DVMM Weekly Report

Yueen Ma

November 27, 2020

1 Experiment

The first step of this project is to add addition temporal information from VisualCOMET [1] dataset as input and see whether the extra pieces of information help improve the accuracy of VL-BERT [3] on the VCR dataset [4]. The VCR dataset itself provides the question and multiple answer choices from which the model should pick the correct one. The VuisualCOMET dataset provides three addition kinds of sentences: "intent", "before", "after". We currently simply append those sentences to the question sentence and separate them with periods. For example:

```
{"question": ["Where", "are", [0, 2], "?"]
"intent": ["ask the woman on a date", "get over his shyness"],
"before": ["approach 3 at an event", "introduce himself to 3",
           "be invited to a dinner party", "dress in formal attire"],
"after": ["ask 3 to dance", "try to make a date with 3",
            "greet her by kissing her hand", "order a drink from the server"]}
```

Is converted to a list of tokens: (Note that even though this is a single non-nested list, tokens from each sentence are separated with a period '.' and the following example is formatted so that the period is in the end of each line):

```
['Where', 'are', [0, 2], '?', '.',
'ask', 'the', 'woman', 'on', 'a', 'date', '.',
'get', 'over', 'his', 'shyness', '.',
'approach', '3', 'at', 'an', 'event', '.',
'introduce', 'himself', 'to', '3', '.',
'be', 'invited', 'to', 'a', 'dinner', 'party', '.',
'dress', 'in', 'formal', 'attire', '.',
'ask', '3', 'to', 'dance', '.',
'try', 'to', 'make', 'a', 'date', 'with', '3', '.',
'greet', 'her', 'by', 'kissing', 'her', 'hand', '.',
'order', 'a', 'drink', 'from', 'the', 'server', '.',]
```

Because VisualCOMET does not have one annotation for every example in VCR, we remove those examples from VCR that does not have a corresponding VisualCOMET annotation.

2 Results

After VL-BERT is trained for 20 epochs, we can see that both VL-BERT models reached their limit in terms of training accuracy, as shown in Figure 1. There is not a big training-accuracy gap between the model trained with VisualCOMET annotations and the model trained with only VCR annotations. But when the two models are tested on the validation set, there exists a constant $3\% - 4\%$ validation-accuracy gap throughout the 20-epoch training, as shown in Figure 2.

In addition, we can see that the best model is reached at epoch 7 and epoch 15 for the models trained with and without VisualCOMET annotations, respectively, which suggests the models are starting to overfit

after those epochs. Therefore, we can confirm that the VisualCOMET does improve the performance of VL-BERT on the VCR task.

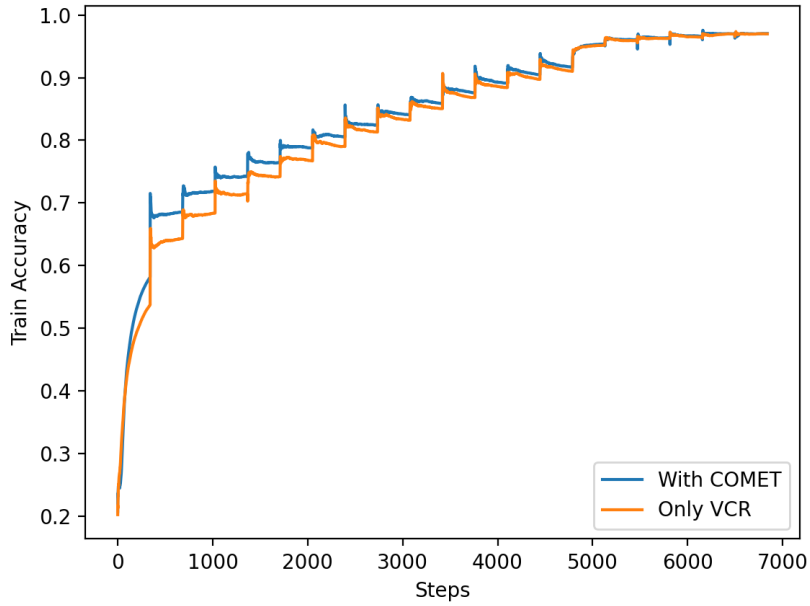

Figure 1: Training Accuracy

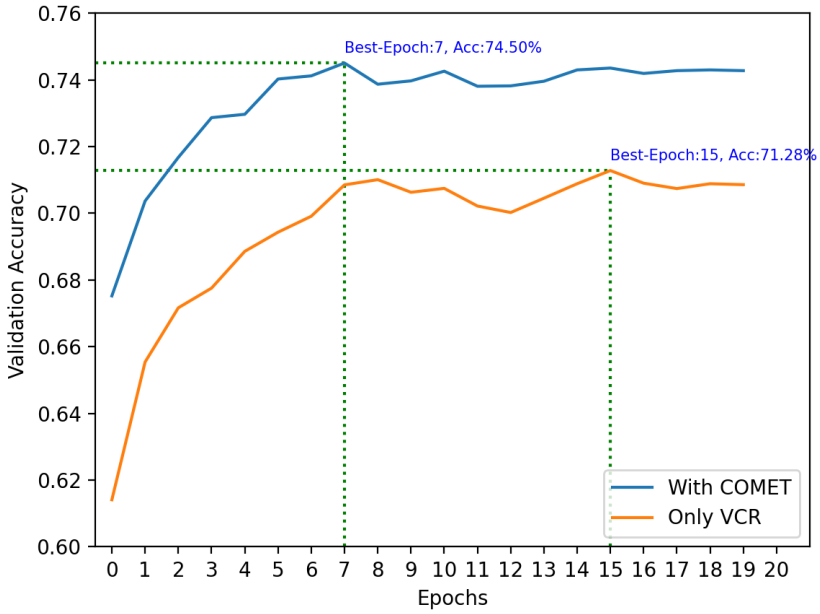

Figure 2: Validation Accuracy

3 Next Steps

I got the GPT-2 model up and running [2] to infer VisualCOMET annotations using the pretrained parameters provided on `https://github.com/jamespark3922/visual-comet`. But the current result is not very good. The validation accuracy using the predicted VisualCOMET annotations is only 58.97%, which suggests that those annotations are misleading the VL-BERT model.

I am aware of some existing problems that I will try to fix:

- In the VisualCOMET dataset, there could be multiple annotations for one image. But I am not only reading one of them.
- I have also calculated the validation accuracy using my own code. But I found there is a slight difference between my validation accuracy and the validation accuracy from the VL-BERT model itself. But the difference is very small (within 1%), so the conclusions I drew above still hold. I will also investigate the cause of this disparity.
- The annotations use number to refer to object/person in the image. But now I am extract the number as a textual token whereas the correct way is to extract the number into an integer.

References

[1] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image, 2020.

[2] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[3] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations, 2020.

[4] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.