Universität der Bundeswehr München

Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model

Fabian Deuser, Konrad Habel, Philipp J. Rösch, Norbert Oswald

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

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.

https://www.unibw.de/vis-en/cvpr2022

TTA: weighted mean of features

Crop			Resize		
0°	-90°	+90°	0°	-90°	+90°
0.250	0.125	0.125	0.250	0.125	0.125

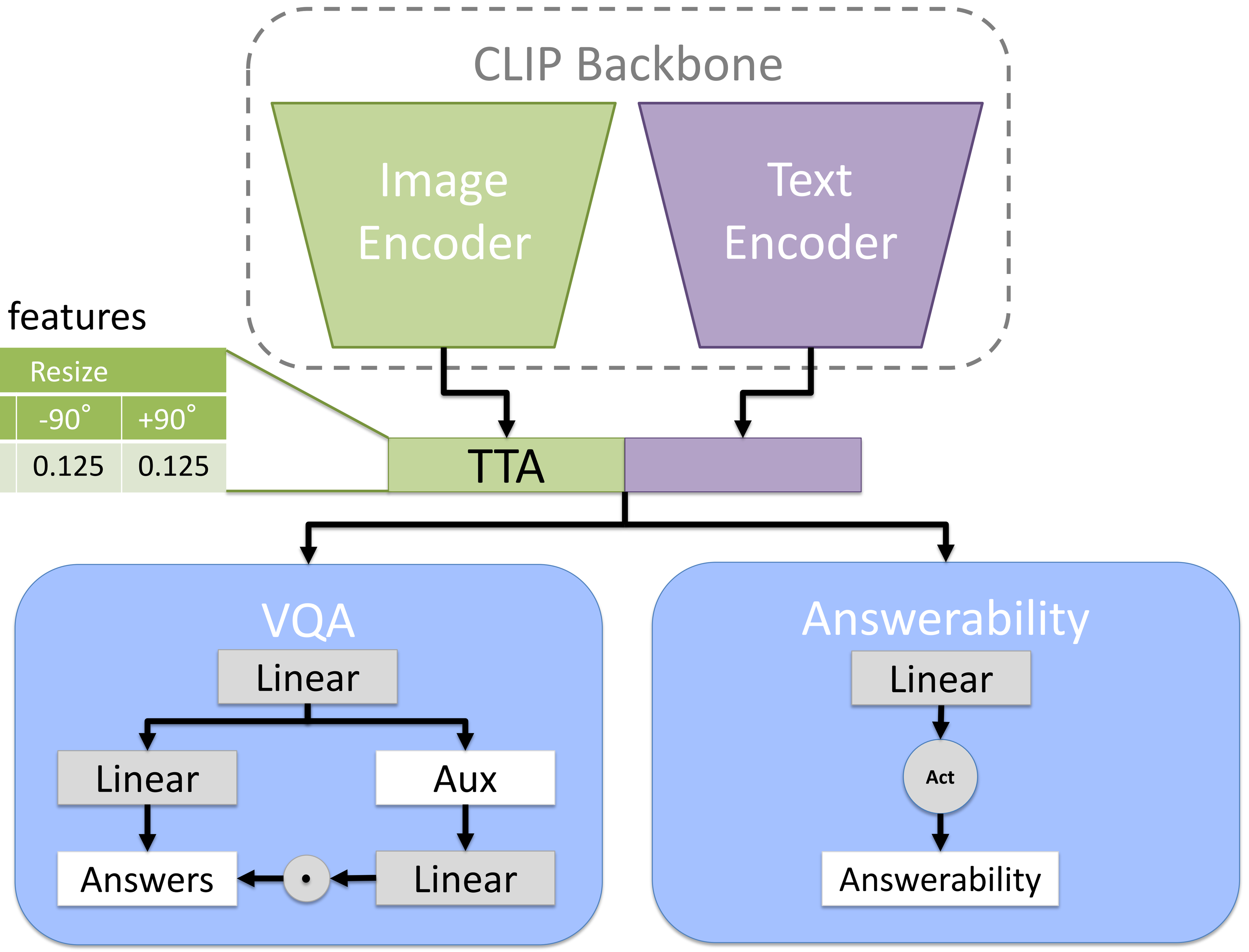

Figure 2.1: Our CLIP-based model for the VizWiz Challenge.

Question Encoding	Image Encoding	TTA	Answer Type Gate	test-dev VQA [Acc]	test-dev Answerability [AP]
X				36.98 %	56.58 %
	X			53.35 %	69.42 %
X	X			59.84 %	82.27 %
X	X	X		60.26 %	**82.74 %**
X	X	X	X	**60.73 %**	-

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

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".

	VQA [Acc]		Answerability [AP]	
CLIP Backbone	test-dev	test-std	test-dev	test-std
RN50x64	60.73 %	59.40 %	82.74 %	82.54 %
ViT-L/14@336px	60.66 %	59.01 %	83.50 %	82.86 %
Ensemble	**61.64 %**	**60.15 %**	**84.13 %**	**83.78 %**

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

overall	unanswerable	other	yes/no	number
60.15 %	88.15 %	49.27 %	64.95 %	33.33 %

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

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021.