Deakin2022 Simpson Challenge: VQA Task
A solution based on Bottom-Up and Top-Down Attention Method

Jiaheng Wei
Deakin University
Melbourne, AU
weijiah@deakin.edu.au
wjheng1999@gmail.com

Langning Bao
Deakin University
Melbourne, AU
Baola@deakin.edu.au

ABSTRACT

The Deakin Simpsons challenge 2022 is a VQA task. In this task, we used bottom-up and top-down attention method to develop our model by Tensorflow. And our trained model achieve 55.20%(#3) in final test phase. But the result did not achieve our expectation, for existing gaps between natural images and simpson image dataset. At the same time, we used simplified object detection head which may reduce performance. Thus, there are still many problems waiting us to solve. The code is avaliable on *This Github Repo*

KEYWORDS

VQA, deep learning

1 INTRODUCTION

The Deakin Simpsons challenge 2022 is a computer vision competition and the task is to provide an accurate natural language answer using machine learning and deep learning when giving an image of simpsons and a natural language question about the image. Obviously, the task is a visual question answering(VQA) task. We should build a model which accepts two inputs: a image and a question, then outputs an answer. In our opinion, the task is challenging, for a model is hard to align image and text, and learn the relationship between them. As we looked up many papers, we found attention mechanism is a good way to build correlation between image and text. After considering objective device and time limitation, we chose BuTd model[1] as our preference.

In the solution, we make following contributions:

- We implement Bottom-Up and Top-Down Attention Method by Tensorflow based on original paper[1]
- We changed bottom-up visual head, current model can support different object detection algorithm including Faster-RNN, YoLo and so on.
- For computation limitations, we design a simplified visual head to replace object detection head.
- Our model achieved accuracy around 64.00% in validation dataset, 52.40% in phase one, and 55.20% in final phase.

2 RELATED WORK

As convolutional neural network achieved good performance on image tasks and recurrent neural network handled text tasks well, researchers began to think about combining images and text together and made machine learn concepts. Visual Question Answering normally proposed in this paper[2], and they also provided VQA dataset and baselines. Then, some researcher found that in VQA task, language features are easier to learn than vision features, which cause models learn too much language prior and break generalization of the model. So in the paper[3], they proposed a balanced VQA dataset and tried to force model learnt more visual features and make images matter in VQA task. About image and language, how to find relationship between them was a important question, so people found attention mechanism may help this. SAN[8]model had been proposed which applied attention mechanism via stacked attention network and used language features to query image multiple times to infer correct answers progressively. After that, researchers found that objects in the image matters in the question, which means smaller image parts which contain key objects as visual features can help model make right decisions. Thus, bottom-up and top-down attention method has been developed[1]. This method used object detection algorithm to capture important visual features from images. Similarly, if extract key nouns from sentences, these key words may help image and text align. So Microsoft team developed Oscar[4] method, the method used object word tags as anchor points to significantly ease the learning of alignments. Surely, VQA is just a kind of image-language tasks. Researcher tried to use pretrained large model and applied weights in all image-language tasks, which calls multi-modality. Following pre-trained method and attention mechanism, VinVL[9],VLMo[6] and OFA[5] had been proposed, these large models can be applied in almost multi-modality tasks. Besides these large models, some easy methods also had been developed like SimVLM[7], the method just used transformer and built a seq2seq model. This model also achieved good performance.

3 SOLUTION

Our solution based on the BuTd method[1], the VQA model is similar as original model, which showed in Figure 1.

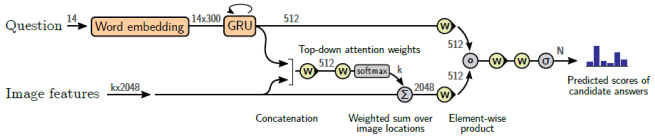

Figure 1: Original BuTd VQA model

Deakin2022AIChallenge, August 2022, Mel,VIC,AU

3.1 Implement Details

Our implement based on Tensorflow. The model has two input heads, one for language and another head is visual head. The language head has two layers, the first one is word embedding layer which produced by FastText CBOW model, the second layer is GRU layer, which can produce the question embedding matrix. About the visual head, the original implement is object detection head with Fast-RCNN and Resnet101. In our implement, the object detection part can be any available algorithm. We used Faster-RCNN to capture objects and example like Figure 2 .The object detection layer produce object boxes, then these boxes will be passed to roipooling layers. When we get same size object image parts, we will use convolutional base to process them and finally gain visual features. With question embedding and visual features, the model would try to align image and text with joint attention module, then the attention weights would apply in visual features and produce visual embedding. Finally, the model use visual embedding and question embedding do element-wise operation and send the joint representation to classifier with softmax function to the answer probability distribution.

But in the real-environment, we found if we use standard object detection visual head, a epoch needs 14h run-time, which is unacceptable. So we tried a simplified visual head, for we assumed imprecise object parts can achieve similar effort. The simplified head just slices one entire image to 9 parts evenly as visual features. Though the performance may reduce but run speed improved.

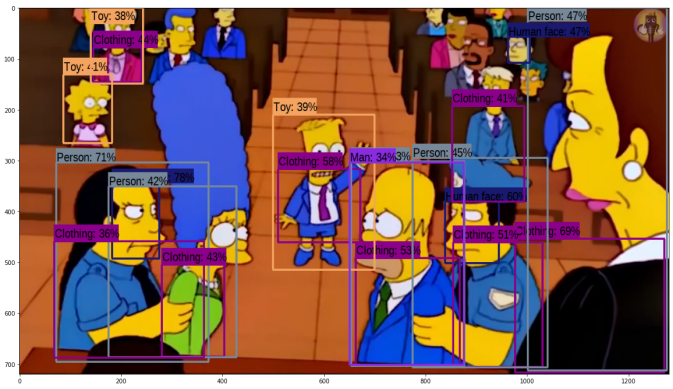

Figure 2: Object Detection Results on a Simpson image

3.2 Training

We firstly applied our model in abstract scene VQA dataset, but unfortunately, we found the model cannot converge in the dataset. We guess that the dataset cannot provide enough useful features. Thus, we change train dataset to COCO VQA dataset. With new dataset, our model successfully converged and as epoch increasing, the model began to overfit.

In experiment setting, we used SGD optimizer and set learning rate 0.01, momentum 0.9, epoch 20. We also used learning rate scheduler, learning rate will be 0.001 in 10th epoch.

4 RESULTS

Our experiment results are shown in the Table 1

	Accuracy	F1-Score	Precision	Recall
On Validation Set	64.00%	-	-	-
On Development Phase Test Dataset	52.40%	59.93%	51.74%	71.20%
On Final Phase Test Dataset	55.20%	65.64%	53.23%	85.60%

Table 1: BuTd model Solution Results

5 EVALUATION AND FUTURE WORK

Actually, our solution did not achieve our expectation. The performance showed the model may not learn too much useful features, and still randomly guess. Through the analysis of the results and model, we though reasons are following:

(1) **Gaps between train dataset and test dataset**
Simpons images are very different from natural images, the visual gaps cause bad performance

(2) **We use simplified visual head**
Object detection head would spend too much computation resources, so we just used simplified head in the competition

If we try to get better performance, we will collect simpons images and make same distribution dataset. Then use the dataset to produce visual features. Surely, the BuTd method may not the advancest method, the VQA task has been a part of multimodality. As developing of transformer attention mechanism and pretrained large model, there are many available solutions that we can reference and apply in this task.

6 CONCLUSION

We used BuTd method in this completion, but the results did not achieve our expectation. There are still some problems waiting us to solve.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Tech. rep. arXiv:1707.07998. arXiv:1707.07998 [cs] type: article. arXiv, (Mar. 2018). DOI: 10.48550/arXiv.1707.07998.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *arXiv:1612.00837 [cs]*, (May 2017). arXiv: 1612.00837. Retrieved Apr. 23, 2022 from http://arxiv.org/abs/1612.00837.

[4] Xiujun Li et al. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *arXiv:2004.06165 [cs]*, (July 2020). arXiv: 2004.06165 version: 5. Retrieved Apr. 23, 2022 from http://arxiv.org/abs/2004.06165.

[5] Peng Wang et al. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. Tech. rep. arXiv:2202.03052. arXiv:2202.03052 [cs] type: article. arXiv, (June 2022). DOI: 10.48550/arXiv.2202.03052.

[6] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. *arXiv:2111.02358 [cs]*, (Nov. 2021). arXiv: 2111.02358 version: 1. Retrieved Apr. 23, 2022 from http://arxiv.org/abs/2111.02358.

[7] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. Tech. rep. arXiv:2108.10904. arXiv:2108.10904 [cs] type: article. arXiv, (May 2022). DOI: 10.48550/arXiv.2108.10904.

[8] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked Attention Networks for Image Question Answering. Tech. rep. arXiv:1511.02274. arXiv:1511.02274 [cs] type: article. arXiv, (Jan. 2016). DOI: 10.48550/arXiv.1511.02274.

[9] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. en. In 5579–5588. Retrieved June 10, 2022 from https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.html.