2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) | 978-1-6654-3328-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCE-TW52618.2021.9603079

BEHAVIOR DETECTION IN SUPERMARKET BASED ON IMAGE CAPTION WITH UNITY

Li-Min Chen and Wei-Liang Lin
National Chung Hsing University, Taichung City, Taiwan, ROC
kd199731@gmail.com and wllin@nchu.edu.tw

ABSTRACT

This paper uses image caption to identify human behavior in supermarkets. In addition to observing the human posture to judge the current human behavior, we can reconstruct various human postures and scenes in Unity, and collect the human activity data set in the supermarket through the above methods.

INTRODUCTION

With the development of deep learning technology, more and more researchers are studying related fields, but in the real world, it is difficult to collect many dangerous scenes, such as sudden encounters of pedestrians when training autonomous driving or special behavior recognition training case, even when the data set is not enough.

But in Unity [1], a commercial game engine, user can create different virtual world in this software, you can easily simulate various situations, which is not only convenient for observation, but also can solve the problem of insufficient data set, and there is no privacy issue.

We can simulate the behavior of supermarket people through Unity, and use image caption to describe the current situation.

In Figure 1, it is mainly to detect whether there are children playing in the supermarket, the crowd chatting, and whether anyone falls down.

Fig.1 Simulate behavior of supermarket people in virtual world.

RELATED WORK

Transformer: Transformer uses a lot of special layers like self-attention [2] in the seq-to-seq [3] model. The only thing a self-attention layer needs is Attention [4]. Each input is multiplied by 3 different matrices, which are K(key), Q(query), V(value). Each Q pays attention to each K.

Attention uses two vectors to obtain a score(α), then through the soft-max layer to obtain $\hat{\alpha}$. Finally, through $\sum_i \hat{\alpha}_{j,i}\ v^i$ can get the output matrix. Figure 2-1 shows that through self-attention, input is a sequence and output a sequence.

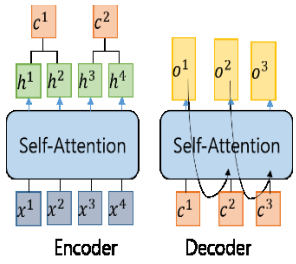


Fig.2-1 self-attention.

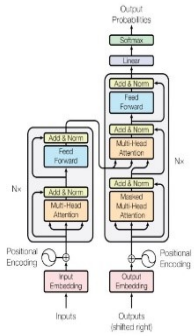


Fig.2-2 Transformer Architecture [4].

In brief, self-attention can be processed in parallel to speed up the calculation. Figure 2-2 is the transformer architecture [4].

Image Caption: Image caption [5][6] is a problem of scene understanding, the challenge is to identify the object, but also to understand the relationship between objects. The model includes two parts, Encoder and Decoder. Encoder uses CNN to extract image features. Decoder uses Attention to decode the features of the picture into natural language sentences.

APPLICATION

Currently, we have a possible application: global navigation environment. Image caption can identify the current human behavior. A navigation tool can use the related information of human environment, and use this information to plan an appropriate path, which can not only find a convenient and fast path but also avoid dangerous situations.

978-1-6654-3328-0/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: National Chung Hsing Univ.. Downloaded on October 27,2022 at 07:08:22 UTC from IEEE Xplore. Restrictions apply.

EXPERIMENT

We use Unity game engine to generate virtual world datasets, there were more than 16,000 training samples, and 1,200 verification samples.

The experiments have the input image size of 299*299*3. A total of 100 epochs was executed. Find the image features through the restnet-101 network structure, and match related sentences through the JSON file. Figure 3 shows the network architecture.

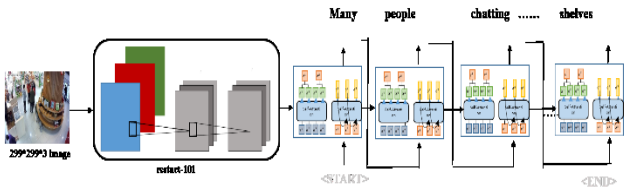


Fig.3 Network architecture.

Figure 4. shows that training loss: 0.0108, validation loss: 0.045, training process spent 26 hours.

Fig.4 Training & Validation loss trends.

After completion of the training, the virtual world datasets and the real world datasets were tested separately, in Figure 5-1 and Figure 5-2, use confusion matrix to calculate accuracy rate.

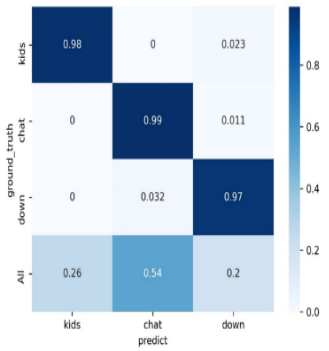


Fig.5-1 Virtual world data test.

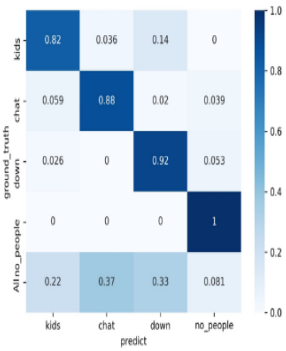
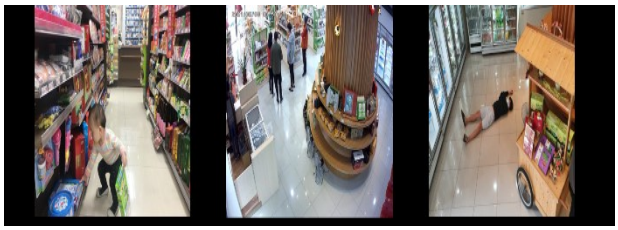


Fig.5 -2 Real world data test.

RESULT AND CONCLUSIONS

With this behavior recognition, we can not only judge the current human behavior, but also generate higher-quality picture descriptions. Then we can use this information effectively. The results are shown in Figure 6, input is a picture of different conditions, and output is the corresponding description.



(a) (b) (c)

Fig.6 Input image can get corresponding description, (a) "A kid are playing on the walkway." (b) "M any people are chatting beside the shelves." (c) "A person collapsed on the ground."

ACKNOWLEDGEMENTS

The study is supported by Ministry of Science and Technology (MOST) under MOST 109-2218-E005-011.

REFERENCES

[1] Sa Wang, Zhengli Mao, Changhai Zeng, Huili Gong, Shanshan Li, Beibei Chen, "A new method of virtual reality based on Unity3D", In 2010 18th International Conference on Geoinformatics.

[2] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena, "Self-Attention Generative Adversarial Networks", in Proceedings of the 36th International Conference on Machine Learning, PMLR 97:7354-7363, 2019.

[3] Harshitha Katpally; Ajay Bansal, "Ensemble Learning on Deep Neural Networks for Image Caption Generation", In 2020 IEEE 14th International Conference on Semantic Computing (ICSC).

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need.", In NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing.

[5] Wei Zhang; Wenbo Nie; Xinle Li; Yao Yu, "Image Caption Generation With Adaptive Transformer", In 2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC).

[6] Chaorui Deng, Ning Ding, Mingkui Tan, Qi Wu, "Length-Controllable Image Captioning", In ECCV 2020.

Authorized licensed use limited to: National Chung Hsing Univ.. Downloaded on October 27,2022 at 07:08:22 UTC from IEEE Xplore. Restrictions apply.