



# 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.

## 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?

## METHODOLOGY/MODEL ARCHITECTURE

### GRU Cell Dynamics

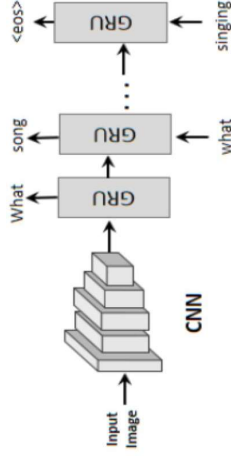
$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) & (1) \text{ (update gate)} \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) & (2) \text{ (reset gate)} \\ \tilde{h}_t &= \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) & (3) \text{ (candidate state)} \\ h_t &= (1 - z_t) \circ h_{t-1} + z_t \circ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

### Output Distribution

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

### Training Loss (Negative Log-Likelihood)

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



- 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

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

Human Evaluation		Human <sub>conversations</sub>	Human <sub>freedoms</sub>	GRNN <sub>x</sub>	GRNN <sub>all</sub>
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
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
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
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

## 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?

GRNN  
- Where was this picture taken ?



HUMAN  
- Is the dog looking to take a shower?

GRNN  
- Why is this dog in a bathroom ?



HUMAN  
- Was this explosion an accident?

GRNN  
- 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. (2016) Generating natural questions about an image, arXiv.org. Available at: <https://arxiv.org/abs/1603.06059> (Accessed: 05 May 2025).