V-GlórIA: Customizing Large Vision and Language Models to European Portuguese

Afonso Simplício, David Semedo, João Magalhães
NOVA LINCS, NOVA School of Science and Technology, Portugal
{am.simplicio}@campus.fct.unl.pt
{df.semedo, jmag}@fct.unl.pt

Abstract

Generative Vision and Language models have obtained remarkable results recently, thanks to the use of robust pre-trained Visual encoders and Large Language Models (LLMs), together with efficient model adaptation training strategies, requiring minimal architectural modifications, while preserving LLMs' original capabilities. With these advances focusing mainly on the English language, there is a gap in customization methodologies for other languages. In this paper, we propose a customization methodology that adapts existing state-of-the-art vision and language architectures to European Portuguese (PT-PT). As a result of applying this methodology, we introduce V-GlórIA , the first Large Vision and Language generative model specifically customized for European Portuguese. V-GlórIA supports multimodal tasks such as image captioning, retrieval, and dialogue. To deliver V-GlórIA, we leverage state-of-the-art V&L architectures, and contribute with PT-PT machine translated pre-training (CC3M PT-PT) and benchmark (MSCOCO PT-PT and VisDial PT-PT) datasets. Our experiments show that V-GlórIA delivers promising performance in text-image retrieval and downstream tasks in a zero-shot setting, such as image captioning and visual dialogue tasks, highlighting the effectiveness of our customization approach. [1]

1 Introduction

Vision and Language are two of the main communication and information perception mediums, serving as fundamental channels through which humans interpret and interact with the world around them. Devising Vision and Language (V&L) models that can seamlessly combine these two modalities is paramount to delivering AI systems capable of addressing tasks such as image captioning and visual question-answering, essential tasks to aid visually impaired individuals, and Image-to-Text and Text-to-Image retrieval, for general multimodal information seeking. Recently, there have been notable advances in vision and language models (Liu et al., 2023; Koh et al., 2023; Kim et al., 2021), which leverage Large Language Models as backbones (Touvron et al., 2023; Brown et al., 2020; Zhang et al., 2022) (LLMs). Most of these advances have been made with models in English or other high-resource languages, leaving behind other lower-resource languages, as is the case of European Portuguese (PT-PT). This evidences the urgent need of having effective customization methologies to deliver V&L LMs, openly available, for PT-PT speakers. This customization process raises two complementary challenges: 1) how to overcome the limited availability of PT-PT multimodal datasets and resources, and 2) how to train a Large Visiwon and Language model, capable of addressing multiple V&L tasks, in PT-PT.

Most LLMs are trained with text-only web scraped data, achieving great performance on a myriad of natural language tasks, but lack an overall understanding of images, thus not having visual reasoning capabilities. Pioneering vision and language models, adopted fully multimodal Transformer-based models (Lu et al., 2019; Yu et al., 2022; Wang et al., 2022), with either single-stream or dual-stream architectures (Bugliarello et al., 2021), pre-trained on image-text pairs. More recently, towards generalizing high-performing large LMs to the visual domain, it is common practice to leverage text-only LLMs as the backbone and equip them with a visual encoder (Radford et al., 2021; Dosovitskiy et al., 2021). Then, LLMs are augmented with a visual projection component that aligns visual tokens with the LLM token-space (Koh et al., 2023; Liu et al., 2023).

In this paper, we seek to establish a V&Lcustomization methodology to European Portuguese, and as a result, deliver the first European

[1] Code and data are available in https://github.com/amsimplicio/V-GlorIA.

Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 317–326

Portuguese vision and language LM, **V-GlórIA**. To this extent, we make two major contributions: **1)** we create and make available both large-scale image-text pre-training datasets as well as well-known V&L benchmarks in European Portuguese. In particular, CC3M (Sharma et al., 2018a) PT-PT (3 million image-caption pairs) for pre-training, MSCOCO (Lin et al., 2014) PT-PT (image-caption pairs) and VisDial (Das et al., 2017) PT-PT (visual dialogs) for benchmarking on downstream V&L tasks. An extensive assessment of available machine translation approaches is carried out. **2)** following the V&L LMs state-of-the-art, we adapt the FROMAGe (Koh et al., 2023) model to support PT-PT. Its flexible decoder-based architecture, augmented with multimodal specialized layers, gives the model the capacity to process and produce interleaved multimodal inputs and outputs. Given that a key step is to replace the original LLM by a PT-PT LLM, we leverage a recent PT-PT text-only decoder, GlórIA (Lopes et al., 2024), and conduct extensive experiments, in a zero-shot setting, on image caption and visual dialog tasks.

2 Related Work

Most Generative Vision and Language models consist of decoder-only Transformers. GPT-3 (Brown et al., 2020) showed that when trained with a lot of data, language models can generalize and solve new (unseen) tasks. This is very useful since although the training is expensive and requires a lot of data, once they are pre-trained, they can be applied to a myriad of tasks with reduced adaptation costs. LLaVA (Liu et al., 2023) takes advantage of this by creating a general Vision and Language model using a frozen LLM as decoder and a frozen visual encoder to encode the images, training a linear layer that transforms the image embeddings into the LLM embedding space. This simple linear transformation has the advantage of introducing a very small number of parameters to be learned, allowing for efficient large-scale training, while leveraging the generalization capabilities of the backbone text LLM. Different ways of mapping image embeddings to the LLM token subspace have been tried, such as a Q-Former (Li et al., 2022) consisting of a Query transformer that learns query embeddings that will interact with the image encoding through cross attention, and CogVLM (Wang et al., 2024) where although the part of the LLM that processes the text input will still be frozen, it trains the weights used to compute the queries, keys, and values relative to the image embeddings. FROMAGe (Koh et al., 2023) takes a step further by extra linear transformations that enable to model to generatively retrieve images/texts. This is accomplished by introducing a special retrieval token, that is then trained under a multimodal contrastive learning of cross-modal mappings.

Most of these models are in English or other high-resource languages. Very recently, open European Portuguese LLMs have been made available. In particular, GlórIA (Lopes et al., 2024) is a European Portuguese LLM Decoder based on GPT-Neo (Black et al., 2021) - which approximates the GPT3 architecture - trained on a 35B token corpus, from a diverse set of domains, including web content, news pieces, encyclopedic knowledge, news articles, and dialogs. Gervásio (Santos et al., 2024) is another relevant European Portuguese LLM decoder which is based on a pre-trained LLaMA 2 7B (Touvron et al., 2023) model, fine-tuned on Portuguese instruction datasets, comprising around 83M tokens. Regarding V&L approaches, literature is scarce. CAPIVARA (dos Santos et al., 2023) trains a Brazilian Portuguese CLIP model, while performing data augmentation through image captioning and machine translation. In this work, and using recent developments in the LLM PT-PT, we seek to narrow this gap, by introducing a European Portuguese V&L model.

3 PT-PT Datasets for Vision and Language AI

Due to the lack of European Portuguese V&L datasets, we embraced the task of translating core vision and language datasets from English into European Portuguese. Given the size of existing datasets (millions scale), translating the datasets with human experts would be too costly, hence we considered three distinct automatic machine translation models: first, we considered a) **MADLAD-400** (Kudugunta et al., 2023), a model trained on a 3T token dataset based on CommonCrawl, created by Google, covering text data from over 400 languages; b) **Narrativa**[2], which is an mBART-50 (Tang et al., 2020) model fine-tuned on the opus-100 (Zhang et al., 2020) dataset for English to Portuguese Translation, c) **DeepL**[3] a

[2] https://huggingface.co/Narrativa/mbart-large-50-finetuned-opus-en-pt-translation

[3] https://www.deepl.com/translator

Table 1: Translation statistics, for CC3M and COCO, with different machine translation approaches. # Samples - total number of samples, # Tokens - total number of tokens, # Avg. Tokens/Sample - average number of tokens per sample. * Stands for the original captions.

	Statistic	English*	MADLAD	Narrativa	DeepL
CC3M	# Samples	2 709 383	2 709 383	2 287 769	2 709 383
	# Tokens	27 919 393	26 558 075	24 257 997	29 844 147
	# Tokens/Sample	10.30	9.80	10.60	11.02
COCO	# Samples	25 014	25 014	23 614	25 014
	# Tokens	282 297	282 172	267 893	292 626
	# Tokens/Sample	11.29	11.28	11.34	11.70

Original*: plenty of space : at square feet the property would have ample room for actor and her daughter

MADLAD-400: abundância de espaço: em pés quadrados a propriedade teria amplo espaço

Narrativa: plenty of space : at square feet the property would have ample room for actor and her daughter

DeepL: muito espaço: em metros quadrados, a propriedade teria muito espaço para o ator e a sua filha

Original*: people waiting for the bus in snow storm

MADLAD-400: pessoas à espera do ônibus na tempestade de neve

Narrativa: Pessoas à espera do autocarro em tempestade de neve

DeepL: pessoas à espera do autocarro numa tempestade de neve

Original*: person serves lunch to two of her daughters at their farm.

MADLAD-400: uma mulher serve o almoço para duas de suas filhas em sua fazenda

Narrativa: A pessoa serve o almoço a duas filhas da fazenda dela.

DeepL: uma pessoa serve o almoço a duas das suas filhas na sua quinta.

Figure 1: Translation results of sample captions from the CC3M dataset, using each of the three considered translation approaches. The original caption is shown for reference.

commercial translation service.

We started by pre-assessing the performance of each of the three approaches, using a subset of CC3M, comprising both shorter and longer captions. Table 1 illustrates some of the translated examples of the CC3M dataset (Sharma et al., 2018a). First, although **MADLAD-400** seems to give good translations, most are in Brazilian Portuguese. **Narrativa** translations are in European Portuguese, but for many captions, the model output is the original English caption, rather than its translation. **DeepL** seems to solve these problems, by providing high-quality European Portuguese translations, with the disadvantage of being a commercial solution. For example, for the first image, Narrativa outputted the original caption, and in the second image MADLAD-400 uses a Brazilian Portuguese lexicon in its translations (e.g. "ônibus" instead of "autocarro", the word bus). Something we also notice is that MADLAD-400 often does not translate the full caption (as in the first image).

Given these observations, we translated the CC3M (Sharma et al., 2018b), MSCOCO (Lin et al., 2014), and VisDial (Das et al., 2017) datasets, using **DeepL** and MADLAD-400(Kudugunta et al., 2023). Given the higher effectiveness of DeepL, we will take them as the main datasets/benchmarks, and refer to them as CC3M PT-PT, MSCOCO PT-PT and VisDial PT-PT, respectively. The CC3M PT-PT dataset was used as the pre-training dataset, and both MSCOCO PT-PT and VisDial PT-PT were used for benchmarking retrieval, image-captioning and visual dialog tasks. Table 1 shows the aggregated statistics of these datasets. It is important to note that the lower number of total tokens in the Narrativa translation stems from the fact that some captions are not actually translated. DeepL translations have higher token numbers than the original

English dataset, which despite corroboration with the increased verbosity of Portuguese vs. English, will have an impact on the models' performance.

4 Method

In this section we present V-GlórIA, an European Portuguese Large V&L model, capable of flexibly interleaving the two modalities, images and text, and therefore generalize to different NLP and CV tasks such as multimodal retrieval, image captioning, and visual dialog. Therefore, we adapt the FROMAGe model (Koh et al., 2023) architecture, which leverages a text LLM and adds a set of projection layers to align images with the LLM input subspace, and support generative retrieval. Specifically, it allows us to use an European Portuguese LLM, that will be frozen during training, with lightweight training strategies aimed at equip V-GlórIA with visual and linguistic reasoning capabilities.

4.1 V-GlórIA Architecture

4.1.1 PT-PT Language Model Backbone.

V-GlórIA uses a Portuguese large language model decoder originally trained with text-only data with a causal language modeling task. V-GlórIA is based on a PT-PT open and top performing LLM, GlórIA (Lopes et al., 2024). In the experiments, we compare it with alternative LLM backbones, such as Gervásio (Santos et al., 2024).

4.1.2 Visual Encoder Model.

Images are encoded using a pre-trained CLIP ViT-L/14 (Radford et al., 2021), such that given an image y, the visual model outputs $v(y) \in \mathbb{R}^m$, corresponding to the [CLS] token embedding. Both θ and ϕ, both LLM and visual encoder parameters will be frozen.

4.1.3 Visual Projection Layer.

With the LLM and the visual encoder frozen, a projection layer is used to map the encoded images to the embedding subspace of the LLM token. Namely, a linear layer, $v(y)^T \cdot \mathbf{W}_c \in \mathbb{R}^d$, where d corresponds to the LLM hidden dimension. This transformation makes it possible for our Portuguese decoder to understand the contents of the image it receives.

4.1.4 Multimodal Retrieval.

In order to support retrieving images, conditioned either on text or images, a special token [RET] is added to the model vocabulary, so that at any point in the decoding, the model can decode this token and its embedding (which will be learned) can be used for retrieval. During training, a [RET] token is appended to the end of the input captions. In practice, two linear mappings are trained, $\mathbf{W}_t \in \mathbb{R}^{d \times q}$ and $\mathbf{W}_i \in \mathbb{R}^{m \times q}$, which will map the hidden representation of [RET] obtained from the last hidden layer of the LLM and the visual embeddings, respectively, into a common q dimensional space.

4.2 Training

The training tasks are specifically designed to equip the model vision and language reasoning capabilities: describing visual content; processing intervealed images and text in its context; and third matching images to text and vice versa. The model is trained with a multitask objective $\mathcal{L}$ comprising image captioning and image-text retrieval, with

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r (\mathcal{L}_{i2t} + \mathcal{L}_{t2i}), \tag{1}$$

with $\lambda_c = \lambda_r = 0.5$, as illustrated in Figure 2.

4.2.1 Image Captioning.

For captioning, the model is trained to autoregressively predict the next token, with a Cross-entropy loss conditioned on the image representation, i.e.

$$l_c(x, y) = \sum_{t=1}^{T} \log p_\theta(s_t | v(y)^T \mathbf{W}_c, s_1, \ldots, s_{t-1}), \tag{2}$$

where s_t represents the t-th token of the caption x, $\mathbf{W}_c$ the weights of the visual projection layer, and θ the frozen parameters of the LLM.

4.2.2 Image-text Retrieval.

For bidirectional multimodal retrieval, given a caption x_i and its corresponding image y_i[4], the InfoNCE (van den Oord et al., 2018) loss for multimodal contrastive learning is used as

$$\mathcal{L}_{t2i} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(x_i \cdot y_i / \tau)}{\sum_{j=1}^{N} \exp(x_j \cdot y_j / \tau)} \right), \tag{3}$$

where $x_i \cdot y_i$ corresponds to the cosine similarity between embeddings. The loss in the opposite direction, $\mathcal{L}_{i2t}$, is defined reciprocally, with x_i and y_i swapped.

[4]For the sake of notation simplification, $x_i \in \mathbb{R}^q$ and $y_i \in \mathbb{R}^q$ correspond to the [RET] token output of the retrieval mapping $\mathbf{W}_t \in \mathbb{R}^{d \times q}$, and to the outputs of the visual mapping $\mathbf{W}_i \in \mathbb{R}^{m \times q}$, respectively.

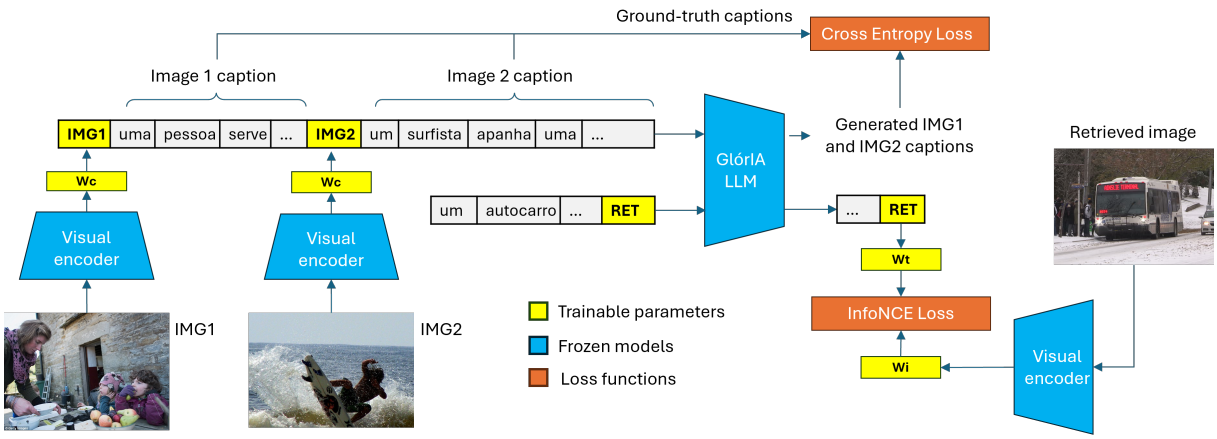


Figure 2: Overview of the V-GlórIA architecture. The model is trained on image-text pairs for image captioning and image-text retrieval. The LLM and visual encoder are frozen, while the three projection layers (in yellow), with weight matrices $\mathbf{W}_c$, $\mathbf{W}_i$, and $\mathbf{W}_t$, are learned.

5 Experimental Setup

We assess the performance of our model in both image retrieval and image-and-text generation tasks. The models were trained on the CC3M PT-PT dataset, originally comprising 3.3 million image-text pairs, which after filtering out missing and corrupted images resulted in a total of 2.7M samples. We consider both GlórIA 1.3B [5] and Gervásio 7B [6] as the PT-PT LLM backbones.

Multimodal retrieval and image captioning are evaluated in both the CC3M PT-PT (full-shot) and MSCOCO PT-PT (zero-shot) evaluation sets. Models are also evaluated in the Visual Dialog task (Das et al., 2017), in a zero-shot setting. To establish a comparison between English and European Portuguese, we consider the architectural twin of GlórIA 1.3B, GPTNeo 1.3B (Black et al., 2021), an English-only LLM.

Training details. For training, we set a batch size of 180 and train for a total of 20000 iterations, taking about 24 hours on 1x NVIDIA A100 40GB GPU. We used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0003 and a warmup of 100 steps.

The loss weight λ_c and λ_r are set to 1, the visual prefix length of $k = 1$. As for the embedding dimensions, we set the retrieval embedding dimension to $q = 256$, and model inner embedding dimension to $d = 2048$.

As most of the model parameters are frozen, our method is memory and compute-efficient: we backpropagate through the frozen LLM and visual model but only compute gradient updates for the 3 trainable linear mappings and [RET] embedding.

[5] https://huggingface.co/NOVA-vision-language/GlorIA-1.3B

[6] https://huggingface.co/PORTULAN/gervasio-7b-portuguese-ptpt-decoder

6 Results and Discussion

In this section, we discuss the experimental results in the image captioning and cross-modal retrieval tasks. We start by evaluating our model in cross-modal retrieval, in both Image to Text (I2T) and Text to Image (T2I) settings, and then in image captioning. We follow related work, and for cross-modal retrieval experiments adopt as metrics Recall@5 (R@5), and Recall@10 (R@10), and for image captioning BLEU and METEOR. Finally, we consider the challenging task of Visual Dialog, in a zero-shot setting. For the three tasks, we follow the established task evaluation protocols.

6.1 Cross-Modal Retrieval Results

The cross-modal retrieval results are shown in Table 2, where, for reference, we show (in gray) the performance of an English model (GPT-Neo 1.3B), trained and evaluated on the corresponding original English datasets. We can observe that V-GlórIA, using GlórIA as LLM, trained with data translated with DeepL, has the best results, significantly outperforming Gervásio in both directions, although the latter has more than five times the number of GlórIA parameters. In the MSCOCO validation set (unseen data), we observe a similar trend, where GlórIA shows to be preferable to Gervásio. However, in MSCOCO, we observe that higher performance is achieved when training

Table 2: Cross-modal Retrieval results for CC3M PT-PT and MSCOCO PT-PT datasets.

			I2T		T2I	
	LLM Backbone	**Data Language**	**R@5**	**R@10**	**R@5**	**R@10**
	GPT-Neo 1.3B	English	13.7	31.3	11.9	29.0
CC3M	GlórIA 1.3B	PT-PT MADLAD-400	22.5	44.9	22.0	44.1
	GlórIA 1.3B	PT-PT DeepL	**23.4**	**45.9**	**23.3**	**45.9**
	Gervásio 7B	PT-PT MADLAD-400	15.5	33.8	15.3	34.4
	Gervásio 7B	PT-PT DeepL	16.6	34.8	16.1	35.5
	GPT-Neo 1.3B	English	21.0	30.7	21.1	29.6
MSCOCO	GlórIA 1.3B	PT-PT MADLAD-400	**34.7**	**46.8**	**35.7**	**47.2**
	GlórIA 1.3B	PT-PT DeepL	30.2	41.1	30.1	40.7
	Gervásio 7B	PT-PT MADLAD-400	16.6	25.5	16.5	24.2
	Gervásio 7B	PT-PT DeepL	16.7	24.5	15.7	22.3

Table 3: Image Captioning results on the validation split of the CC3M PT-PT and MSCOCO PT-PT datasets.

	LLM Backbone	**Data Language**	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**METEOR**
	GPT-Neo 1.3B	English	18.5	9.9	6.0	4.0	17.6
CC3M	GlórIA 1.3B	PT-PT MADLAD-400	11.9	6.0	3.5	2.3	13.9
	GlórIA 1.3B	PT-PT DeepL	11.8	5.7	3.3	2.2	13.7
	Gervásio 7B	PT-PT MADLAD-400	9.6	5.4	3.4	2.3	12.3
	Gervásio 7B	PT-PT DeepL	10.8	6.1	3.8	2.6	13.1
	GPT-Neo 1.3B	English	42.8	24.1	12.9	7.0	13.1
MSCOCO	GlórIA 1.3B	PT-PT MADLAD-400	29.7	16.2	8.8	4.7	13.8
	GlórIA 1.3B	PT-PT DeepL	25.8	12.7	6.8	3.6	12.1
	Gervásio 7B	PT-PT MADLAD-400	21.6	12.3	7.0	3.9	13.4
	Gervásio 7B	PT-PT DeepL	23.8	13.3	7.9	4.7	12.9

and evaluating using the dataset translations obtained with MADLAD-400. This might be because although these models are European Portuguese LLMs, some of the data they were trained on may be in Brazilian Portuguese allowing the model to better understand the latter variety present in the MADLAD-400 translation.

When comparing the performance between the two languages, i.e. PT-PT (GlórIA 1.3B) and English (GPT-Neo 1.3B), we observe that performance is higher in PT-PT. This shows the robustness of our training procedure and hints at the promising capabilities of PT-PT vision and language models.

Table 4: Zero-shot results on VisDial (Das et al., 2017), for image-and-text-to-text (IT2T) and text-to-image (T2I) retrieval. Unlike previous methods, is capable of generating free-form text interleaved with image outputs through text-to-image retrieval.

	IT2T		T2I	
Backbone	**R@5**	**R@10**	**R@5**	**R@10**
GlórIA 1.3B	**4.2**	**14.1**	**17.3**	**25.2**
Gervásio 7B	4.0	13.9	8.2	14.0

6.2 Image Captioning Results

Table 3 shows the results of the image captioning. Again, for reference, we show (in gray) the performance of an English model (GPT-Neo 1.3B), trained and evaluated on the corresponding original English datasets.

Query: "*Uma mota Honda preta estacionada em frente a uma garagem.*"
Query in English - *"A dark Honda motorbike parked in front of a garage."*

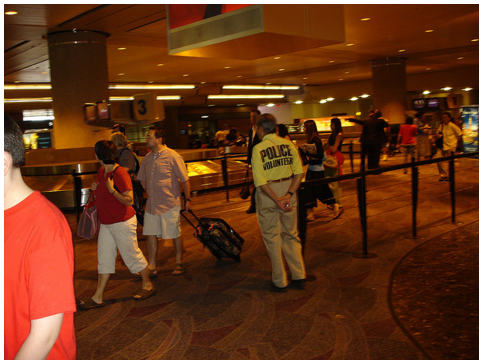
Retrieved images:

(a) Image retrieval

Ground truth:
Um homem numa prancha de surf na água.

V-GlórIA generated caption:
Um surfista a surfar a onda!

Gervásio generated caption:
um surfista a saltar de uma onda[RET] surfista a saltar de uma onda[RET] surfista ...



Ground truth caption:
Várias pessoas caminham pelo aeroporto enquanto esperam pelas suas malas.

V-GlórIA generated caption:
A fila de pessoas que se encontram a caminho do aeroporto.

Gervásio generated caption:
pessoas a descer a passagem de nível.

(b) Image captioning.

Question: Quantas pessoas estão na foto?
Answer (GT): 5
Answer (V-GlórIA): 13

Question: Estão virados para a câmara?
Answer (GT): sim
Answer (V-GlórIA): sim

Question: Estão a usar casacos?
Answer (GT): sim
Answer (V-GlórIA): sim

Question: Existem árvores visíveis?
Answer (GT): sim
Answer (V-GlórIA): branco

(c) Visual dialog.

Figure 3: V-GlórIA can solve core vision and language tasks.

First, we observe the same trend in which our model, V-GlórIA, using GlórIA 1.3B as its LLM backbone, consistently achieves superior performance, compared to the Gervásio LLM backbone. Second, it can be seen that the task is much more challenging on CC3M-PT, with all models obtaining a lower performance. These low BLEU scores on CC3M, might be explained by the fact that since CC3M captions are collected from the web, and not manually annotated like in MSCOCO, making them more prone to being unaligned with the image. This is evidenced in the first example of Table 1, where the caption mentions "actor and her daughter" which cannot be guessed from the picture. However, in MSCOCO, higher BLEU and METEOR scores are obtained. This is explained by three aspects: 1) the captions' lexicon diversity in MSCOCO is significantly lower when compared to CC3M, and 2) the connection between images and the captions is much tighter in MSCOCO, and 3) captions have a more predictable format. It should be noted that for MSCOCO, models are evaluated in a zero-shot setting, evidencing that V-GlórIAis capable of generalizing to unseen data.

When comparing a full English setup (gray lines) vs. a PT-PT model trained on PT-PT data, we observe that the former achieves higher performance in both datasets. Given the proximity of the image captioning task to the original LLM loss, and the fact that GPT-Neo 1.3B was pre-trained on a significantly larger text corpus, compared to GlórIA and Gervásio, this is not surprising, and we believe that this can be countered with an improved PT-PT LLM.

6.3 Visual Dialog Results

To assess our model performance on a more challenging vision and language task, we evaluate it on the Visual Dialog (VisDial) (Das et al., 2017) task, in zero-shot, in two different settings: **a)** IT2T (image and text to text) where given an image, a dialog about it, and a question, the model has to select the correct answer from a pool of 100 candidate answers, and **b)** T2I (text to image), where given a dialog about an image, the model has to retrieve the correct image. Given that V-GlórIA is an autoregressive decoder, we follow the protocol of (Koh et al., 2023) for IT2T, and given a question and answer sequence, we select the answer with the lowest perplexity, among the candidate answer options.

Table 4 shows the results. We observe that all models exhibit low performance, regardless of the PT-PT LLM backbone. Performance is, however, higher in T2I, compared to IT2T, which is consistent with the fact that the T2I task is closer to the vision and language tasks considered in training. Notwithstanding, V-GlórIA , using the GlórIA PT-PT LLM, demonstrates better generalization capabilities to new tasks, significantly outperforming the model using the Gervásio PT-PT LLM. We believe that part of these results can be dramatically improved by using a stronger PT-PT LLM. That is, despite the higher effectiveness of the GlórIA PT-PT LLM, it was not trained on instructions. This makes the model struggle when instructed to answer questions.

7 Conclusions

In this paper, we proposed a methodology to efficiently customize a Vision and Language LLM to European Portuguese. In particular, we introduced V-GlórIA, the first European-Portuguese Vision and Language model, capable of addressing multimodal tasks such as retrieval, image captioning, and visual dialogs illustrated in Figure 3. Experiments, leveraging current best performing open PT-PT LLMs as backbones, reveal performances that are competitive with the English counterpart setting (i.e. English pre-training and benchmarks), on these tasks. V-GlórIA demonstrated to be capable of generalizing to unseen data, especially in multimodal retrieval. For more challenging tasks, such as Visual Dialog, the proposed approach is still not on par with English models. However, we believe that as better PT-PT models arise, including instruction-tuned ones, the performance gap can be narrowed down by employed our devised customization methodology, and leveraging our contributed PT-PT data resources. We will release the PT-PT high-quality translations of the most popular V&L datasets to foster research in this area.

8 Ethical Considerations

This research presents a methodology for customizing and adapting vision and language models to European Portuguese. In alignment with principles of transparency and ethical responsibility, we exclusively utilized publicly available research datasets and benchmarks. No private or sensitive information, whether personal or proprietary, was used in this work.

Acknowledgements

This work has been partially funded by the FCT project NOVA LINCS Ref. UIDP/04516/2020, by CMU|Portugal project iFetch, Ref. CMUP LISBOA-01-0247-FEDER-045920, and by the Google Cloud Grant Ref. Nº CPCA-IAC/AV/594875/2023.

References

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Preprint*, arXiv:2011.15124.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gabriel Oliveira dos Santos, Diego Alysson Braga Moreira, Alef Iury Ferreira, Jhessica Silva, Luiz Pereira, Pedro Bueno, Thiago Sousa, Helena Maia, Nádia Da Silva, Esther Colombini, Helio Pedrini, and Sandra Avila. 2023. CAPIVARA: Cost-efficient approach for improving multilingual CLIP performance on low-resource languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 184–207, Singapore. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17283–17300. PMLR.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

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 *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ricardo Lopes, Joao Magalhaes, and David Semedo. 2024. GlórIA: A generative and open large language model for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 441–453, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International*

Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024. Advancing generative ai for portuguese with open decoder gervásio pt*. *Preprint*, arXiv:2402.18766.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018a. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018b. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Preprint*, arXiv:2008.00401.

Hugo Touvron, Louis Martin, and et. al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *Preprint*, arXiv:2004.11867.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.