Does Text Matter?
Extending CLIP with OCR and NLP in Image Classification and Retrieval

Author
Jordan Sassoon

Supervisor
Zilong Zhao

Responsible Professor
Lydia Y. Chen

Research Project for the Technical University of Delft, June 28th, 2023

Introduction

1

This paper proposes a novel architecture: **OSBC** (**O**CR **S**entence **B**ERT **C**LIP), which leverages the text contained within images as an additional feature when performing image classification and retrieval.

OSBC combines two architecures:

1. CLIP [1], a popular zero-shot computer vision model.
2. OCR-SBERT, a novel pipeline which focuses on text extraction.

The aim is to create an architecture that can support CLIP when inner text is important, as CLIP struggles on this.
OSBC was tested on multiple datasets for image classification and retrieval [Fig. 1], occasionally outperforming CLIP, while maintaining finetunability, and improving model robustness.

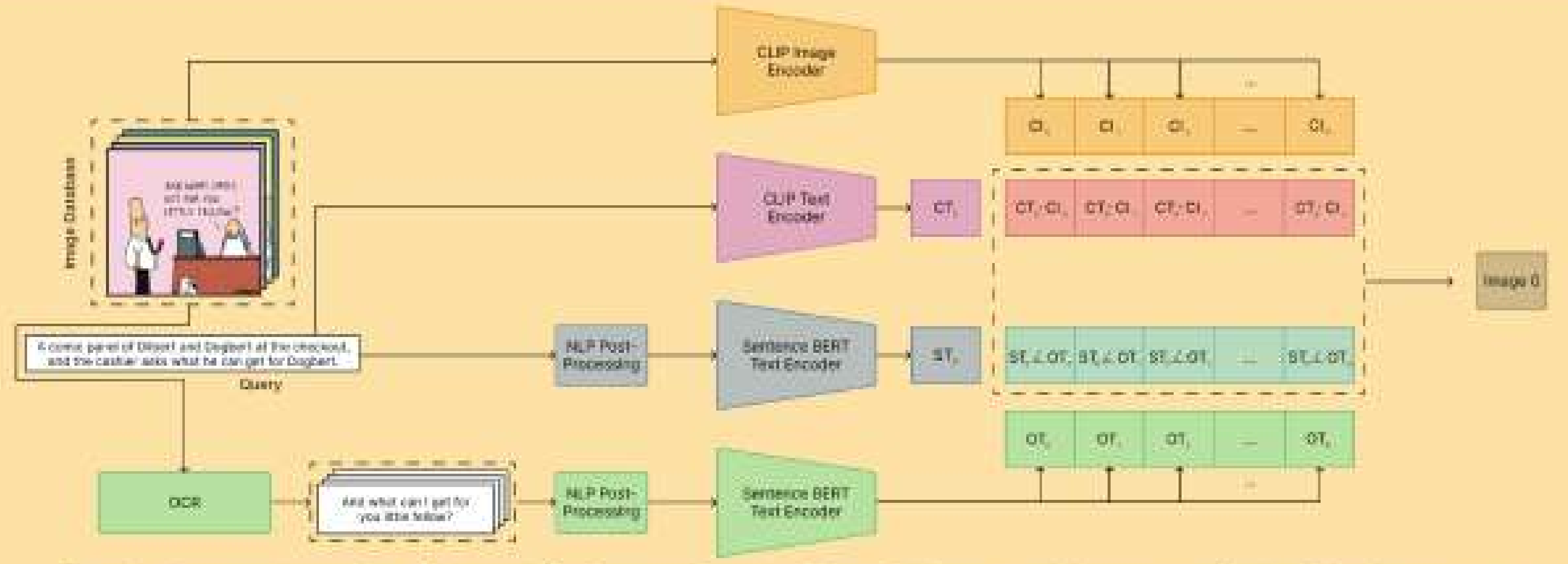

Figure 1: An example overview of the model for image retrieval on Dilbert. The images and the text contained within them are pre-encoded and ready for inference. The query is then encoded by CLIP and SBERT, and the similarity scores are aggregated.

Research Questions

2

1. Does OSBC outperform CLIP and the OCR-SBERT pipeline on image retrieval and classification?
2. Does OSBC maintain zero-shot generalizability overtasks and datasets?
3. Does OSBC maintain finetunability?
4. Do the results hold with larger, newer CLIP versions?

Methodology

3

The aim is to represent a triplet of features: images, descriptions, and the text within images. To achieve this, we need to support CLIP with a text extraction pipeline.

For this reason, OSBC is composed of three models:

1. An **OCR** (Optical Character Recognition) model. As OCR models we chose TrOCR [2] or PyTesseract [3]. They extract the text within the image in raw natural language.
2. An **SBERT** [4] model. This model receives the extracted text from the OCR model and embeds it. SBERT is tailored for calculating sentence similarity.
3. A **CLIP** model. This model instead focuses on encoding images and descriptions. Depending on the query, CLIP can either act as an image-to-text classifier, or a text-to-image retriever.

Together, the OCR and SBERT models form the **OCR-SBERT** text extraction pipeline.

Both the OCR-SBERT pipeline and the CLIP pipeline output a similarity vector between the query, and the pre-populated search space [Fig 2]. Their similarity scores are normalized and added together.

In the case the OCR-SBERT pipeline does not extract any text, or if the similarity between the query and the embeddings computed by SBERT is lower than 70% (threshold), its predictions are ignored.

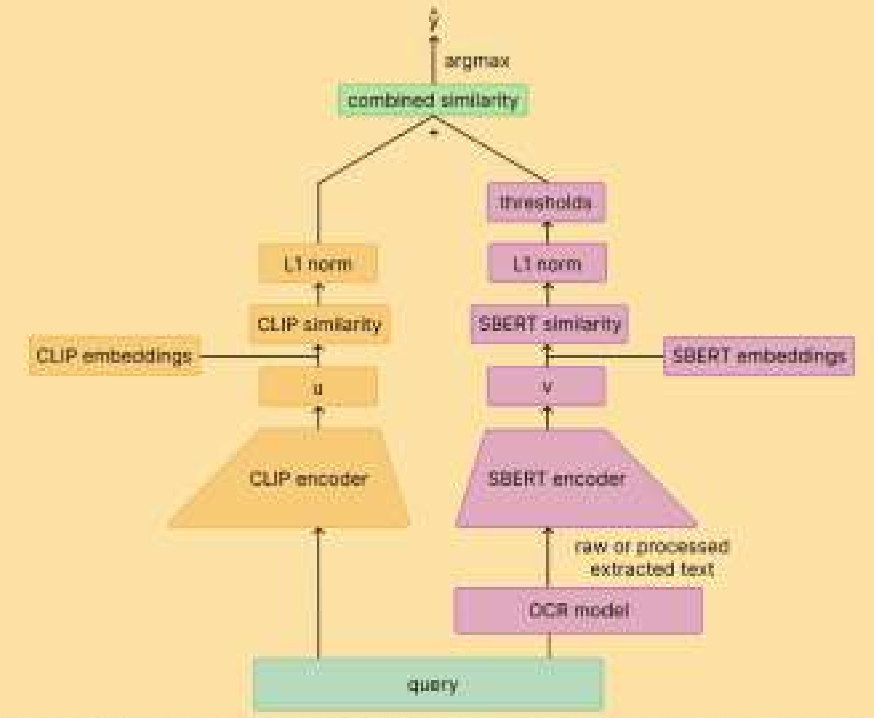

Figure 2: A representation of OSBC during inference. In purple the OCR-SBERT pipeline, in orange the CLIP pipeline. After encoding the query, similarity scores with previous embedding are normalized, post-processed in OCR-SBERT, and added together.

References

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[2] Li, Minghao, et al. "Trocr: Transformer-based optical character recognition with pre-trained models." arXiv preprint arXiv:2109.10282 (2021).

[3] https://github.com/madmaze/pytesseract

[4] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

Experiments and Results

4

We tested the model on three image classification datasets (MNIST [5], Characters from the standard OCR dataset [6], CIFAR-10 [7]) and two image retrieval datasets (Flickr8k [8], Dilbert [9]) [Fig. 3].

	ViT-B/16	ViT-B/32	ViT-L/14	-
TrOCR Printed	77.55	75.84	**79.92**	78.86
TrOCR Handwritten	54.89	54.45	70.16	49.97
PyTesseract 'psm 10'	36.45	31.87	66.06	27.92
-	29.51	24.37	71.02	-

Table 1: zero-shot classification on Characters

	ViT-B/16	ViT-B/32	ViT-L/14	-
TrOCR Printed	55.940	47.029	**78.217**	4.950
TrOCR Handwritten	56.930	47.524	**78.217**	0.0
PyTesseract 'psm 6'	66.831	62.871	75.247	65.346
-	57.425	48.019	**78.217**	-

Table 2: zero-shot retrieval on Dilbert

	ViT-B/16	ViT-B/32	ViT-L/14	-
TrOCR Printed	42.306	37.658	48.833	0.002
TrOCR Handwritten	40.704	36.412	48.229	0.007
PyTesseract 'psm 6'	42.365	37.717	48.793	1.087
-	42.494	37.865	**48.872**	-

Table 3: zero-shot retrieval on Flickr8k

In scenarios involving textless data, the OCR-SBERT pipeline is be combined in a way that is disruptive.

The datasets containing text within images instead have varied performances. The OCR-SBERT pipeline combined with CLIP through OSBC is beneficial in 11 cases, disruptive in 14, and inconsequential in 2.

The model was applied to two tasks, and five datasets, though the performance highly depends on the choice of OCR model.
The CLIP component of the architecture is successfully tuned, and the overall accuracy of the model increases [Table 6].

	ViT-B/16 - Finetuned	ViT-B/32 - Finetuned	ViT-L/14 - Finetuned	-
TrOCR Printed	96.565	97.527	97.527	90.521
TrOCR Handwritten	90.247	90.024	91.071	70.604
-	99.175	99.862	**100.00**	-

Table 4: finetuned CLIP based models on Characters

	CLIP (ViT-L/14)	OSBC (ViT-L/14, TrOCR Printed)
"an image of the letter _"	94.780	94.505
"an image of the letter: _"	66.346	87.912

Table 5: robustness to prompt engineering if we add a ":" in the prompt for Characters

There is only one combination where OSBC outperforms the ViT-L/14 image encoder based CLIP model, which could indicate that larger CLIP architectures inherently improve in OCR.
The OCR-SBERT pipeline makes the overall architecture more robust to prompt engineering changes [Table 5].

Conclusion and Limitations

5

OSBC **occasionally vastly outperformed CLIP**, especially smaller CLIP architectures, but **more often slightly underperformed** it.

The tests showed that OSBC is highly dependent on the OCR model selection, resulting in a **loss of generalizability**.

On the other hand, OSBC was **successfully partly finetuned,** and showed far **more resilience than CLIP on prompt engineering**.

Given a more general OCR method, and a more stable overall architecture, OSBC could consistently outperform CLIP on images containing text, and match CLIP when the data is textless.

Contacts

jordisassoon.github.io/site
jsassoon@student.tudelft.nl