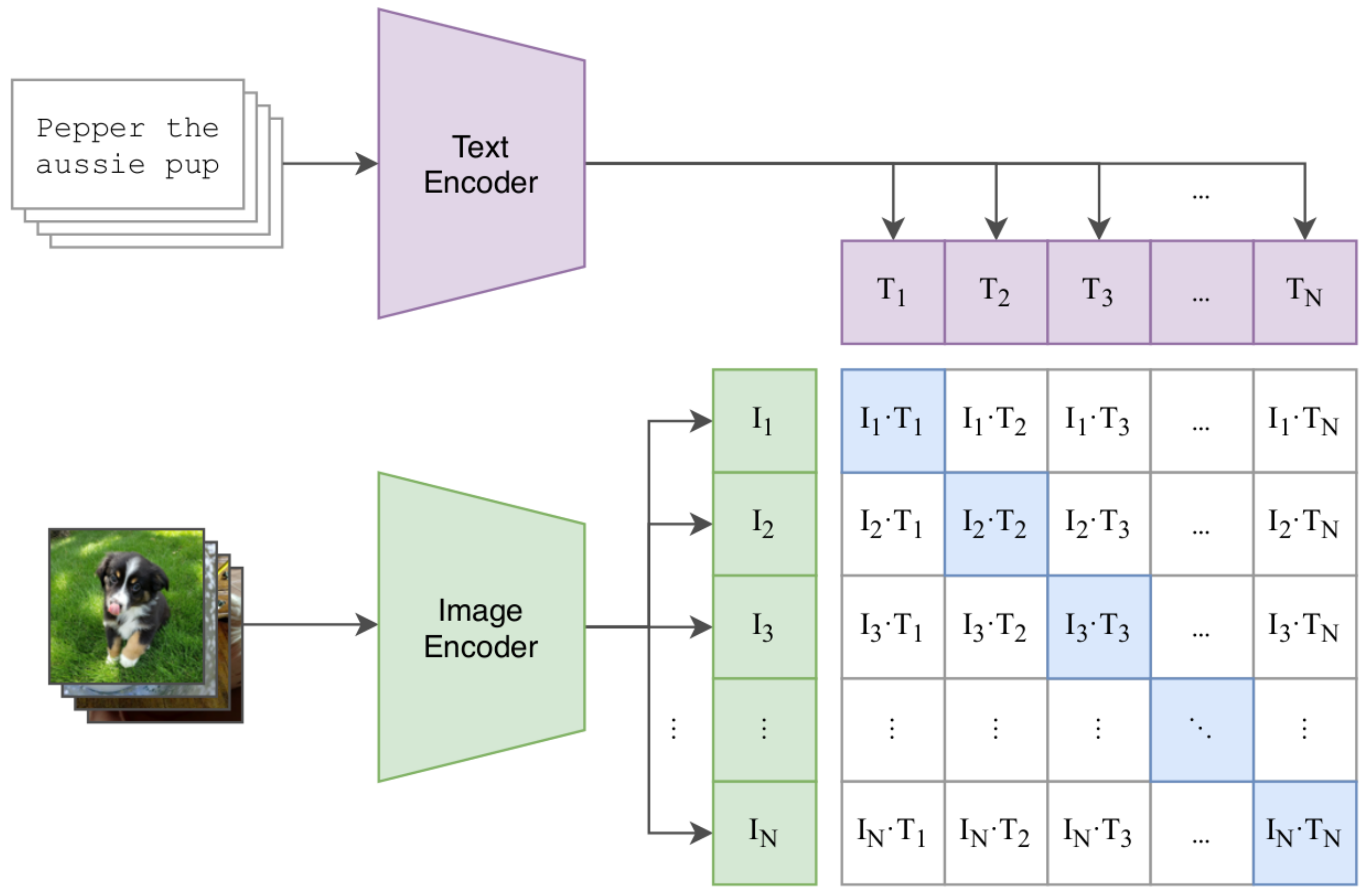
When and why vision-language models behave like bags-of-words, and what to do about it?

by Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, James Zou. [1] reviewed by Téo Guichoux, Candice Moyet, Joris Savin [2]

[1]Stanford University [2]Master DAC, Sorbonne Université

Context

Vision Language Models (**VLMs**) are very successful in various downstream tasks. However, their ability to encode **relations, order and attributes** of words can be improved. In this article, the authors propose the **ARO benchmark** (Attributions, Relations and Order) to evaluate this encoding capacity. In doing so, they highlight the *"bag-of-words"* behaviour of VLMs. Mert Yuksekgonul et al. also propose a method to improve CLIP: NegCLIP which consists of a **fine-tuning** with perturbed captions to focus learning on word relations, attributes and order.



	T_1	T_2	T_3	...	T_N
I_1	$I_1{\cdot}T_1$	$I_1{\cdot}T_2$	$I_1{\cdot}T_3$	...	$I_1{\cdot}T_N$
I_2	$I_2{\cdot}T_1$	$I_2{\cdot}T_2$	$I_2{\cdot}T_3$	...	$I_2{\cdot}T_N$
I_3	$I_3{\cdot}T_1$	$I_3{\cdot}T_2$	$I_3{\cdot}T_3$	...	$I_3{\cdot}T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N{\cdot}T_1$	$I_N{\cdot}T_2$	$I_N{\cdot}T_3$	...	$I_N{\cdot}T_N$

Figure 1: Illustration of CLIP similarity matrix

ARO benchmark

The ARO benchmark proposed by the authors is composed of 4 image-caption datasets. They are generated by **shuffling the words** of the captions of existing datasets (GQA and MS-COCO) to obtain new negative captions.

- **VGR** is the dataset that tests the understanding of **relations**: the object on the left of the relation is swapped with the one on the right.
- **VGA** is the dataset that tests the understanding of **attributes**: the adjective of object 1 is exchanged with the adjective of object 2.
- **COCO-order & Flickr-order** are the datasets that test the understanding of the **order** of words : the words are mixed in 4 different ways.

VGA

✔ *the paved road and the white house*
X *the white road and the paved house*

VGR

✔ *the person is riding the motorcycle*
X *the motorcycle is riding the person*

COCO - Order

✔ a brown cat is looking at a gray dog and sitting in a white bathtub
X (shuffle adjective/noun) a gray bathtub is looking at a white cat and sitting in a brown dog
X (shuffle all but adjective/noun) at brown cat a in looking a gray dog sitting is and a white bathtub
X (shuffle words within trigrams) cat brown a at is looking a gray dog in and sitting bathtub a white
X (shuffle trigrams) a brown cat a white bathtub is looking at a gray dog and sitting in

Figure 2: Example of caption perturbation

The test of the models on this benchmark consists in **choosing the right caption**, among 2 captions for VGA and VGR and among 4 for COCO-order.

NegCLIP

The proposed model is a fine-tuning of CLIP focused on two elements:

1. **Strong alternative images:** for each image, the **most similar image** in the whole dataset is added to the batch with its associated caption. The aim is to train the model to better discriminate between similar images and captions.
2. **Negative captions :** For each caption, a **shuffled version of the text** is added to the batch. This teaches the model to distinguish the original captions from the perturbed ones, and thus to encode the composition and order of the words in the captions.

As shown in figure 3, the **image-captions similarity matrix** is constructed as for CLIP training:

$$\text{logit\_scale} \odot (\text{encoded}_{\text{image}} \bullet \text{encoded}_{\text{text}})$$

Then the loss is computed with two **cross-entropy**, the row-wise loss is computed from the whole matrix and the column-wise from the matrix without the permuted captions (because they don't have associated images).

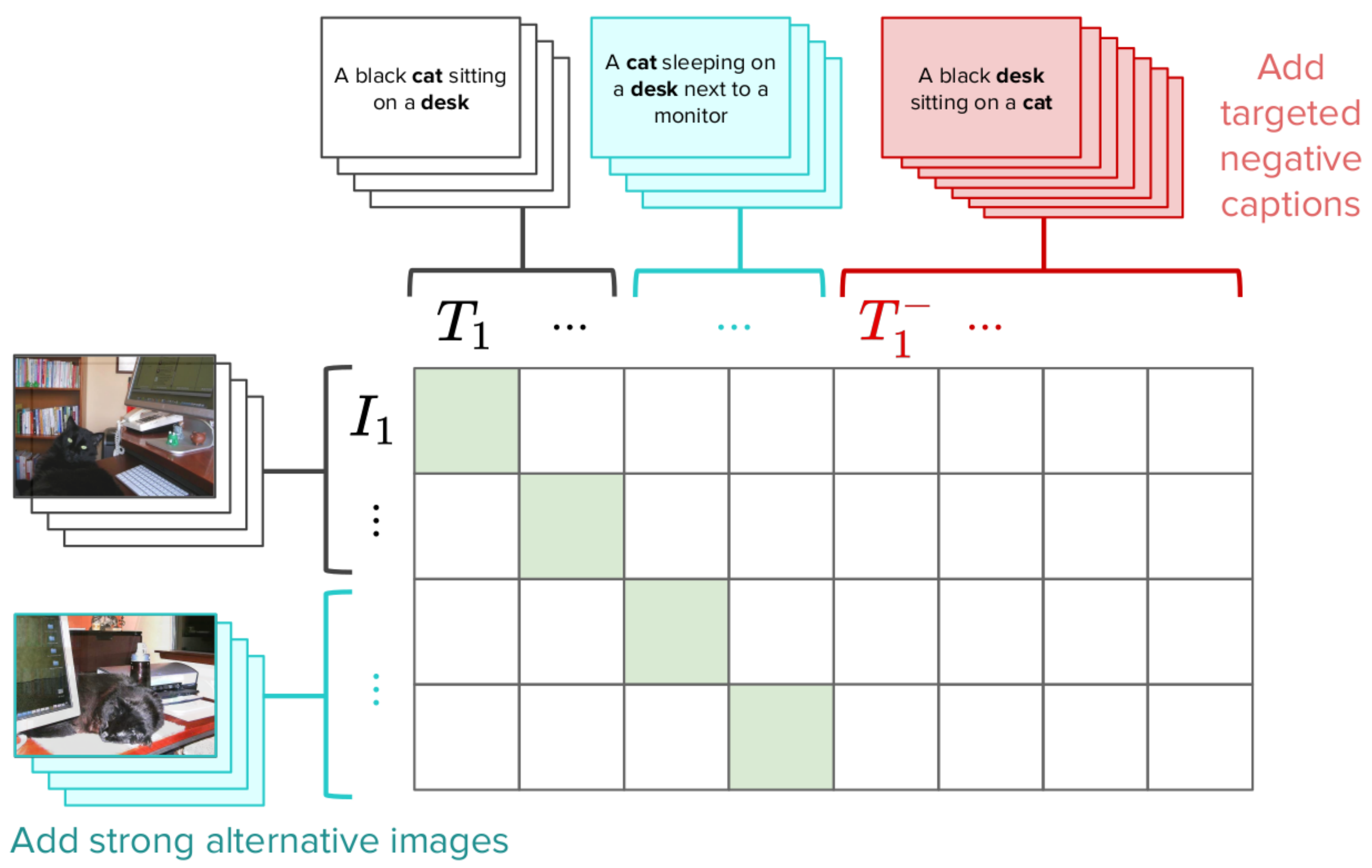

A black cat sitting on a desk
A cat sleeping on a desk next to a monitor
A black desk sitting on a cat
Add targeted negative captions
T_1
...
...
T_1^-
...
I_1
⋮
⋮
Add strong alternative images

Figure 3: Illustration of NegCLIP similarity matrix

Comparison of our results

	Theirs		Ours			
	CLIP	NegCLIP	CLIP	OurCLIP	NegCLIP	NegCLIP_FTXT
VGR	0.63	0.81(+0.18)	0.52	0.52	0.65(+0.13)	0.64(+0.12)
VGA	0.62	0.71(+0.09)	0.62	0.62	0.79(+0.17)	0.75(+0.13)
COCO-ord	0.46	0.86(+0.40)	0.50	0.50	0.83(+0.33)	0.82(+0.32)
CIFAR100	0.80	0.79(-0.01)	0.62	0.62	0.49(-0.13)	0.38(-0.24)

Table 1: Accuracy measures on ARO and CIFAR100

Values in parenthesis are performance difference between NegCLIP and CLIP

- **Similar performance** to authors with CLIP and NegCLIP on ARO
- **Increase in performances** on ARO after NegCLIP fine-tuning
- **Lower performance** on CIFAR100 (retrieval benchmark) with our implementation of NegCLIP

Additional study: Contrastive loss function on text

New fine-tuning : NegCLIP_FTXT

Based on the assumption that encoded images play no role in the encoding of the caption composition and word ordering, we propose a method derived from NegCLIP to fine-tune CLIP (NegCLIP_FTXT), which would allow **better discrimination between captions** and their permuted version. To do so, we construct the caption-caption similarity matrix (see figure 4) in addition to the NegCLIP image-caption similarity matrix. We then compute a cross-entropy on each of the matrices.

Similarity matrices

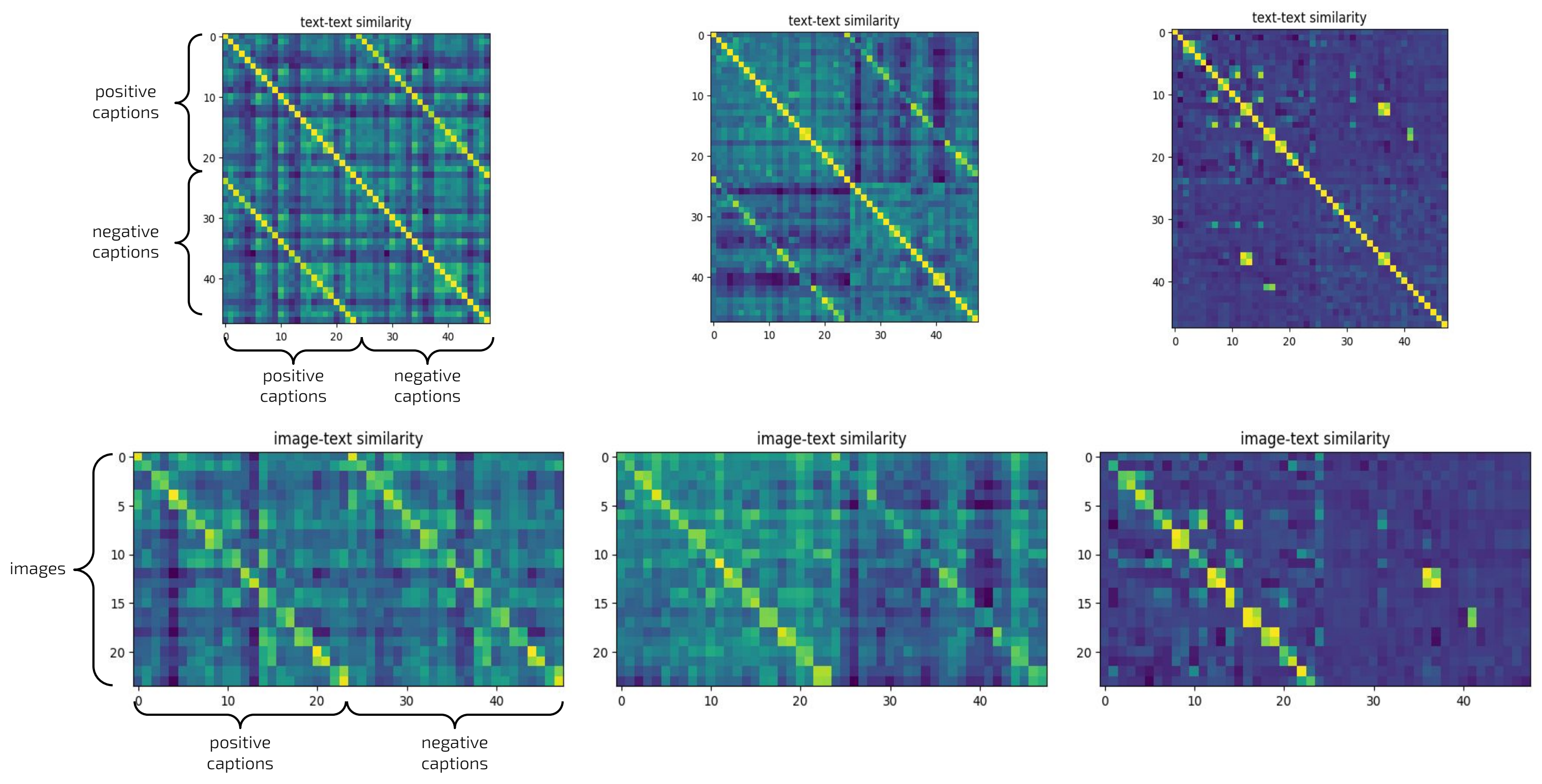


Figure 4: Similarity matrices of CLIP (left), NegCLIP (center) and NegCLIP_FTXT (right)

The diagonals between positive captions, images and negative captions fade with the fine tuning proposed by NegCLIP and disappear almost completely with that of NegCLIP_FTXT. This translates a **lower positive-negative similarity** and thus a **a shift away from their encoding**.
Our results on ARO and CIFAR100 with NegCLIP_FTXT are shown in the table 1.

Critical analysis

- **Results critical analysis :** We obtain **comparable performance** with NegCLIP on ARO, but at the expense of retrieval performance. We believe that this **drop in performance** is related to the size of the batches used (20), which is much smaller than the one proposed in the article (1024).
- **Paper critical analysis :**
  - **Pros** This article addresses a **well-posed problem** and proposes an evaluation method to identify it, as well as an **efficient model** to remedy it. It also presents many interesting and detailed experiences.
  - **Cons** The authors only study the impact of permutations on captions and not on images. There is also a missing ablation study where the disrupted captions are removed.
- **Possible improvements :** Similar to NegCLIP, it would be possible to use a **contrastive loss on the similarity matrices between images** to move encodings further away from similar images (strong alternatives).

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al.
Learning transferable visual models from natural language supervision.
In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[2] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou.
When and why vision-language models behave like bag-of-words models, and what to do about it?
arXiv preprint arXiv:2210.01936, 2022.