From CLIP to FLIP: A Case Study on Interpretability

Rochelle Choenni[1], Simone Alberto Peirone[2], Roman Pflugfelder[3,7],
Giuseppe Cartella[4], Amir Hamza[5], Tingyu Qu[6]

[1]University of Amsterdam, [2]Politecnico di Torino, [3]TU Munich,
[4]University of Modena and Reggio Emilia, [5]Fondazione Bruno Kessler, [6]KU Leuven, [7]Technion

1. Introduction

Deep neural networks have proven their usefulness in a range of applications from computer vision, NLP, and augmented reality to robotics and health. Recently, the vision-and-language paradigm has been gaining much popularity in the field of computer vision. Large vision-and-language models such as CLIP [10], have already shown to learn meaningful representations from multi-modal sources of data, and have successfully been applied to several downstream tasks from image captioning, and object detection to semantic and instance segmentation.

Yet, the training procedure for CLIP [10] is expensive, hence much effort has been put towards developing efficient techniques for scaling up the pretraining process. As such, FLIP [8] was proposed that uses random masking of image patches during pretraining to speed up the process while retaining performance on the downstream tasks.

At the same time, developing models that guarantee both faithfulness and interpretability has become a widely discussed research topic [1]. As increasingly more importance is placed on producing interpretable models, this raises an important question about the trade-off between efficiency and interpretability.

In this project we study whether an efficiency/interpretability trade-off can emerge by taking the CLIP and FLIP models as a case study. In particular, we study the effect that scaling language-image pretraining via masking, as is performed in FLIP, has on the inherent interpretability of the model.

To study this, we enhance the ViT architecture [5] with the B-Cos transform method [2] to force an alignment between the model weights and task-relevant input patterns during optimization. As a result, these transforms are explicitly designed to yield explanations that highlight task-relevant neuron activation patterns in the input. By alternating the amount of masking used during pretraining, starting from 0% (i.e, CLIP) and scaling up to 75% (i.e., FLIP), we systematically test how the interpretability of our explanations changes as an effect of masking during pretraining. We hypothesize that randomly masking patching during training would negatively affect the interpretability.

2. Methodology

Following [2], we replace the output projection in the self attention module and the linear layer in the feedforward module in the conventional ViT with the B-cos layer as:

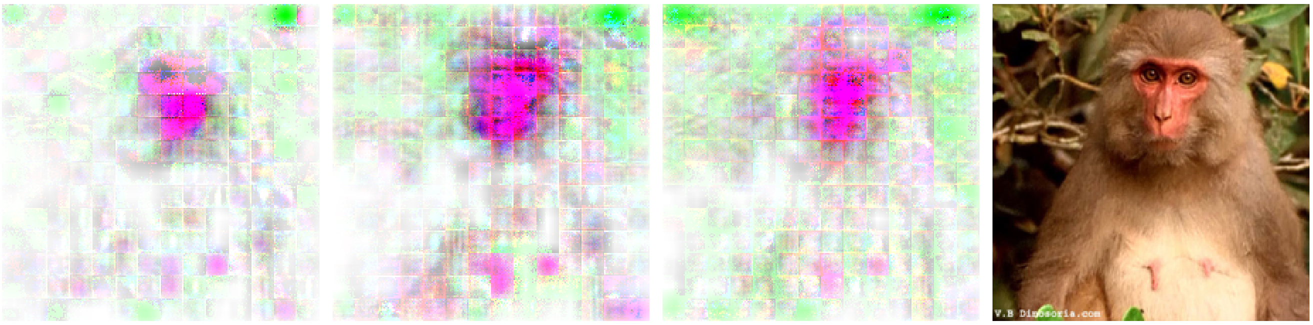
$$\text{B-cos}(x; w) = \hat{w}^T x \times cos(x, \hat{w})^{B-1} \tag{1}$$

where $\hat{w}$ scales w to unit norm, B is the scaler.
Instead of using the original implementation of FLIP written entirely in JAX[1], we use the Pytorch implementation V2.20.0 of OpenCLIP[2]. This version of OpenCLIP allows as FLIP to using masking during the training and let us replace the ViT network of OpenCLIP with the new B-cos ViT [2]. As provided by Böhle[3], the ViT B/16 architecture (entrypoint: `simple_vit_b_patch16_224`) is used for all our experiments.
We train our models for 32 epochs on CC3M [3] with learning rate of 5e-4 using AdamW optimizer [9]. All models are trained on 16×A100 GPUs with a batch size of 4096. We warmup the training for 2000 steps, and apply weight decay equal to 0.2. For the B-cos network, we adopt B=2 as in [2]. We report the zero-shot accuracy on ImageNet-1K [4].
To test and then compare the interpretability of B-cos CLIP and B-cos FLIP, the following experiments are conducted: 1) **Baseline:** Train vanilla CLIP and B-cos CLIP on the CC3M dataset. Report and compare then zero-shot accuracy of both models on the ImageNet dataset. Both models should perform similarly. 2) **FLIP-50:** Train CLIP and B-cos CLIP with 50 % dropout (masking) on the CC3M dataset respectively. Report and compare zero-shot accuracy as for the baseline experiment. Report results on the interpretability of the networks. 3) **FLIP-75:** As the FLIP-50 experiment, except using 75 % dropout.

3. Experimental Results

Classification Performance We report the zero-shot accuracy of our model on ImageNet-1k classification in Table 1. The classification results are less satisfactory due to limited amount of training data, which is also observed

[1] https://github.com/facebookresearch/flip
[2] https://github.com/mlfoundations/open_clip
[3] https://github.com/B-cos/B-cos-v2



(a) 0% dropout (CLIP)

(b) 50% dropout (FLIP)

(c) 75% dropout (FLIP)

(d) Original image

Figure 1. The effect of random masking on the interpretability of the explanations for 0 to 75% masking during pretraining on a *correctly* classified image with the B-cos variant of CLIP/FLIP.

Model	Acc@1 (%)	Acc@5 (%)
CLIP	13.01	27.50
FLIP-50	12.21	26.25
FLIP-75	10.14	23.52
CLIP w/ B-cos (ours)	11.60	25.44
FLIP-50 w/ B-cos (ours)	10.18	23.02
FLIP-75 w/ B-cos (ours)	8.28	19.68

Table 1. Zero-shot results on ImageNet-1K reported as top-1 and top-5 accuracies. 50 and 75 indicate the applied masking dropout.

in [6]. By masking 50% of the patches as in FLIP-50, we observe slight drops in performance as compared to the conventional CLIP. By replacing the conventional ViT in CLIP/FLIP with the B-cos variant, the zero-shot accuracies in all scenarios decrease. However, we gain the interpretability of the learned representations, which we further elaborate in the next paragraph.

Qualitative analysis of the explanations To qualitatively analyze and compare the explanations produced by our models we constrain ourselves to only evaluating explanations for images that were *correctly* classified by all three models. Given the low performance of our models, this leaves us with only a limited amount of viable images for analysis. In Figure 1, we show the effect of random masking on the explanations. For this particular example, we do see that, in line with our hypothesis, using 75% dropout reduces the interpretability of the explanation. Yet, we do not believe that based on our own judgement of a few input images we can draw any hard conclusions. In Section 4, we outline a few of the lessons learned from the project, and highlight the limitations of our study. Moreover, we propose some ideas that, given the time constraints, we leave for future work.

4. Conclusion and future work

While we found some preliminary evidence for the hypothesis that random masking during pretraining can reduce the inherent interpretability of language-vision models, there are a few limiting factors that do not allow us to draw any hard conclusions from these results:

1. The accuracy of all three models remains relatively low which leaves us with relatively little correctly predicted images for analysis.

2. The low performance on all models suggests that we do not have a sufficient amount of data for pretraining. FLIP especially seems to suffer from this, making the comparison somewhat skewed.

3. Evaluating the explanations is not trivial. Comparing the interpretability of the explanations amongst models seems to leave much room for subjectivity. In order to more concretely study systematic shifts in the explanations, we would need to come up with a quantitative measure for analysis.

Following up on 3., in future work, we suggest to use a similarity metric, e.g., Jaccard Similarity [7], that allows us to compute similarity scores between matrices. We could use this metric to compare how much, on average, our explanations change across the models. Note that while this would not tell us *how* the explanations change (i.e. for better or for worse), it would allow us to properly test to what extent the explanations do really change as an effect of masking during pretraining.

References

[1] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *CVPR*, 2022. 1

[2] Moritz Böhle, Mario Fritz, and Bernt Schiele. Holistically explainable vision transformers, 2023. 1

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1

[6] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023. 2

[7] GI Ivchenko and SA Honov. On the jaccard similarity test. *Journal of Mathematical Sciences*, 88:789–794, 1998. 2

[8] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023. 1

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

A. Project Team

From left to right: Giuseppe Cartella, Tingyu Qu, Roman Pflugfelder, Rita Cucchiara (scientific organiser), Amir Hamza, Simone Alberto Peirone, Rochelle Choenni