Cross-Lingual Representation Alignment Through Contrastive Image-Caption Tuning

Nathaniel Krasner, Nicholas Lanuzo, Antonios Anastasopoulos
Department of Computer Science, George Mason University; Fairfax, VA.
Correspondence: nkrasner@gmu.edu

Abstract

Multilingual alignment of sentence representations has mostly required bitexts to bridge the gap between languages. We investigate whether visual information can bridge this gap instead. Image caption datasets are very easy to create without requiring multilingual expertise, so this offers a more efficient alternative for low-resource languages. We find that multilingual image-caption alignment can implicitly align the text representations between languages, languages unseen by the encoder in pretraining can be incorporated into this alignment post-hoc, and these aligned representations are usable for cross-lingual Natural Language Understanding (NLU) and bitext retrieval.[1]

1 Introduction

Encoder language models are very popular and widely used for extracting semantic information from text to be used downstream for natural language understanding (NLU) tasks. In general, an encoder language model (LM) is pretrained on a large corpus using self-supervision and then a smaller component is fine-tuned on annotated data using the representations produced by the pretrained LM. For widely spoken data-rich languages, this is no problem and the existence of task-specific, annotated data is a given (Joshi et al., 2020; Blasi et al., 2022). For low-resource languages this is rarely the case, and collecting data for each task in such languages is expensive and time consuming. Thus, cross-lingual knowledge transfer is a more practical direction for low-resource languages than further data collection.

The internal representations of encoder models trained on multilingual data tend to be disjoint, so the representation of a sentence in language A may not be similar to the representation of its translation in language B. Most likely, this is the result of pretraining data imbalance and domain mismatch across the languages included in their pretraining. If these internal representations were aligned such that representations of translations *were* similar, cross-lingual transfer for NLU tasks should be much easier to achieve, as Hu et al. (2021) showed. This cross-lingual transfer of task knowledge can greatly benefit speakers of low-resource languages by giving them access to NLP tools without the difficulty of annotating task-specific data in their language. As an additional benefit, these aligned representations can be used to mine bitexts from large scraped corpora to build parallel translation datasets (Team et al., 2022).

In this work, we explore whether one could encourage multilingual representation alignment *without any parallel data*, by relying instead on images as shared modality across languages. This is a worthwhile direction to pursue for two reasons. First, parallel text curation through expert translation is time-consuming, expensive, and requires bilingual annotators. In contrast, it is easy for an annotator to describe an image to produce a caption regardless of which language(s) they speak (Madaan et al., 2020). Second, language documentation efforts often produce media accompanied with monolingual audio or text in the language of interest. Developing techniques which leverage such materials could enable the creation of technologies for these otherwise under-served languages.

To summarize, we (1) show that a multilingual text-image contrastive learning setup can produce multilingually aligned text representations; (2) focus specifically on Quechua, as an example of a language unseen during pretraining that may benefit from such approaches; and (3) show that the addition of an unseen language does not degrade representation quality in other languages.

[1] Data and code will be publicly released at `https://github.com/nkrasner/cl-clip-align`.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1193–1199
July 27 - August 1, 2025

2 Related Work

Previous endeavors in multilingual alignment in the absence of parallel-text supervision have predominantly concentrated on the alignment of static word-embeddings through adversarial techniques (Zhang et al., 2017; Chen and Cardie, 2018). Approaches that extend multilingual alignment to sentence-level representations have generally necessitated a bitext signal (Feng et al., 2022; Escolano et al., 2021; Artetxe and Schwenk, 2019), with limited exceptions employing adversarial methodologies (Aghajanyan et al., 2019; Tien and Steinert-Threlkeld, 2022). Even though multilingual alignment may extend to languages not encountered during fine-tuning (Tien and Steinert-Threlkeld, 2022), we hypothesize that a more direct fine-tuning strategy using some pivot (even if not textual) could potentially produce superior alignment for languages with limited bitext resources.

Contrastive methods have been used for text-text (Feng et al., 2022) encoder alignment as well as text-image encoder alignment in both monolingual (Radford et al., 2021) and multilingual (Muraoka et al., 2023; Bianchi et al., 2023) settings. One such text-image alignment work introduces an image representation into the input sequence of NLU tasks leading to improved cross-lingual transfer (Muraoka et al., 2023). This offers additional support to our hypothesis that visual information can act as a semantic bridge between languages.

3 Method, Experiments, and Results

Our approach strings together a text encoder with a vision encoder. These two produce representations for each modality input, which are then used in a contrastive learning setup. In particular, given pairs of image representations E_i and caption representations E_c we use the following, simple contrastive loss function:

$$S = E_c \cdot E_i^\top * t$$
$$L(E_i, E_c) = \text{CrossEntropy}(S, I),$$

where I is the identity matrix and t is a learned temperature parameter.

This is similar to what CLIP (Radford et al., 2021) used for text-image alignment and LaBSE (Feng et al., 2022) for text-text alignment.

3.1 Experimental Setup

Datasets We work with the MS-COCO dataset (Lin et al., 2014), which provides 118k

Figure 1: A demonstration of the data sampling methods. Orange boxes highlight how our multi-modal approaches sample data. Blue boxes highlight how the `Eng-Pivot` approach samples data.

English Image-Caption pairs. Using Google Translate, we translate the English captions into Spanish, Japanese, Hindi, and Quechua. From this 5-way parallel image caption dataset, we derive 4 datasets for various experiments:

1. `Eng-Pivot`: The English captions from MS-COCO paired with one translation each from a rotation of Spanish, Japanese, and Hindi.
2. `Eng-only`: The English MS-COCO dataset without translations to other languages.
3. `Multilingual`: The MS-COCO dataset but each caption is from a rotation of English, Spanish, Japanese, and Hindi with only one language paired with each image.
4. `Multilingual+Quechua`: The same as the `Multilingual` dataset but with Quechua added into the rotation of languages.

While most of these datasets are designed for use with text-image alignment, the `Eng-Pivot` dataset is used for text-text alignment to create a model similar to LaBSE (Feng et al., 2022) with a comparable data size to our other models. This is the only dataset which contains parallel text data. Figure 1 gives a visual representation of this distinction.

Training We fine-tune an XLM-Roberta-Large (XLM-R) (Conneau et al., 2020) text encoder and a VIT-Base-patch16-224-in21k (Dosovitskiy et al., 2021) image encoder for 10 epochs with early stopping.

The token-level representations are mean pooled to create a sentence-level representation. Since the hidden dimensions of these encoders do not match, we add a linear layer to their outputs to adapt them to a matching dimensionality of 512. Following existing approaches to text-image alignment under these circumstances (Bianchi et al., 2023), we allow these linear layers to warm up for a certain number of steps before fine-tuning the encoders themselves. In our case, we chose to begin training the encoders halfway through the first epoch since the learning curves had flattened out by that point.

3.2 Experiment 1: Does multilingual text-image alignment lead to text-text alignment?

We hypothesize that text-image alignment involving multiple languages will implicitly align text representations between languages.

With the exception of the `Eng-Pivot` encoder (which is trained on bitext alignment), our encoders are only fine-tuned to align the text representations to the image representations, but we evaluate them on their alignment between text representations. Specifically, we use the FLoRes-200 dataset (Team et al., 2022), which contains 1012 204-way parallel sentences including all of our test languages. We perform a formal analysis using the task of bitext retrieval (Heffernan et al., 2022; Duquenne et al., 2023) as well as a visual analysis via t-SNE.

We compare against a baseline of the off-the-shelf XLM-R encoder, as well as one fine-tuned on text-image alignment using the English only (`Eng-Only`) dataset, and another trained directly on contrastive text-text alignment with an English pivot similarly to LaBSE (Feng et al., 2022).

For each sentence in each language, we search the English sentences in FLoRes-200 for the minimum cosine distance to find the corresponding English translation. If the true translation is selected, we count that sentence as correct. We calculate the retrieval accuracy over each language and then aggregate using the mean over all languages to produce a final score. Since XLM-R has not seen all of these languages in pretraining, we report the retrieval accuracy over the disjoint subsets of languages on which it was pretrained (or not). Table 1 contains the results.

While not quite matching the `Eng-Pivot` text-text aligned encoder, the `Multilingual` text-image aligned encoder is still very capable in the bi-text retrieval task. The `Eng-Only` text-image alignment improves on the abysmal results of the plain XLM-R model, but does not compare with

Encoder	All (203 langs)	in XLM-R (92 langs)	not in XLM-R (111 langs)	Quechua
XLM-R	0.5	0.6	0.4	0.5
Eng-Pivot	**62.2**	**92.6**	**37.1**	13.1
Eng-Only	18.3	27.5	10.7	7.2
Multilingual	55.7	82.2	33.7	18.0
+ Quechua	50.4	76.6	28.6	**29.2**

Table 1: Bitext retrieval accuracy on All of FLoRes-200, on the subset of languages in/not in XLM-R's pretraining, and just on Quechua.

the `Multilingual` alignment. This is likely because the pretraining of XLM-R does not scale to sentence level tasks well (Reimers and Gurevych, 2019). The text-image alignment, on its own, may expand the existing knowledge of XLM-R to the sentence level.

To further visualize the multilingual alignment of our encoders, we generate sentence-level representations for all sentences in the FLoRes-200 dataset and use t-SNE to project them down to 2 dimensions while preserving relative distances. We plot these embeddings in Figure 2 for the 4 fine-tuning languages with lines connecting parallel cliques of translated sentences. This way we can visualize whether an encoder produces language-specific clusters or whether certain sentences are encoded far from their translations.

Figure 2 shows that the original XLM-R representations are not aligned at all. Tuning only on the English image-caption data leads to better alignment than the untuned model, but the languages still form distinct clusters. Our `Multilingual` approach falls just short of the text-text aligned model in terms of the number of misaligned translations and adding Quechua into the mix does not make it that much worse. Interestingly, the text-image aligned models have tighter inter-sentence clusters indicating that the image alignment may have drawn connections between sentences that are not captured by a text-only semantic space.

Real versus Synthetic Captions To control for external factors, our experiments rely on synthetic captions generated by translating the MS-COCO English captions. To measure the effects of this synthetic data, we trained two additional models using the English and German captions from the Multi30k dataset (Elliott et al., 2016). For the first, we alternate between these real English and German captions in the training data. We will refer to this approach as `Multi30k`. For the second, we replace the german captions from `Multi30k` with the translations of their English counterparts. We

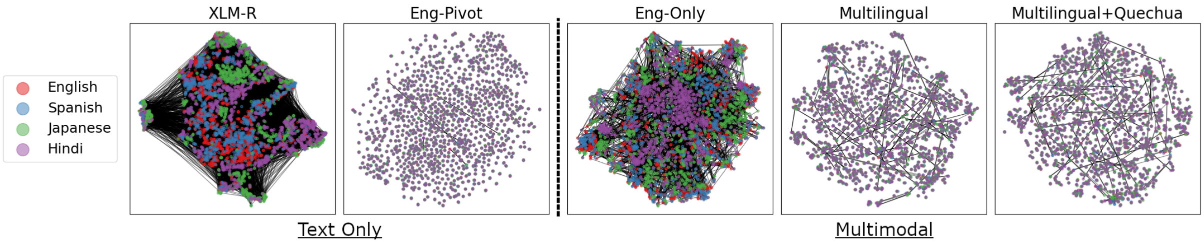

Figure 2: t-SNE embeddings for the outputs of each encoder over FLoRes-200 sentences. Translations are shown as cliques with lines connecting them. Visible lines, such as those in the XLM-R and `Eng-Only` panels, indicate that representations of translated sentences are far from each other, ie., *poor* alignment. While not as clear as parallel text alignment (`Eng-Pivot`), multilingual image-text alignment (two rightmost panels) shows promising results.

will refer to this as `Translated-Multi30k`.

Using the same bitext retrieval procedure from before, we find that the retrieval accuracy for `Multi30k` across all languages in FLoRes-200 is 40.4% and in German the retrieval accuracy is 94.6%. Similarly, `Translated-Multi30k` scores a retrieval accuracy of 38.7% across all of FLoRes-200 and 94.1% in German. Using translated captions does not significantly change the quality of the alignment.

3.3 Experiment 2: Can a language unseen in the encoder's pretraining be added using only image caption tuning?

Here, we turn to investigating the possibility of using only image-caption data to obtain good representations for a language *unseen* during pretraining, without any parallel text data. This approximates a real setting where we could ask an annotator to write image captions in a low-resource language which we want to add to our aligned language encoder for use in downstream tasks in a zero-shot cross-lingual transfer setting (Madaan et al., 2020).

We find that languages not included in the pretraining or fine-tuning still benefit from some alignment. But as one would expect, not to the same degree as those which have been already included in the model's training data.

We retrained the encoder from Experiment 1, but now with a dataset that also mixes in Quechua captions. Indigenous Latin American languages, including Quechua, are not included in the pretraining data of XLM-R. Quechua is also typologically distinct from all other pretraining languages.

We calculate the retrieval accuracy on FLoRes-200 from Quechua to English as well as the overall X→English accuracy to determine how well Quechua has been integrated into the encoder and aligned with other languages.

When Quechua is added to the image-caption dataset, the overall performance goes down slightly, but the performance on Quechua is greatly improved (cf last two rows of Table 1) from 18% to 29.2%. Importantly, the average accuracy for all other languages remains largely unaffected – we attribute the small drop in performance to the fact that we reduced the data in the other four languages to ensure experimental data-size comparability; in practice, this is not a requirement in the real world.

3.4 Experiment 3: Are the downstream qualities of the representations preserved and is cross-lingual transfer possible?

Here, we go beyond intrinsic evaluation to test our embeddings for a downstream task: natural language inference (NLI). Since images and text contain different types of semantic information, we want to ensure that aligning a text encoder to an image encoder does not overwrite the features which are useful for downstream NLU tasks.

We train simple feed-forward NLI models on frozen representations from each of the models in the previous experiments using the combined MultiNLI (Williams et al., 2018) training and dev sets.

We train using the MultiNLI train and dev datasets which only contain English samples. Any samples marked by the authors as lacking agreement were discarded. For evaluation of downstream NLI quality, we use the XNLI (Conneau et al., 2018) and AmericasNLI (Ebrahimi et al., 2022) test sets to measure both English NLI and cross-lingual transfer performance.

For each encoder, we train identical NLI models with input features ($\oplus$ stands for concatenation):

$$x_i = e(p_i) \ \oplus \ e(h_i) \oplus |e(p_i) - e(h_i)| \\ \oplus \ e(p_i) * e(h_i)$$

where e is the encoder and p_i and h_i are a premise and hypothesis respectively (Conneau et al., 2017).

Encoder	XNLI															AmericasNLI										Avg
	en	es	hi	ar	bg	de	el	fr	ru	sw	th	tr	ur	vi	zh	quy	aym	bzd	cni	gn	hch	nah	oto	shp	tar	
XLM-R	50	44	44	45	43	43	43	47	44	37	42	43	42	46	44	34	33	33	35	35	34	35	32	34	33	40
Eng-Only	53	50	46	47	49	49	48	51	50	42	47	47	45	48	48	40	38	35	35	38	39	36	35	37	36	44
Eng-Pivot	**67**	**65**	**60**	**61**	**63**	**64**	**63**	**65**	**62**	**52**	**61**	**61**	**58**	**63**	**62**	39	**40**	37	**41**	40	**40**	**42**	**39**	**44**	39	**53**
Multilingual	55	52	51	51	53	52	52	53	52	45	51	51	48	52	51	37	35	36	37	37	37	37	37	39	**40**	46
+ Quechua	56	53	51	51	53	53	53	53	53	45	50	51	49	52	51	**41**	36	**39**	**41**	**41**	36	40	**39**	41	**40**	47

Table 2: Rounded XNLI and AmericasNLI accuracy. Languages seen for alignment fine-tuning are underlined. NLI models are only trained on English data with frozen encoders; results in other languages require cross-lingual transfer.

The NLI models are a simple feed-forward architecture with 2 hidden layers and a hidden size of 2048. They are trained using the Adam optimizer and a learning rate of $2 * 10^{-5}$ for 100 epochs with early stopping.

The results in Table 2 show that the alignment of the text encoder with the space of the image encoder does not damage the quality of the text representations for downstream use, but actually improves them. Comparing the `Multilingual` image aligned model before and after adding Quechua, downstream performance is somewhat uncoupled from bitext retrieval performance. The addition of Quechua matched or exceeded the performance without it across nearly all languages, suggesting that NLI performance benefits from increased language coverage regardless of individual language data size. English represents $\frac{1}{4}$ of the `Multilingual` dataset and $\frac{1}{5}$ after adding Quechua, but the addition of Quechua increased the NLI score on English! Additionally, fine-tuning the encoder on the `Eng-Only` dataset only made a minimal improvement to the XLM-R performance even though it saw the largest portion of English data.

Additionally, the Quechua captions lead to improved results across the AmericasNLI languages even matching or outperforming the `Eng-Pivot` results in many of those languages. Adding a language from an unseen family not only improves representation quality for that language, but also improves cross-lingual transfer to unseen languages in that family.

4 Conclusion

The task of multilingual text-image contrastive alignment implicitly aligns text from multiple languages into the same space. This alignment carries over into unseen languages, and performance on a particular unseen language can be improved by collecting image-caption pairs in that language.

While this technique does not outperform SOTA methods, it performs remarkably well considering the non-reliance on parallel corpora. For low resource languages, this method could act as a bootstrapping step to scrape higher quality bitexts for use in further alignment.

Limitations

With the addition of Quechua to the training set, the drop in overall bitext retrieval performance could be due to the decrease in data for the other languages to accommodate the Quechua data. Whether this is the case is not captured by our experiment, but can be taken into account in a follow-up work.

Acknowledgments

This work was partially supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Award Number IIS-2327143).

References

Armen Aghajanyan, Xia Song, and Saurabh Tiwary. 2019. Towards language agnostic universal representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4033–4041, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Dario Balestri. 2023. Contrastive language–image pre-training for the Italian language. In *Proceedings of the 9th Italian*

Conference on Computational Linguistics (CLiC-it 2023), pages 78–85, Venice, Italy. CEUR Workshop Proceedings.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world‘s languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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 *Proceedings of International Conference on Learning Representations*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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

Aman Madaan, Shruti Rijhwani, Antonios Anastasopoulos, Yiming Yang, and Graham Neubig. 2020. Practical comparable data collection for low-resource languages via images. In *Practical ML for Developing Countries Workshop, ICLR*.

Masayasu Muraoka, Bishwaranjan Bhattacharjee, Michele Merler, Graeme Blackwood, Yulong Li, and Yang Zhao. 2023. Cross-lingual transfer of large language model by visually-derived supervision toward low-resource languages. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 3637–3646, New York, NY, USA. Association for Computing Machinery.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Chih-chan Tien and Shane Steinert-Threlkeld. 2022. Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8696–8706, Dublin, Ireland. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.