Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6848–6854
Marseille, 20-25 June 2022

Cross-lingual and Multilingual CLIP

Fredrik Carlsson[⋆], Philipp Eisen[◦], Faton Rekathati[⊗], Magnus Sahlgren[•]
[⋆]RISE, [◦]Depict, [⊗]Kungliga Biblioteket, [•]AI Sweden
Fredrik.Carlsson@ri.se, Philipp@Depict.ai, Faton.Rekathati@kb.se Magnus.Sahlgren@ai.se

Abstract

The long-standing endeavor of relating the textual and the visual domain recently underwent a pivotal breakthrough, as OpenAI released CLIP. This model distinguishes how well an English text corresponds with a given image with unprecedented accuracy. Trained via a contrastive learning objective over a huge dataset of 400M of images and captions, it is a work that is not easily replicated, especially for low resource languages. Capitalizing on the modularization of the CLIP architecture, we propose to use cross-lingual teacher learning to re-train the textual encoder for various non-English languages. Our method requires no image data and relies entirely on machine translation which removes the need for data in the target language. We find that our method can efficiently train a new textual encoder with relatively low computational cost, whilst still outperforming previous baselines on multilingual image-text retrieval.

1. Introduction

Bridging the gap between Computer Vision and Natural Language Processing (NLP) has long been a goal within Artificial Intelligence(AI) research. Recently, OpenAI released CLIP (Radford et al., 2021) which pushed State-Of-The-Art in multimodal text and image representations. CLIP has since received mainly positive attention from the research community, as it has been proved useful for a large variety of different tasks (see Section 2.1.).

However, CLIP and the majority of related work focus exclusively on English. This has created a performance vacuum for other low-resource languages, where there is often a lack of high-quality data. A problem further exacerbated by the computational resources, and hence cost, required to train large high-performing models. This language bias is a common trend in current AI research and is the cause of much concern in the discourse regarding fairness and inclusivity (Bender et al., 2021).

To mitigate this problem we propose a teacher learning approach where we train a non-English language encoder for CLIP. Our approach capitalizes on the clear text and image modularization of the CLIP architecture, where the language and vision encoder is only connected via the loss function. This allows us to discard the CLIP vision encoder during training, and train the student encoder to mimic the original CLIP encoder when given language parallel data. Our approach relies entirely on machine translation, hence effectively side-stepping the data required in the target language. Finally, our approach is significantly less computationally demanding than the original CLIP pretraining.

To evaluate our approach we focus on Image-Text retrieval. We train a multilingual encoder in multiple languages simultaneously, along with a Swedish-only encoder. Our multilingual CLIP encoder outperforms previous baselines in 11 languages, and the monolingual Swedish model outperforms its multilingual counterpart for Swedish. Finally, we vary the number of data examples for the Swedish-only encoder, and find that the number of translated captions has a noticeable effect on performance.

The source code and pre-trained models are available at the following link: `https://github.com/FreddeFrallan/Multilingual-CLIP`

2. Related Work

2.1. CLIP

CLIP incorporates two disconnected encoder models, one for text data and one for image data. The text encoder is based on the Transformer (Vaswani et al., 2017) architecture, and the visual encoder is based on ResNet (He et al., 2016), or Image Transformer (Parmar et al., 2018). Both encoders output a fixed size embedding that is compared via a Contrastive Loss (Hadsell et al., 2006), where the goal is to maximize the cosine-similarity of embeddings from matching image-text pairs while minimizing the cosine-similarity of non-matching image-text pairs.

Applying this method to a large dataset with vast computational resources resulted in an expressive joint image-text representation space. Although the initial main contribution was image classification without predefined labels, CLIP representations have since proved useful for various other tasks such as Image-Text retrieval (Radford et al., 2021), Visual Question-Answering (Shen et al., 2021), and Automatic Image Captioning (Mokady et al., 2021). Notably, OpenAI decided to only release the smaller versions of CLIP, meaning that any succeeding work (including this work) only works with smaller CLIP models.

2.2. Multilingual Image-Text Retrieval

Multilingual image-text retrieval has seen significantly less attention than its English-only counterpart. This is most likely due to the limited number of languages for which there exist high-quality image-text datasets.

Aside from CLIP, previous work has investigated using language-aligned word embeddings and a multimodal contrastive learning objective, to train a language encoder and an image encoder (Portaz et al., 2019). Also pre-training a single BERT (Devlin et al., 2019) like model, using both images and texts from multiple languages (Fei et al., 2021). Finally, there has been work that fine-tunes multilingual sentence embedding models, such as mUSE (Yang et al., 2020) and LASER (Artetxe and Schwenk, 2019) to match a pre-trained vision embedding space (Aggarwal et al., 2021).

Unfortunately, the majority of the mentioned methods have not released their trained models, hindering us from evalu-

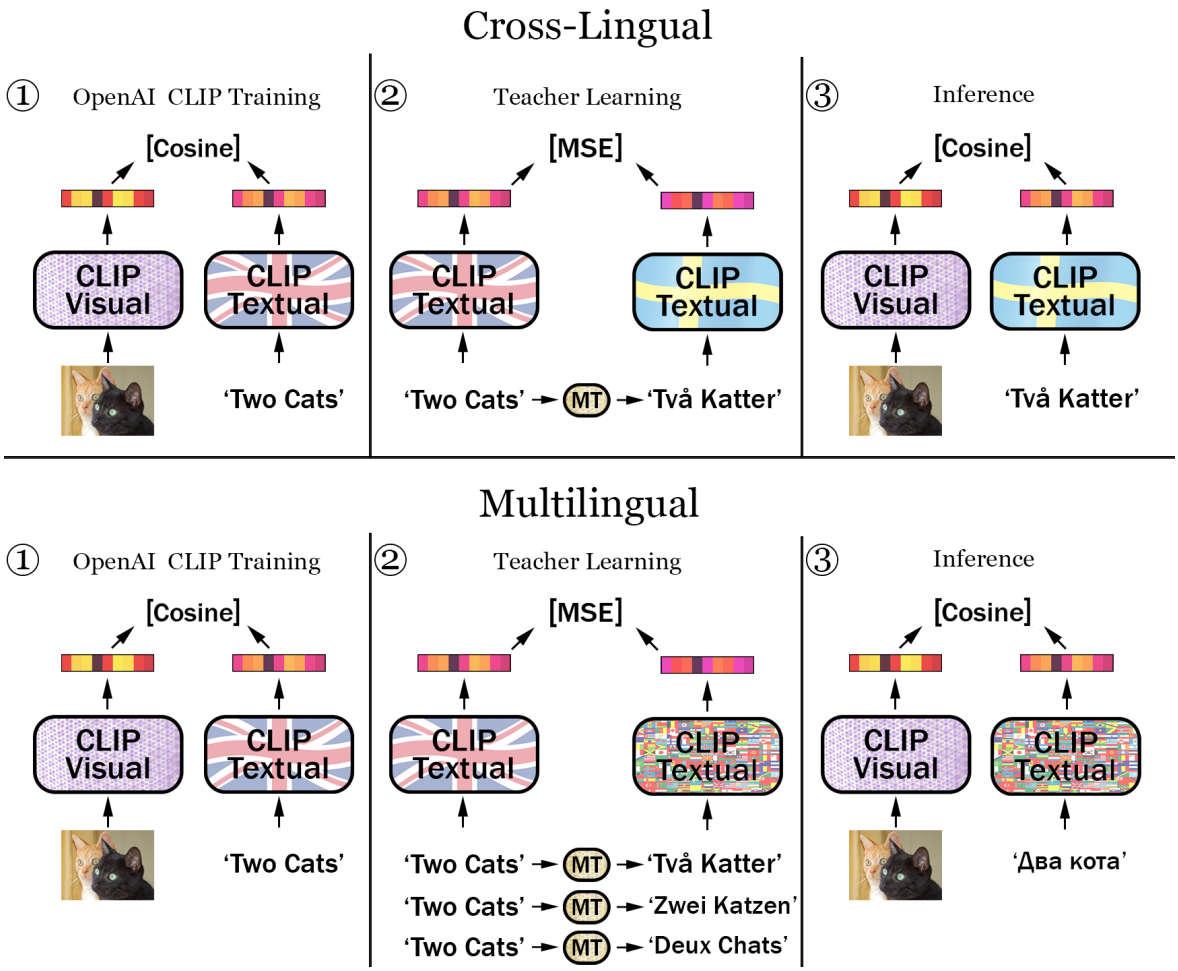

Figure 1: Overall training procedure. The original CLIP training in step number 1 is to train the teacher model, and is the most expensive. In step number 2 we temporarily replace the CLIP visual encoder and train a student model to mimic the English CLIP encoder. At the inference step 3 we now use the student model together with the CLIP visual encoder.

ating against them on recently released datasets.

2.3. Teacher Learning

Teacher Learning (Hinton et al., 2015) is a domain agnostic Machine Learning method for transferring the knowledge of an already trained teacher model into a new student model. Although often used to train a smaller and efficient student model (Sanh et al., 2019; Tang et al., 2019), previous work in NLP has used utilized this method for other ends. Such as cross-lingual teacher learning of Sentence Embeddings (Reimers and Gurevych, 2020) and multimodal transfer learning of Text-To-Speech (Jiang et al., 2021). To the extent of our knowledge, there has however been no previous work investigating cross-lingual teacher learning in a multimodal setting.

3. Method

Working from the assumption that the original training of the CLIP text and image encoders has led them to produce similar embeddings for matching text-image pairs, we can ignore the image encoder and only focus on mimicking the CLIP text encoder. This effectively alleviates the need to incorporate any images into the training loop. Instead, we apply teacher learning directly between the original English text encoder, and a student model pre-trained in a different language.

Using language parallel data, created via machine translation, the student model is trained to generate matching embeddings to that of the teacher model. The teacher CLIP text encoder is kept unaltered, and only the parameters of the student language encoder are updated during training. The overall training procedure is summarized in Figure 1, and additional details regarding hyperparameters are available in Appendix A.

3.1. Text Encoder

All investigated text encoders are BERT transformer models, pre-trained in either a monolingual or multilingual setting. This is unlike the original CLIP paper, which found no performance gain by using pre-trained language models. However, the original work utilized a far larger dataset and computational resources (see Section 5.3.).

Following previous work on sentence embeddings using Transformer language models, we generate fixed-size text embedding by mean pooling the model's output tokens. (Reimers and Gurevych, 2019; Wang and Kuo, 2020; Carlsson et al., 2021). Additionally, we apply a linear transformation to this representation to match the dimensionality

	En	De	Fr	It	Es	Ru	Ja	Zh	Pl	Tr	Ko
mUSE PATR	83.6	71.2	75.6	76.9	76.1	73.4	64.3	73.6	71.8	66.9	69.4
mUSE m3	85.3	73.5	78.9	78.9	76.7	73.6	67.8	76.1	71.7	70.9	70.7
CLIP RN50x4	90.2	52.4	64.4	48.1	55.8	3.8	4.3	3.1	14.1	12.9	2.6
CLIP ViT-B	**90.3**	43.7	57.8	40.9	51.0	3.1	13.9	4.5	11.9	9.3	1.8
Our contributions											
M-CLIP RN50x4	**90.3**	85.8	87.1	**89.1**	**88.3**	84.0	81.2	86.6	**88.3**	**87.6**	84.7
M-CLIP ViT-B	90.1	**86.5**	**87.7**	88.5	87.6	**84.5**	**82.6**	**87.7**	88.3	86.5	**85.1**

Table 1: Text-To-Image recall@10 for the XTD dataset.

of the image encoder. During training, both the model's parameters and the linear transformation are updated.

3.2. Cross-lingual Teacher Learning

The overall teacher learning can be summarized as feeding the teacher and student model language parallel texts, and minimizing the Mean Squared Error (MSE) between their respective output embeddings. Notably, this differs from the original CLIP training objective, where the model is instead trained to correlate the cosine-similarity between image-text pairs. But while it's possible to use cosine-similarity directly in the teacher learning, previous work has found that minimizing MSE provides a richer learning signal (Carlsson and Sahlgren, 2021).

$$\begin{aligned} \text{Translation}(X) &= X^* \\ \text{Teacher}(X) &= E_T \\ \text{Student}(X^*) &= E_S \\ Loss &= MSE(E_T,\ E_S) \end{aligned} \tag{1}$$

For completeness, the full training objective is formalized in equation 1. First, a set of captions X are machine translated into the target language, creating X^*. The teacher CLIP model encodes the original texts X and generates the set of embeddings E_T. The student model encodes the translated texts X^* and generates E_S. Finally, the loss is calculated as the mean squared error between E_T and E_S.

4. Model Training

4.1. Training Data

We create a text corpus by compiling captions from multiple English datasets. Although our method could hypothetically work with any text data, the intuition is to have text data that is strongly linked with the visual domain. These English captions are then translated into their target language using the corresponding MarianMT (Junczys-Dowmunt et al., 2018) model.

Our corpus is comprised of the training data from MS-COCO (Lin et al., 2014), Google Conceptual Captions(GCC) (Sharma et al., 2018), and VizWiz (Bigham et al., 2010). The size of each dataset is available in Table 2.

4.2. Model Versions

Starting from a pre-trained Multilingual-BERT (M-BERT) (Devlin et al., 2019), we train one model against the CLIP ViT-B vision encoder, and one model against the CLIP RN50x4 encoder. Resulting in two different M-CLIP (Multilingual-CLIP) models. Although the original model M-BERT was trained towards 101 languages we limit ourselves to 68 languages (see Appendix B). Throughout the training, we uniformly sample captions and which language to translate them to.

Additionally, we train two versions of a Swedish-only encoder by starting from the Swedish KB-BERT (Malmsten et al., 2020). For one version we limit the number of sampled captions to 500k, and for the other, we limit the number of sampled captions to 2M.

5. Experiments

To test the proficiency of our method we evaluate on the task of image-text retrieval in multiple different languages. This is achieved by encoding the images using the corresponding official OpenAI vision encoder, and embedding the captions with our non-English text encoder. The encoded images and texts are then ranked in descending order according to their cosine-similarity, and the goal is thus to minimize the rank between an image and its corresponding text.

5.1. Multilingual Image Retrieval

To evaluate our multilingual models we use the recent XTD dataset (Aggarwal et al., 2021), which provides translation of the 1K MS-COCO test set for 11 different languages. MS-COCO, and hence XTD, provide 1 caption per image, and the task is to retrieve the matching image given that caption. Following convention, we report Recall@10 and compare against the baseline mUSE trained with both a Positive aware Triplet Ranking Loss(PATR) and mUSE trained with a Multi-modal Metric Loss (M3L) (Aggarwal et al., 2021). The results of these experiments are is available in Table 1.

Most notably, both M-CLIP models outperform the previous baselines on all languages, often by a large margin. This is interesting considering that the M-CLIP models are trained in 69 languages, unlike the baselines which specialize directly in the 11 evaluation languages. Discounting the original English CLIP, Japanese is the language where all

Dataset	Quantity	Percentage
MS-COCO	118 k	3.4
GCC	3.3 M	95.9
VizWiz	23.4 k	0.68

Table 2: Training data distribution

			Image2Text				Text2Image			
Text Encoder	Vision Encoder	Language	Median	R@1	R@5	R@10	Median	R@1	R@5	R@10
OpenAI	ViT-B/32	En	1	82.3	95.2	97.5	1	61.4	85.9	91.7
OpenAI	RN50x4	En	1	84.8	96.6	98.9	1	64.6	87.3	92.7
Our contributions										
M-BERT	ViT-B/32	En	1	77.2	93.5	97	1	55.7	82.4	89.4
M-BERT	RN50x4	En	1	67.6	92.1	96.8	1	57.6	82.8	88.9
M-BERT	ViT-B/32	Sv	1	72.3	90.6	95.3	2	48	75.1	83.7
M-BERT	RN50x4	Sv	1	56.5	87.3	94.3	1	50.2	75.6	83.5
KB-BERT 500k	RN50x4	Sv	1	69.9	92.9	96.9	1	55.7	81.4	88.1
KB-BERT 2M	RN50x4	Sv	1	76.5	94.4	97.6	1	62.8	85.4	91.4

Table 3: Unilingual Flickr30 image-text retrieval results.

models performed the worst, and English is the language where all models performed the best.

There does not seem to be any clear winner between the two M-CLIP models, as they perform nearly equally well in all languages. Finally, we note that there is nearly no difference in performance in English when compared to the original CLIP models.

5.2. Unilingual Text-Image Retrieval

Unfortunately, there exists no image-caption dataset for Swedish. Therefore we translate the test data of Flickr30K (Young et al., 2014) using MarianMT, and using the test set split of (Plummer et al., 2017). Unlike MS-COCO, Flickr30K provides 5 captions per image, and the task is both to retrieve images from text, and vice-versa. Following the convention, we report the median recall along with the recall at the thresholds 1, 5, and 10. These results are available in Table 3.

Starting by comparing the M-CLIP against the original CLIP in English, we find that M-CLIP performs noticeably worse, although R@10 remains fairly close. This is unlike the scores seen in Section 5.1., and seemingly indicates that increasing the number of captions per image aggravates the M-CLIP performance.

Interestingly, the Swedish model trained with 2M captions outperforms M-CLIP using the original English captions, and performs on par with the original CLIP model in the Text-To-Image setting. The Swedish model trained with 500k captions performs more similarly to M-CLIP for Swedish, but significantly worse than its 2M counterpart.

Finally, we note that all models using the ViT-B/32 encoder perform better than the models using RN50x4 on Image-To-Text, and worse on Text-To-Image.

5.3. Computational Comparison

As detailed in the hyperparameter section in Appendix A, we train each model with a batch size of 64, and perform 53772 update steps. The number of update steps was chosen so that the M-CLIP models, which don't limit the number of unique samples, would perform 1 full epoch of the dataset. This means that for each of our models, there were exactly 3,441,408 samples propagated through the training process.

The original CLIP training used a dataset consisting of 400M image-text pairs, where each model trained for 32 epochs. This entails that the original CLIP training propagated in total 12.8 billion samples. As displayed in Table 4 this means that the original CLIP training performed 3719 times more computations.

Finally, a very important distinction between the original CLIP training and our approach is the batch size, and sample independence within each batch. The contrastive CLIP loss compares each sample in the batch against all other samples within the batch. This entails that unlike most loss functions (including ours), the samples in each CLIP batch are not independent, making it expensive to compute large batch sizes. Notably, CLIP was trained using a large batch size of 32,768 samples.

6. Discussion & Conclusion

We have presented a method for training new text encoders for existing CLIP models using a teacher learning setup that utilizes machine translation. Results for image-text retrieval for multiple languages indicate that one can efficiently train new text encoders by starting from pre-trained language models. Utilizing a model pre-trained in a multilingual setting we outperform previous benchmarks on the newly released XTD dataset. A hint that machine translation is becoming a ripe technology, and a potential tool for overcoming language bias.

In our experiments, we notice that the number of unique captions have a noticeable effect on the final results. Indicating that future work could most likely achieve better results by increasing this quantity. Other hyperparameters such as batch size and the size of the text encoder are also interesting investigations for future work.

It is our belief that cutting-edge AI should be equally available to people, independent of what language they speak. And indeed, the models presented here were released in the spring 2021 and have, as of writing this paper, been downloaded roughly two million times, and amassed close to 300 Github stars.

	CLIP	Our Method	Factor
#Samples	12.8 B	$\sim$3.44 M	$\sim$3719
Batch Size	32,768	64	512

Table 4: Training statistics comparison between the original CLIP training and our method. Factor denotes how many times larger the corresponding CLIP value is.

7. Bibliographical References

Aggarwal, P., Tambi, R., and Kale, A. (2021). Towards zero-shot cross-lingual image retrieval and tagging. *ArXiv*, abs/2109.07622.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. (2010). Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, page 333–342, New York, NY, USA. Association for Computing Machinery.

Carlsson, F. and Sahlgren, M. (2021). Sentence embeddings by ensemble distillation.

Carlsson, F., Gyllensten, A. C., Gogoulou, E., Hellqvist, E. Y., and Sahlgren, M. (2021). Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Fei, H., Yu, T., and Li, P. (2021). Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3650, Online, June. Association for Computational Linguistics.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. cite arxiv:1503.02531Comment: NIPS 2014 Deep Learning Workshop.

Jiang, Y., Sharma, B., Madhavi, M., and Li, H. (2021). Knowledge distillation from bert transformer to speech transformer for intent classification.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, L. (2014). Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of sweden – making a swedish bert.

Mokady, R., Hertz, A., and Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). Image transformer. In Jennifer Dy et al., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4055–4064. PMLR, 10–15 Jul.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93.

Portaz, M., Randrianarivo, H., Nivaggioli, A., Maudet, E., Servan, C., and Peyronnet, S. (2019). Image search using multilingual texts: a cross-modal learning approach between image and text.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November. Association for Computational Linguistics.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. (2021). How much can clip benefit vision-and-language tasks? *ArXiv*, abs/2107.06383.

Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., and Lin, J. (2019). Distilling task-specific knowledge from bert into simple neural networks.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones,

L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, B. and Kuo, C.-C. (2020). Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP:1–1, 07.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strope, B., and Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online, July. Association for Computational Linguistics.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). 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.

A Hyperparameters

We train all models using the same hyperparameters and for 53772 update steps. The number of update steps was chosen so that the M-CLIP models, which don't limit the number of unique samples, would perform 1 full epoch of the dataset. The Swedish-only models thus perform the same amount of weight updates, but do this over a smaller number of samples. Finally, we used the Adam optimizer with a final learning rate of 5^-5, and a linear warm-up schedule for the first 1000 updates.

B Multilingual Training Languages

The 68 languages included in the M-CLIP training are available in Table 5.

Afrikaans
Albanian
Amharic
Arabic
Armenian
Azerbaijani
Bengali
Bosnian
Bulgarian
Catalan
Chinese Simplified
Chinese Traditional
Croatian
Czech
Danish
Dutch
English
Estonian
Finnish
French
Georgian
German
Greek
Gujarati
Haitian
Hausa
Hebrew
Hindi
Hungarian
Icelandic
Indonesian
Italian
Japanese
Kannada
Kazakh
Korean
Latvian
Lithuanian
Macedonian
Malay
Malayalam
Maltese
Mongolian
Norwegian
Persian
Polish
Pushto
Portuguese
Romanian
Russian
Serbian
Sinhala
Slovak
Slovenian
Somali
Spanish
Swahili
Swedish
Tagalog
Tamil
Telugu
Thai
Turkish
Ukrainian
Urdu
Uzbek
Vietnamese
Welsh

Table 5: Languages included in the M-CLIP training procedure.