1. The paper needs to state what exactly has been proposed in new terms vis-a-vis existing models like LXMERT or Pixel-BERT for VQA.

2. The validation accuracy is low indicating the possibility of overfitting, and the authors should explore the application of regularization techniques or potential validation cross-dataset for improvement of generalization.

3. Scene graphs can be richly illustrated with better pictorials and an extensive breakdown of how they improve performance on question answering.

4. Evaluation should not be limited to accuracy alone, but must include some VQA-standard metrics, such as consensus scores and precision-recall assessment.

5. It is very interesting the CBMS approach, but one can expand it with failures, some ablation studies, and comparisons with other model fusion approaches.

6. Such related work is covered in a comprehensive manner, but it should be relatively focused and complemented by a comparative table summarizing model architectures and certain critical differences.

7. There are issues regarding data leakage because of the overlap between VQA v2.0 and Visual Genome datasets; these need to be countered to ensure a fair evaluation.

8. Improvement should be made in the figures and illustrations in the paper by having clear labeling and color coding along with an exemplary illustration that shows the entire pipeline for VQA with scene graphs.

|  |  |  |
| --- | --- | --- |
| **Model** | **Datasets** | **Accuracy** |
| VisualBERT | VQA v2.0 | 71.00 |
|  | NLVR2 | 67.00 |
|  | VCR | 71.60 |
|  | Flickr30K | 71.33 |
| ViLBERT | VQA v2.0 | 70.55 |
|  | VCR | 72.42 |
| ViLBERT | VQA v2.0 | 70.55 |
| Pixel-BERT (x152) | VQA 2.0 | 74.45 |
| MMFT | TVQA | 74.97 |
| ViLT | Visual Genome | 70.33 |
| VLC-BERT | A-OKVQA | 38.05 |
| UNITER | COCO, Visual Genome, Conceptual Captions, and SBU Captions | 74.02 |