Can CLIP Read?

Blessing Itoro Bassey
Machine Intelligence
African Institute for Mathematical Sciences (AIMS)
bbassey@aimsammi.org

Yonatan Bisk
Language Technology Institute (LTI)
Canegie Mellon University (USA)
ybisk@cs.cmu.edu

1 Introduction

So far, in most computer vision task, we can take advantage of the knowledge a neural network has gained from task A (trained with a lot of annotated data) and apply that knowledge to learn a separate task B (usually scarce of annotations). This what we call a transfer learning approach. Examples are Imagenet [1]. However, the representations learned from neural network A might be limited in generalizing to other task since additional labeled data is needed to specify any other visual concept.

In the bit to solve the above challenge of the the zero-shot transfer of model to downstream tasks, CLIP model [2] trained on a dataset of 400 million (image, text) pairs collected from the internet. Given a set of (image, text) pair $\{(x_i, c_i)\}$ where $i \in \mathcal{R}$, CLIPS trained I and T representing the Image and Text encoder respectively. 1 was with ResNet or Vision Transformer while T was with CBOW or Text Transformer. The cosine similarity between related pairs of $(x_i), T(c_i)$ was maximize while minimizing unrelated pairs.

2 Motivation and Research Question

CLIP is much more efficient at zero-shot transfer than our image caption baseline. However, **Can CLIP Read?** i.e. given an image alongside questions, with potential possible answers can CLIP read the question alongside the image to select the best possible answer? Contrary to what it's originally trained for just (image,captions). Can CLIP be used for Visual Question Answering task? where we have (image and questions, captions). Using the TextVQA [3] comprising of 34,602 questions and answers for training set, 21,953 training and validation set images, and 5000 validation questions and answers. The distributions of the questions and answers can be seen in figure 1 to 4.

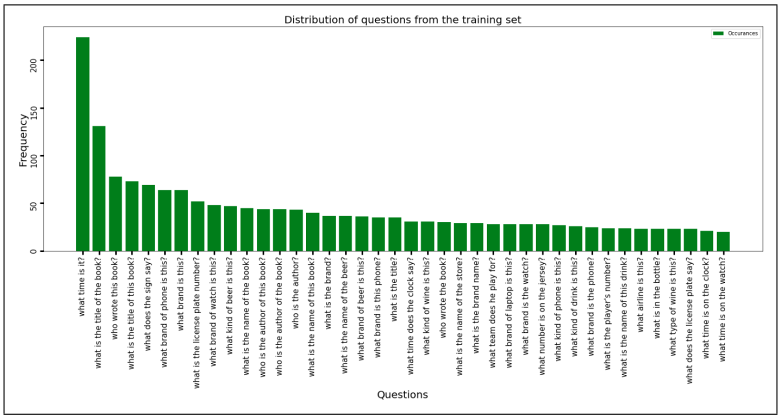

Figure 1: Distribution of the 40 top most occurring questions in the TestVQA training set

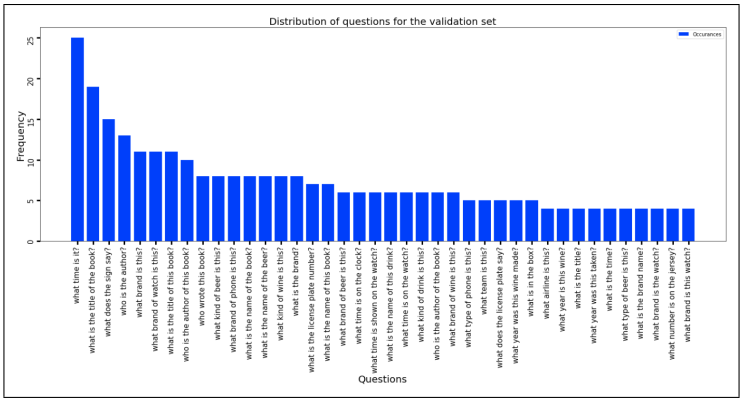

Figure 2: Distribution of the 40 top most occurring questions in the TestVQA test set

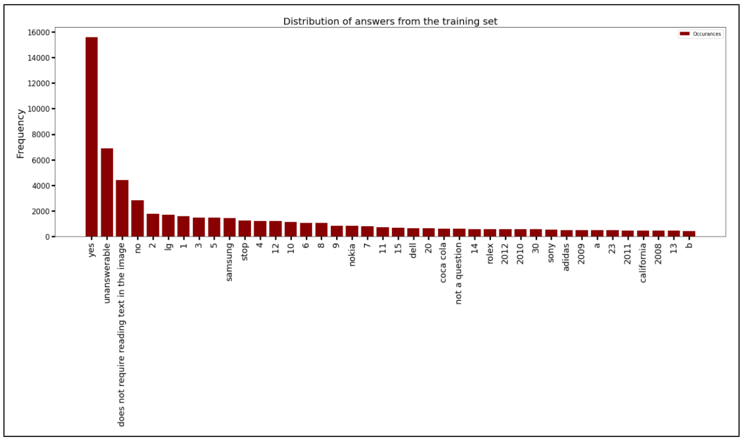

Figure 3: Distribution of the 40 top most occurring answers in the TestVQA training set

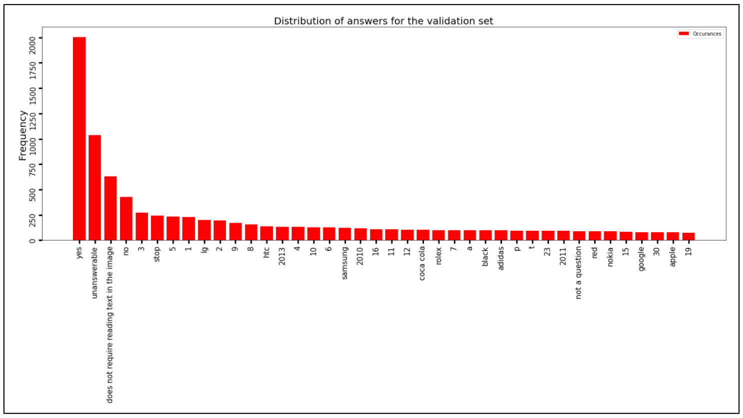

Figure 4: Distribution of the 40 top most occurring answers in the TestVQA validation set

As seen from the overall distributions, most of the questions in training set can also be found in the validation set, likewise the answers too.

3 Experimentation Ideas

1. Using CLIP directly on TextVQA dataset for Zeroshoot: Here we only passed in (image and questions, answer) rather than the original clip format of (image, caption) pair.

2. Fune-tuning CLIP by training the embeddings gotten from CLIP again: The same was done as above, but after training. Bellow were the different method tried already:
    - The Naive training approach: In this case all we did was training the images, question and answer pair using the embeddings from CLIP. For this first CLIP experimentation, we want to see how well the clip model will theoretically perform on validation part, given answers from the training.
    - Using the negative hard mining: Due to the performance gotten above, we decided to make the model learn hard negative sets of images. I.e allowing the model to choose it's batch itself. The first step was to find the similarities between an image and all other images in the dataset, afterwards, we cluster this similarities so that images with high similarities are passed into the model to be trained as batch size. This would allow the model to learn very had. So as to Identify distinct and clear difference between an image and and another (even though they might look so similar). See Figure 5 to 7.

4 Results

variables	training	validation
Image	21,953	3,166
Questions	34,602	5000
Answers	34,602	5000

Table 1: Summary of TextVQA dataset

Model	Accuracy
CLIP	3.033
Naive trained CLIP	3.395
Negative mining of images	0.221
Negative mining of text	0.321
Negative mining of images and text	0.22

Table 2: Using the image and questions from the validation seta and the top 10,000 answers from the training set, we see that the trained CLIP had a slightly better performance compared to the ordinary CLIP model.

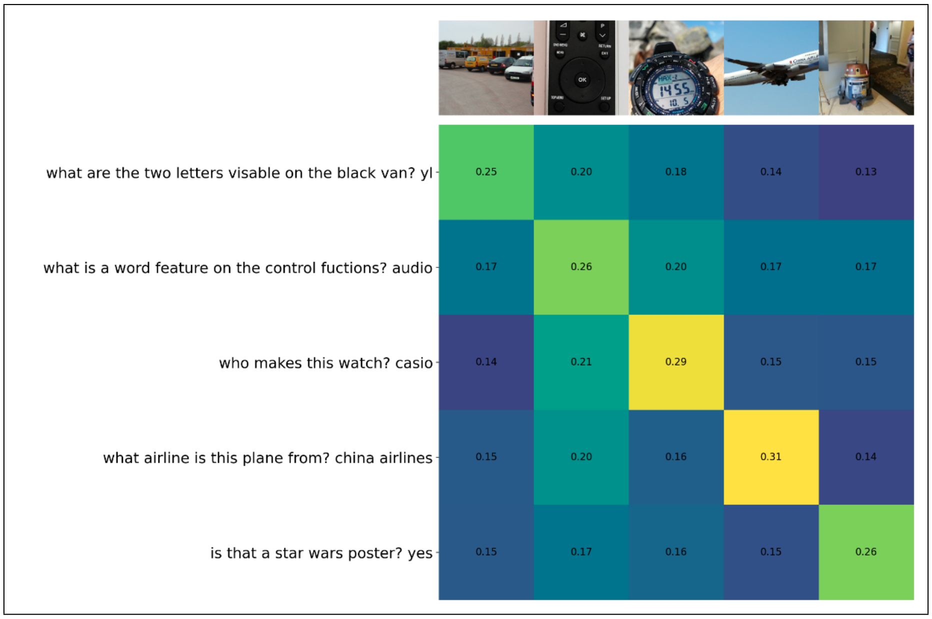

Figure 5: Zero shoot prediction from CLIP model

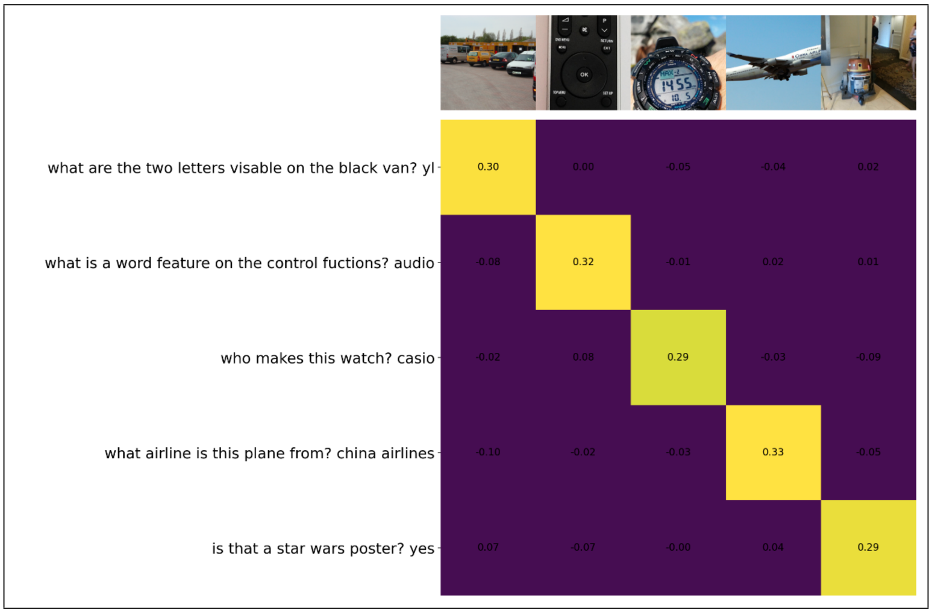

Figure 6: Zero shoot prediction from naively trained CLIP model

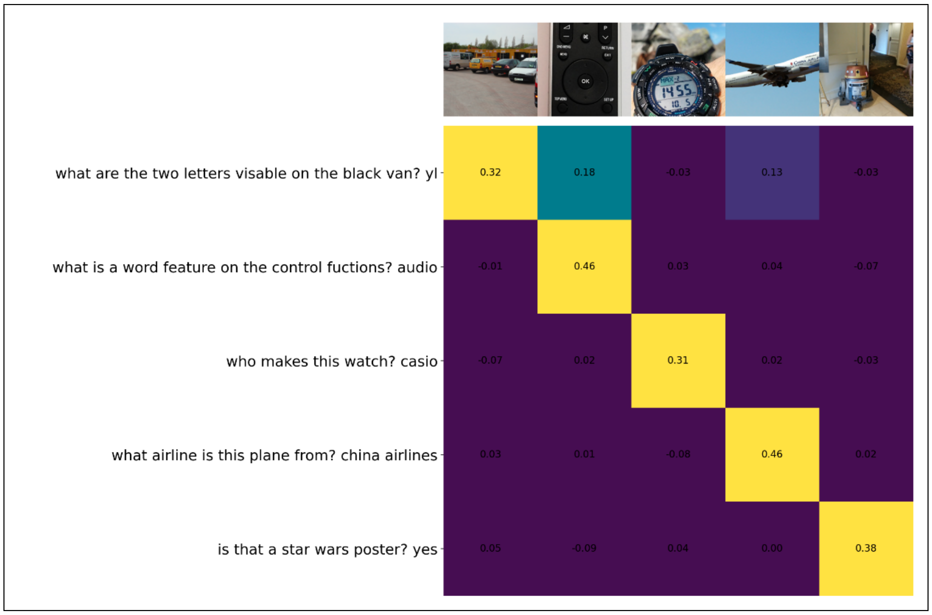

Figure 7: Zero shoot prediction from hard negative mining of images

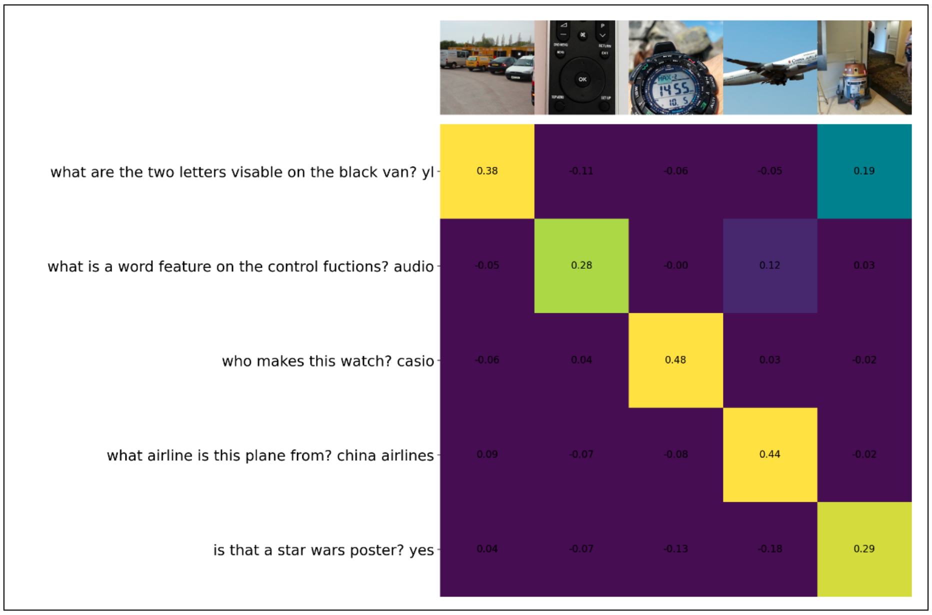

Figure 8: Zero shoot prediction from hard negative mining of text

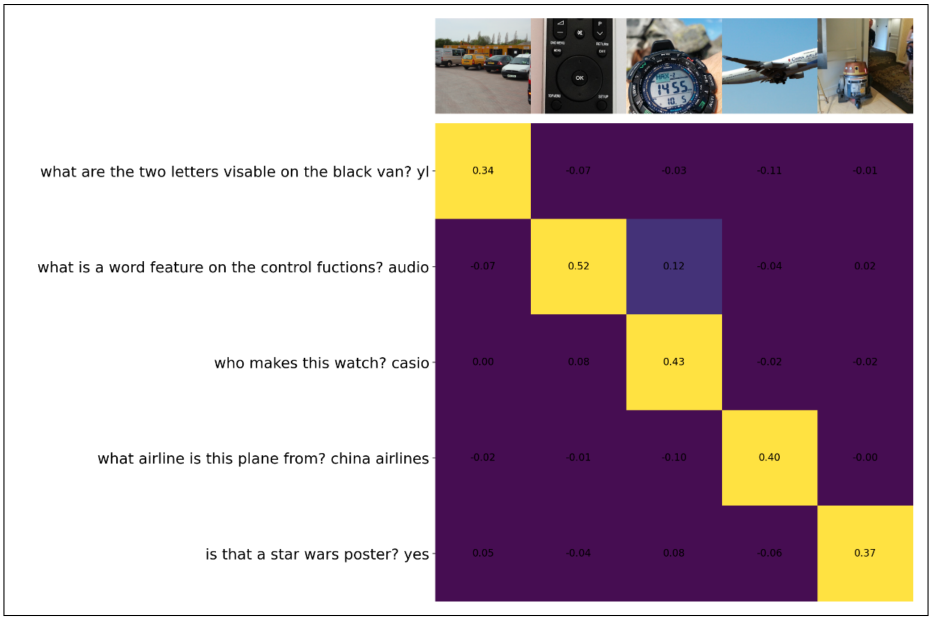

Figure 9: Zero shoot prediction from hard negative mining of images and text

As seen in figure 7, there is an improvement in the probabilities of image captioning compared to that of CLIP in figure 5. However, there are still some limitations to the model, especially for "extreemly hard images" e.g in figure 8 bellow.

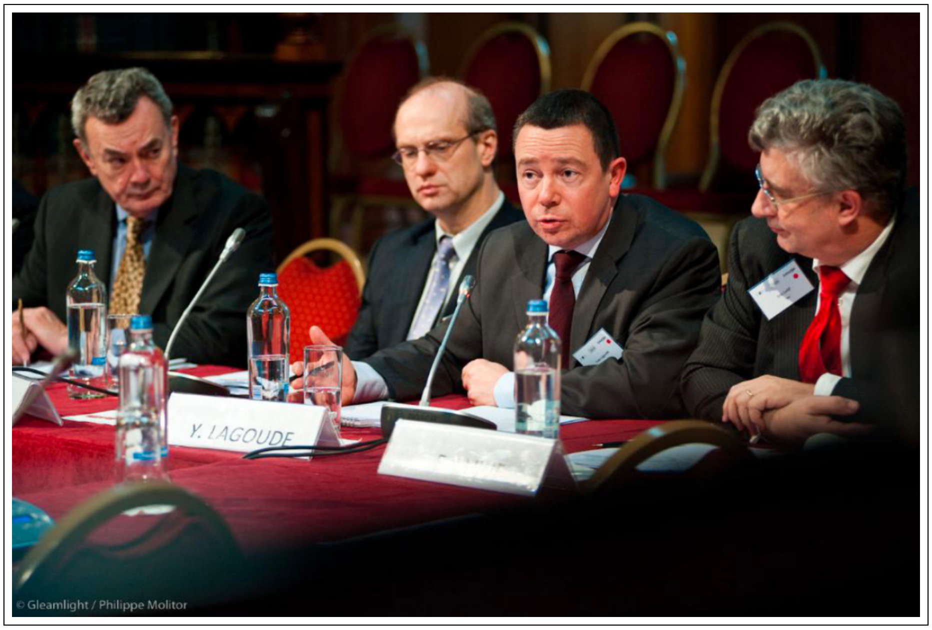

Figure 10: This image is very challenging for the model to caption. Given a question like, "who is the photographer?"

The image above is very had for the model to reason that the answer it should predict is at the lower left corner of the image. The model trained predicted the answer to be "American Politicians" while the CLIP model predicted the answer to be "Europe China forum". The predicted answers might bot be right based on the question asked, but to some extent, it is reasonable enough, as all the model did was to capture the people present in the image rather than focusing on the text. Hence, In order to try addressing this mistake bellow are the proposed directions to look at.

1. Does CLIP know where to look? I.e can it predict where the answer is? If it had understaning that the answer to the above image was at the left coner, maybe it would have predicted something much more better. We might thing of cropping a single image into different segment and feeding the model with it to train, with this, we might have segment that only has text (the require answer) and segments with images (not needed answer) based on the question. An idea is seen bellow:

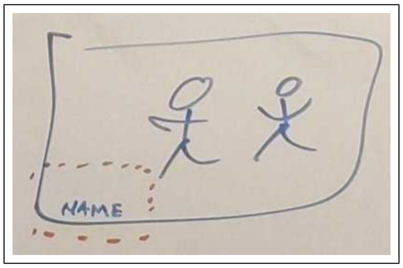

Figure 11: This image is very challenging for the model to caption. Given a question like, "who is the photographer?"

2. Creating synthetic dataset of images of words in the dictionary. Can CLIP read this?

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[3] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.