Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model
Fabian Deuser, Konrad Habel, Philipp J. Rösch, Norbert Oswald

1. Motivation
Easy-to-train approach, without a massive ensemble. We use a simple model on top of extracted CLIP [1] features. CLIP is trained with 400 million image-text pairs and therefore has powerful representations for both modalities.

2. Methods
Our model leverages CLIP as feature extractor for the image and question encoding. The pre-trained CLIP backbone is kept frozen and is not fine-tuned on the VizWiz data set. We encode six different versions of the image and combine these vision features with a weighted mean (noted as TTA). Both feature vectors of the image and question encoding are concatenated and passed to the VQA and Answerability module.

2.1 VQA
Answer Vocabulary Generation:
- Selection of the most common answer per sample
- If this selection yields in several answers, the answer which appears most often in the whole training set is used
- With this selection process the remaining number of answer candidates for training decreases to 5726 classes

Answer Type Gate:
- We create eight answer types for the auxiliary loss, based on the best selected answer using regular expressions
- The resulting answer types are linear projected to a vector with the same size as the possible answer classes
- After a sigmoid layer this vector is multiplied with the logits of the answer vocabulary

2.2 Answerability
Simple classifier with linear layers and activation function SiLU.

3. Discussion
- CLIP is trained on texts from webpages. Questions are semantically different from typical image descriptions, but nevertheless performance is surprisingly good.
- We do not have a separate OCR module. But in line with the original CLIP model, our model also shows some OCR capabilities.
- There are "unsuitable" and "unsuitable image" as ground truth answers. We merge both into the single class "unsuitable".

Table 2.1: Impact of model components on test-dev with CLIP RN50x64 backbone.

<table>
<thead>
<tr>
<th>Question Encoding</th>
<th>Image Encoding</th>
<th>TTA</th>
<th>Answer Type Gate</th>
<th>VQA [Acc]</th>
<th>Answerability [AP]</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>36.98 %</td>
<td>56.58 %</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>53.35 %</td>
<td>69.42 %</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>59.84 %</td>
<td>82.27 %</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>60.26 %</td>
<td>82.74 %</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>60.73 %</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.1: Final results on test-dev and test-std.

<table>
<thead>
<tr>
<th>VQA [Acc]</th>
<th>Answerability [AP]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP Backbone test-dev</td>
<td>CLIP Backbone test-std</td>
</tr>
<tr>
<td>RN50x64</td>
<td>60.73 %</td>
</tr>
<tr>
<td>ViT-L/14@336px</td>
<td>60.66 %</td>
</tr>
<tr>
<td>Ensemble</td>
<td>61.64 %</td>
</tr>
</tbody>
</table>

Table 3.2: VQA grouped final results on test-std.

4. Conclusion
- Simple and small VQA and Answerability module enables fast training without high computational resources
- No fine-tuning of the backbone needed
- Utilizing the advantages of pre-trained CLIP model
- Novel way of using CLIP for VQA tasks

5. References