VQA on Simpsons scenes: Transfer learning from pre-trained Vision-and-Language Transformer

Jessie Xiaojuan He
hexiaoju@deakin.edu.au
Deakin University
Burwood, Victoria, Australia

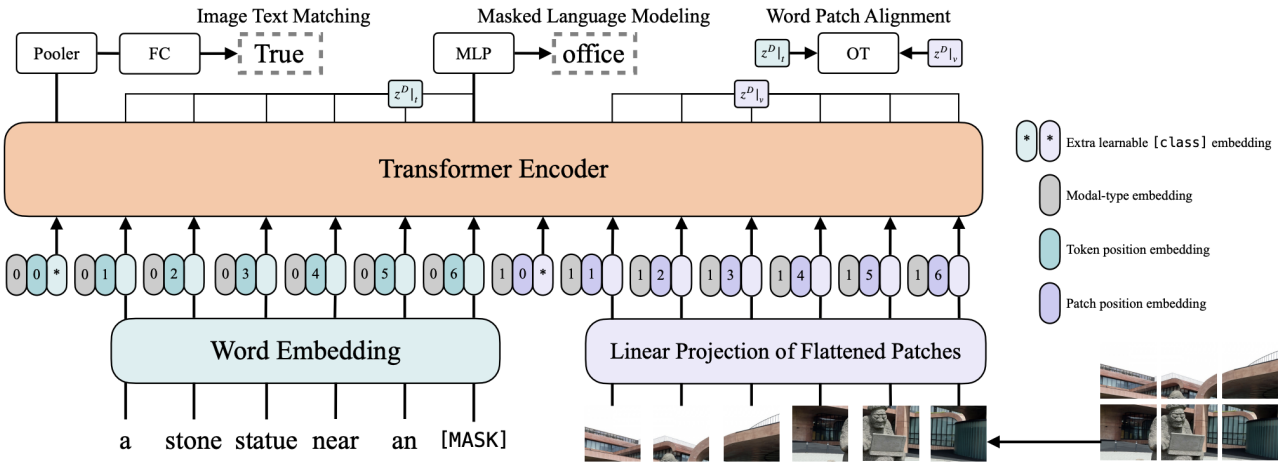

Figure 1: Vision-and-Language Transformer overview [2]

ABSTRACT

Transfer learning is a prevalent technique in deep learning, where pre-trained models are employed as a starting point for computer vision and natural language processing tasks due to the extensive computational and time resources required to create neural network models for these tasks. The ViLT for VQA model was pre-trained on a vast number of images with captions and VQAv2 dataset, so it could effectively serve as a generic model of VQA tasks. We fine-tuned the model on relatively small Simpsons data. On the validation dataset, the fine-tuned model achieves an accuracy score of 72%, while on the test dataset, it is 70.4%.

KEYWORDS

transfer learning, vision and language, transformer

ACM Reference Format:
Jessie Xiaojuan He. 2022. VQA on Simpsons scenes: Transfer learning from pre-trained Vision-and-Language Transformer. In *Proceedings of Aug 03–05, 2022 (Deakin Simpsons Challenge).* ACM, New York, NY, USA, 3 pages.

Deakin Simpsons Challenge, Burwood, VIC,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

1 INTRODUCTION

VQA is a multimodal task in which the desired result is an accurate natural language response given an image and a natural language question connected to the image. Typically, a VQA model's architecture consists of three major components: word embedding, image embedding, and modality interaction. The modality interaction combines the word embedding and the image embedding into a single representation. Given the complexity of the task and the limited data available, transfer learning is an effective option. Using transfer learning, we want to fine-tune a pre-trained VQA model to answer a natural language Yes or No question given an image of the Simpsons. Our code is available at https://github.com/JessieHex/SimpsonsVQA

2 PRE-TRAINED MODEL

A transformer is a deep learning model that uses the attention mechanism to differentially weight the importance of each input data segment. It has been demonstrated that Transformers operates not only in the respective domains of Natural Language Processing but also with Computer Vision[1]. It enables the two VQA modalities to be handled in a unified manner. ViLT[2] turns an image into a vector by dividing the image into fixed-size patches and linearly projecting the flattened patches without needing a visual embedder. The transformer encoder combines the word and image embeddings to provide a unified representation. Initialize the interaction transformer with the pre-trained ViT. Such an initialization leverages the power of interaction layers to interpret visual

information without a separate deep visual embedder. The ViLT model used for transfer learning was pre-trained on millions of captioned photos and fine-tuned with VQAv2. ViLT achieves an accuracy of 59.21% on the local test dataset, which is significantly higher than the random model's accuracy of 50%, despite not being tuned with Simpsons images. The model does not recognize the Simpsons character and their facial features prior to fine-tuning with Simpsons data, but it can infer the answer given the question's context. Figure 2 illustrates an example of the outcome.

```
Q: Is there a man ?
A: no

Q: Is the man wearing glasses ?
A: yes

Q: Is the man smiling ?
A: yes

Q: Is the man holding a box ?
A: yes

Q: Is the man holding an apple ?
A: no
```

Figure 2: Prior to fine-tuning with Simpsons data, the model does not recognize the Simpsons character and facial traits; nonetheless, it is able to deduce what he is holding based on the query context.

3 DATA

The dataset used to fine-tune the pre-trained model consists of 23 questions and answers related to Simpsons characters and their features. The data augmentation is implemented during fine-tuning. We employ a subset of the ViLT's policies by omitting the color-related operation, as Simpsons graphics have less color variation than real images. Each question has the same number of images with the answers 'yes' and 'no' in an attempt to eliminate dataset bias when constructing the dataset[3]. The number of questions must also be comparable; otherwise, the questions with more data may be overfitted while the other questions remain underfitted.

```
Q: Is there a man ?
A: yes

Q: Is the man wearing glasses ?
A: no

Q: Is the man smiling ?
A: yes

Q: Is the man holding a box ?
A: yes

Q: Is the man holding an apple ?
A: no
```

Figure 3: After fine-tuning, the model is able to recognize the character and certain characteristics without losing its previous knowledge.

4 FINE-TUNING

The ViLT model outputs the possibility of 3,120 answer classes containing the "yes" and "no" answers required for the task. It is usual practice to remove the pre-trained model's output layer and train a new one. However, with the small dataset we have, we discovered that the model with the newly added output layer lost most of its capacity for inference. As a result, we opt to leave the model intact and modify the "yes" and "no" indices outside of the model. We set the learning rate to a relatively low value (1e-5) to prevent the model from losing its prior knowledge.

Different questions appear to be of varying degrees of difficulty. The answers to inquiries such as "Is there a man/woman?" may be tuned in five epochs. However, the character's facial expression appears more subtle and challenging to train. The capacity of the

original model to classify the facial expressions of real images is likewise limited.

After fine-tuning, the model can recognize the character and certain characteristics without losing previous knowledge, and the accuracy on the local test dataset is improved to 68.29%.

5 CONCLUSION AND FUTURE WORK

Transfer learning for VQA demonstrates that if a model is trained on a sufficiently large and general dataset, it could effectively function as a generic model for a specific category of images. We can take advantage of these learned feature maps without having to train a vast model on a large dataset from begin.

Relatively restricted data was obtained, further data on a variety of images and questions could enhance the accuracy. Also, the model has been trained on VQAv2, which has more open-ended questions. It may be possible to increase the model's performance by exploiting the pre-trained model's power by asking open-ended and closed questions.

REFERENCES

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[2] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[3] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5014–5022.