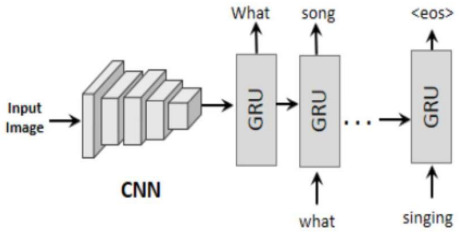
$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \text{(update gate)} \quad (1)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad \text{(reset gate)} \quad (2)$$
$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) \quad \text{(candidate state)} \quad (3)$$
$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \quad \text{(new hidden state)} \quad (4)$$

Output Distribution

$$y_t = \text{softmax}(W_o h_t + b_o) \quad (5)$$

Training Loss (Negative Log-Likelihood)

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(w_t \mid w_{<t}, I) \quad (6)$$



- 3 datasets, MS COCO, Flickr, and Bing
- COCO dataset limited in terms of concepts covered
- Flickr dataset images appear as middle of a photo album
- Bing dataset queried a search engine with 1,200 event-centric query terms
- 5,000 images per each dataset, total of 15,000 images and 75,000 questions

CONCLUSION

		$Human_{consensus}$	$Human_{random}$	$GRNN_X$	$GRNN_{all}$
Human Evaluation					
	Bing	2.50	2.36	1.38	**1.81**
	COCO	2.50	2.40	1.62	**1.97**
	Flickr	2.33	2.28	1.27	**1.58**
BLEU	Bing	87.3	83.6	**12.4**	11.0
	COCO	86.1	83.8	**13.8**	14.3
	Flickr	84.5	83.4	**10.0**	9.8
MET.	Bing	62.0	59.0	**16.0**	15.6
	COCO	60.7	58.5	**18.2**	18.3
	Flickr	59.5	58.0	**14.1**	14.0
ABLEU	Bing	63.0	57.5	**11.5**	10.7
	COCO	61.0	56.9	**12.3**	12.4
	Flickr	62.0	57.2	**9.4**	9.2

Our evaluation results for the GRNN model using BLEU 1-4 metrics n-gram overlap

BLEU Scores	Bing	COCO	Flickr
	12.1	13.6	10.2

RESEARCH CHALLENGES

- **Corrupted image dataset, removal and data augmentation**
- **Dataset limitations, creating datasets with truly natural questions**
- **Question diversity - pursuing more complex event-centric questions**
- **Text Augmentation - use back translation**
- **Evaluation metrics, how do we evaluate the quality of a generated question?**

RESULTS



HUMAN	- How long did it take to make that ice sculpture?	- Is the dog looking to take a shower?	- Was this explosion an accident?
GRNN	- Where was this picture taken ?	- Why is this dog in a bathroom ?	- What caused this explosion ?

NEXT STEPS

- **Question generation within a conversation system?**
- **While our models learn to generate promising questions, large gap to match humans still exists**

ACKNOWLEDGEMENTS

- Mostafazadeh, N., et al. Generating Natural Questions about an Image. arXiv, 2016. [https://arxiv.org/abs/1603.06059]
- Hudson, D. A., and Manning, C. D. GQA: A New Dataset for Real-World Visual Reasoning. CVPR, 2019. [https://cs.stanford.edu/people/dorarad/gqa/]
- He, K., et al. Deep Residual Learning for Image Recognition. CVPR, 2016. [https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py]
- Radford, A., et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv, 2021. [https://github.com/openai/CLIP]
- Sennrich, R., Haddow, B., and Birch, A. Improving Neural Machine Translation Models with Monolingual Data. arXiv, 2015. [https://arxiv.org/abs/1511.06709]