CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages

Gabriel Oliveira dos Santos[1*], Diego A. B. Moreira[1*], Alef Iury Ferreira[2],
Jhessica Silva[1], Luiz Pereira[1], Pedro Bueno[1], Thiago Sousa[2], Helena Maia[1],
Nádia da Silva[2], Esther Colombini[1], Helio Pedrini[1], Sandra Avila[1]

[1]Instituto de Computação, Universidade Estadual de Campinas (UNICAMP), Brasil
[2]Instituto de Informática, Universidade Federal de Goiás (UFG), Brasil

Abstract

This work introduces CAPIVARA, a cost-efficient framework designed to enhance the performance of multilingual CLIP models in low-resource languages. While CLIP has excelled in zero-shot vision-language tasks, the resource-intensive nature of model training remains challenging. Many datasets lack linguistic diversity, featuring solely English descriptions for images. CAPIVARA addresses this by augmenting text data using image captioning and machine translation to generate multiple synthetic captions in low-resource languages. We optimize the training pipeline with LiT, LoRA, and gradient checkpointing to alleviate the computational cost. Through extensive experiments, CAPIVARA emerges as state of the art in zero-shot tasks involving images and Portuguese texts. We show the potential for significant improvements in other low-resource languages, achieved by fine-tuning the pre-trained multilingual CLIP using CAPIVARA on a single GPU for 2 hours. Our model and code is available at https://github.com/hiaac-nlp/CAPIVARA.

1 Introduction

The challenge of learning a joint multimodal representation for vision and language has developed various pre-trained models in recent years (Wang et al., 2021; Gao et al., 2021; Yang et al., 2022b; Geng et al., 2022; Li et al., 2023). Remarkably, CLIP (Radford et al., 2021) has gained attention for achieving state of the art on zero-shot vision-language tasks through contrastive learning to align images and text within a multimodal embedding.

Training models such as CLIP requires massive data and computational resources despite their good generalization capacity. These models are

Figure 1: Improving multilingual CLIP Performance in Low-Resource Languages: Xhosa, Hindi, and Portuguese. This figure illustrates CAPIVARA's effectiveness in enhancing the performance of pre-trained multilingual CLIP models, the OPEN-CLIP baseline (B), for low-resource languages. The percentage point increase in mean recall for text-to-image (txt2img) and image-to-text (img2txt) retrieval with low-resource languages on Flickr30k and MS COCO datasets is highlighted above the respective bars. CAPIVARA significantly improves the model's baseline performance with only 2 hours of training and 8.5 GB of GPU memory.

typically trained with datasets containing hundreds of millions of image-text pairs, often collected from the web. However, many datasets only provide images paired with English descriptions; as a result, the research community focuses excessively on English texts, whereas other languages are neglected, reinforcing cultural, regional, and linguistic biases (Bender et al., 2021). While recent advancements include approaches for languages beyond English (Bianchi et al., 2021; Yang et al., 2022a; Ko and Gu, 2022) and multilingual methods (Carlsson et al., 2022; Chen et al., 2023), they primarily focus on high-resource languages. There is a scarcity of approaches considering low-resource languages, and even models including them show performance disparities in tasks involving these languages compared to tasks with English texts.

---

*Equal contribution. Corresponding authors: G.O.S. (gabriel.santos@ic.unicamp.br), D.A.B.M. (diego.moreira@ic.unicamp.br) and S.A. (avila@unicamp.br).
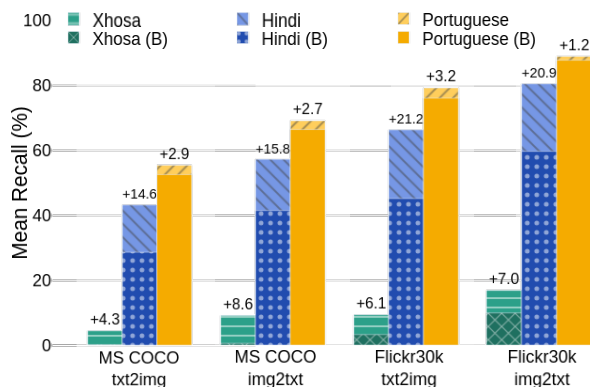
We propose a cost-efficient approach for improving multilingual CLIP performance in low-resource languages (CAPIVARA), addressing the performance gap with English and reducing computational requirements. Our approach relies on the assumption that datasets may contain images annotated with noisy descriptions. In this way, our framework utilizes BLIP2 (Li et al., 2023) to generate multiple synthetic captions for each image, addressing noisy annotations and limited language diversity challenges. Using the re-annotated dataset, we translate both the original and generated captions into the target language and conduct fine-tuning on the multilingual model. To mitigate the computational cost associated with CLIP model training, we propose to optimize the training pipeline with LiT strategy (Zhai et al., 2022), wherein the image encoder remains frozen during training, gradient checkpointing (Chen et al., 2016) and LoRA (Hu et al., 2021). Figure 1 demonstrates that substantial improvements in low-resource language can be achieved by fine-tuning the pre-trained multilingual CLIP with CAPIVARA.

Our main contributions are as follows:

- We introduce CAPIVARA, a low-cost data-centric framework that leverages image captioning models to enhance the annotation of existing datasets to improve the performance of pre-trained multilingual CLIP in low-resource languages. We report the carbon footprint of our method.

- To the best of our knowledge, we are the first to employ LoRA for language adaptation in CLIP models, considerably reducing the number of trainable parameters.

- We show that augmenting text data, by generating multiple image-conditioned captions with image captioning models, can boost CLIP performance in low-resource language.

- We achieve state of the art in many zero-shot tasks involving images and Portuguese texts. This work aims to push forward the multimodal learning literature in the Portuguese-speaking community[1].

- We make available the re-annotated CC3M with descriptions in Portuguese and English

for seamless utilization by other researchers as a data augmentation resource. We also provide the annotations translated to Portuguese for Flickr30k, MS COCO, CC3M, ImageNet-1k, and ELEVATER datasets.

2 Related Work

CLIP. The multimodal vision and language model known as CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) rapidly gained attention for its simplicity, scalability, and impressive results. It is pre-trained on 400 million image-text pairs to learn a contrastive representation of images and texts in a multimodal space.

OpenCLIP (Ilharco et al., 2021) is an open-source initiative that provides CLIP models trained on large datasets. It offers well-trained and robust models for pre-training purposes. Based on the original CLIP architecture, OpenCLIP maintains similar accuracy when trained on the same dataset. However, it extends its training to datasets like LAION-400M, LAION-2B, and DataComp-1B. Unlike the original CLIP, OpenCLIP introduces various image and text encoder configurations, including the OPENCLIP VIT-B/32 XLM-ROBERTA BASE used in this work.

Non-English CLIPs. Bianchi et al. (2021) introduce the first non-English CLIP-based models. The Italian CLIP model, unlike the original CLIP model, is trained using networks previously pre-trained in text and image tasks. It employs 1.4 million samples from translated datasets.

The Chinese CLIP (Yang et al., 2022a) explores different training approaches. The most effective architecture combines a pre-trained model with the LiT (Locked-image text Tuning) strategy (Zhai et al., 2022), freezing the text encoder until stability and extensive parameter training. Training data comprises 200 million image-text pairs.

The Korean CLIP (KELIP) model (Ko and Gu, 2022) focuses on training from scratch using substantial data and language-specific techniques. It involves self-supervised pre-training of the image encoder and alignment with the English CLIP version. The training dataset comprises 1.1 billion examples, including 708 million Korean samples.

Multilingual CLIPs. M-CLIP (Multilingual CLIP) (Carlsson et al., 2022) builds on the pre-trained CLIP model, using its text encoder while discarding the visual encoder. It employs a teacher-learning technique to transfer knowledge

---

[1]Portuguese, despite being ranked fifth among world languages in the number of native speakers, is a low-resource language from a machine-learning perspective.

from a pre-trained teacher network to new language models. M-CLIP is applied to 68 languages, translated versions of datasets by the MarianMT model (Junczys-Dowmunt et al., 2018).

AltCLIP (Altering the Language Encoder in CLIP) (Chen et al., 2023) introduces a bilingual model for Chinese and a multilingual one for 11 languages. Like M-CLIP, the teacher-learning technique uses only the textual model across various languages. However, AltCLIP differs by incorporating English text distillation, human-curated translations, and a final fine-tuning phase. It also uses the LiT strategy to freeze the image encoder.

Data-Centric Approaches. Multimodal learning has been mainly explored through algorithmic designs, often treating datasets as monolithic. Santurkar et al. (2023) reveal that CLIP's performance depends on three pre-training datasets properties: dataset size, caption descriptiveness, and caption variability for each image. They employ BLIP (Bootstrapping Language-Image Pre-training) (Li et al., 2022b) to generate new captions to address limited text diversity, improving CLIP performance. Similarly, Fan et al. (2023) propose LaCLIP (Language augmented CLIP) that uses LLM (Large Language Model) to rewrite captions to increase the text diversity within text-image pairs in the pre-training dataset. However, the decoupled text-generation process might limit effectiveness in datasets with non-descriptive captions (Nguyen et al., 2023).

Our work is related to Fan et al. (2023) and Nguyen et al. (2023). However, their studies focus on English captions during training and require extensive computational resources. In contrast, our research addresses a constrained scenario with limited computational power — a single GPU — and a lack of annotated datasets in the target language. We leverage multilingual OpenCLIP and English-annotated open datasets to enhance model performance in Portuguese. Our method, centered on Portuguese-translated captions, can be extended to other languages, making it well-suited for low-resource language challenges.

3   Method

This section details our approach, including generating captions, translating them into Portuguese, and integrating these new captions into the training pipeline. It also describes optimization through LoRA and gradient checkpointing, effectively reducing the computational resources for CLIP model training. Figure 2 illustrates the main components of CAPIVARA.

3.1   Model Architecture

We use the pre-trained multilingual model OPENCLIP VIT-B/32 XLM-ROBERTA BASE[2] (OPENCLIP for short). This model utilizes XLM-RoBERTa Base (Conneau et al., 2020) and ViT Base (Dosovitskiy et al., 2020) with 32×32 resolution as text and image encoder, respectively. The model was pre-trained on LAION-5B (Schuhmann et al., 2022) for 12.8B steps and a batch size of 90k. We employ base versions of the encoders, as larger models would demand significantly greater computational resources for both training and inference. This consideration is crucial when addressing the low-resource language community.

3.2   Datasets

We use CC3M (Sharma et al., 2018) and modifications over it to fine-tune the OPENCLIP model to improve its performance in Portuguese. For zero-shot text-to-image and image-to-text retrieval tasks, we use PraCegoVer (dos Santos et al., 2022), which is composed of images annotated originally with Portuguese texts, and our Portuguese-translated versions of MS COCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2017). We also translate the labels from ImageNet (Deng et al., 2009) and the ELEVATER benchmark datasets (Li et al., 2022a) for image classification.

3.3   Dataset Filtering

Similar to Schuhmann et al. (2022); Gadre et al. (2023), we apply CLIP score filtering. Thus, we discard examples where the cosine similarity, computed by OPENCLIP VIT-B/32 XLM-ROBERTA BASE, between the image and text embeddings is lower than 0.20. We employ this method to CC3M, naming the resulting dataset as CC3M-Filtered. We also apply this method to PraCegoVer[3], used as a test set, to remove unrelated image-text pairs.

3.4   Dataset Re-annotation & Translation

CLIP is a framework based on contrastive learning to train a multimodal model. In its pipeline, a large batch of image-text pairs (x_I, x_T) is sampled at each training step. Then, the image and
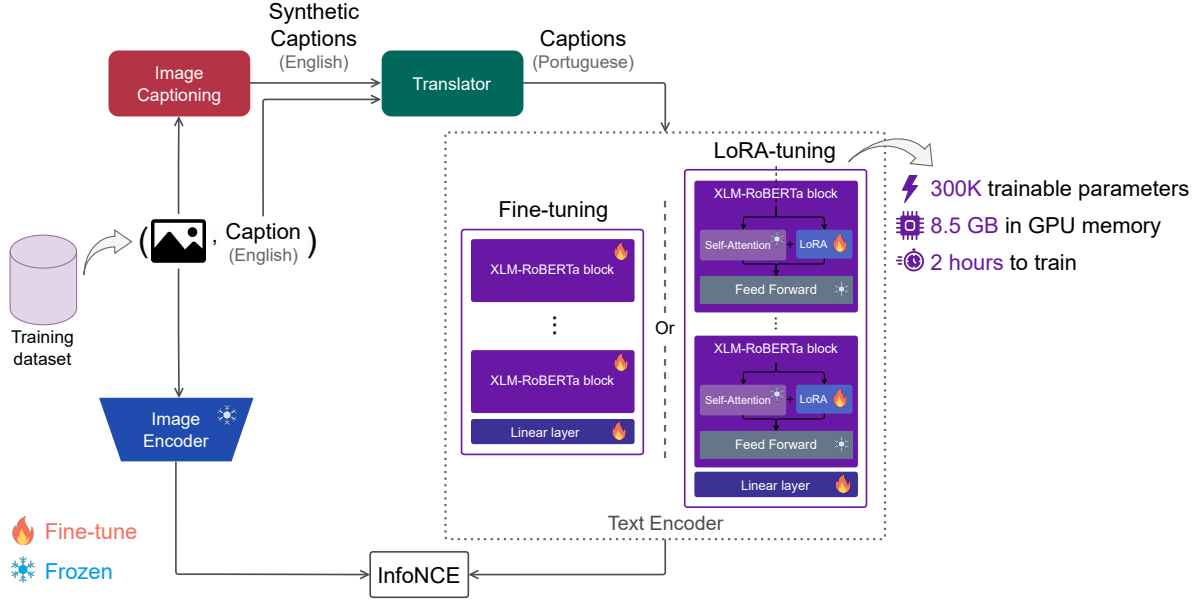

Figure 2: CAPIVARA overview. In our framework, the training dataset comprehends images annotated with English captions. To enhance the annotations, we use an image captioning model to generate synthetic captions for the images. Then, both original and synthetic captions are translated from English to the target language, in our case, Portuguese. We freeze the image encoder and fine-tune the text encoder using the translated captions to align the visual representation by optimizing the InfoNCE loss. While it is possible to fine-tune the entire text encoder, such an approach is resource-intensive. Thus, we propose an optimization method based on LoRA-tuning that can significantly reduce the associated computational cost and speed up the training time.

text features are extracted by the respective encoders f_T and f_I and are used to compute InfoNCE loss (Oord et al., 2018) as follows:

$$L_{\text{InfoNCE}}(x, y) = -\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(x^i, y^i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(x^i, y^j))/\tau}, \quad (1)$$

$$L_{\text{CLIP}} = L_{\text{InfoNCE}}(f_I(\text{aug}(x_I)), f_T(x_T)), \quad (2)$$

where B is the batch size, τ is a learnable temperature to scale the logits, $\text{sim}(\cdot)$ and $\text{aug}(\cdot)$ stands for cosine similarity and augmentation operation, respectively.

In the original proposal, only images are augmented as indicated in Equation 2, which might limit the text guidance to the image encoder. Fan et al. (2023) propose to use LLM to augment texts in addition to the image augmentation, as shown in Equation 3. However, this text-generation process does not consider the image content.

$$L_{\text{text aug}} = L_{\text{InfoNCE}}(f_I(\text{aug}(x_I)), f_T(\text{aug}(x_T))). \quad (3)$$

We propose to use BLIP2[4] to generate new captions conditioned on the images from CC3M. In

contrast to Nguyen et al. (2023), and drawing inspiration from LaCLIP (Fan et al., 2023), we propose to generate multiple captions for each image in the dataset by passing different prefixes to BLIP2. Our approach addresses the limitation of LaCLIP and has the advantage of generating multiple captions per image, which is a drawback of Nguyen et al. (2023). Still, as BLIP2 is a monolingual model, we decided to generate the captions in English and then translate them into Portuguese using Google Translate[5]. Therefore, our text augmentation comprehends generating English captions with BLIP2 and translating them into Portuguese. During training, for each image, we randomly sample a caption among the original and the generated ones to fine-tune the text encoder. Hence, at each epoch, a different text can be selected for each image. For evaluation, we translate the annotations from Flickr30k and MS COCO, and the labels from ImageNet and ELEVATER.

3.5 Training

This work takes place within the context of limited computational resources. We apply many techniques to reduce the cost of fine-tuning the OPEN-

CLIP. First, we use Gradient Checkpointing (Chen et al., 2016), which reduces the memory usage to $O(\sqrt{n})$ when training n layers. This method removes the layers' activation after the forward pass and recalculates them during the backward pass if necessary. Using this technique, we achieved a considerable reduction in GPU memory usage.

Another method contributing to memory reduction is LiT (Zhai et al., 2022), which only trains the text encoder while keeping the image encoder frozen. The motivation for training only the text encoder is that the image encoder has already undergone extensive pre-training and can produce good representations for images. Hence, we train the text encoder with captions in Portuguese so that this model learns to align the text embeddings to fixed image features, producing a multimodal embedding space. This strategy speeds up training and reduces memory since the image encoder does not compute gradients.

Finally, we also apply LoRA (Hu et al., 2021) to reduce the number of trainable parameters, reducing the memory needed to train the models and the training time. LoRA involves a re-parameterization of the dense layers as follows:

$$h = W_o x + \frac{\alpha}{r} B A x, \qquad (4)$$

where $W_o \in \mathbb{R}^{d_1 \times d_2}$ is the frozen pre-trained weight matrix, h is the result of the re-parameterization, $A \in \mathbb{R}^{r \times d_2}$ and $B \in \mathbb{R}^{d_1 \times r}$ are decomposition matrices and $r < \min(d_1, d_2)$ is the low-dimensional rank of the decomposition, an α is a hyperparameter for scale. Similar to Hu et al. (2021), we use LoRA in the query (Q) and value (V) self-attention modules from the text encoder.

The original OPENCLIP consists of 366M parameters. Applying LiT strategies reduces this number to 88M trainable parameters (24% of the total). Further integration of LoRA reduces the trainable parameters to only 0.1% (300k). We report all the training hyperparameters in Appendix A.1.

3.6 Evaluation

To evaluate the proposed framework's generalization capacity, we follow the typical procedure of evaluating pre-trained models (Radford et al., 2021; Yang et al., 2022a; Ko and Gu, 2022) in zero-shot cross-modal retrieval (text-to-image and image-to-text retrieval) and zero-shot image classification.

Zero-shot Cross-modal Retrieval: We evaluate our methods on three cross-modal retrieval datasets: PraCegoVer, MS COCO, and Flickr30k.

PraCegoVer is a multimodal dataset with native Portuguese captions based on Instagram posts. We built upon the conventional MS COCO and Flickr30k datasets, using Google Translate to translate all captions to Portuguese. To assess cross-modal retrieval performance, we adopted the recall@K evaluation metric, where $K = \{1, 5, 10\}$, and the mean recall, representing the average recall value across the recall@K instances.

Zero-shot Image Classification: We evaluate our pre-trained models on ImageNet-1k (Deng et al., 2009) and on ELEVATER image classification toolkit (Li et al., 2022a). It contains 20 datasets designed for image classification tasks across various domains and an easy-to-use toolkit to evaluate pre-trained language-augmented visual models. To accommodate evaluation in the Portuguese language, we manually translated the labels for each dataset, as well as the templates, following the methodology outlined in (Radford et al., 2021). In the evaluation process, ImageNet-1k employs the top-1 accuracy metric. Appendix A.2 provides the specific metrics for each dataset in ELEVATER benchmark.

4 Experiments and Results

This section presents a comprehensive set of experiments designed to investigate the effects of dataset filtering and the specific influence of each module within our framework, CAPIVARA. To reduce the effects of randomness, we ran each experiment setup three times. We also focus on zero-shot tasks involving images and Portuguese texts. Since no CLIP model is publicly available for Portuguese, we adopt as baseline the pre-trained multilingual model OPENCLIP due to its state-of-the-art performance in many tasks with Portuguese captions.

Dataset Filtering & CAPIVARA. We investigate two data-centric approaches: filter the training set by selecting promising samples capable of removing noise, and annotation enhancement with our proposed framework. Using CAPIVARA, for each image in CC3M, we add 10 synthetic captions, generated with BLIP2, besides the original caption. We comprehensively analyze the impact of the dataset filtering presented in Sec. 3.3 and the effectiveness of CAPIVARA in cross-modal retrieval tasks on Flickr30k, MS COCO (with Portuguese-translated captions), and PraCegoVer datasets.

Table 1 shows the results of the text-to-image (txt2img) and image-to-text (img2txt) retrieval

tasks conducted on OPENCLIP. These results encompass models fine-tuned and trained using the CAPIVARA framework on the original CC3M dataset and its filtered version, CC3M-Filtered. In Table 1, the columns "Synth." and "Trans." indicate which settings include synthetic captions and whether or not the captions are translated.

Employing the CC3M with translated captions, fourth row in Table 1, for fine-tuning increases the mean recall score by roughly 2 percentage points (*pp.*) in text-to-image and image-to-text retrieval tasks on Flickr30k and MS COCO, compared to the baseline, OPENCLIP. However, for the PraCegoVer dataset, a decline of 1.6 *pp.* in text-to-image retrieval and a more significant drop of 9.3 *pp.* in image-to-text retrieval are observed. Comparing the fine-tuning using CC3M and CC3M-Filtered, one can note an average enhancement of 0.9 *pp.* in mean recall score for text-to-image retrieval and a 0.4 *pp.* improvement for image-to-text retrieval across all three datasets.

In addition, as an intermediate step in our architecture, we employ synthetic captions to mitigate noise in the training data. To illustrate the performance gains, we compare the results of only translating the training set and translating and generating synthetic captions (CAPIVARA), fourth and sixth rows in Table 1, respectively. For the Flickr30k dataset, we observe a 1.1 *pp.* improvement in text-to-image retrieval with synthetic captions, with no significant difference in image-to-text retrieval. On the MS COCO dataset, we note a 1.5 *pp.* increase in text-to-image retrieval and a 1.2 *pp.* gain in image-to-text retrieval. Additionally, when evaluating the PraCegoVer dataset under the same conditions, we find a 2.6 *pp.* improvement in text-to-image retrieval and a 4.7 *pp.* gain in image-to-text retrieval. Thus, in most cases, using synthetic data as a means of data augmentation and noise reduction yields a positive impact. Details about the impact of the number of synthetic captions in the performance are shown in Table A6 (Appendix A.3).

The most significant performance gains over the baseline are achieved using CAPIVARA. For instance, the model trained on CC3M with CAPIVARA, sixth row, yields a 3.5 *pp.* improvement in text-to-image retrieval for Flickr30k and MS COCO and 1 *pp.* enhancement on PraCegoVer. Notably, in image-to-text retrieval, CAPIVARA (CC3M) increases 2 *pp.* on Flickr30k and it has a remarkable 4.7 *pp.* gain on MS COCO over the baseline. Also, models trained on CC3M and CC3M-Filtered with CAPIVARA demonstrate similar performance levels. These experiments demonstrate the effectiveness of our proposal, CAPIVARA, in enhancing multilingual CLIP performance in Portuguese.

Caption Translation. We also investigate the impacts of automatic translations of captions in the final model performance for Portuguese texts. We conducted experiments training the model on datasets containing only English annotations (i.e., CC3M + no-translation and CC3M + no-translation + synthetic captions), and their counterparts translated into Portuguese using Google Translate (i.e., CC3M + translation and CC3M + translation + synthetic captions). The evaluation comprehends Flickr30k, MS COCO, and PraCegoVer datasets with only Portuguese captions, particularly images in PraCegoVer that are originally annotated in Portuguese. We present the results in Table 1.

One can note a substantial improvement when translating annotations within the training dataset. Specifically, models trained on datasets containing Portuguese annotations exhibit an average increase of 2.5 *pp.* in text-to-image mean recall scores compared to their English-trained counterparts. Similarly, employing Portuguese-translated captions leads to a mean recall improvement of 1.6 *pp.* for image-to-text retrieval on both the Flickr30k and MS COCO datasets. Fine-tuning with the original CC3M (i.e., CC3M + no-translation) hampers text-to-image performance across all three datasets and drops notable 7 *pp.* the mean recall in image-to-text on PraCegoVer. By training the model on translated synthetic captions, CAPIVARA consistently outperformed all the other settings. Our method increases the average performance in 3.2 *pp.* compared to fine-tuning on the original CC3M dataset. This experiment highlights the importance of including the automatic translation of captions into the target language, Portuguese, in our training pipeline.

Training Pipeline Optimization. This work is inserted in a context of restricted computational resources, in which only a single RTX Quadro 8000 GPU is available. In this way, we propose a method to optimize our training pipeline, detailed in Sec. 3.5. It combines LiT, Gradient Checkpointing (G. Checkpt), and LoRA techniques. In this section, we investigate the impacts of this optimization in terms of model performance and cost

Table 1: Impact analysis of synthetic captions (Synth.) and translation (Trans.) on our framework. This table compares the performance of CLIP fine-tuning on English and Portuguese-translated texts, both with and without the addition of synthetic captions. It shows the experimental results in cross-modal retrieval on Flickr30k and MS COCO with captions translated into Portuguese, and PraCegoVer. We report the average and standard deviation of mean recall for text-to-image (txt2img) and image-to-text (img2txt) retrieval tasks. Our CAPIVARA achieves the best performance across datasets, highlighting its efficacy in enhancing pre-trained multilingual CLIP.

Method/Model	Training dataset	Synth.	Trans.	Flickr30k		MS COCO		PraCegoVer	
				txt2img	img2txt	txt2img	img2txt	txt2img	img2txt
OPENCLIP (Baseline)				76.23	87.93	52.62	66.55	65.36	**69.43**
OPENCLIP + Fine-tuning	CC3M	✗	✗	75.78 ± 0.02	88.78 ± 0.04	52.28 ± 0.01	68.18 ± 0.01	61.41 ± 0.00	62.35 ± 0.01
	CC3M	✓	✗	77.08 ± 0.02	89.01 ± 0.03	53.87 ± 0.01	70.04 ± 0.02	64.01 ± 0.01	66.43 ± 0.01
	CC3M	✗	✓	78.42 ± 0.02	**90.02 ± 0.05**	54.77 ± 0.01	70.06 ± 0.01	63.79 ± 0.01	60.10 ± 0.00
	CC3M-Filtered	✗	✓	79.02 ± 0.01	89.49 ± 0.02	55.46 ± 0.01	69.52 ± 0.02	65.11 ± 0.01	62.29 ± 0.01
🦫 CAPIVARA	CC3M	✓	✓	**79.56 ± 0.01**	89.95 ± 0.04	56.27 ± 0.01	**71.24 ± 0.01**	66.40 ± 0.01	64.75 ± 0.01
	CC3M-Filtered	✓	✓	**79.67 ± 0.01**	89.97 ± 0.04	**56.32 ± 0.01**	71.06 ± 0.01	**66.55 ± 0.01**	65.06 ± 0.01

Table 2: Impact of optimization techniques. We evaluate training models on CC3M with CAPIVARA combined with many optimization techniques. We report the experimental results in terms of mean recall in text-to-image (txt2img), and image-to-text (img2txt) and memory (M) and training time cost (T). Our optimization method leads to the best training time and computational cost while performing similarly to other approaches.

	Optimization	Flickr30k		MS COCO		PraCegoVer		M (GB)	T (h)
		txt2img	img2txt	txt2img	img2txt	txt2img	img2txt		
	OPENCLIP (Baseline)	76.23	87.93	52.62	66.55	65.36	69.43	-	-
CAPIVARA	LiT + G.Checkpt	79.56 ±0.01	89.95 ±0.04	56.27 ±0.01	71.24 ±0.01	66.44 ±0.01	66.57 ±0.01	38	31
	LiT + G.Checkpt + LoRA	79.51 ±0.04	89.50 ±0.03	55.56 ±0.01	69.63 ±0.04	67.07 ±0.02	68.14 ±0.01	21.5	16
	LiT + G.Checkpt + LoRA + 1500 steps + BS=1000	79.39 ±0.05	89.13 ±0.08	55.49 ±0.06	69.26 ±0.05	66.89 ±0.04	67.93 ±0.01	8.5	2

Table 3: Summary of the models and resources invested in their training, considering the dataset size, the GPU/TPU used, and the required training time.

Model	Language	# Dataset size	GPU/TPU	Training time
Italian CLIP	Italian	1.4M	2 TPUs	14 days
Chinese CLIP	Chinese	200M	128 V100 (2048 GB)	7.5 days
Korean CLIP	Korean	708M	80 A100 (640 GB)	15.7 days
LaCLIP	English	365M	32 V100 (512 GB)	–
AltCLIP	Multilingual	38M/115M	–	–
M-CLIP	Multilingual	3.3M	–	–
CAPIVARA	**Portuguese**	**3.3M**	**1 Quadro RTX 8000 (48 GB)**	**2 hours**

formance similar to the one that fine-tunes the entire text encoder on Flickr30k, but it decreases by 1.2 pp. the average performance on MS COCO.

In addition, the model trained with our optimization technique LiT + G. Checkpt + LoRA + 1500 steps + BS=1000 presented a decline of 0.2 pp. compared to LiT + G. Checkpt + LoRA. Using our optimization method can remarkably reduce the GPU memory (from 38 GB to 8.5 GB) and training time (from 31h to 2h), yet outperform the baseline by 2.5 pp. across the tasks. Our training pipeline requires very modest computational resources compared to the literature, as shown in Table 3. These experiments demonstrate that our optimization method can effectively reduce the cost of fine-tuning CLIP, allowing researchers with restricted computing resources to conduct experiments.

Low-resource Languages. To demonstrate the effectiveness of CAPIVARA in improving pretrained multilingual CLIP performance on low-resource languages, we expand our investigation to include Xhosa and Hindi. Figure 1 compares the performance between the pre-trained OPENCLIP (baseline) and the models trained by employing the whole CAPIVARA optimized pipeline, which refers to the setting LiT + G. Checkpt + LoRA + 1500 steps + BS=1000, named CAPIVARA + Opt., for text-to-image and image-to-text retrieval tasks on Flickr30k and MS COCO. This experiment em-

reduction. All experiments include LiT and gradient checkpointing, otherwise, we could not run the training in our infrastructure. In addition, we conducted experiments to assess the impact of including LoRA in our training pipeline. To compare the computational cost among the settings, we fixed the GPU architecture, and we trained the models with batch size (BS) equal to 2816 for 5863 steps, except for LiT + G. Checkpt + LoRA + 1500 steps + BS=1000, trained with a batch size of 1000 samples for only 1500 steps. Still, we demonstrate that it is possible to reduce the batch size and the number of training steps and reach a competitive performance.

Table 2 shows experimental results. Our initial attempt to fine-tune the complete CLIP model encountered infrastructure limitations, hindering its execution. We overcame this constraint by utilizing LiT and gradient checkpointing, which enabled the training process. Comparison between the setups, namely LiT + G. Checkpt and LiT + G. Checkpt + LoRA, reveals that LoRA substantially reduces memory usage by over 40% and cuts training time in half. The model trained with LoRA had a per-

ploys our optimized training pipeline (Sec. 4), training models for 2 hours on a single GPU Quadro RTX 8000 with a memory usage of 8.5 GB.

The baseline presents the weakest performance in Xhosa across all tasks, with mean recall close to zero in MS COCO and 3 and 10 in text-to-image and image-to-text on Flickr30k, respectively. CAPIVARA increases the average performance in this language by 6.5 *pp.* on Flickr30k and MS COCO. The most significant improvement can be noted in Hindi. A remarkable increase of 15 *pp.* on MS COCO and 21 *pp.* on Flickr30k is obtained with CAPIVARA. This experiment shows that CAPIVARA effectively boosts the pre-trained multilingual CLIP's performance in other low-resource languages with a low computational cost.

Image Classification. In addition to zero-shot cross-modal retrieval tasks, we also evaluate our models in zero-shot image classification across 21 datasets. The results are presented in Table 4. In the context of ELEVATER, training CLIP with CAPIVARA yielded an average improvement of 0.6 *pp.* over our baseline. We plot the bar chart in Figure A1 to thoroughly analyze the performance gap between the baseline and the model trained with CAPIVARA for each dataset within ELEVATER. Our method consistently surpassed the baseline across most datasets, yielding substantial accuracy improvements of 5.53 *pp.*, 5.15 *pp.*, and 3.07 *pp.* for KITTI-Distance, MNIST, and GTSRB, respectively. Regarding ImageNet-1k, CAPIVARA exhibited a performance gain of 0.2 *pp.* compared to the baseline. In addition, the model's performance trained with CAPIVARA + Opt. is close to our baseline. Hence, LoRA-tuning for 1500 steps keeps the average performance on zero-shot image classification, whereas it improves considerably the performance on zero-shot cross-modal retrieval.

Carbon Footprint. Despite the remarkable achievements of large language models, their deployment requires substantial computational power, resulting in significant energy usage. For instance, models such as GPT-3 and BLOOM consumed approximately 1,287 MWh and 433 MWh, respectively, in their training, corresponding to 502 tonnes of CO_2 and 25 tonnes of CO_2 emissions (Maslej et al., 2023). The BLOOM model's carbon footprint alone surpasses an average American's annual carbon emissions by 1.4 times. The energy consumed during BLOOM's training could power a

Table 4: Zero-shot image classification performance on ELEVATER and ImageNet-1k.

Dataset	OPENCLIP (Baseline)	CAPIVARA	CAPIVARA + Opt.
Caltech-101	84.53 ±0.00	82.97 ±0.03	83.68 ±0.02
CIFAR-10	93.99 ±0.00	93.85 ±0.00	93.93 ±0.03
CIFAR-100	68.44 ±0.00	69.37 ±0.01	68.87 ±0.01
Country-211	17.82 ±0.00	17.61 ±0.00	17.32 ±0.02
DTD	41.17 ±0.00	42.34 ±0.04	41.79 ±0.07
EuroSAT	47.16 ±0.00	47.77 ±0.02	48.85 ±0.12
FER-2013	48.65 ±0.00	46.68 ±0.05	46.85 ±0.13
FGVC-Aircraft	26.30 ±0.00	25.49 ±0.01	25.54 ±0.09
Food-101	65.06 ±0.00	64.58 ±0.01	64.46 ±0.00
GTSRB	43.27 ±0.00	46.34 ±0.01	44.66 ±0.06
Hateful-Memes	56.50 ±0.00	56.17 ±0.00	56.81 ±0.03
KITTI-Distance	28.41 ±0.00	33.94 ±0.13	28.27 ±0.11
MNIST	54.99 ±0.00	60.14 ±0.04	55.00 ±0.10
Oxford Flowers-102	50.88 ±0.00	49.93 ±0.02	51.99 ±0.12
Oxford-IIIT Pets	81.56 ±0.00	79.37 ±0.00	80.90 ±0.09
PatchCamelyon	50.96 ±0.00	51.71 ±0.01	52.39 ±0.07
Rendered-SST2	54.20 ±0.00	54.82 ±0.03	52.94 ±0.04
RESISC-45	58.51 ±0.00	59.71 ±0.01	56.93 ±0.01
Stanford-Cars	84.93 ±0.00	85.10 ±0.02	84.90 ±0.06
PASCAL VOC-2007	82.09 ±0.00	82.29 ±0.00	81.99 ±0.02
Average	56.97 ±0.00	**57.51 ±0.02**	56.90 ±0.06
ImageNet-1k	45.84 ±0.00	**46.06 ±0.01**	45.65 ±0.02

Table 5: Average costs per trained model in terms of energy consumption and equivalent CO_2 emissions (CO_2-eq), compared with the number of trainable parameters (# Param.). All the models were trained with a batch size (BS) of 2816 for 5863 steps, except for CAPIVARA + LoRA + 1500 steps / BS=1000.

Model	# Param.	Energy	CO_2-eq
Gopher	280 B	1,066 MWh	352 tonnes
BLOOM	176 B	433 MWh	25 tonnes
GPT-3	175 B	1,287 MWh	502 tonnes
OPT	175 B	324 MWh	70 tonnes
CAPIVARA	278 M	6.49 kW	0.50 kg
CAPIVARA + LoRA	1.9 M	5.67 kW	0.43 kg
CAPIVARA + LoRA +1500 steps / BS=1000	1.9 M	0.22 kW	0.017 kg

household in the United States for up to 41 years.

To compare energy consumption between our model and larger language models, we employed the codecarbon tool (Courty et al., 2023). The results are shown in Table 5. As other CLIP-like models do not provide energy and carbon expenditure data, we present a comparison with other large language models for which such data is available in the literature (Maslej et al., 2023). For the baseline model, the energy usage amounted to 6.4 kW, resulting in 0.5 kg of CO_2 equivalent emissions. Applying LoRA and reducing the number of training steps decreased energy consumption to 5.6 kW and 1.8 kW, respectively, resulting in 0.4 kg and 0.1 kg of CO_2 equivalent emissions. These calculations are based on Brazil's energy mix, where hydropower is the primary energy source. This calculation does not include the carbon footprint of the initial pre-training performed by OPENCLIP,

but only the training with CAPIVARA. We aim to advance sustainable AI systems development by employing these techniques and optimizing training times.

5 Conclusion

This work demonstrates the potential challenges of fine-tuning multilingual CLIP models within low-resource languages due to noisy annotations. To address this issue, we introduce 🦫 CAPIVARA, a cost-effective framework that leverages image captioning models to enhance the dataset annotations. We conducted extensive experiments involving dataset filtering, re-annotation, and automatic translation. CAPIVARA effectively boosts OPEN-CLIP performance for Portuguese texts, achieving state-of-the-art results in many zero-shot tasks. Our findings show the importance of dataset re-annotation and automatic translation.

We also propose optimizing our training pipeline using LiT, including LoRA and gradient checkpointing. Our results show a substantial improvement in Portuguese performance by fine-tuning the pre-trained OPENCLIP in a single GPU for 2 hours, and only 8.5 GB of memory — considerably modest compared to literature. Moreover, we demonstrate that our framework is readily extensible to other low-resource languages.

A direction for future research involves investigating the scalability of the proposed approach in terms of dataset and model size, building upon its success with base models. We also plan to explore different image captioning models and text decoding methods. Due to the cost of generating synthetic captions and translating them to Portuguese, there is interest in automating the process, possibly by improving BLIP2's performance in Portuguese. Besides, due to the success of LoRA, other parameter-efficient fine-tuning can be explored. Lastly, an interesting research question remains open: "how many examples annotated in a low-resource language are necessary to achieve a performance comparable to English?".

Limitations

Model. Unlike other studies that compare models with varying architectures and sizes (Radford et al., 2021; Yang et al., 2022a; Li et al., 2022c; Mu et al., 2022), our research focuses on specific choices: the ViT-B/32 as image encoder and the XLM-Roberta Base as text encoder. Future work will explore different model sizes within our budget and consider alternative fine-tuning approaches, such as Parameter-Efficient Fine-Tuning (PEFT) (Liao et al., 2023).

Data. Recent efforts to adapt CLIP for specific languages (Ko and Gu, 2022; Yang et al., 2022a; Bianchi et al., 2021) have typically used datasets much larger than our study. Investigating scalability using training datasets could reveal the optimal trade-off between cost and performance.

Generating captions in languages such as Portuguese involves two steps: caption generation and machine translation; due to the lack of robust non-English image captioning models. Hence, future research could focus on fine-tuning image captioning models for target languages to streamline the process and improve accuracy. Our study used the BLIP2 model for caption generation, but exploring alternative models could enhance results.

An additional limitation is the prevalent use of machine-translated datasets in various multilingual datasets (Carlsson et al., 2022; Jain et al., 2021; Yang et al., 2022a; Bianchi et al., 2021). However, these datasets may not effectively capture unique expressions, cultural nuances, and proper nouns, leading to bias over-amplification, where biases from the source text become exaggerated in the translated output (Hovy and Prabhumoye, 2021; Prabhumoye et al., 2021; Hovy et al., 2020).

Ethics Statement

CAPIVARA is a cost-efficient framework designed to enhance the performance of multilingual CLIP models in low-resource languages. For this purpose, CAPIVARA augments text data using image captioning and machine translation to generate multiple synthetic captions in low-resource languages, and the training pipeline is optimized with LiT, LoRA, and gradient checkpointing to alleviate the computational cost. Intended to be used for general tasks, the model learns to represent in a joint space texts and images. It can be employed in text-to-image, image-to-text retrieval, and image classification tasks. The developed model is particularly intended for scientific researchers.

Based on known problems with image and language models, the model may present lower performance for under-represented and minority groups (Bender et al., 2021). To adapt the model to low-resource languages, we use texts translated from English; thus, the model does not represent the cul-

tural and local aspects of the countries that speak these target languages. This can lead to linguistic biases and a lack of representativeness for the target groups.

The datasets used comprehend texts from the internet and carry biases; thus, the model may perform differently for data collected from other sources. Also, the datasets may contain data with cultural, political, or religious positioning.

Furthermore, CAPIVARA does not generate any type of data that could pose a risk to human life. However, our model can be adapted for other specific tasks, e.g., image or text generation, which could contribute to generating false information and harming people. CAPIVARA is a framework that aims to improve performance for low-resource languages. However, our results show that despite the significant improvements achieved with CAPIVARA, there is still a considerable gap between the model performance with English texts and texts in low-resource languages. Further research is needed to improve performance across languages and incorporate cultural and linguistic elements into the model.

Since language models require large computational, environmental, and financial resources (Bender et al., 2021), CAPIVARA optimizes its training pipeline, resulting in a smaller carbon footprint than traditional fine-tuning. More details about ethical considerations can be found in Model Cards (Appendix A.6).

Acknowledgements

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pretraining for the italian language. *arXiv:2108.08688*. ArXiv:2108.08688 [cs].

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Language Resources and Evaluation Conference*, page 6848–6854.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023. AltCLIP: Altering the language encoder in CLIP for extended language capabilities. In *Findings of the Association for Computational Linguistics*, pages 8666–8682. Association for Computational Linguistics.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Benoit Courty, Victor Schmidt, Goyal-Kamal, Marion-Coutarel, Boris Feld, Jérémy Lecourt, SabAmine, Kngoyal, Mathilde Léval, Alexis Cruveiller, Ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Inimaz, Amine Saboni, Hugues De Lavoreille, Niko Laskaris, Edoardo Abati, LiamConnell, Douglas Blank, Ziyao Wang, Armin Catovic, Michał Stęchły, Alencon, JPW, MinervaBooks, SangamSwadiK, Christian Bauer, and M. Hervé. 2023. *mlco2/codecarbon: v2.3.0*. Zenodo.

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

Li Deng. 2012. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2022. #PraCegoVer: A Large Dataset for Image Captioning in Portuguese. *Data*, 7(2).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338.

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

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE.

Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. 2013. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems*, pages 1693–1700. IEEE.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, and Jieyu Zhang. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. 2022. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. In *The Eleventh International Conference on Learning Representations*.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, and Dong-Hyun Lee. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multi-task representations across languages. In *Findings of the Association for Computational Linguistics*, page 3449–3463. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *ACL 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Byungsoo Ko and Geonmo Gu. 2022. Large-scale bilingual language-image contrastive learning. *arXiv:2203.14463*. ArXiv:2203.14463 [cs].

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pages 554–561.

Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical report, Toronto, ON, Canada.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, and Yong Jae Lee. 2022a. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*. ArXiv:2301.12597 [cs].

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022c. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*.

Baohao Liao, Yan Meng, and Christof Monz. 2023. Parameter-efficient fine-tuning without introducing new latency. In *Annual Meeting of the Association for Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. 2023. The AI Index 2023 Annual Report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. Case study: Deontological ethics in NLP. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. 2023. Is a caption worth a thousand images? a study on representation learning. In *The Eleventh International Conference on Learning Representations*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, and Mitchell Wortsman. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned,

hypernymed, image alt-text dataset for automatic image captioning. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The german traffic sign recognition benchmark: a multi-class classification competition. In *International Joint Conference on Neural Networks*, pages 1453–1460. IEEE.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *21st International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 210–218. Springer.

Jue Wang, Haofan Wang, Jincan Deng, Weijia Wu, and Debing Zhang. 2021. Efficientclip: Efficient cross-modal pre-training by ensemble confident learning and language modeling. *arXiv preprint arXiv:2109.04699*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. LVLM-eHub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022a. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv:2211.01335*. ArXiv:2211.01335 [cs].

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022b. Vision-language pre-training with triple contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 18102–18112, New Orleans, LA, USA. IEEE.

A Appendix

Authors' Contributions

G.O.S., D.A.B.M., and A.I.F. collaborated on dataset translation, designing and implementing the proposed pipeline, analyzing the results, and writing the manuscript. G.O.S. also conducted experiments related to dataset filtering, re-annotation, translation, and low-resource languages. D.A.B.M. worked on constructing training datasets, focused on experiments to optimize the pipeline and conducted a carbon footprint analysis. A.I.F. executed inferences for zero-shot image classification. In collaboration with G.O.S., D.A.B.M., and A.I.F., J.S. wrote the Ethics Statement and Model Cards sections. L.P. helped in the result analysis. P.B. contributed to dataset translation. T.S. helped in constructing training datasets. H.M. contributed to the discussion with the team and the result analysis. N.S. advised A.I.F. and T.S. throughout all tasks. E.C. advised G.O.S. throughout all tasks. H.P. advised the team on all tasks and contributed to the writing process. S.A. served as the principal advisor of the team, providing guidance on all tasks and contributing to the writing process. All authors reviewed the manuscript and provided critical feedback to enhance its quality.

A.1 Hyperparameters

To facilitate the reproducibility of the work, we present Tables A1 and A2. These tables contain the hyperparameters used for the best models evaluated in the different experiments. Table A1 contains only the hyperparameters used in the fine-tuning of the OPENCLIP model for Portuguese. Table A2 considers the hyperparameters with the LoRA-tuning for the models with optimizations and 1500 steps, in Portuguese, Hindi and Xhosa.

Table A1: Hyperparameters used in the fine-tuning.

Hyperparameters	Value
Batch size	2816
Maximum token length	77
Optimizer	Adam
Weight decay	0.2
Adam ϵ	1e-8
Adam β	[0.9, 0.98]
Learning rate schedule	CosineWarmupLR
Maximum learning rate	5e-7
Minimum learning rate	1e-7
# Steps	5863

Table A2: Hyperparameters used in LoRA-tuning.

Hyperparameters	Value
LoRA r	8
LoRA Alpha	8
LoRA dropout	0
bias	None
Target modules	(query, value)
Modules to save	projection
Batch size	1000
Maximum token length	77
Optimizer	Adam
Weight decay	0.2
Adam ϵ	1e-8
Adam β	[0.9, 0.98]
Learning rate schedule	CosineWarmupLR
Maximum learning rate	1e-5
Minimum learning rate	1e-6
# Steps	1500

A.2 Results on ELEVATER and ImageNet-1k

In our supplementary experiments on ELEVATER and ImageNet-1k benchmarks, summarized in Table A3, we consistently observe that our approach outperforms the baseline model across various setups, with the exception of CAPIVARA + Opt. This suggests that more training steps might be necessary to fully leverage LoRA's potential in fine-tuning. Furthermore, Table A3 reveals the effect of caption generation and filtering on the efficacy of our method. By analyzing the scenarios with synthetic captions, one can note that training with multiple captions per image outperforms training on only OPENCLIP + Fine-tuning both with or without filtering. Notably, the optimal configuration involves training with CAPIVARA on CC3M-Filtered, resulting in a performance boost of 0.6 *pp.* over the baseline. Still, similar to the cross-modal retrieval in Sec. A.3.1, we do not observe a significant performance gain by augmenting the number of generated captions. Table A4 provides the specific metrics for each dataset in ELEVATER benchmark.

Figure A1 presents the difference in performance between fine-tuning with CAPIVARA and the baseline, OPENCLIP. It can be noted that the majority of datasets exhibit positive differences in performance, indicating a favorable improvement over the baseline with CAPIVARA. Notably, the model trained with CAPIVARA led to substantial improvements of 5.53 and 5.15 *pp.* in two datasets, namely KITTI-Distance and MNIST, respectively. However, it is important to acknowledge instances where the performance of our model under this configuration falls short. Noteworthy cases include the Oxford-IIIT Pets dataset, encompassing 37 distinct

Table A3: Results on ELEVATER benchmark. Ablation without LoRA and with LoRA.

Dataset	OPENCLIP (Baseline)	OPENCLIP + Fine-tuning	OPENCLIP + Fine-tuning (CC3M-Filtered)	CAPIVARA (CC3M-Filtered)	CAPIVARA	CAPIVARA + 5 synth. captions	CAPIVARA + 1 synth. caption	OPENCLIP + Fine-tuning + LoRA	CAPIVARA + LoRA	CAPIVARA + Opt.
Caltech-101	84.53 ±0.00	82.50 ±0.01	82.23 ±0.01	82.90 ±0.00	82.97 ±0.03	82.66 ±0.00	82.87 ±0.01	83.06 ±0.07	83.70 ±0.01	83.68 ±0.02
CIFAR-10	93.99 ±0.00	94.10 ±0.00	93.93 ±0.00	93.94 ±0.00	93.85 ±0.00	93.87 ±0.00	93.96 ±0.00	94.05 ±0.01	93.96 ±0.01	93.93 ±0.03
CIFAR-100	68.44 ±0.00	69.13 ±0.01	68.98 ±0.01	69.33 ±0.01	69.37 ±0.01	69.37 ±0.01	69.27 ±0.01	69.07 ±0.01	68.97 ±0.01	68.87 ±0.01
Country-211	17.82 ±0.00	17.80 ±0.01	17.73 ±0.01	17.63 ±0.01	17.61 ±0.00	17.79 ±0.00	17.78 ±0.00	17.63 ±0.00	17.36 ±0.02	17.32 ±0.02
DTD	41.17 ±0.00	42.36 ±0.03	42.59 ±0.03	42.59 ±0.05	42.34 ±0.04	42.62 ±0.03	42.61 ±0.00	41.52 ±0.05	41.95 ±0.05	41.79 ±0.07
EuroSAT	47.16 ±0.00	50.45 ±0.04	50.51 ±0.02	48.14 ±0.03	47.77 ±0.02	49.19 ±0.05	50.03 ±0.03	48.21 ±0.02	48.53 ±0.08	48.85 ±0.12
FER-2013	48.65 ±0.00	46.08 ±0.01	46.78 ±0.02	46.93 ±0.03	46.68 ±0.05	46.80 ±0.01	46.44 ±0.01	47.93 ±0.01	47.00 ±0.06	46.85 ±0.13
FGVC-Aircraft	26.30 ±0.00	25.56 ±0.02	25.70 ±0.01	25.52 ±0.04	25.49 ±0.01	25.74 ±0.02	25.70 ±0.01	26.45 ±0.01	26.23 ±0.03	25.54 ±0.09
Food-101	65.06 ±0.00	63.83 ±0.00	64.27 ±0.01	64.54 ±0.01	64.58 ±0.01	64.52 ±0.00	64.21 ±0.02	64.52 ±0.01	64.67 ±0.00	64.46 ±0.00
GTSRB	43.27 ±0.00	46.06 ±0.02	46.95 ±0.01	46.81 ±0.03	46.34 ±0.01	46.33 ±0.03	46.62 ±0.02	44.64 ±0.01	44.88 ±0.06	44.66 ±0.06
Hateful-Memes	56.50 ±0.00	56.06 ±0.01	56.25 ±0.01	56.09 ±0.01	56.17 ±0.00	55.98 ±0.01	56.03 ±0.01	57.01 ±0.01	56.64 ±0.02	56.81 ±0.03
KITTI-Distance	28.41 ±0.00	30.80 ±0.00	30.24 ±0.11	33.19 ±0.11	33.94 ±0.13	32.21 ±0.00	29.96 ±0.00	26.30 ±0.00	28.36 ±0.07	28.27 ±0.11
MNIST	54.99 ±0.00	53.64 ±0.04	54.83 ±0.02	61.86 ±0.03	60.14 ±0.04	59.57 ±0.01	56.06 ±0.03	55.68 ±0.04	55.37 ±0.06	55.00 ±0.10
Oxford Flowers-102	50.88 ±0.00	49.98 ±0.02	49.72 ±0.03	49.74 ±0.02	49.93 ±0.02	50.03 ±0.02	50.07 ±0.01	51.26 ±0.01	51.91 ±0.04	51.99 ±0.12
Oxford-IIIT Pets	81.56 ±0.00	79.52 ±0.02	80.69 ±0.02	79.60 ±0.03	79.37 ±0.00	79.24 ±0.02	79.46 ±0.02	81.29 ±0.02	81.24 ±0.03	80.90 ±0.09
PatchCamelyon	50.96 ±0.00	57.15 ±0.01	55.70 ±0.01	51.93 ±0.00	51.71 ±0.01	52.56 ±0.03	55.49 ±0.02	52.86 ±0.02	52.23 ±0.01	52.39 ±0.07
Rendered-SST2	54.20 ±0.00	53.05 ±0.04	53.82 ±0.09	53.67 ±0.03	54.82 ±0.03	54.35 ±0.03	53.03 ±0.03	53.47 ±0.03	53.14 ±0.07	52.94 ±0.04
RESISC-45	58.51 ±0.00	58.78 ±0.01	58.92 ±0.02	59.56 ±0.01	59.71 ±0.01	59.25 ±0.02	58.88 ±0.01	57.06 ±0.00	57.21 ±0.02	56.93 ±0.01
Stanford-Cars	84.93 ±0.00	85.00 ±0.01	85.04 ±0.01	85.10 ±0.00	85.10 ±0.02	85.08 ±0.01	85.08 ±0.01	85.35 ±0.02	84.99 ±0.03	84.90 ±0.06
PASCAL VOC-2007	82.09 ±0.00	82.73 ±0.00	82.31 ±0.01	82.24 ±0.01	82.29 ±0.00	82.39 ±0.00	82.67 ±0.01	82.35 ±0.00	82.00 ±0.01	81.99 ±0.02
Average	56.97 ±0.00	57.23 ±0.02	57.36 ±0.02	57.57 ±0.02	57.51 ±0.02	57.48 ±0.02	57.31 ±0.01	56.99 ±0.02	57.02 ±0.03	56.90 ±0.06
ImageNet-1k	45.84 ±0.00	46.23 ±0.01	46.32 ±0.02	46.09 ±0.00	46.06 ±0.01	46.19 ±0.00	46.33 ±0.01	45.89 ±0.01	45.90 ±0.01	45.65 ±0.02

Table A4: Details of the image classification datasets on the ELEVATER benchmark.

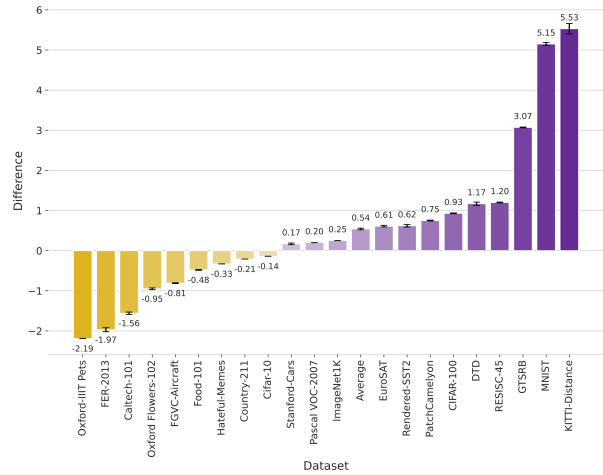| Dataset	#Labels	Test Size	Metric
Caltech-101 (Fei-Fei et al., 2004)	101	6,084	Mean-per-class
CIFAR-10 (Krizhevsky and Hinton, 2009)	10	10,000	Accuracy
CIFAR-100 (Krizhevsky and Hinton, 2009)	100	10,000	Accuracy
Country-211 (Radford et al., 2021)	211	21,100	Accuracy
DTD (Cimpoi et al., 2014)	47	1,880	Accuracy
EuroSAT (Helber et al., 2019)	10	5,000	Accuracy
FER-2013 (Goodfellow et al., 2013)	7	3,589	Accuracy
FGVC-Aircraft (Maji et al., 2013)	100	3,333	Mean-per-class
Food-101 (Bossard et al., 2014)	101	25,250	Accuracy
GTSRB (Stallkamp et al., 2011)	43	12,630	Accuracy
Hateful-Memes (Kiela et al., 2020)	2	500	ROC AUC
KITTI-Distance (Fritsch et al., 2013)	4	711	Accuracy
MNIST (Deng, 2012)	10	10,000	Accuracy
Oxford Flowers-102 (Nilsback and Zisserman, 2008)	102	6,149	Mean-per-class
Oxford-IIIT Pets (Parkhi et al., 2012)	37	3,669	Mean-per-class
PatchCamelyon (Veeling et al., 2018)	2	32,768	Accuracy
Rendered-SST2 (Radford et al., 2021)	2	1,821	Accuracy
RESISC-45 (Cheng et al., 2017)	45	25,200	Accuracy
Stanford-Cars (Krause et al., 2013)	196	8,041	Accuracy
Pascal VOC-2007 (Everingham et al., 2010)	20	4,952	11-point mAP
Total	1,151	192,677	-

Figure A1: Difference between the OPENCLIP fine-tuned with CAPIVARA on CC3M, and the baseline (OPENCLIP), considering the ELEVATER benchmark and ImageNet-1k.

breeds of cats and dogs, and the FER-2013 dataset, featuring a range of human emotional expressions. Also, our model presented a performance decline on these datasets, with respective decrements of 2.19 and 1.97 *pp.* in comparison to the baseline.

Figures A2 to A5 offer a deeper dive into these observations, presenting normalized confusion matrices that provide granular insights into datasets where CAPIVARA underperformed the baseline. Specifically, Figures A2 and A3 unveil nuances in accurate and erroneous predictions within the Fer-2013 dataset. Notably, the baseline model excels in recognizing neutral expressions, while the fine-tuned model performs well in identifying expressions of sadness. However, the fine-tuned model is also more likely to confound emotions such as sadness and neutral expressions. Figures A4 and A5 present normalized confusion matrices for the Oxford-IIIT Pets Dataset, highlighting the fine-tuned model's tendency to amplify confusion be-

tween cat breeds British Shorthair and Russian Blue, as well as dog breeds Leonberger and Newfoundland, leading to reduced overall correctness.

A.3 Ablation Study

A.3.1 Impact of Multiple Captions & Generated Caption Selection

To further validate the contributions of synthetic captions, we analyze the influence of multiple captions per image and how to select proper captions for each image. This latter aspect is related to BLIP2's hallucination, i.e., the model generates a text that does not match the associated image (Xu et al., 2023). The use of these synthetic annotations can introduce noise to the dataset. To address this issue, we implement the Captioning and Filtering
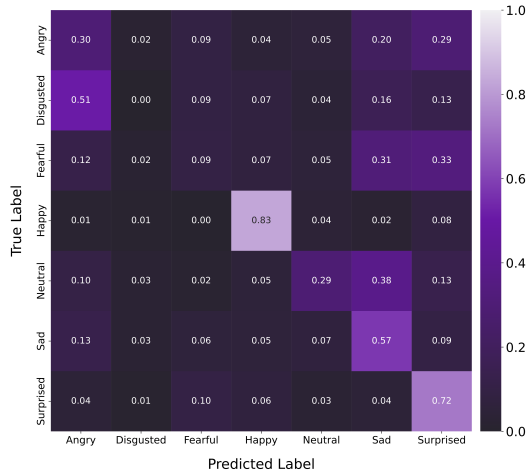
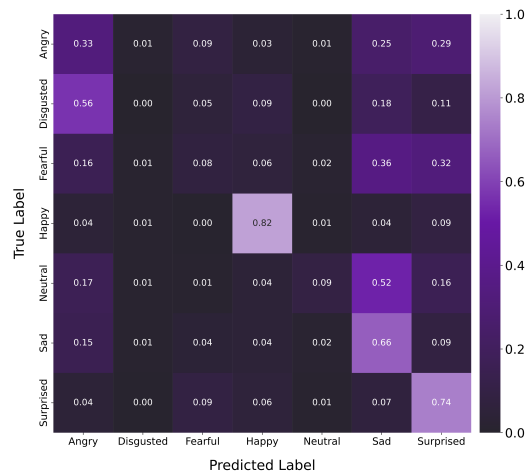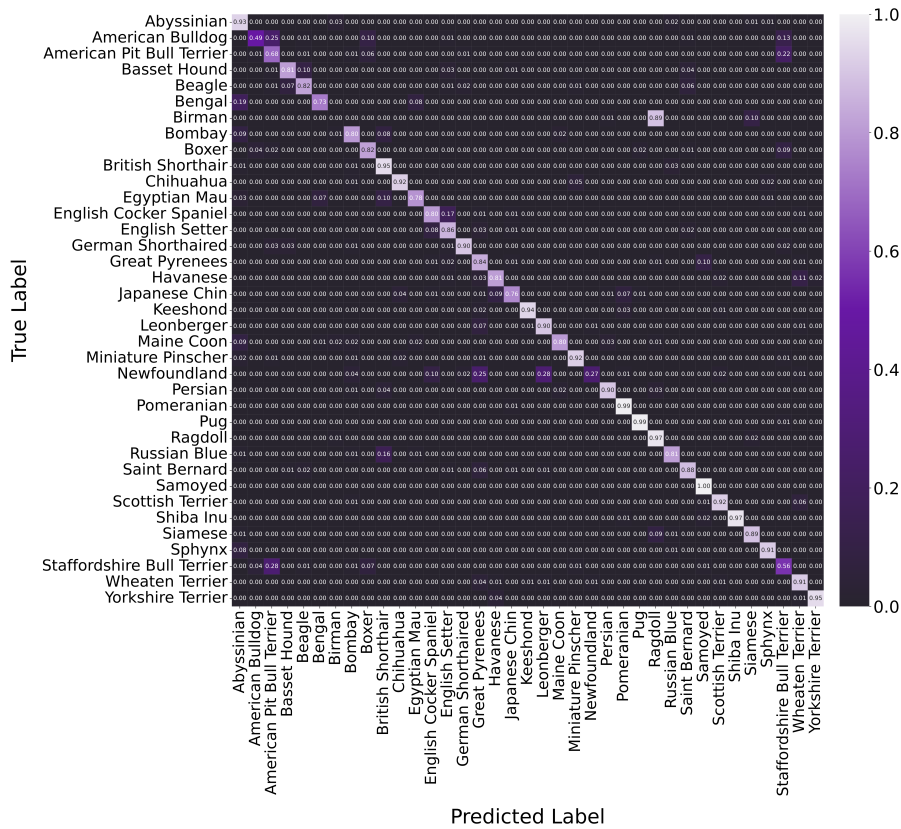Figure A2: Normalized confusion matrix of the FER-2013 dataset for the OPENCLIP baseline model.

Figure A3: Normalized confusion matrix of the FER-2013 dataset for CAPIVARA.

(CapFilt) (Li et al., 2022b, 2023) method with three different selection strategies: rank-based, threshold-based, and threshold-based + near-duplication removal. All strategies rely on similarity scores produced by OPENCLIP VIT-B/32 XLM-ROBERTA BASE model.

Rank-based: We rank the synthetic captions along with the original descriptions based on the image-text similarity and select the top-k examples; in our tests, we adopted $k = 5$.

Threshold-based: We select the texts among the original and generated captions based on their similarity to the associated image. Then, a caption is selected if the similarity between it and the image is greater than or equal to a given threshold; in this case, the threshold is 0.15.

Threshold-based + near-duplication removal: We first apply the threshold-based filter, and then we remove the near-duplicate captions using the algorithm described in Algorithm 1, keeping a minimum of $k_{min} = 3$ captions per image. Algorithm 1 first computes the text similarity matrix. Then, it computes the cost of removing a text t_i as $c(t_i) = \sum_{j=1}^{B} sim(t_i, t_j), \forall i \neq j$. At each step, it removes the text with the highest cost and updates the cost array. The algorithm stops when all similarity scores are lower than a given threshold or the minimum number of captions is reached. In this way, the algorithm can keep the maximum diversity among the texts.

Figure A4: Normalized confusion matrix of the Oxford-IIIT Pets dataset for OPENCLIP baseline model.
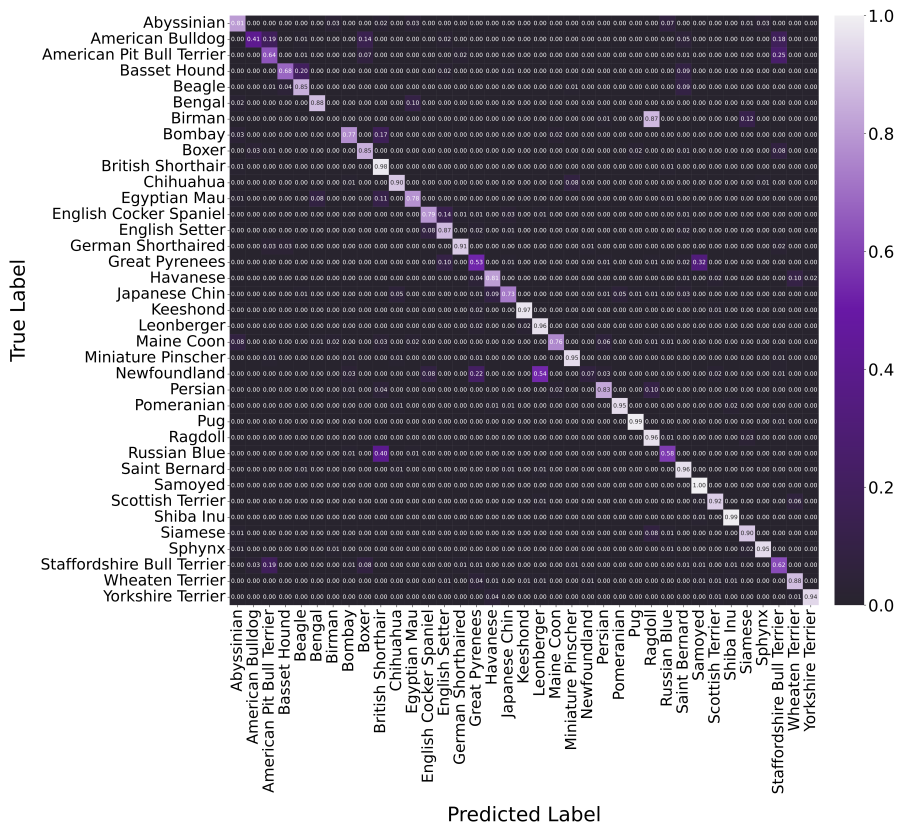

Figure A5: Normalized confusion matrix of the Oxford-IIIT Pets dataset for the OPENCLIP + Fine-tuning model with 10 generated annotations.

200

```python
# captions: image captions
# k_min: minimum number of texts to keep
# thr: maximum similarity between texts
#      allowed

# Remove similar texts keeping the
# maximum diversity among them
def remove_similar(captions, k_min=3,
    thr=0.3):
  if len(captions) < k_min:
    return captions

  sim_matrix = text_similarity(captions)
  n_texts = sim_matrix.shape[0]
  # set the cost in the diagonal to zero
  sim_matrix -= np.eye(n_texts)
  while not (sim_matrix <= thr).all()
        and n_texts > k_min:
    # compute the cost to remove each
    # text as sum of the similarity
    # between that text and all others.
    cost = sim_matrix.sum(axis=0)

    # remove the text with the highest
    # cost
    i = np.argmax(cost)

    # set the cost of the texts to be
    # removed to zero
    sim_matrix[i, :] = 0
    sim_matrix[:, i] = 0
    n_texts -= 1

  # compute the final cost for all texts
  cost = sim_matrix.sum(axis=0)
  # all texts whose cost is zero will be
  # removed
  remove_indices = np.where(cost==0)[0]
  # return the filtered texts
  return
   [caption
    for i,caption in enumerate(captions)
             if i not in remove_indices]
```

Algorithm 1: Python-like pseudocode of near-duplicate text removal algorithm.

From a thorough analysis of the results exhibited in Table A5, we note that none of the caption selection strategies significantly impacted the model performance. All strategies performed similarly to CAPIVARA with no caption selection. Specifically, the threshold-based caption selection strategy performed slightly better than the others but still in pair with CAPIVARA. This result suggests that BLIP2 is effective in generating captions related to images and, because of this, the caption selection methods did not affect the final performance. Nevertheless, Figure A8 and the results in Table A6 reveal that BLIP2 produces slightly different texts. Therefore, generating multiple captions per image has a limited effect on text augmentation. Note that adding 10 captions slightly improved compared to

Table A5: Experimental results for caption selection strategies. In this table, "threshold-based near-duplication", "threshold-based", and "rank-based" refer to caption selection methods, whereas CAPIVARA does not consider any caption selection strategy. For each setting, we report the average and the standard deviation of mean recall.

Method	Flickr30k		MS COCO		PraCegoVer	
	txt2img	img2txt	txt2img	img2txt	txt2img	img2txt
OPENCLIP (Baseline)	76.23	87.93	52.62	66.55	65.36	69.43
OPENCLIP + Fine-tuning	78.42 ± 0.02	90.02 ± 0.05	54.77 ± 0.01	70.06 ± 0.01	63.79 ± 0.01	60.10 ± 0.00
Threshold-based near-duplication	79.59 ± 0.01	90.02 ± 0.02	56.37 ± 0.01	71.14 ± 0.01	66.72 ± 0.01	65.33 ± 0.01
Threshold-based	79.65 ± 0.03	89.72 ± 0.02	56.39 ± 0.01	71.11 ± 0.02	66.77 ± 0.01	65.47 ± 0.01
Rank-based	79.60 ± 0.01	89.13 ± 0.04	56.32 ± 0.01	70.64 ± 0.02	66.85 ± 0.00	65.96 ± 0.01
CAPIVARA	79.56 ± 0.01	89.95 ± 0.04	56.27 ± 0.01	71.24 ± 0.01	66.40 ± 0.01	64.75 ± 0.01

Table A6: Impact of multiple captions. This table presents the results of models trained with different numbers of synthetic captions translated into Portuguese. We report the average and the standard deviation of mean recall for each setting across Flickr30k, MS COCO, and PraCegoVer datasets.

Method	Flickr30k		MS COCO		PraCegoVer	
	txt2img	img2txt	txt2img	img2txt	txt2img	img2txt
OPENCLIP (Baseline)	76.23	87.93	52.62	66.55	65.36	**69.43**
OPENCLIP + Fine-tuning	78.42 ± 0.02	90.02 ± 0.05	54.77 ± 0.01	70.06 ± 0.01	63.79 ± 0.01	60.10 ± 0.00
CAPIVARA + 10 synth. captions	**79.56 ± 0.01**	**89.95 ± 0.04**	**56.27 ± 0.01**	**71.24 ± 0.01**	**66.40 ± 0.01**	64.75 ± 0.01
CAPIVARA + 5 synth. captions	79.17 ± 0.02	90.72 ± 0.02	55.62 ± 0.01	70.95 ± 0.00	65.18 ± 0.01	62.14 ± 0.01
CAPIVARA + 1 synth. caption	79.46 ± 0.01	90.02 ± 0.05	56.26 ± 0.01	71.27 ± 0.01	66.09 ± 0.01	63.95 ± 0.01

adding just one caption per image. Therefore, it is necessary to explore methods for generating more diverse texts, for instance, testing different sampling methods and other image captioning models, because we only used BLIP2 with default parameters.

A.3.2 Impact of Increasing the Batch Size

Among the different hyperparameters used to train the model, batch size has significant potential to improve model results. As batch size increases, more examples are observed per training step, and more examples might be discriminated by contrastive learning. Therefore, to determine the optimal batch size to use in our method, we conducted experiments fixing the number of steps in 5863 and varying this value considering our GPU memory limitation. We experimented three different batch sizes: 1000, 2816, and 4300. Each setting was tested with traditional fine-tuning and with CAPIVARA, the results are presented in Table A7.

Overall, we do not observe a significant gain in increasing the batch size. Intriguingly, in the

Table A7: Comparison between different batch sizes in fine-tuning and CAPIVARA settings.

Method	Batch size	Flickr30k		MS COCO		PraCegoVer	
		txt2img	img2txt	txt2img	img2txt	txt2img	img2txt
OpenCLIP + Fine-tunning	1000	78.68 ±0.02	90.02 ±0.02	54.45 ±0.01	69.06 ±0.01	66.38 ±0.01	66.49 ±0.02
	2816	78.71 ±0.02	89.85 ±0.02	54.57 ±0.00	69.17 ±0.03	66.44 ±0.01	66.57 ±0.01
	4300	78.70 ±0.01	89.86 ±0.02	54.62 ±0.04	69.22 ±0.02	66.42 ±0.05	66.76 ±0.19
CAPIVARA + Opt. (5863 steps)	1000	79.71 ±0.03	90.51 ±0.05	55.36 ±0.03	69.58 ±0.03	67.00 ±0.03	68.01 ±0.01
	2816	79.81 ±0.03	90.65 ±0.02	55.56 ±0.01	69.64 ±0.04	67.07 ±0.02	68.14 ±0.01
	4300	79.87 ±0.01	90.63 ±0.04	55.63 ±0.01	69.70 ±0.04	67.08 ±0.01	68.19 ±0.01

context of CAPIVARA, the performance slightly improves across the datasets as we increase the batch size from 1000 to 2816. However, it declines when we use a batch size of 4300. For this reason, the CAPIVARA models were trained with an average batch size of 2816, while the optimized CAPIVARA models were trained with a batch size of 1000. This study shows that using smaller batches to train the optimized models does not result in significant loss. At the same time, it saves memory and training time.

A.4 Qualitative Analysis

We conducted experiments on Flickr30k for a qualitative analysis of the model's ability in cross-modal retrieval tasks, the outcomes are presented in Figures A6 and A7. Figure A6 shows the result of the image-to-text retrieval task, where the five texts in Portuguese more similar to a given image are retrieved by our model. For the first example, all the texts retrieved describe correctly the image content, which consists of a group of women running in a race. However, in the second example, none of the retrieved text matches the input image. It illustrates the limitations of our model.

Similarly, we analyze qualitatively our model in text-to-image retrieval. In Figure A7, we present four examples of texts and the top-5 images more similar to each of them. We can see that overall the model ranks the correct images on the top. Regarding the other images, although the scene representations match the texts, there is still a lack of details in the images that are not considered by the model, such as the number of people, objects, and colors. This can happen because there are no images that contain all elements from the text within the dataset, and it tries to retrieve the most similar images, or by model limitations. Thus, in the last example, we present an instance in which the model fails. Given the text "Woman and man walking across wooden rope bridge with a caution sign beside it.", the model does not rank the expected
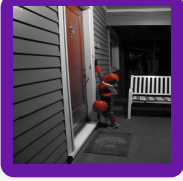
image among the top-5 most similar.

A.5 Synthetic Captions Generated by BLIP2

In the process of text augmentation, the BLIP2 model (Li et al., 2023) was used to generate new captions for the images. However, this model presents some issues regarding text generation. For example, it may generate text that does not match the image and repeat words. Several strategies have been used to mitigate these problems in our work. They are best described in Sec. 3. Figure A8 shows three images from CC3M along with their original caption and 10 captions generated with BLIP2.

The first image represents an example where the generated captions are good and diverse, as all captions correctly describe the image, there are no repeated words, and there is a high diversity of words used to describe the scene. The captions generally describe the image and add new elements to the description, although they still contain repetitive structures. In the second example, we present a scenario of good caption and low textual diversity. The captions describe the image, but there is a high level of repetition in the sentence structures. In the third example, we illustrate a case of badly generated captions and low textual diversity. In this example, the model not only shows a lot of word repetition, but also fails to represent the image, hallucinating.

Image-to-Text Retrieval

Text-to-Image Retrieval

Figure A6: Examples of image-to-text retrieval using CAPIVARA + Opt.

Figure A7: Examples of text-to-image retrieval using CAPIVARA + Opt.

Original Caption:

a toting airship flying around the blue skies

Generated Captions:

the foreground features 's a large airship with an american flag painted on it - stock image

a photo of airship blimp flying in the sky with palm trees in the background - stock image

a picture of airship flying in the sky with palm trees in the background - stock image

this is a scene depicting airship blimp flying in the sky with palm trees in the background - stock image

an image of airship blimp flying in the sky with palm trees in the background - stock image

portrait of a airship flying in the sky with palm trees in the background - stock image

this image captures a moment of a blimp flying in the sky with palm trees in the background - stock image

a painting of airship with american flag flying in the sky - stock image

an art of airship flying in the sky with palm trees in the background - stock image

the picture shows the american flag blimp flying in the sky with palm trees in the background - stock image

Good captions, high diversity

Original Caption:

a teenager from the rural town of person , was shocked to find a koala behind the wheel.

Generated Captions:

the foreground features koalas sitting in the driver's seat of a vehicle

a photo of koala sitting on the steering wheel of a vehicle

a picture of a koala sitting on the steering wheel of a vehicle

this is a scene depicting of a koala sitting on the steering wheel of a vehicle

an image of koala sitting on the steering wheel of a vehicle

portrait of a a koala sitting on the steering wheel of a vehicle

this image captures a moment of a koala sitting in the driver's seat of a vehicle

a painting of of a koala sitting on the steering wheel of a vehicle

an art of koala sitting on the steering wheel of a vehicle

the picture shows a koala sitting in the driver's seat of a vehicle
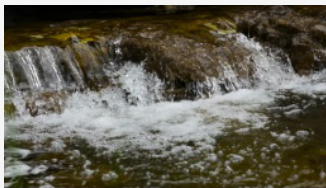
Good captions, low diversity

Original Caption:

waterfall on a small stream.

Generated Captions:

the foreground features a man holding a bird in his hand

a photo of of a man holding a bird in his hand

a picture of a man holding a bird in his hand

this is a scene depicting of a person holding a bird

an image of of a man holding a bird in his hand

portrait of a of a man holding a bird in the water

this image captures a moment of a man holding a bird in the water

a painting of a man holding a bird in the water

an art of of a man holding a bird in the water

the picture shows a man holding a bird in the water

Bad captions, low diversity

Figure A8: Examples of images with synthetic captions generated by BLIP2.

A.6 Model Cards

This section was done using the Model Cards for Model Reporting (Mitchell et al., 2019) tool.

Model Details

- Developed by researchers from the Natural Language Processing Group of the Artificial Intelligence and Cognitive Architectures Hub – H.IAAC.

- CAPIVARA, 2023, v1.

- CAPIVARA is a cost-efficient framework designed to enhance the performance of multilingual CLIP models in low-resource languages.

- CAPIVARA augments text data using image captioning and machine translation to generate multiple synthetic captions in low-resource languages. The training pipeline is optimized with LiT, LoRA, and gradient checkpointing to alleviate the computational cost.

- More information can be found on CAPIVARA's official GitHub `https://github.com/hiaac-nlp/CAPIVARA`.

- For further information or questions, please contact Sandra Avila `avilas@unicamp.br`.

Intended Use

- Intended to be used for general tasks focused on finding a representation in a common space for texts and images. Examples of tasks are image-to-text and text-to-image retrieval and image classification.

- Particularly intended for scientific researchers.

- Not intended to be used with aspects, positions, and cultural values from an under-represented region (e.g., Brazilian memes) due to the lack of representativeness of the datasets used for training. It cannot be used with long texts (more than 77 tokens).

Factors

- Based on known problems with image and language models, potential relevant factors include groups for under-represented and minority people. In order to adapt the model to languages with low resources, texts were initially translated from English; thus, the model does not represent the cultural and geographical aspects of the countries that speak these target languages. The datasets used are made of texts collected from the Internet; therefore, the model may not perform as well for data collected from other sources and may carry biases from the original texts.

Metrics

- Evaluation metrics include Mean Recall, representing the average recall value across the recall@K instances, where K = 1, 5, 10, for cross-modal retrieval, which is the main task of CAPIVARA, and top-1 accuracy metrics for image classification task on ImageNet-1k. Moreover, the ELEVATER benchmark was used for the image classification task, and Appendix A.2 provides the specific metrics used (see Table A4).

- Each experiment was run three times, and the mean and standard deviation were reported for all experiments performed (see Section 4).

Quantitative Analyses

- Quantitative Analyses can be seen in Figure 1 and Section 4.

Evaluation Data

- Evaluation data include Flickr30k, MS COCO, and PraCegoVer datasets for cross-modal retrieval task, and all 20 datasets from ELEVATOR benchmark and ImageNet-1k for image classification task (see Table A3).

- These datasets were chosen because they are the most widely used datasets in the literature, except for PraCegoVer. PraCegoVer is a dataset with images and texts originally in Portuguese that was used precisely to evaluate linguistic and cultural aspects present in the Portuguese language. (NOTE: Data originally in English that has been translated into the target language will be made available with the model).

- See Section 3.2 for more details about data preprocessing.

Training Data

- Training data was CC3M dataset.

- This dataset was chosen because of the amount of example data provided and the better quality of the data. In addition, our limited computing infrastructure for training the model was considered.

- See Section 3.2 for more details about data preprocessing.

- It is possible that the model was trained with data where group distributions are not homogeneous and, therefore, encoded some type of bias.

Ethical Considerations

- CAPIVARA does not deliberately use sensitive data in training. However, since it uses data collected from the Internet consisting of images and annotations about the image's content, it is possible that data with political, religious, or cultural positioning have been used.

- CAPIVARA does not generate any type of data that could pose a risk to human life. However, our model can be adapted for other specific tasks, e.g., image or text generation, which could contribute to generating false information and harming people.

- The model's training data was translated via Google Translate from English into the target language. This can lead to linguistic biases and a lack of representativeness for the target groups.

- CAPIVARA adopts training time optimizers, resulting in a smaller carbon footprint than traditional fine-tuning. Therefore, it presents a better financial and environmental alternative to improve the performance of pre-trained models.

Caveats and Recommendations

- Further work is needed to assess the impact of adding more samples from the target language and how much this brings the performance of the target language closer to English, which currently has the best performances. See Section 5 for more future works.

- People and groups who do not have access to the Internet and, therefore, do not produce digital content were under-represented in the training set. However, CAPIVARA is intended to be applied to languages with low digital resources. CAPIVARA offers the technique to improve performance for low-resource languages, however there is still a gap in performance between English texts and texts in low-resource languages. Future studies are required to improve performance for different languages and include cultural and linguistic aspects of the target language in the model.

- An ideal evaluation dataset would additionally include annotations made in the target language, which also considers cultural and linguistic aspects and has a background of minority and under-represented groups.

- Current literature is constantly evaluating the ethical risks and impacts that vision and language models can have on society. Keeping up with this work is extremely important, as these studies can point to risks and negative impacts that have not yet been considered in this current version of Model Cards.

- Ideally, when using CAPIVARA as a base model for other applications, a study of the ethical impacts of the application should be carried out before it is implemented.

- It is highly recommended to read this Model Cards in conjunction with the article that introduces CAPIVARA, as the article contains detailed information on the entire life cycle of the proposed model.