Appendix

Table 1. Evaluation of different language-only Transformer models on the VQA-CP v2 test set.

Model	Yes/No	Number	Other	Overall
BERT base	42.66	16.19	45.66	40.29
BERT large	42.72	17.43	48.47	42.06
XLNET base	43.49	17.61	48.45	42.30
XLNET large	44.58	**20.67**	**50.52**	**44.23**
RoBERTa base	44.39	17.46	48.74	42.70
RoBERTa large	**45.13**	20.06	49.33	43.64

1. Language Transformers

We compare the performance of prevailing language-only Transformer models: BERT [1], XLNET [6], and RoBERTa [4], both in their base and large versions. All the models are exposed to the same input, consisting of the question, the narrative, and the five captions.

Results are shown in Table 1. All the models present a similar behavior. XLNET large has the best performance, and its accuracy is higher 0.59% compared to RoBERTa large. However, while the improvement is minor, the computational time of XLNET large is about 2.7 times that RoBERTa. Considering this, it is reasonable to use RoBERTa large to save the training time while obtaining relatively similar results.

2. Qualitative Analysis

We show some qualitative examples in Figure 1. We compare our model's predictions with the predictions of BAN [2] and VisualBERT [3]. Example (1) asks what the man is doing with his hand. The image description contains the expression of what the man is doing (*"A man pointing to pots..."*), result in our language-only model answering correctly. On the other hand, BAN and VisualBERT fail to make the correct prediction, which they answer *"cooking"* nevertheless the man is not cooking. The object detector can detect the cooking tools, so the models may cause this mistake guess from the utensils, not from the man's move. In example (2), the image description also describes the critical information to answer the question (*"Six snowboards"*), while the object detector cannot detect all objects. On the contrary, examples (3) and (4) show the limitations of our language-only model. Example (3) requires reading the letters on the side of the plane. The image description contains the necessary word (*"AirFrance"*) to answer the question but fails to make a correct prediction. This result poses the lack of language-only Transformer models' ability to understand the contents and relations between the question and description. Additionally, example (4) shows the importance of the image descriptions' quality. Our model fails to correctly answer because there is no information to answer the question. From these observations, utilizing well-described text as image representation has the advantage against the deep visual features when answering the questions that deep visual features do not work well. On the other hand, we identify the limitation of our language-only model for understanding the text input.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[2] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018.

[3] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, 2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

[5] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[6] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.

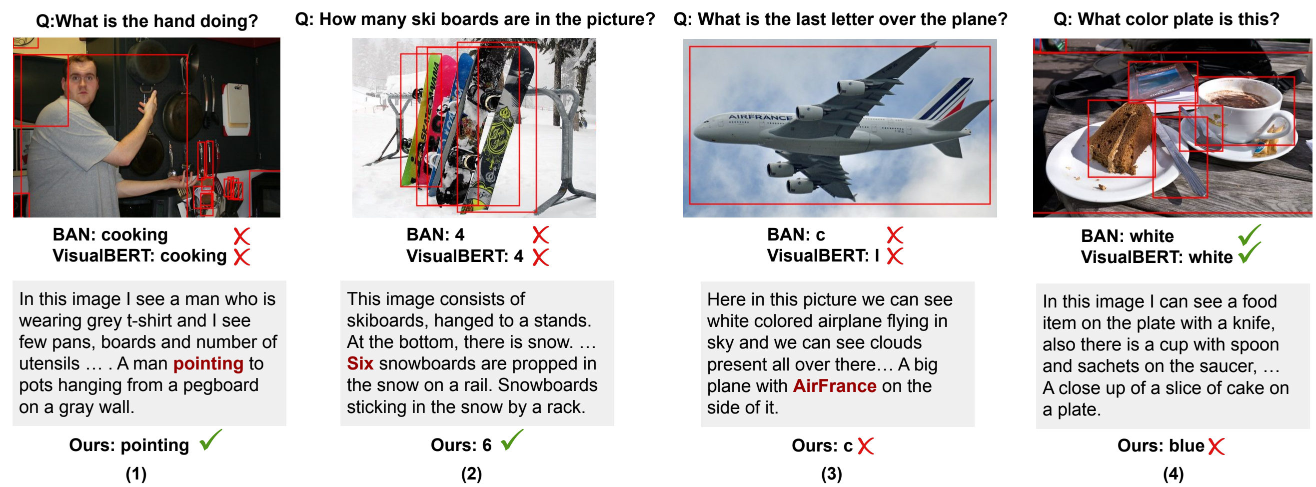

Figure 1. Qualitative Comparison. The red boxes in the images denote the result of detection by Faster R-CNN [5]. Only bounding boxes with a confidence score greater than $\mathbf{0.5}$ are shown. The shown descriptions are extraction of the actual descriptions. The red highlighted words are the relevant words to the answers.