Visual Question Answering in Fashion

Seif Maged (*Media Engineering Technology*) and **Prof. Mohammed Salem** (*Digital Media Engineering*)
The German university in Cairo

seifeldin.ibrahim@student.guc.edu.eg

Introduction

Visual Question Answering (VQA) is an AI problem that given an image and a question about the image, the system should output an answer without the interference of a human. An interesting problem that combines different subdomains together, such as Computer Vision (CV) to understand the given image, Natural Language Processing (NLP) to understand the question and provide a reasonable answer.

Literature Review

Devi Parikh et al. [1] contributed mainly with two aspects. First, they offer a dataset which was the beginning of the era of enhancing datasets in VQA [2]. Second, experimenting with several methods to obtain the best accuracy. After several experiments, their best model consisted of LSTM to encode questions, l2 normalization for the last hidden layer in VGGNET (a convolutional neural network architecture) to encode images, and finally fusing both feature by element-wise multiplication to be passed over a fully connected layer followed by a softmax layer for obtaining distribution over the top K most frequent answers (K=1000). The previous model gained an accuracy of 0.58 for open-ended questions / 0.63 for multiple-choice questions, where the accuracy metric is defined as follows, $\min((no.of\ humans)/3,1)$ Their work was astonishing by the time (2015-2016).

Fig 1.1 [1]

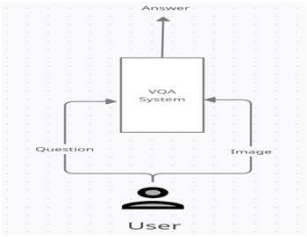
What color are her eyes?
What is the moustache made of?

Fig 1.2 [1]

How many slices of pizza are there?
Is this a vegetarian pizza?

Aller et al. [3][4] focused on fine tuning existing models to be applied on ecommerce use cases. First of all, they provided a review of the implementation of the 2018 VQA challenge winner model, Pythia. This is the model they chose to build upon (fine tune) on the VQA- V2 dataset, which is another version of the previously mentioned VQA dataset. The implementation of Pythia is open-source available on a modular framework called MMF that belongs to Facebook AI Research (FAIR). The authors try several approaches to fine tune the model, the best one was using a schedular in the training, but not only that, the schedular usage in training didn't provide a much better accuracy, however, using this model as being pre-trained model and then apply two regularization methods (weight decay and dropout) as a post training method.

Methodology



- A dataset containing fashion products images with labels partially describing each image, Kaggle's fashion product images [6].
- Modifying the dataset mentioned in the previous point to add Q&A pairs, as the VQA model needs to be fine-tuned on them.
- Use Vilt (Vision and Language Transformer) model and instead of fine-tuning it on the VQA-v2 dataset, finetune it on the newly modified fashion dataset

Results

Here are some samples run by our model

Fig 4.1

Q: What's in the image?
A: Sunglasses

Fig 4.2

Q: What style is this Tshirt for?
A: Casual

Evaluation

There is an existing Vilt model fine-tuned on VQA-v2, which is generic dataset not specialized in the fashion domain. We compared our model with theirs by choosing random 1000 non-trained images and their corresponding questions as a validation set. For each image, we did a sentence similarity between each question and the ground truth answer. We have 6 questions per image, so we calculated the mean of all questions per image for the predicted answers of our models and comparing it with the mean score of the other model. 99.3% of the images had the mean score of our model higher than theirs

Conclusion

The role of images is greater than before in social media and chatbots. However, not all people perceive images the same way as we do, like visually impaired people, here is where Visual Question Answering come for the rescue. In the future, we wish such system can be integrated in a real time application that allows people to capture a photo for any fashion related item, and ask a certain question, the application is expected to answer in real time. Added to this, It can be involved in a chatbot system that the user can upload images to ask questions about them. How amazing it would be to provide for the society such a big favor

References

1. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in Proceedings of the IEEE international conference on computer vision, pp. 2425–2433, 2015.
2. https://visualqa.org/download.html
3. M. E. Ortiz, L. M. Bergasa, R. Arroyo, S. Álvarez, and A. Aller, "Towards fine-tuning of vqa models in public datasets," in Workshop of Physical Agents, pp. 256–273, Springer, 2020
4. R. Arroyo, S. Álvarez, A. Aller, L. M. Bergasa, and M. E. Ortiz, "Fine-tuning your answers: a bag of tricks for improving vqa models," Multimedia Tools and Applications, pp. 1–25, 2022
5. Kim, W., Son, B., & Kim, I. (2021, July). Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning* (pp. 5583-5594). PMLR.
6. https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset

Prepared for Thesis Poster Display Conference
11th -12th June 2022
Faculty of Management Technology
THESIS Poster Display Conference 2022