Parallel Multilingual Pre-Training for Multi-Modal Representations

Nathaniel Krasner
George Mason University
nkrasner@gmu.edu

Nicholas Lanuzo
George Mason University
nlanuzo@gmu.edu

Abstract

Multi-modal representations are useful in many downstream tasks such as image-grounded caption generation and text-grounded image generation. The majority of the datasets available for training these representations are in English or are heavily skewed to certain languages. We extend the CLIP technique from Radford et al. (2021) by incorporating artificial parallel data from three additional diverse languages and find that this not only improves the multilingual performance of the CLIP model, but also improves its performance in English.

1 Introduction

Contrastive Language-Image Pre-training (CLIP) is a technique by Radford et al. (2021) for learning a shared Image-Language embedding space such that an embedded caption of an image should occur near the embedding for the image. Though there are several multilingual datasets with image-caption pairs for training CLIP models such as LIAON-5B (Schuhmann et al., 2022), these datasets skew heavily to certain languages. In the case of LIAON-5B, 10% of their data comes from Russian and the distribution quickly drops off with only a very small percentage left for languages originating outside of Europe and East Asia (Schuhmann et al., 2022).

We would like to see access to multi-modal languages technologies such as image captioning and grounded image generation brought to other regions with less task specific data. For this reason, we propose a simple technique for removing this imbalance from existing datasets.

2 Methods

2.1 CLIP Pre-training

The CLIP model aligns two encoders, one for images and one for text, into a single latent space. This is accomplished through a contrastive learning task with the goal of minimizing cosine distance between the matched image and text encodings while maximizing the cosine distance between all of the non-matching pairs of encodings (Radford et al., 2021). We needed the model to understand a diverse array of languages, so we used XLM-Roberta for our text encoder since it has been pre-trained with 100 languages (Conneau et al., 2020). The image encoder selection was less important since images are a language-agnostic medium. We used Google's vision transformer for image encoding, since it has been shown to produce SOTA results (Dosovitskiy et al., 2021).

Since these encoders project into differently sized spaces, we add an additional layer onto the end of each in order to match their projected dimensions. The encoders are pre-trained with language and image knowledge which we want to be careful not to heavily overwrite. For this reason, we freeze the pre-trained layers for the first half-epoch, to give the interface layers as head start at convergence. This freezing technique was also used by Bianchi et al. (2021).

As a baseline, we train our CLIP model on the Microsoft COCO dataset (Lin et al., 2015) which contains only English captions for each image.

2.2 Parallel Multilingual CLIP Pre-training

Using Google Translate, we produced translations for each of the captions in the COCO dataset (Lin et al., 2015) into Spanish, Japanese, and Hindi. We chose these languages since they cover a broad array of language families. During each step of training, the captions were encoded in parallel across all four languages. We only encode the image once and reuse this encoding for all languages. We treat this as a multi-class classification task and treat the four translations as the targets for cross entropy.

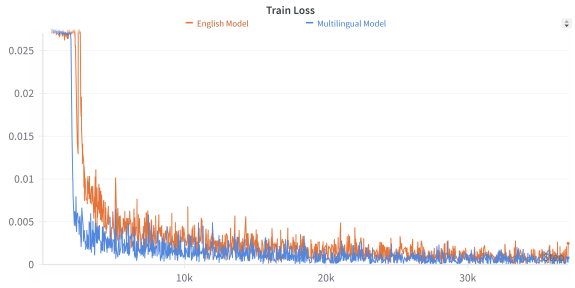

Figure 1: The training loss curves for the English and Multilingual models.

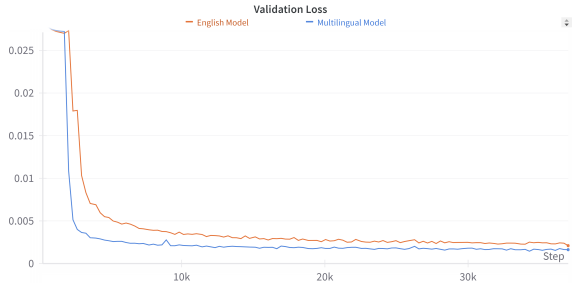

Figure 2: The validation loss curves for the English and Multilingual models.

3 Results

We trained both models for the same number of steps and with the same hyper parameters. Their training and validation loss curves can be seen in Figures 1 and 2. The multilingual model converged in fewer steps and reached a lower loss than the model trained with only English captions. This could possibly be due to the increased usage of other parts of the model. For example, the English only model may focus on updating the pre-trained English sub-network while the multilingual model may be updating a more broad or language agnostic sub-network.

To test the zero-shot performance of these models, we translated the validation data into an additional four languages which have same-family correspondents in the multilingual training set: German, French, Chinese, and Bengali. Since the pre-training of these models used a classification type task, we analyze how the model performs in a sort of classification. We encoded the captions and the images and then sort the images by their distance from the caption.

The multilingual model outperformed the English only model in all cases, even the English result. The zero-shot results are comparable but slightly lower than the results on seen languages.

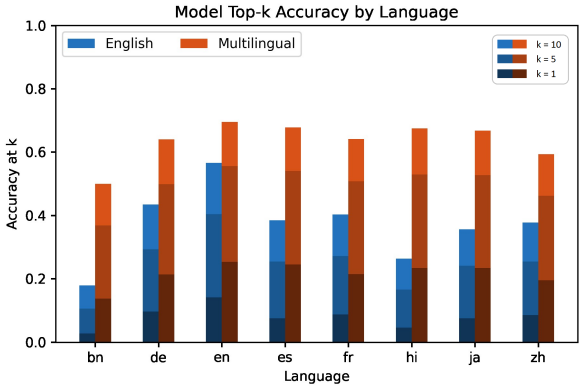

Figure 3: The accuracy@k per language for selecting the image given the caption. (Larger is better.)

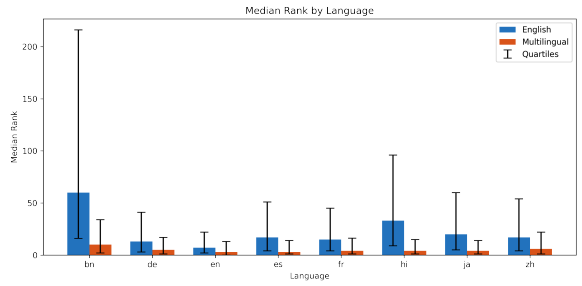

Figure 4: The median and quartiles for the ranks of images given the caption. (Smaller is better.)

4 Limitations

Unfortunately, COCO does not offer a test set, so we tested on the validation set. In the future, we would like to test this on a separate set of data. We only used the translated data in our testing. In future work, we would test with organic data which would fit a realistic setting. We also only tested the zero-shot performance on languages where the family was seen in training. Future work could add a few more languages from an unseen family.

5 Conclusion

Our initial motivation for this area of research was to make CLIP more accessible to other languages. Our hope was that training it on more languages would at least improve the results for those languages. We see that not only were our hopes met, but they were exceeded in that unseen language scores were also drastically increased. This shows that multilingual training for CLIP models can bring their capabilities to an even broader audience than we initially expected. In the future, more work could be done by adding even more languages to the training set to see how performance scales with an increased language count.

References

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pre-training for the italian language.

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.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

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.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models.