2022 Deakin Simpsons Challenge: Visual Question Answering Task

Fenglu Cai

Deakin University
Geelong Victoria Australia
caife@deakin.edu.au

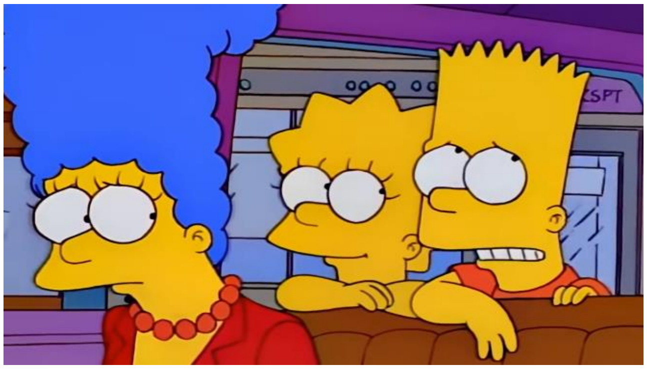


Figure 1 : one possible question for the image could be : is the lady wearing a red neckless ?

ABSTRACT

Visual Question Answering (VQA) is a challenging task. The submitted model is designed for binary answer (yes or no) VQA problems. The model comprises image input and question input. Image input is flattened before being processed by a 64-unit Dense layer, and question input is represented by a 64-dimension embedding vector. Both interim results are multiplied and fed into a 2-unit Dense layer. The highest accuracy is 54%, which is slightly higher than random guess for a problem requiring yes or no answer.

KEYWORDS

Deep learning, Visual Question Answering, Computer Vision, Natural Language Processing

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Visual Question Answering (VQA) has become possible with the development of deep learning domains, such as computer vision and natural language processing (NLP). VQA systems attempt to answer questions regarding images, which mimic how humans can answer questions when they see pictures. The models are usually complex because image-based models as well as natural language-based models are involved (Goyal et al., 2019). It is usually difficult to get a high accuracy compared with classification models because questions can be arbitrary and subjective (Zhang et al.,2016). Another reason could be the model only gives an answer from language input and never really learn to recognize the image content.

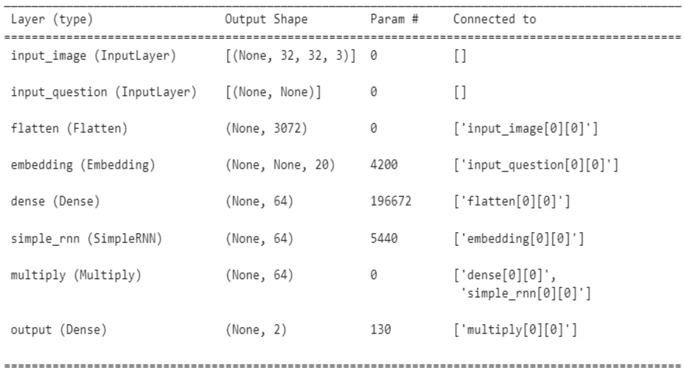
2. Model Development

2.1 Training data

The training data used the original validation dataset provided by deakin-ai-challenge team. The training data is a total of 150 images with one question for each image.

2.2 Model summary

The model includes an image input and a question input. The image input is flattened, processed by a Dense layer. The question input is transformed into 20-dimension embedding vectors and processed by 64-unit SimpleRnn. The two branches are multiplied as the input into a 2-unit Dense layer, which is the classifier of yes or no answer.



```
_____________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
=============================================================================================
 input_image (InputLayer)       [(None, 32, 32, 3)]  0           []

 input_question (InputLayer)    [(None, None)]       0           []

 flatten (Flatten)              (None, 3072)         0           ['input_image[0][0]']

 embedding (Embedding)          (None, None, 20)     4200        ['input_question[0][0]']

 dense (Dense)                  (None, 64)           196672      ['flatten[0][0]']

 simple_rnn (SimpleRNN)         (None, 64)           5440        ['embedding[0][0]']

 multiply (Multiply)            (None, 64)           0           ['dense[0][0]',
                                                                  'simple_rnn[0][0]']

 output (Dense)                 (None, 2)            130         ['multiply[0][0]']

=============================================================================================
```

Figure 2: the model architecture

2.3 Training Parameters

Batch_size:128

Image width:32
Image length:32
Adam optimizer learning rate:0.001
Epochs:13

3. Conclusion

This final submitted model produces the highest accuracy in the development phase (55.2%) and proved to produce similar accuracy (54%) although slightly lower.

ACKNOWLEDGMENTS

Thank you to all competition facilitators involved, especially to Dr. Mohamed Reda Bouadjenek for providing timely support when I have difficulties for model submissions

Thank you to Deakin University School of Information Technology, Faculty of Science Engineering & Built Environment for making this com for hosting the competition

Thank you to Community Bank at Deakin University as competition sponsor

REFERENCES

Goyal, Y. et al. (2019) 'Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering', International Journal of Computer Vision, 127(4), pp. 398–414. doi:10.1007/s11263-018-1116-0.
Zhang, P. et al. (2016) 'Yin and Yang: Balancing and Answering Binary Visual Questions', 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, pp. 5014–5022. doi:10.1109/CVPR.2016.542.

Github link for code:

https://github.com/Xiaolumang/deakin-ai-challenge2022/blob/main/deakin_Simpsons_VQA_2022.ipynb