Aligning Cross-lingual Sentence Representations with Dual Momentum Contrast

Liang Wang and **Wei Zhao** and **Jingming Liu**
Yuanfudao AI Lab, Beijing, China
{wangliang01,zhaowei01,liujm}@yuanfudao.com

Abstract

In this paper, we propose to align sentence representations from different languages into a unified embedding space, where semantic similarities (both cross-lingual and monolingual) can be computed with a simple dot product. Pre-trained language models are fine-tuned with the translation ranking task. Existing work (Feng et al., 2020) uses sentences within the same batch as negatives, which can suffer from the issue of easy negatives. We adapt MoCo (He et al., 2020) to further improve the quality of alignment. As the experimental results show, the sentence representations produced by our model achieve the new state-of-the-art on several tasks, including Tatoeba en-zh similarity search (Artetxe and Schwenk, 2019b), BUCC en-zh bitext mining, and semantic textual similarity on 7 datasets.

1 Introduction

Pre-trained language models like BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018) have achieved phenomenal successes on a wide range of NLP tasks. However, sentence representations for different languages are not very well aligned, even for pre-trained multilingual models such as mBERT (Pires et al., 2019; Wang et al., 2020). This issue is more prominent for language pairs from different families (e.g., English versus Chinese). Also, previous work (Li et al., 2020) has shown that the out-of-box BERT embeddings perform poorly on monolingual semantic textual similarity (STS) tasks.

There are two general goals for sentence representation learning: cross-lingual representations should be aligned, which is a crucial step for tasks like bitext mining (Artetxe and Schwenk, 2019a), unsupervised machine translation (Lample et al., 2018b), and zero-shot cross-lingual transfer (Hu et al., 2020) etc. Another goal is to induce a metric space, where semantic similarities can be computed with simple functions (e.g., dot product on L_2-normalized representations).

Translation ranking (Feng et al., 2020; Yang et al., 2020) can serve as a surrogate task to align sentence representations. Intuitively speaking, parallel sentences should have similar representations and are therefore ranked higher, while non-parallel sentences should have dissimilar representations. Models are typically trained with in-batch negatives, which need a large batch size to alleviate the *easy negatives* issue (Chen et al., 2020a). Feng et al. (2020) use *cross-accelerator negative sampling* to enlarge the batch size to 2048 with 32 TPU cores. Such a solution is hardware-intensive and still struggles to scale.

Momentum Contrast (MoCo) (He et al., 2020) decouples the batch size and the number of negatives by maintaining a large memory queue and a momentum encoder. MoCo requires that queries and keys lie in a shared input space. In self-supervised vision representation learning, both queries and keys are transformed image patches. However, for translation ranking task, the queries and keys come from different input spaces. In this paper, we present *dual momentum contrast* to solve this issue. Dual momentum contrast maintains two memory queues and two momentum encoders for each language. It combines two contrastive losses by performing bidirectional matching.

We conduct experiments on the English-Chinese language pair. Language models that are separately pre-trained for English and Chinese are fine-tuned using translation ranking task with dual momentum contrast. To demonstrate the improved quality of the aligned sentence representations, we report state-of-the-art results on both cross-lingual and monolingual evaluation datasets: Tatoeba similarity search dataset (accuracy $95.9\% \rightarrow 97.4\%$), BUCC 2018 bitext mining dataset (f1 score $92.27\% \rightarrow 93.66\%$), and 7 English STS datasets (average Spearman's correlation

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3807–3815
November 7–11, 2021. ©2021 Association for Computational Linguistics

77.07% → 78.95%). We also carry out several ablation studies to help understand the learning dynamics of our proposed model.

2 Method

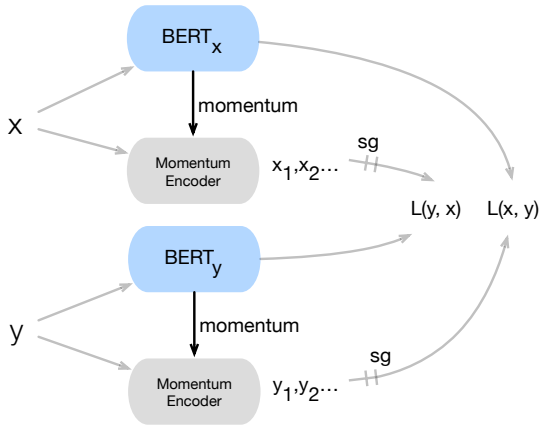

Figure 1: Illustration of dual momentum contrast. *sg* denotes "stop gradient". x and y are sentences from two different languages.

Dual Momentum Contrast is a variant of the MoCo proposed by He et al. (2020). Our method fits into the bigger picture of contrastive learning for self-supervised representation learning (Le-Khac et al., 2020). Given a collection of parallel sentences $\{x_i, y_i\}_{i=1}^n$, as illustrated in Figure 1, we first encode each sentence using language-specific BERT models (base encoder), then apply mean pooling on the last-layer outputs and L_2 normalization to get the representation vector $\mathbf{h}_{x_i}, \mathbf{h}_{y_i} \in R^{768}$.

Each BERT encoder has a momentum encoder, whose parameters $\boldsymbol{\theta}$ are updated by exponential moving average of the base encoder as follows:

$$\boldsymbol{\theta}_t \leftarrow m\boldsymbol{\theta}_{t-1} + (1-m)\boldsymbol{\theta}_{\text{base}} \quad (1)$$

Where t is the iteration step. Two memory queues are maintained for each language to store K vectors encoded by the corresponding momentum encoder from most recent batches. The oldest vectors are replaced with the vectors from the current batch upon each optimization step. The momentum coefficient $m \in [0, 1]$ is usually very close to 1 (e.g., 0.999) to make sure the vectors in the memory queue are consistent across batches. K can be very large (>10^5) to provide enough negative samples for learning robust representations.

To train the encoders, we use the InfoNCE loss (Oord et al., 2018):

$$\mathrm{L}(x, y) = -\log \frac{\exp(\mathbf{h}_x \cdot \mathbf{h}_y/\tau)}{\sum_{i=0}^{K} \exp(\mathbf{h}_x \cdot \mathbf{h}_{y_i}/\tau)} \quad (2)$$

τ is a temperature hyperparameter. Intuitively, Equation 2 is a (K+1)-way softmax classification, where the translation sentence $y = y_0$ is the positive, and the negatives are those in the memory queue $\{y_i\}_{i=1}^K$. Note that the gradients do not back-propagate through momentum encoders nor the memory queues.

Symmetrically, we can get $\mathrm{L}(y, x)$. The final loss function is the sum:

$$\min \ \mathrm{L}(x, y) + \mathrm{L}(y, x) \quad (3)$$

After the training is done, we throw away the momentum encoders and the memory queues, and only keep the base encoders to compute the sentence representations. In the following, our model is referred to as MoCo-BERT.

Application Given a sentence pair (x_i, y_j) from different languages, we can compute cross-lingual semantic similarity by taking dot product of L_2-normalized representations $\mathbf{h}_{x_i} \cdot \mathbf{h}_{y_j}$. It is equivalent to cosine similarity, and closely related to the Euclidean distance.

Our model can also be used to compute monolingual semantic similarity. Given a sentence pair (x_i, x_j) from the same language, assume y_j is the translation of x_j, if the model is well trained, the representations of x_j and y_j should be close to each other: $\mathbf{h}_{x_j} \approx \mathbf{h}_{y_j}$. Therefore, we have $\mathbf{h}_{x_i} \cdot \mathbf{h}_{x_j} \approx \mathbf{h}_{x_i} \cdot \mathbf{h}_{y_j}$, the latter one is cross-lingual similarity which is what our model is explicitly optimizing for.

3 Experiments

3.1 Setup

Data Our training data consists of English-Chinese corpora from UNCorpus [1], Tatoeba, News Commentary [2], and corpora provided by CWMT 2018 [3]. All parallel sentences that appear in the evaluation datasets are excluded. We sample 5M sentences to make the training cost manageable.

[1] https://conferences.unite.un.org/uncorpus
[2] https://opus.nlpl.eu/
[3] http://www.cipsc.org.cn/cwmt/2018/

Hyperparameters The encoders are initialized with *bert-base-uncased* (English) for fair comparison, and *RoBERTa-wwm-ext* [4](Chinese version). Using better pre-trained language models is orthogonal to our contribution. Following Reimers and Gurevych (2019), sentence representation is computed by the mean pooling of the final layer's outputs. Memory queue size is 409600, temperature τ is 0.04, and the momentum coefficient is 0.999. We use AdamW optimizer with maximum learning rate 4×10^{-5} and cosine decay. Models are trained with batch size 1024 for 15 epochs on 4 V100 GPUs. Please checkout the Appendix A for more details about data and hyperparameters.

3.2 Cross-lingual Evaluation

Model	Accuracy
$\text{mBERT}_{\text{base}}$ (Hu et al., 2020)	71.6%
LASER (Artetxe and Schwenk, 2019b)	95.9%
VECO (Luo et al., 2020)	82.7%
$\text{SBERT}_{\text{base}}$-p †	95.0%
$\text{MoCo-BERT}_{\text{base}}$ (zh→en)	**97.4%**
$\text{MoCo-BERT}_{\text{base}}$ (en→zh)	96.6%

Table 1: Accuracy on the test set of Tatoeba en-zh language pair. †: Reimers and Gurevych (2020).

Model	F1
$\text{mBERT}_{\text{base}}$ (Hu et al., 2020)	50.0%
LASER (Artetxe and Schwenk, 2019b)	92.27%
VECO (Luo et al., 2020)	78.5%
$\text{SBERT}_{\text{base}}$-p†	87.8%
LaBSE (Feng et al., 2020)	89.0%
$\text{MoCo-BERT}_{\text{base}}$	**93.66%**

Table 2: F1 score on the en-zh test set of BUCC 2018 dataset. †: Reimers and Gurevych (2020).

Tatoeba cross-lingual similarity search Introduced by Artetxe and Schwenk (2019b), Tatoeba corpus consists of 1000 English-aligned sentence pairs. We find the nearest neighbor for each sentence in the other language using cosine similarity. Results for both forward and backward directions are listed in Table 1. MoCo-BERT achieves an accuracy of 97.4%.

BUCC 2018 bitext mining aims to identify parallel sentences from a collection of sentences in two languages (Zweigenbaum et al., 2018). Following Artetxe and Schwenk (2019a), we adopt the margin-based scoring by considering the average cosine similarity of k nearest neighbors ($k = 3$ in our experiments):

$$\text{sim}(x, y) = \text{margin}(\cos(x, y), \sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2k}) \quad (4)$$

We use the distance margin function: $\text{margin}(a, b) = a - b$, which performs slightly better than the ratio margin function (Artetxe and Schwenk, 2019a). All sentence pairs with scores larger than threshold λ are identified as parallel. λ is searched based on the validation set. The F1 score of our system is 93.66%, as shown in Table 2.

3.3 Monolingual STS Evaluation

We evaluate the performance of MoCo-BERT for STS without training on any labeled STS data, following the procedure by Reimers and Gurevych (2019). All results are based on $\text{BERT}_{\text{base}}$. Given a pair of English sentences, the semantic similarity is computed with a simple dot product. We also report the results using labeled *natural language inference* (NLI) data. A two-layer MLP with 256 hidden units and a 3-way classification head is added on top of the sentence representations. The training set of SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) are used for multi-task training. See Appendix B for the detailed setup.

As pointed out by Gao et al. (2021), existing works follow inconsistent evaluation protocols, and thus may cause unfair comparison. We report results for both "weighted mean" (wmean) and "all" settings (Gao et al., 2021) in Table 3 and 8 respectively.

When training on translation ranking task only, MoCo-BERT improves the average correlation from 67.67 to 76.50 (+8.83). With labeled NLI supervision, MoCo-BERT+NLI advances state-of-the-art from 77.07 to 78.95 (+1.88).

3.4 Model Analysis

We conduct a series of experiments to better understand the behavior of MoCo-BERT. Unless explicitly mentioned, we use a memory queue size 204800 for efficiency.

Memory queue size One primary motivation of MoCo is to introduce more negatives to improve

[4] https://github.com/ymcui/Chinese-BERT-wwm

Model	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B	SICK-R	Avg
w/o labeled NLI supervision								
Avg GloVe†	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ [CLS]†	20.16	30.01	20.09	36.88	38.08	16.05	42.63	29.19
$\text{BERT}_{\text{base}}$-flow	59.54	64.69	64.66	72.92	71.84	58.56	65.44	65.38
$\text{IS-BERT}_{\text{base}}$	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
$\text{BERT}_{\text{base}}$-whitening♠	61.46	66.71	66.17	74.82	72.10	67.51	64.90	67.67
$\text{MoCo-BERT}_{\text{base}}$	**68.85**	**77.52**	**75.85**	**83.14**	**80.15**	**77.50**	**72.48**	**76.50**
w/ labeled NLI supervision								
InferSent	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
$\text{SBERT}_{\text{base}}$-NLI†	68.70	74.37	74.73	79.65	75.21	77.63	74.84	75.02
$\text{BERT}_{\text{base}}$-flow	67.75	76.73	75.53	80.63	77.58	79.10	78.03	76.48
$\text{BERT}_{\text{base}}$-whitening♠	69.87	77.11	76.13	82.73	78.08	79.16	76.44	77.07
$\text{MoCo-BERT}_{\text{base}}$+NLI	**71.66**	**79.42**	**76.37**	**84.08**	**80.81**	**82.15**	**78.19**	**78.95**

Table 3: Spearman's correlation for 7 STS datasets downloaded from SentEval (Conneau and Kiela, 2018). We report "weighted mean" (wmean) from SentEval toolkit. Baseline systems include $\text{BERT}_{\text{base}}$-flow (Li et al., 2020), $\text{IS-BERT}_{\text{base}}$ (Zhang et al., 2020), $\text{BERT}_{\text{base}}$-whitening♠ (Su et al., 2021), and InferSent (Conneau et al., 2017). †: from Reimers and Gurevych (2019).

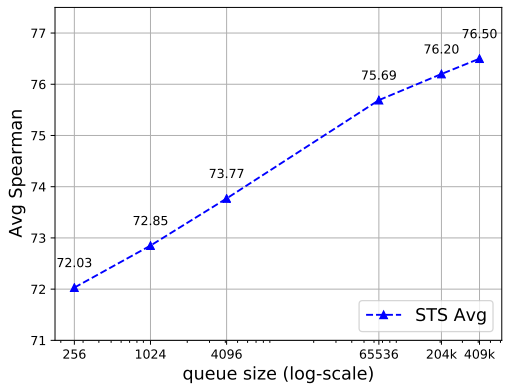

Figure 2: Average Spearman's correlation across 7 STS datasets for different memory queue sizes. The performance does not seem to saturate with queue size as large as $409k$. We do not run experiments $> 409k$ as it reaches the GPU memory limit.

the quality of the learned representations. In Figure 2, as expected, the performance consistently increases as the memory queue becomes larger. For visual representation learning, the performance usually saturates with queue size $\sim$ 65536 (He et al., 2020), but the ceiling is much higher in our case. Also notice that the model can still reach 72.03 with a small batch size 256, which might be because the encoders have already been pre-trained with MLM.

Temperature A lower temperature τ in InfoNCE loss makes the model focus more on the hard negative examples, but it also risks over-fitting label noises. Table 4 shows that τ could dramatically affect downstream performance, with $\tau = 0.04$ getting the best results on both STS and BUCC bitext mining tasks. The optimal τ is likely to be task-specific.

Temperature	0.01	0.04	0.07	0.1
STS Avg	74.80	**76.20**	74.23	69.81
BUCC F1	90.76	**93.14**	90.42	77.04

Table 4: Performance of our proposed MoCo-BERT under different temperatures.

Model	STS Avg	BUCC F1
MoCo-BERT	**76.20**	**93.14**
w/o momentum	-0.01	0.00

Table 5: Ablation results for momentum update mechanism. *w/o momentum* shares the parameters between the momentum encoder and the base encoder.

Momentum Update We also empirically verify if the momentum update mechanism is really necessary. Momentum update provides a more consistent matching target but also complicates the training procedure. In Table 5, without momentum update, the model simply fails to converge with the training loss oscillating back and forth. The resulting Spearman's correlation is virtually the same as random predictions.

Pooling	STS Avg	BUCC F1
mean pooling	**76.20**	**93.14**
max pooling	75.90	92.78
[CLS]	75.97	92.47

Table 6: Performance comparison between different pooling mechanisms for MoCo-BERT.

Pooling mechanism Though the standard practices of fine-tuning BERT (Devlin et al., 2019) directly use hidden states from [CLS] token, Reimers and Gurevych (2019); Li et al. (2020) have shown that pooling mechanisms matter for downstream STS tasks. We experiment with mean pooling, max pooling, and [CLS] embedding, with results listed in Table 6. Consistent with Reimers and Gurevych (2019), mean pooling has a slight but pretty much negligible advantage over other methods.

In Appendix C, we also showcase some visualization and sentence retrieval results.

4 Related Work

Multilingual representation learning aims to jointly model multiple languages. Such representations are crucial for multilingual neural machine translation (Aharoni et al., 2019), zero-shot cross-lingual transfer (Artetxe and Schwenk, 2019b), and cross-lingual semantic retrieval (Yang et al., 2020) etc. Multilingual BERT (Pires et al., 2019) simply pre-trains on the concatenation of monolingual corpora and shows good generalization for tasks like cross-lingual text classification (Hu et al., 2020). Another line of work explicitly aligns representations from language-specific models, either unsupervised (Lample et al., 2018a) or supervised (Reimers and Gurevych, 2020; Feng et al., 2020).

Contrastive learning works by pulling positive instances closer and pushing negatives far apart. It has achieved great successes in self-supervised vision representation learning, including SimCLR (Chen et al., 2020a), MoCo (He et al., 2020; Chen et al., 2020b), BYOL (Grill et al., 2020), CLIP (Radford et al., 2021) etc. Recent efforts introduced contrastive learning into various NLP tasks (Xiong et al., 2020; Giorgi et al., 2020; Chi et al., 2021; Gunel et al., 2020). Concurrent to our work, SimCSE (Gao et al., 2021) uses dropout and hard negatives from NLI datasets for contrastive sentence similarity learning, Sentence-T5 (Ni et al., 2021) outperforms SimCSE by scaling to larger models, and xMoCo (Yang et al., 2021) adopts a similar variant of MoCo for open-domain question answering.

Semantic textual similarity is a long-standing NLP task. Early approaches (Seco et al., 2004; Budanitsky and Hirst, 2001) use lexical resources such as WordNet to measure the similarity of texts. A series of SemEval shared tasks (Agirre et al., 2012, 2014) provide a suite of benchmark datasets that is now widely used for evaluation. Since obtaining large amounts of high-quality STS training data is non-trivial, most STS models are based on weak supervision data, including conversations (Yang et al., 2018), NLI (Conneau et al., 2017; Reimers and Gurevych, 2019), and QA pairs (Ni et al., 2021).

5 Conclusion

This paper proposes a novel method that aims to solve the *easy negatives* issue to better align cross-lingual sentence representations. Extensive experiments on multiple cross-lingual and monolingual evaluation datasets show the superiority of the resulting representations. For future work, we would like to explore other contrastive learning methods (Grill et al., 2020; Xiong et al., 2020), and experiment with more downstream tasks including paraphrase mining, text clustering, and bilingual lexicon induction etc.

Acknowledgements

We would like to thank three anonymous reviewers for their valuable comments, and EMNLP 2021 organizers for their efforts. We also want to thank Yueya He for useful suggestions on an early draft of this paper.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012:*

The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 385–393.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *NAACL-HLT*.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources*, volume 2, pages 2–2.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *ArXiv*, abs/1803.05449.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

J. Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and M. Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *ArXiv*, abs/2003.11080.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. *ArXiv*, abs/1804.07755.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2020. Veco: Variable encoder-decoder pre-training for cross-lingual understanding and generation. *arXiv preprint arXiv:2010.16046*.

L. V. D. Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Jianmo Ni, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, Yinfei Yang, et al. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

T. Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *ArXiv*, abs/1906.01502.

A. Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP*.

Nils Reimers and Iryna Gurevych. 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.

Nuno Seco, Tony Veale, and Jer Hayes. 2004. An intrinsic information content metric for semantic similarity in wordnet. In *Ecai*, volume 16, page 1089.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *ArXiv*, abs/2103.15316.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and J. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. *ArXiv*, abs/1910.04708.

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.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Nan Yang, Furu Wei, Binxing Jiao, Daxin Jiang, and Linjun Yang. 2021. xmoco: Cross momentum contrastive learning for open-domain question answering.

Yinfei Yang, Daniel Matthew Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, G. Ábrego, Steve Yuan, C. Tar, Yun-Hsuan Sung, B. Strope, and R. Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *ACL*.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.

A Details on Training Data and Hyperparameters

Dataset	# of sents	# of sampled
Tatoeba	46k	46k
News Commentary	320k	320k
UNCorpus	16M	1M
CWMT-neu2017	2M	2M
CWMT-casia2015	1M	1M
CWMT-casict2015	2M	1M

Table 7: List of parallel corpora used. *# of sampled* are randomly sampled subset from the corresponding dataset to make the training cost manageable. Duplicates are removed during preprocess.

We list all the parallel corpora used by this paper in Table 7. Hyperparameters are available in Table 9. We start with the default hyperparameters from MoCo (He et al., 2020) and use grid search to find the optimal values for several hyperparameters. The specific search ranges are $\{10^{-5}, 2 \times 10^{-5},$

Model	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B	SICK-R	Avg
w/o labeled NLI supervision								
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
MoCo-$\text{BERT}_{\text{base}}$	**70.99**	**76.51**	**73.17**	**82.09**	**78.32**	**77.50**	**72.48**	**75.87**
w/ labeled NLI supervision								
$\text{SBERT}_{\text{base}}$-NLI	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{BERT}_{\text{base}}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{BERT}_{\text{base}}$-whitening	69.65	77.57	**74.66**	82.27	78.39	79.52	76.91	77.00
MoCo-$\text{BERT}_{\text{base}}$+NLI	**76.07**	**78.33**	74.51	**84.19**	**78.74**	**82.15**	**78.19**	**78.88**

Table 8: Spearman's correlation for 7 STS datasets under the "all" evaluation setting (Gao et al., 2021). We use the official script from SimCSE.

Hyperparameter	value
# of epochs	15
# of GPUs	4
queue size	409k
temperature τ	0.04
momentum coefficient	0.999
learning rate	4×10^{-5}
gradient clip	10
warmup steps	400
batch size	1024
dropout	0.1
weight decay	10^{-4}
pooling	mean

Table 9: Hyperparameters for our proposed model.

$4 \times 10^{-5}\}$ for learning rate, $\{102k, 204k, 409k\}$ for queue size, $\{0.01, 0.04, 0.07, 0.1\}$ for temperature, and $\{0.9999, 0.999, 0.99\}$ for momentum coefficient. The entire training process takes approximately 15 hours with 4 V100 GPUs and automatic mixed precision support from PyTorch.

B Multi-task with NLI

Given a premise x_p and a hypothesis x_h, the sentence representations are computed as stated in the paper. Then, a two-layer MLP with 256 hidden units, ReLU activation, and a 3-way classification head is added on top of the sentence representations. Dropout 0.1 is applied to the hidden units. The loss function $\text{L}_{\text{nli}}(x_p, x_h)$ is simply the cross-entropy between gold label and softmax outputs. The model is jointly optimized with the following:

$$\min \quad \text{L}(x, y) + \text{L}(y, x) + \alpha \text{L}_{\text{nli}}(x_p, x_h) \qquad (5)$$

Where α is used to balance different training objectives, we set $\alpha = 0.1$ empirically. The batch size for NLI loss is 128. The training set is the union of SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) dataset (~1M sentence pairs).

C Visualization of Sentence Representations

To visualize the learned sentence representations, we use t-SNE (Maaten and Hinton, 2008) for dimensionality reduction. In Figure 3, we can see the representations of parallel sentences are very close, indicating that our proposed model is successful at aligning cross-lingual representations.

In Table 10, we illustrate the results of monolingual sentence retrieval. Most top-ranked sentences indeed share similar semantics with the given query, this paves the way for potential applications like paraphrase mining.

Figure 3: t-SNE visualization of the representations of 15 random parallel sentences from Tatoeba test set. For visualization purpose, if two points are too close, we move them a little bit far apart. Enlarge the graph for better views.

query: *I am willing to devote my life to education career.*	
0.853	He dedicated his life to the cause of education.
0.776	He devoted his whole life to education.
0.764	She has dedicated herself to the cause of education.
query: *The Committee resumed consideration of the item.*	
0.928	The Committee continued consideration of the item.
0.843	The Committee resumed its consideration of this agenda item.
0.686	The Committee began its consideration of the item.
query: *There are a great many books on the bookshelf.*	
0.837	There are many books on the bookcase.
0.690	There is a heap of books on the table.
0.655	The bookshelf is crowded with books on different subjects.
query: *Everyone has the privilege to be tried by a jury.*	
0.718	They have the right to have their case heard by a jury.
0.647	Every defendant charged with a felony has a right to be charged by the Grand Jury.
0.580	Everyone has the right to be educated.

Table 10: Examples of sentence retrieval using learned representations. Given a query, we use cosine similarity to retrieve the 3 nearest neighbors (excluding exact match). The first column is the cosine similarity score between the query and retrieved sentences. The corpus is 1M random English sentences from the training data.