NLLB-CLIP – train performant multilingual image retrieval model on a budget

Alexander Visheratin
alexvish91@gmail.com

Abstract

Today, the exponential rise of large models developed by academic and industrial institutions with the help of massive computing resources raises the question of whether someone without access to such resources can make a valuable scientific contribution. To explore this, we tried to solve the challenging task of multilingual image retrieval having a limited budget of $1,000. As a result, we present NLLB-CLIP – CLIP model with a text encoder from the NLLB model. To train the model, we used an automatically created dataset of 106,246 good-quality images with captions in 201 languages derived from the LAION COCO dataset. We trained multiple models using image and text encoders of various sizes and kept different parts of the model frozen during the training. We thoroughly analyzed the trained models using existing evaluation datasets and newly created XTD200 and Flickr30k-200 datasets. We show that NLLB-CLIP is comparable in quality to state-of-the-art models and significantly outperforms them on low-resource languages.

1 Introduction and model description

Contrastive Language-Image Pre-Training (CLIP) [20] is a powerful architecture that allows achieving high-quality results on a variety of tasks, such as zero-shot classification, text-image, and image-text extraction. It uses vision and text transformers [9, 25] to encode information from images and texts into the common latent space. CLIP can be applied to new tasks without any fine-tuning [24, 13] or be extended to solve more complex tasks as semantic segmentation [16].

CLIP has been adapted to languages other than English, like Italian [2] and Chinese [28]. Numerous works also demonstrated that CLIP can be extended to multiple languages at once [3, 6]. But to date, there have been no capable CLIP-like models for low-resource languages [19, 12].

Recently, the "No Language Left Behind" (NLLB) model [7] was introduced to enable translation between more than 200 languages. The model shows great performance in both high- and low-resource languages. Notably, the model has encoder-decoder architecture and was trained in different sizes (600M, 1.3B, and 3.3B parameters) that allow a wider variety of deployments.

The core question we investigated is whether we can use a pre-trained text encoder from NLLB models to extend CLIP capabilities to the languages of the Flores-200 dataset and stay within a limited budget of $1,000. To answer this, we replaced the standard text encoder of the OpenAI CLIP [20] with the text encoder from the NLLB model. We left the rest of the model and loss functions the same as we wanted to understand the impact of this specific change.

For the experiments, we used various backbone models. For the image encoder, we used the original CLIP ViT base (denoted as "b" in experiments) and large ("l") and CLIP ViT huge ("h") trained by LAION. For the text encoder, we used respective parts of three NLLB variants – base ("b"), large ("l"), and huge ("h").

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Datasets

2.1 Training dataset

Since there are no image-text datasets available for all Flores-200 languages, we had to create a new, sufficiently large dataset to train the model. We used a random subset of the LAION COCO dataset[1] that contains automatically generated captions in MS COCO [15] style. We used the LAION aesthetic predictor model[2] to filter the images during the processing and preserved only the pictures with an aesthetic score higher than 4.5. The threshold score was obtained empirically by manually analyzing 500 scored images from the dataset. We aimed to collect enough data but not too much so that we would stay within our budget when training the models. As a result, we got 106,246 images. English captions were translated into 200 languages of the Flores-200 dataset using the NLLB-3.3B model. We left 15% of the dataset (15,937 samples) for validation. The result is the LAION-COCO-NLLB dataset [26], which we make publicly available. To the best of our knowledge, this is the largest image-text dataset in terms of languages covered. The key property of this dataset is that all 201 languages are presented equally, which greatly affects the model performance for low-resource languages, as shown in Section 5.

2.2 Evaluation datasets

We used two existing evaluation datasets for the experiments. (1) XTD10 [1] – 1,000 image-text pairs from COCO 2014 dataset translated into 10 languages. It extends previous works [21, 29] with 7 new languages. (2) Crossmodal-3600 [23] – 3,600 images annotated with captions in 36 languages. The main advantage of this dataset is that it covers many languages, including five low-resource ones – Bengali, Cusco Quechua, Maori, Swahili, and Telugu.

To test the performance of the models using all Flores-200 languages, we created two new datasets. (1) XTD200 – 1,000 English captions from XTD10 dataset translated to 200 languages using NLLB-3.3B model. (2) Flickr30k-200 – 1,000 English captions from the test part of the Flickr30k dataset translated to 200 languages using NLLB-3.3B model.

3 Training

Since we already have a high-quality pre-trained text encoder, unlike other works [3, 6], we performed only fine-tuning on the collected dataset. This allowed us to minimize training costs per model and run many experiments within a limited budget.

On each epoch of training, we used only one randomly selected caption per image. It makes the training epochs shorter, and we don't need to run validation multiple times within an epoch to see the progress. Since the ratio between the training set size and the number of languages is 447/1, we are confident that every language goes through the model in each epoch.

All experiments were performed on a single Nvidia H100 GPU with 80GB VRAM. Large memory capacity enabled using large batch sizes, which is crucial for contrastive loss [20]. We used Lion optimizer [5] as we found that it makes the model converge significantly faster and to better performance than weighted Adam [18]. To save the GPU memory on the optimizer state, we used an 8-bit version of Lion [8]. This allowed us to have a 25% larger batch size.

4 Experiments

4.1 Freezing different parts of NLLB-CLIP

In this series of experiments, we explored the effect of freezing different parts of the model on its performance. We tried three training regimes: (1) freeze nothing (denoted as "full"); (2) freeze only image encoder and train text encoder and projection layers ("text encoder"); (3) freeze both image and text encoders and train only projection layers ("projection"). We tested all combinations of image and text encoder sizes and found that freezing only the image encoder produces significantly better

[1] `https://laion.ai/blog/laion-coco/`

[2] `https://laion.ai/blog/laion-aesthetics/`

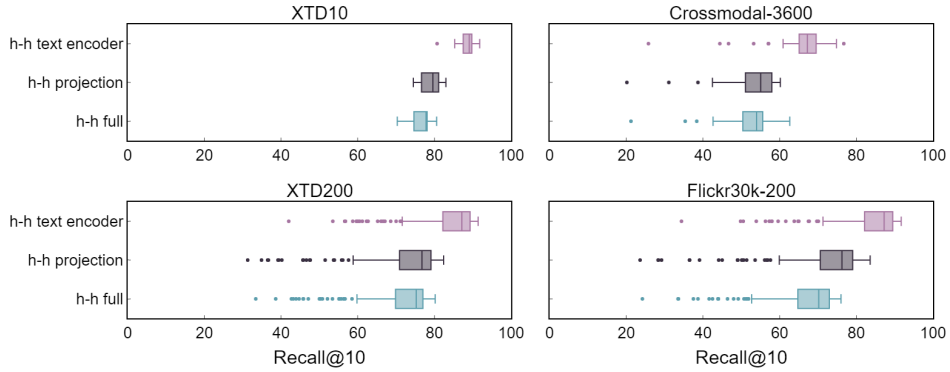

Figure 1: Performance on evaluation datasets for the same NLLB-CLIP model with different training regimes

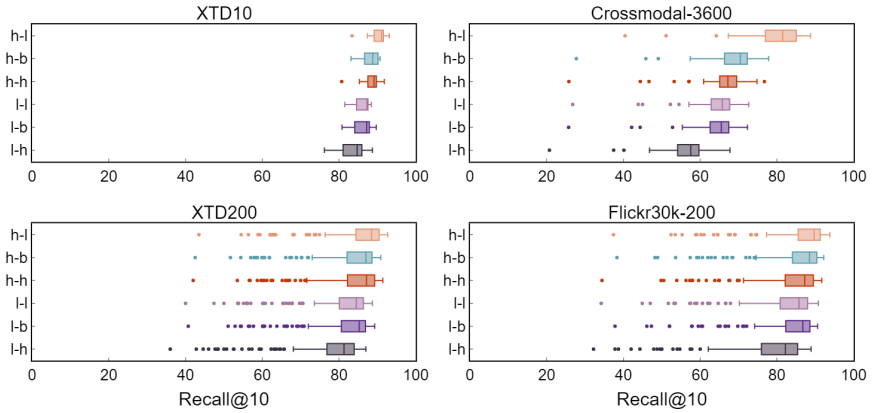

Figure 2: Performance on evaluation datasets for different combinations of image and text encoders

results in all cases. Interestingly, training only projection layers gives better performance than full training. Exemplar results for the model with CLIP ViT huge and NLLB huge (h-h) are presented in Figure 1. The smallest model variants (e.g., b-b and l-b) could not converge to get R@10 higher than 20% when we performed training of the full model.

Our results are consistent with the results from the LiT paper [31], where the authors found that freezing the image encoder is the best regime for fine-tuning the model for the new tasks. Also, freezing the image encoder makes even more sense for our task, where for the same image, there are 201 different captions in the training dataset. The best training scenario in this case is to use high-quality image representations from the pre-trained image encoder to adjust the text encoder and align the text representations in the same latent space. To make this process faster, we train both visual and text projection layers along with the text encoder backbone.

4.2 Encoder backbones size effect

In the next part of the experiments, we investigated the effect of backbone sizes on the model performance. For all models, we kept the image encoder frozen and trained only the text encoder with visual and text projection layers. We examined six models – l-b, l-l, l-h, h-b, h-l, and h-h (refer to Section 1 for abbreviation sources and respective backbone sizes). From the experiments (Figure 2), we can see that for all datasets, the models with larger text encoders perform worse than the ones with smaller text encoders. For the large image encoder, the base text encoder performs better than the large one.

The results demonstrate that a larger (and higher quality) image encoder enables better results for any text encoder. Regarding the worse performance of larger text encoders, we attribute this phenomenon to a lack of data to fully align the largest text encoder with the image encoder in the latent space. The

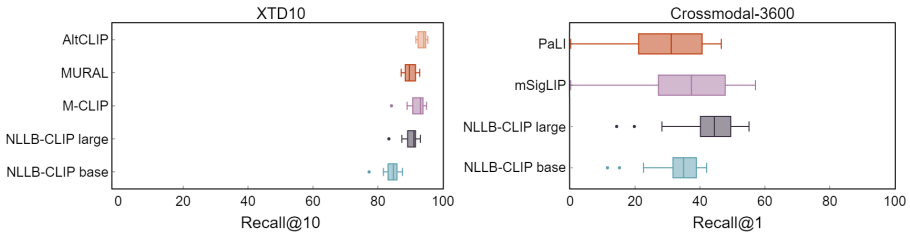

Figure 3: Evaluation results on multilingual datasets

need for a sufficient amount of data for model training has been discussed in the literature [22, 17, 11]. We plan to collect more data and train the models to validate this hypothesis.

5 Evaluation

For comparison with existing works, we used two variants of NLLB-CLIP: (1) CLIP ViT base with NLLB base – NLLB-CLIP base (501M parameters) and (2) CLIP ViT huge with NLLB large – NLLB-CLIP large (1.4B parameters). Both models were trained with an image encoder frozen. We chose NLLB-CLIP large because it is the best-performing variant across all experiments. NLLB-CLIP base was chosen to evaluate the capabilities of the smallest model.

5.1 XTD10

For the XTD10 dataset, we used Recall@10 as a comparison metric, as it is used in other works. We compared NLLB-CLIP with state-of-the-art models – Multilingual CLIP [3], MURAL [10], and AltCLIP [6]. Figure 3 (left) shows the results of the experiments. Although NLLB-CLIP did not outperform state-of-the-art models, the large model is not very far behind - 90.1% vs. 93.7% on average. It is worth mentioning that other models were trained on significantly larger datasets (up to 1,000x).

5.2 Crossmodal-3600

For the Crossmodal-3600 dataset, we used Recall@1 as a comparison metric, as it is used in other works. We compared NLLB-CLIP with state-of-the-art models – mSigLIP [32] and PaLI [4]. Figure 3 (right) shows the evaluation results. NLLB-CLIP large sets state-of-the-art results with 42.96% R@1 on average across 36 languages. mSigLIP outperforms NLLB-CLIP in high-resource languages like English, Italian, or Spanish. The contrary is true for lower- and low-resource languages – the advantage of NLLB-CLIP is higher the lower the resourcefulness of the target language. These results demonstrate the advantage of using the LAION-COCO-NLLB dataset, where all languages are represented equally, no matter their real-world resourcefulness.

6 Conclusion

In this paper, we demonstrate that by replacing the text encoder and fine-tuning on a small automatically created dataset, we can create a CLIP model capable of high-quality image retrieval in 201 languages of the Flores-200 dataset. NLLB-CLIP performs better than existing models on low-resource languages, primarily because the training dataset has the same number of captions for both low- and high-resource languages. NLLB-CLIP large sets new state-of-the-art results on the Crossmodal-3600 dataset that includes low- and high-resource languages.

Acknowledgments and Disclosure of Funding

We thank Lambda [3] for providing compute credits for running data collection and experiments.

We thank the ML Collective community for helpful discussions, ideas, and feedback on experiments.

[3] `https://lambdalabs.com/`

References

[1] P. Aggarwal and A. Kale. Towards zero-shot cross-lingual image retrieval, 2020.

[2] F. Bianchi, G. Attanasio, R. Pisoni, S. Terragni, G. Sarti, and S. Lakshmi. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*, 2021.

[3] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, 2022.

[4] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[5] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, Y. Liu, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023.

[6] Z. Chen, G. Liu, B.-W. Zhang, F. Ye, Q. Yang, and L. Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022.

[7] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[8] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*, 2022.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] A. Jain, M. Guo, K. Srinivasan, T. Chen, S. Kudugunta, C. Jia, Y. Yang, and J. Baldridge. Mural: Multimodal, multitask representations across languages. In *Findings of the Association for computational Linguistics: EMNLP 2021*, pages 3449–3463, 2021.

[11] Y. Ji, Y. Deng, Y. Gong, Y. Peng, Q. Niu, L. Zhang, B. Ma, and X. Li. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*, 2023.

[12] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*, 2020.

[13] J. Li, G. Shakhnarovich, and R. A. Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022.

[14] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[16] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, and X. He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023.

[17] T. Linjordet and K. Balog. Impact of training dataset size on neural answer selection models. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 828–835. Springer, 2019.

[18] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[19] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Kolawole, T. Fagbohungbe, S. O. Akinola, S. H. Muhammad, S. Kabongo, S. Osei, et al. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*, 2020.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 171–181, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1021. URL `https://aclanthology.org/N16-1021`.

[22] D. Soekhoe, P. Van Der Putten, and A. Plaat. On the impact of data set size in transfer learning using deep neural networks. In *Advances in Intelligent Data Analysis XV: 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, Proceedings 15*, pages 50–60. Springer, 2016.

[23] A. Thapliyal, J. Pont-Tuset, X. Chen, and R. Soricut. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *EMNLP*, 2022.

[24] V. Thengane, S. Khan, M. Hayat, and F. Khan. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*, 2022.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[26] A. Visheratin. Laion-coco-nllb dataset, 2023. URL `https://huggingface.co/datasets/visheratin/laion-coco-nllb`.

[27] C. Xie, H. Cai, J. Song, J. Li, F. Kong, X. Wu, H. Morimitsu, L. Yao, D. Wang, X. Zhang, D. Leng, X. Ji, and Y. Deng. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework, 2022.

[28] A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, and C. Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

[29] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2066. URL `https://aclanthology.org/P17-2066`.

[30] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[31] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[32] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.