MCSE: Multimodal Contrastive Learning of Sentence Embeddings

Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani,
Michael A. Hedderich, and **Dietrich Klakow**
Spoken Language Systems (LSV)
Saarland Informatics Campus, Saarland University, Germany
{mzhang,mmosbach,didelani}@lsv.uni-saarland.de
{mhedderich,dklakow}@lsv.uni-saarland.de

Abstract

Learning semantically meaningful sentence embeddings is an open problem in natural language processing. In this work, we propose a sentence embedding learning approach that exploits both visual and textual information via a multimodal contrastive objective. Through experiments on a variety of semantic textual similarity tasks, we demonstrate that our approach consistently improves the performance across various datasets and pre-trained encoders. In particular, combining a small amount of multimodal data with a large text-only corpus, we improve the state-of-the-art average Spearman's correlation by 1.7%. By analyzing the properties of the textual embedding space, we show that our model excels in aligning semantically similar sentences, providing an explanation for its improved performance.

1 Introduction

Sentence embedding learning, i.e., encoding sentences into fixed-length vectors that faithfully reflect the semantic relatedness among sentences, is a fundamental challenge in natural language processing (NLP). Despite the tremendous success of pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), it has been shown that the off-the-shelf sentence embeddings of PLMs without fine-tuning are even inferior to averaging Glove embeddings (Pennington et al., 2014) in terms of semantic similarity measure (Reimers and Gurevych, 2019). Hence, recent research (Li et al., 2020; Zhang et al., 2020; Su et al., 2021) focuses on adjusting the original sentence embeddings derived from PLMs in an unsupervised manner. In particular, there has been growing interest in adopting contrastive learning objectives to achieve this goal (Carlsson et al., 2020; Kim et al., 2021; Gao et al., 2021).

Although purely text-based models have led to impressive progress, it remains an open question to what extent they capture the deeper notion of sentence meaning beyond the statistical distribution of texts, which lies outside of the text and is grounded in the real-world (Bender and Koller, 2020; Bisk et al., 2020). As a central part of the human perceptual experience, vision has been shown to be effective in grounding language models and improving performance on various NLP tasks (Zhang et al., 2019; Bordes et al., 2019; Zhao and Titov, 2020). We hypothesize that using vision as supplementary semantic information can further promote sentence representation learning.

In this work, we propose *MCSE*, an approach for multimodal contrastive learning of sentence embeddings. To exploit both visual and textual information, we adopt the state-of-the-art contrastive sentence embedding framework SimCSE (Gao et al., 2021) and extend it with a multimodal contrastive objective. In addition to the textual objective in SimCSE that maximizes agreement between positive sentence pairs, the multimodal objective maximizes agreement between sentences and corresponding images in a shared space. We conduct extensive experiments on standard Semantic Textual Similarity (STS) benchmarks and show the effectiveness of MCSE across various datasets and pre-trained encoders. We find that, using a small amount of multimodal data in addition to a text-only corpus yields significant improvements on STS tasks. By analyzing the alignment and uniformity properties of the embedding space (Wang and Isola, 2020), we show that MCSE better aligns the semantically similar sentences while maintaining uniformity, providing an explanation for its superior performance.[1]

2 Related Work

Sentence Representation Learning. Existing works for learning sentence embeddings can be

[1]Our code and pre-trained models are publicly available at https://github.com/uds-lsv/MCSE.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5959 - 5969

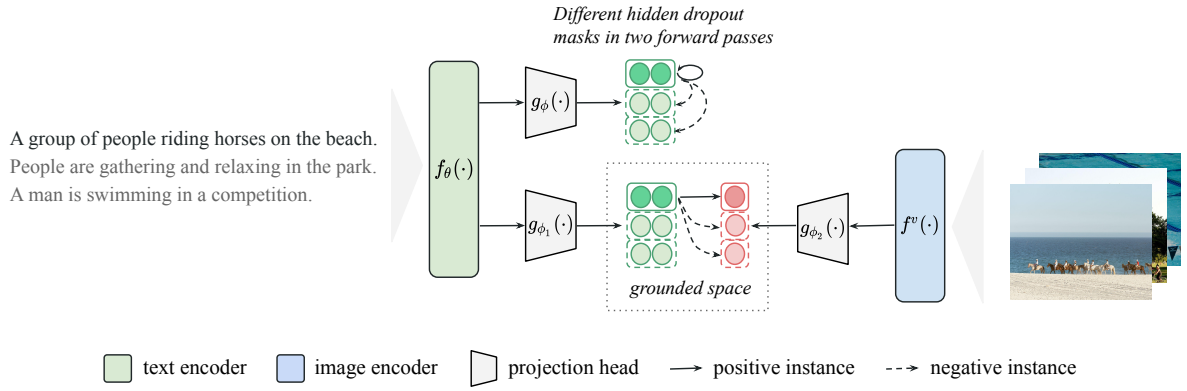


Figure 1: The overall architecture of MCSE. Compared to SimCSE, a new multimodal objective is calculated in the grounded space. For each input sentence, the positive instance is the paired image and the negative instances are all other in-batch images.

categorized into supervised (Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019; Wieting et al., 2020) and unsupervised approaches (Li et al., 2020; Carlsson et al., 2020; Su et al., 2021; Kim et al., 2021; Gao et al., 2021; Liu et al., 2021; Yan et al., 2021). Supervised approaches mostly utilize supervision from annotated natural language inference data or parallel data. Unsupervised approaches are able to make use of the intrinsic semantic information embedded in the natural language text corpus by adjusting the training objective to STS tasks, thereby eliminating the need for a costly annotation process. In particular, contrastive learning objective (Carlsson et al., 2020; Kim et al., 2021; Gao et al., 2021; Liu et al., 2021; Yan et al., 2021) regularizes the embedding space by pulling positive (i.e., semantically similar) sentences closer and pushing apart negatives, showcasing great effectiveness in capturing the semantic similarity among sentences. Our approach adopts the contrastive learning framework and is built on top of the current state-of-the-art approach (Gao et al., 2021), further pushing the frontier of STS by leveraging multimodal semantic information.

Visually Grounded Representation Learning. There are various works showing that grounding NLP models to the visual world can improve textual representation learning. Lazaridou et al. (2015) and Zablocki et al. (2018) learn word embeddings by aligning words to the visual entity or visual context. Kiela et al. (2018) ground sentence embeddings by predicting both images and alternative captions related to the same image. Bordes et al. (2019) enhance the Skip-Thought model (Kiros et al., 2015) by learning a grounded space that preserves the structure of visual and textual spaces. Recently, Tan and Bansal (2020) and Tang et al. (2021) train large scale language models with multimodal supervision from scratch with the goal of improving general language understanding. Different from the aforementioned works, we focus on learning visually grounded sentence embeddings by fine-tuning pre-trained models in a contrastive learning framework.

3 Method

To exploit both visual and textual information, we adopt SimCSE (Gao et al., 2021) as the textual baseline and extend it with a multimodal contrastive learning objective.

3.1 Background: Unsupervised SimCSE

Data augmentation plays a critical role in contrastive self-supervised representation learning (Chen et al., 2020). The idea of unsupervised SimCSE is to use dropout noise as a simple yet effective data augmentation strategy. Given a collection of sentences $\{x_i\}_{i=1}^m$, we construct a positive pair for each input x_i by encoding it twice using different dropout masks: $\boldsymbol{h}_i^z = g_\phi(f_\theta(x_i, z))$ and $\boldsymbol{h}_i^{z'} = g_\phi(f_\theta(x_i, z'))$, where z and z' denote different dropout masks[2], $f_\theta(\cdot)$ is a pre-trained language encoder such as BERT, and $g_\phi(\cdot)$ is a projection head[3] on top of the `[CLS]` token. The training objective is:

$$\ell_i^{\mathrm{S}} = -\log \frac{e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_i^{z_i'})/\tau}}{\sum_{j=1}^N e^{\mathrm{sim}(\boldsymbol{h}_i^{z_i}, \boldsymbol{h}_j^{z_j'})/\tau}}, \quad (1)$$

[2]The standard dropout masks in Transformers are used.

[3]There is a MLP pooler layer over `[CLS]` in BERT's implementation. Gao et al. (2021) use it with re-initialization.

	Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.↑
	BERT (first-last avg.)	39.7	59.4	49.7	66.0	66.2	53.9	62.1	56.7
	RoBERTa (first-last avg.)	40.9	58.7	49.1	65.6	61.5	58.6	61.6	56.6
wiki	SimCSE-BERT$^\diamondsuit$	68.4	82.4	74.4	80.9	78.6	76.9	72.2	76.3
wiki	SimCSE-RoBERTa$^\diamondsuit$	70.2	81.8	73.2	81.4	80.7	80.2	68.6	76.6
wiki	SimCSE-BERT	$67.8_{\pm 1.6}$	$80.0_{\pm 2.1}$	$72.5_{\pm 1.7}$	$80.1_{\pm 0.8}$	$77.6_{\pm 0.8}$	$76.5_{\pm 0.8}$	$70.1_{\pm 0.9}$	$74.9_{\pm 1.1}$
wiki	SimCSE-RoBERTa	$68.7_{\pm 1.0}$	$82.0_{\pm 0.5}$	$74.0_{\pm 1.0}$	$82.1_{\pm 0.4}$	$81.1_{\pm 0.4}$	$80.6_{\pm 0.3}$	$69.2_{\pm 0.2}$	$76.8_{\pm 0.5}$
wiki+flickr	SimCSE-BERT	$69.9_{\pm 1.7}$	$79.8_{\pm 1.5}$	$72.9_{\pm 0.9}$	$81.9_{\pm 0.8}$	$\mathbf{77.8}_{\pm 0.9}$	$76.6_{\pm 1.1}$	$68.4_{\pm 0.8}$	$75.3_{\pm 0.9}$
wiki+flickr	MCSE-BERT	$\mathbf{71.4}_{\pm 0.9}$	$\mathbf{81.8}^*_{\pm 1.3}$	$\mathbf{74.8}^*_{\pm 0.9}$	$\mathbf{83.6}_{\pm 0.9}$	$77.5_{\pm 0.8}$	$\mathbf{79.5}^*_{\pm 0.5}$	$\mathbf{72.6}^*_{\pm 1.4}$	$\mathbf{77.3}^*_{\pm 0.5}$
wiki+flickr	SimCSE-RoBERTa	$69.5_{\pm 0.9}$	$81.6_{\pm 0.5}$	$74.1_{\pm 0.6}$	$82.4_{\pm 0.3}$	$80.9_{\pm 0.5}$	$79.9_{\pm 0.3}$	$67.3_{\pm 0.5}$	$76.5_{\pm 0.4}$
wiki+flickr	MCSE-RoBERTa	$\mathbf{71.7}^*_{\pm 0.2}$	$\mathbf{82.7}^*_{\pm 0.4}$	$\mathbf{75.9}^*_{\pm 0.3}$	$\mathbf{84.0}^*_{\pm 0.4}$	$\mathbf{81.3}_{\pm 0.3}$	$\mathbf{82.3}^*_{\pm 0.5}$	$\mathbf{70.3}^*_{\pm 1.3}$	$\mathbf{78.3}^*_{\pm 0.1}$
wiki+coco	SimCSE-BERT	$69.1_{\pm 1.0}$	$80.4_{\pm 0.9}$	$72.7_{\pm 0.7}$	$81.1_{\pm 0.3}$	$\mathbf{78.2}_{\pm 0.9}$	$73.9_{\pm 0.6}$	$66.6_{\pm 1.2}$	$74.6_{\pm 0.2}$
wiki+coco	MCSE-BERT	$\mathbf{71.2}^*_{\pm 1.3}$	$\mathbf{79.7}_{\pm 0.9}$	$\mathbf{73.8}_{\pm 0.9}$	$\mathbf{83.0}^*_{\pm 0.4}$	$77.8_{\pm 0.9}$	$\mathbf{78.5}^*_{\pm 0.4}$	$\mathbf{72.1}^*_{\pm 1.4}$	$\mathbf{76.6}^*_{\pm 0.5}$
wiki+coco	SimCSE-RoBERTa	$66.4_{\pm 0.9}$	$80.7_{\pm 0.7}$	$72.7_{\pm 1.1}$	$81.3_{\pm 0.9}$	$80.2_{\pm 0.8}$	$76.8_{\pm 0.6}$	$65.7_{\pm 0.7}$	$74.8_{\pm 0.5}$
wiki+coco	MCSE-RoBERTa	$\mathbf{70.2}^*_{\pm 1.7}$	$\mathbf{82.0}^*_{\pm 0.7}$	$\mathbf{75.5}^*_{\pm 1.2}$	$\mathbf{83.0}^*_{\pm 0.6}$	$\mathbf{81.5}^*_{\pm 0.7}$	$\mathbf{80.8}^*_{\pm 1.0}$	$\mathbf{69.9}^*_{\pm 0.6}$	$\mathbf{77.6}^*_{\pm 0.8}$

$*$: difference between SimCSE and MCSE is significant at $\alpha = 0.05$ according to an independent t-test.

Table 1: Performance comparison on STS tasks. STS-B: STS Benchmark, SICK-R: SICK-Relatedness, Avg.: average across 7 tasks. $\diamondsuit$: single seed results from Gao et al. (2021). All other results are from our implementation. Models are trained with 5 random seeds and we report the means and standard deviations.

where N is the size of the mini-batch, τ is a temperature parameter and $\text{sim}(\boldsymbol{h}_1, \boldsymbol{h}_2)$ is the cosine similarity $\frac{\boldsymbol{h}_1^T \boldsymbol{h}_2}{\|\boldsymbol{h}_1\| \cdot \|\boldsymbol{h}_2\|}$. After training, the `[CLS]` token outputs of the language encoder are taken as the sentence embeddings.

3.2 Multimodal Contrastive Learning

Beyond the textual objective in SimCSE, we introduce a multimodal objective within the contrastive learning framework. The overview of our MCSE model is shown in Figure 1. Given a collection of sentence-image pairs $D = \{x_i, y_i\}_{i=1}^m$, firstly we map sentence x_i and image y_i into a shared space:

$$\boldsymbol{s}_i^z = g_{\phi_1}(f_\theta(x_i, z)), \boldsymbol{v}_i = g_{\phi_2}(f^v(y_i)) \ , \quad (2)$$

where $f^v(\cdot)$ is a pre-trained image encoder such as ResNet (He et al., 2016), which is fixed during training. $g_{\phi_1}(\cdot)$ and $g_{\phi_2}(\cdot)$ are distinct projection heads for text and image modality respectively. To pull semantically close image-sentence pairs together and push away non-related pairs, we define the multimodal contrastive learning objective as:

$$\ell_i^{\text{M}} = -\sum_{z \in \{z_i, z_i'\}} \log \frac{e^{\text{sim}(\boldsymbol{s}_i^z, \boldsymbol{v}_i)/\tau'}}{\sum_{j=1}^N e^{\text{sim}(\boldsymbol{s}_i^z, \boldsymbol{v}_j)/\tau'}} \ , \quad (3)$$

where τ' is a temperature parameter. Let λ denote the trade-off hyperparameter between two objectives, we formulate the final loss as:

$$\ell_i = \ell_i^S + \lambda \ell_i^M \ . \quad (4)$$

Our method further regularizes the sentence representation in a way that aligns with the image representation in the grounded space.

4 Experiments

4.1 Setup

Dataset We use Flickr30k (Young et al., 2014) and MS-COCO (Lin et al., 2014) as our multimodal datasets. Flickr30k contains $29,783$ training images and MS-COCO contains $82,783$ training images. Each image is annotated with multiple captions and we randomly sample only one caption to create image-sentence pairs. Following Gao et al. (2021), we use Wiki1M as the text-only corpus, which consists of 10^6 sentences randomly drawn from English Wikipedia.

Implementation Details We use BERT$_{base}$ (Devlin et al., 2019) and RoBERTa$_{base}$ (Liu et al., 2019) as language encoders and ResNet-50 (He et al., 2016) as the image encoder. Distinct single-layer MLPs are applied as projection heads. More details are provided in Appendix A.

Evaluation We evaluate the trained models on seven Semantic Textual Similarity (STS) tasks: STS 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). Each of these datasets consists of a collection of sentence pairs and the goal is to predict a similarity score for each sentence pair. Following Gao et al. (2021), we report the Spearman's correlation ($\times 100$) between

gold annotations and predicted scores in the "all" setting, i.e., for each task, we concatenate all the subsets and report the overall Spearman's correlation.

4.2 Main Results

Augmenting text-only corpus with small scale multimodal data yields significant improvements. To fully utilize different types of data resources, we conduct experiments with a text-only corpus and multimodal data. SimCSE is trained on sentences and captions only, while MCSE additionally computes the multimodal objective for image-caption pairs. As shown in Table 1, averaging the off-the-shelf BERT and RoBERTa embeddings[4] yields poor performance on STS tasks. SimCSE models significantly outperform the average embeddings. MCSE models, which have access to auxiliary visual information, further achieve noticeable improvements even if the amount of multimodal data is relatively small. When MCSE is applied to the combination of Wiki1M and Flickr30k, it improves the state-of-the-art result for BERT (76.3 $\rightarrow$ 77.3) and RoBERTa (76.6 $\rightarrow$ 78.3) by a decent margin. Looking at performance on the individual tasks, we find that MCSE models using BERT encoder perform worse on STS16. This can be attributed to the domain discrepancy, where some subsets that are close to the training distribution benefit more from visually grounding than others (see Appendix B.1).

To further investigate the impact of different datasets, we train models solely on multimodal data and report results in Table 2. We observe that, without the large text-only corpus, the performances decrease considerably compared to results in Table 1. Still, MCSE models consistently surpass SimCSE models (0.9 – 3.8 points improvement). Moreover, replacing the paired images with shuffled images before training MCSE leads to 0.8 – 5.0 points reduction in terms of average Spearman's correlation, further validating the efficacy of visual semantics. We also replace the ResNet encoder with CLIP (Radford et al., 2021) and our results show that different image encoders lead to similar results. Details are shown in Appendix B.2.

Grounding to the visual world improves alignment and maintains uniformity. To dissect the inner workings of MCSE, we use two quantifiable

Model	Trained on	
	flickr	*coco*
SimCSE-BERT	$68.8_{\pm 0.7}$	$67.8_{\pm 0.4}$
MCSE-BERT	$\mathbf{70.6}^{*}_{\pm 0.5}$	$\mathbf{71.6}^{*}_{\pm 0.2}$
w/ shuffling	$67.9_{\pm 0.6}\downarrow$	$66.6_{\pm 0.3}\downarrow$
SimCSE-RoBERTa	$72.9_{\pm 0.3}$	$72.8_{\pm 0.3}$
MCSE-RoBERTa	$\mathbf{73.8}^{*}_{\pm 0.2}$	$\mathbf{74.3}^{*}_{\pm 0.3}$
w/ shuffling	$73.0_{\pm 0.4}\downarrow$	$72.8_{\pm 0.3}\downarrow$

$*$: difference between SimCSE and MCSE is significant.

Table 2: Comparison of the average Spearman's correlation on 7 STS tasks (Avg. column in Table 1). We report the means and standard deviations over 5 seeds.

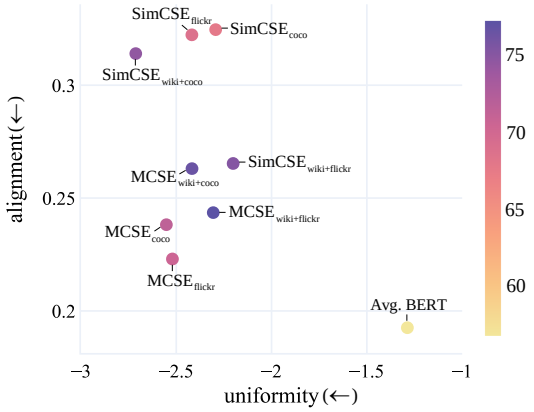


Figure 2: The alignment-uniformity plot of models when using BERT encoder. Colors of dots represent the average Spearman's correlation.

metrics proposed in Wang and Isola (2020): *alignment* and *uniformity*, as measurements of representation quality. Let p_{pos} denote the positive pairs distribution and p_{data} denote the data distribution. The *alignment loss* prefers encoders that assign similar features to semantically similar instances (assuming features have been normalized):

$$\mathcal{L}_{align} \triangleq \underset{(x,x^+) \sim p_{\text{pos}}}{\mathbb{E}} \left\| f(x) - f(x^+) \right\|_2^2 . \quad (5)$$

And the *uniformity loss* prefers a uniform distribution in the hypersphere:

$$\mathcal{L}_{uniform} \triangleq \log \underset{x,y \overset{i.i.d.}{\sim} p_{\text{data}}}{\mathbb{E}} e^{-2\|f(x)-f(y)\|_2^2} . \quad (6)$$

Gao et al. (2021) empirically showed that sentence embedding models with both lower alignment and uniformity achieve better performance in general. Similarly, we calculate the two losses on STS-B[5]

[4]Following (Gao et al., 2021), we take the average of the first and last layers, which is better than only using the last.

[5]We take STS-B pairs with a score higher than 4.0 as p_{pos} and the full STS-B as p_{data}. Since Gao et al. (2021) did not

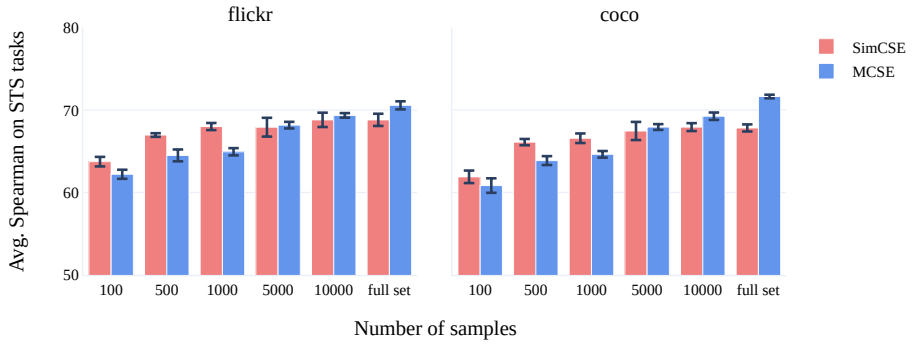


Figure 3: Performances of different data scales. The full set indicates 30K and 87K samples for Flickr30k and MS-COCO respectively.

and results are presented in Figure 2. It shows that MCSE models achieve better alignment scores compared to SimCSE while maintaining uniformity. This analysis provides further support that visually grounding can enhance sentence representation learning by improving the alignment property of the textual embedding space.

4.3 Analysis

For brevity, we take BERT-based models trained merely on caption datasets and investigate the impact of training data scales. More analysis results (sentence retrieval, cross-modal retrieval) are provided in Appendix B.3. We limit the number of training samples to 100, 500, 1000, 5000 and 10000, and compare their performance with the full set performance. In all of these settings, we optimize the models for same number of training steps as the full set setting. The results are shown in Figure 3. SimCSE achieves better performance than MCSE with limited samples, while MCSE starts to outperform SimCSE with the increasing data scale. We conjecture that this phenomenon can be ascribed to the progressive training of weights in multimodal projection heads.

5 Limitations

Despite showing performance improvements on STS benchmarks, MCSE has its limitations as well. We take caption datasets as the source of multimodal information, while these datasets are collected and curated with non-negligible human efforts. It will have great practical value if we can properly leverage noisy image-sentence pairs or even get rid of the explicit alignments between images and sentences. Furthermore, we find that only subsets from related domains can get significant improvements while others suffer from distribution shifts. It is critical to mitigate domain gaps for learning general-purpose sentence embeddings. In addition, the definition of "semantic similarity" is highly task-dependent. Besides STS benchmarks, it is worth exploring the performance gap between text-only models and multimodal models on other benchmarks that can also assess the quality of sentence representations.

6 Conclusion

In this paper, we propose MCSE, a novel approach for sentence embedding learning that applies a multimodal contrastive objective to align sentences and corresponding images in a grounded space. Experiments show that MCSE consistently improves the performance on STS tasks. We also highlight the superiority of our method by analyzing the alignment and uniformity properties of the embedding space. The multimodal objective is generic and can be potentially incorporated into other sentence embedding methods to boost their performance.

Acknowledgements

We thank Dingfan Chen, Fangzhou Zhai and Xiaoyu Shen for their helpful comments on the paper draft. We would also like to thank the reviewers for their valuable feedback. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project-id 232722074 – SFB 1102. This work was also partially funded by the EU Horizon 2020 project ROXANNE under grant number 833635 and COMPRISE grant number 3081705.

release the code for calculating these two losses, the absolute values we obtained might be different from theirs. We make sure our calculation across different models is consistent.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

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, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 497–511.

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.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5185–5198.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. Incorporating visual semantics into sentence representations within a grounded space. 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 696–707.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations (ICLR)*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*, pages 1–14.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 169–174.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607.

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 (EMNLP)*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HIT)*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HIT)*, pages 408–418.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 2528–2540.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems (NeurIPS)*, pages 3294–3302.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 153–163.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1442–1459.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *International Conference on Language Resources and Evaluation (LREC)*, pages 216–223.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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 (ICML)*, pages 8748–8763.

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.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. VidLanKD: Improving language understanding via video-distilled knowledge transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 24468–24481.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9929–9939.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A bilingual generative transformer for semantic sentence embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1581–1594.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5065–5075.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual

denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, pages 67–78.

Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. 2018. Learning multi-modal word representation grounded in visual context. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *International Conference on Learning Representations (ICLR)*.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound PCFGs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4369–4379.

A Implementation Details

Language Encoder Our implementation is based on the Hugging Face Transformers library[6] (Wolf et al., 2020). We start from the checkpoints of `bert-base-uncased` and `roberta-base`, and fine-tune the pre-trained models using a contrastive objective function. We use the 768-dimensional `[CLS]` token outputs before the MLP pooler layer as sentence embeddings for evaluation.

Image Encoder We use ResNet-50 and extract 2048-dimensional feature vectors at the last layer. The image encoder is not fine-tuned.[7]

Projection Heads We use distinct projection heads for different modalities and objectives. All of them are implemented by single-layer MLPs with Tanh activation. We map sentence embeddings to a 768-dimensional space before calculating the textual objective. We map both sentence embeddings and image feature vectors to a 256-dimensional shared space, and normalize them before calculating the multimodal objective.

Parameter Settings We explore 5 training settings in the paper: {*wiki*, *wiki+flickr*, *wiki+coco*, *flickr*, *coco*}. For *wiki+flickr* and *wiki+coco*, we sample mini-batches from either Wiki1M or the caption dataset in proportion to their data size. We adopt most of the parameter settings suggested by Gao et al. (2021). Moreover, temperature parameters τ and τ' are set to 0.05, and other hyperparameters are reported in Table 3. We use the dev set of STS-B to tune the trade-off parameter λ and ablation studies are shown in Table 4. We evaluate models every 125 training steps on STS-B dev set and keep the best checkpoint for final evaluation.

settings:	*wiki*	*wiki+flickr*	*wiki+coco*	*flickr*	*coco*
BERT					
learning rate	3e-5				
batch size	64				
λ	–	0.01	0.01	0.05	0.05
epochs	3	3	3	6	3
RoBERTa					
learning rate	1e-5				
batch size	128				
λ	–	0.01	0.01	0.01	0.01
epochs	3	3	3	6	3

Table 3: The hyperparameters used for different training settings and pre-trained encoders.

[6] https://github.com/huggingface/transformers

[7] In our preliminary results, fine-tuning the image encoder does not have a significant impact on the STS performance.

λ	0.001	0.01	0.05	0.1	0.5
MCSE-BERT	78.38	79.95	**80.41**	80.35	80.01
MCSE-RoBERTa	80.60	**81.48**	81.08	80.73	79.85

Table 4: STS-B performance of MCSE models trained on Flickr30k with different trade-off parameters.

B More Results

B.1 Improvements on Different Subsets

To delve into the performance gap between `MCSE-BERT` and `SimCSE-BERT`, we calculate the Spearman's correlation for different subsets of each year's STS challenge separately. The improvements of MCSE over SimCSE are shown in Figure 4. In STS12, "MSRvid" subset achieves the largest improvement, which is a corpus of video descriptions. "Image" subsets in STS14 and STS15 also get considerable improvements. Meanwhile, the performance of "answers-students" subset in STS15 drops extensively, and none of the subsets in STS16 get noticeable improvement by MCSE. The results indicate that the subsets benefit to different degrees from the visually grounding because of domain discrepancy.

B.2 Ablation Study

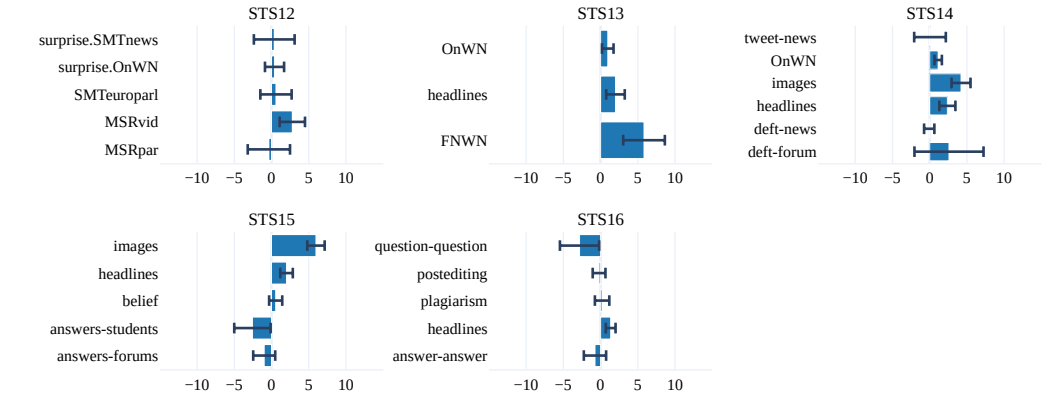
CLIP as Image Encoder We use CLIP (Radford et al., 2021) as an alternative image encoder. The implementation is based on the Sentence Transformer library[8] (Reimers and Gurevych, 2019) and we use the checkpoint `clip-ViT-B-32` to extract 512-dimensional feature vectors. As shown in Table 7, different image encoders lead to very similar results, thus we use ResNet as the default image encoder.

Combining Wiki1M, Flickr30k and MS-COCO We adopt the same parameter setting as *wiki+flickr* and *wiki+coco*, and train models on the combination of Wiki1M, Flickr30k, and MS-COCO. As shown in Table 5, MCSE models achieve 1.9 point and 2.6 point improvements when using BERT and RoBERTa, respectively.

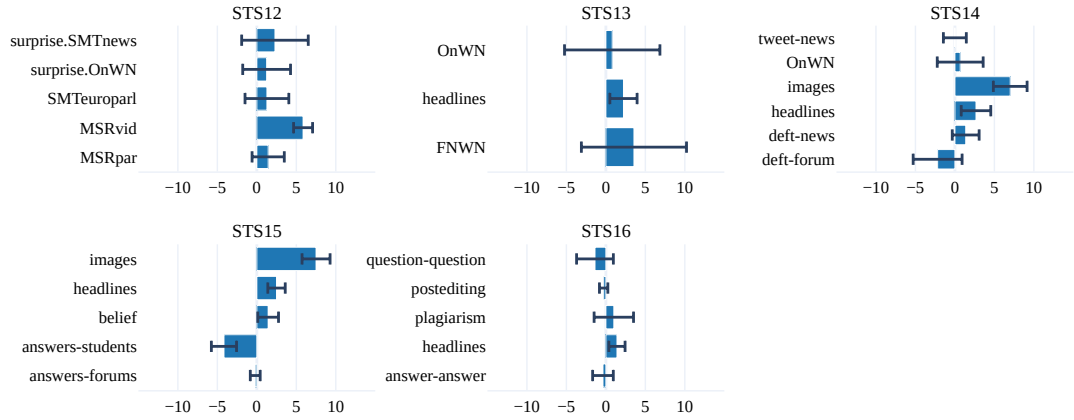
B.3 Analysis

Sentence Retrieval We take BERT-based models trained on the Flickr30k train set (same seed) and conduct a sentence retrieval experiment on Flickr30k test set. Given an input sentence, the nearest neighbor will be retrieved based on cosine

[8] https://github.com/UKPLab/sentence-transformers



(a) training data: *wiki+flickr*



(b) training data: *wiki+coco*

Figure 4: The Spearman's correlation improvements over different subsets.

Model	Trained on *wiki+flickr+coco*
SimCSE-BERT	74.3$_{\pm 1.0}$
MCSE-BERT	**76.2**$_{\pm 0.3}$
SimCSE-RoBERTa	75.3$_{\pm 0.7}$
MCSE-RoBERTa	**77.9**$_{\pm 0.6}$

Table 5: Comparison of the average Spearman's correlation of 7 STS tasks. We report the means and standard deviations over 5 random seeds.

Model	image → text		text → image	
	R@1	R@5	R@1	R@5
MCSE-BERT$_{wiki+flickr}$	16.7	43.5	22.5	50.4
MCSE-BERT$_{flickr}$	20.4	50.2	23.8	52.5
MCSE-BERT$_{wiki+coco}$	8.8	26.6	10.9	31.2
MCSE-BERT$_{coco}$	8.2	25.2	9.0	27.1

Table 6: Multimodal retrieval results on Flickr30k test set (1k) and MS-COCO minival set (5k).

similarity. Some retrieval examples are shown in Table 8. We observe that (1) SimCSE is prone to retrieving sentences with similar syntax, while MCSE can retrieve sentences that vary in syntax and share semantics. Examples: Q1, Q3, Q6. (2) MCSE is better at recognizing similar event scenes and capturing the number of entities. Examples: Q2, Q4, Q5.

Cross-Modal Retrieval We take BERT-based models (same seed) and conduct cross-modal retrieval experiments. We use the metric Recall@K, which is calculated based on if the ground truth of the query image or caption appears in the top-K retrieved captions or images. As results in Table 6 show, MCSE models also achieve a decent level of retrieval performance as a by-product of multimodal contrastive learning.

	Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.↑
flickr	SimCSE-BERT	$62.1_{\pm 0.5}$	$73.8_{\pm 0.9}$	$64.2_{\pm 0.6}$	$74.2_{\pm 0.8}$	$\mathbf{74.8}^{*}_{\pm 0.6}$	$67.1_{\pm 1.1}$	$65.4_{\pm 1.1}$	$68.8_{\pm 0.7}$
	MCSE-ResNet-BERT	$\mathbf{63.6}^{*}_{\pm 0.7}$	$\mathbf{74.0}_{\pm 0.9}$	$65.5_{\pm 1.1}$	$75.5_{\pm 0.2}$	$71.6_{\pm 0.4}$	$74.0_{\pm 0.4}$	$69.8_{\pm 0.3}$	$70.6_{\pm 0.5}$
	MCSE-CLIP-BERT	$63.1_{\pm 0.7}$	$73.9_{\pm 1.0}$	$\mathbf{65.8}^{*}_{\pm 0.9}$	$\mathbf{76.0}^{*}_{\pm 0.7}$	$70.7_{\pm 0.3}$	$\mathbf{74.9}^{*}_{\pm 0.5}$	$\mathbf{70.7}^{*}_{\pm 0.3}$	$\mathbf{70.7}^{*}_{\pm 0.2}$
	SimCSE-RoBERTa	$66.6_{\pm 0.5}$	$78.3_{\pm 0.5}$	$69.7_{\pm 0.6}$	$77.7_{\pm 0.5}$	$\mathbf{76.3}^{*}_{\pm 0.5}$	$75.8_{\pm 0.3}$	$66.2_{\pm 0.4}$	$72.9_{\pm 0.3}$
	MCSE-ResNet-RoBERTa	$\mathbf{67.6}^{*}_{\pm 0.5}$	$\mathbf{78.8}_{\pm 0.4}$	$\mathbf{70.1}_{\pm 0.3}$	$78.5_{\pm 0.2}$	$75.4_{\pm 0.5}$	$\mathbf{77.4}^{*}_{\pm 0.3}$	$68.6_{\pm 0.3}$	$\mathbf{73.8}^{*}_{\pm 0.2}$
	MCSE-CLIP-RoBERTa	$67.0_{\pm 0.5}$	$78.6_{\pm 0.4}$	$69.8_{\pm 0.5}$	$\mathbf{78.7}^{*}_{\pm 0.8}$	$74.9_{\pm 0.5}$	$\mathbf{77.4}^{*}_{\pm 0.4}$	$\mathbf{69.5}^{*}_{\pm 0.5}$	$73.7_{\pm 0.2}$
coco	SimCSE-BERT	$59.3_{\pm 0.9}$	$73.0_{\pm 1.2}$	$62.7_{\pm 0.6}$	$74.7_{\pm 0.7}$	$\mathbf{74.4}^{*}_{\pm 0.4}$	$65.3_{\pm 0.7}$	$65.4_{\pm 0.5}$	$67.8_{\pm 0.4}$
	MCSE-ResNet-BERT	$\mathbf{64.9}^{*}_{\pm 0.5}$	$\mathbf{74.8}^{*}_{\pm 0.9}$	$\mathbf{68.1}^{*}_{\pm 0.6}$	$\mathbf{76.8}^{*}_{\pm 0.6}$	$72.7_{\pm 0.8}$	$\mathbf{74.5}^{*}_{\pm 0.4}$	$69.7_{\pm 0.4}$	$\mathbf{71.6}^{*}_{\pm 0.2}$
	MCSE-CLIP-BERT	$64.8_{\pm 0.6}$	$74.1_{\pm 0.6}$	$68.0_{\pm 0.2}$	$76.2_{\pm 0.5}$	$71.6_{\pm 0.4}$	$\mathbf{74.5}^{*}_{\pm 0.3}$	$\mathbf{70.3}^{*}_{\pm 0.6}$	$71.4_{\pm 0.1}$
	SimCSE-RoBERTa	$64.7_{\pm 0.6}$	$79.2_{\pm 0.4}$	$70.2_{\pm 0.4}$	$79.0_{\pm 0.6}$	$\mathbf{78.2}_{\pm 0.5}$	$73.8_{\pm 0.5}$	$64.6_{\pm 0.3}$	$72.8_{\pm 0.3}$
	MCSE-ResNet-RoBERTa	$\mathbf{67.0}^{*}_{\pm 0.8}$	$\mathbf{79.4}_{\pm 0.4}$	$\mathbf{70.9}^{*}_{\pm 0.4}$	$\mathbf{80.0}^{*}_{\pm 0.4}$	$77.8_{\pm 0.5}$	$\mathbf{76.9}^{*}_{\pm 0.4}$	$67.9_{\pm 0.7}$	$\mathbf{74.3}^{*}_{\pm 0.3}$
	MCSE-CLIP-RoBERTa	$66.0_{\pm 1.0}$	$79.0_{\pm 0.7}$	$70.6_{\pm 0.6}$	$\mathbf{80.0}^{*}_{\pm 0.8}$	$77.6_{\pm 0.5}$	$76.5_{\pm 0.4}$	$\mathbf{68.4}^{*}_{\pm 0.8}$	$74.0_{\pm 0.2}$

$*$: difference between SimCSE and MCSE (ResNet/CLIP) is significant at $\alpha = 0.05$ according to an independent t-test.

Table 7: Performance comparison on STS tasks. STS-B: STS Benchmark, SICK-R: SICK-Relatedness, Avg.: average across 7 tasks. Models are trained with 5 random seeds and we report means and standard deviations.

Model	Result
Query 1: A young girl is washing her teddy bear in the kitchen sink.	
SimCSE:	A middle-aged woman is vacuuming her kitchen floor with a canister vac.
MCSE:	A young girl, blond and wearing a polka-dot shirt, washes a stuffed animal in a vanity sink.
Query 2: Three chefs , wearing white hats and black aprons , are preparing food in a crowded kitchen.	
SimCSE:	Numerous workers with blue shirts and white aprons are preparing fish for sale.
MCSE:	Three men are preparing food in a kitchen setting.
Query 3: A couple kisses in a shady walkway.	
SimCSE:	A couple strolls down a path near benches and water.
MCSE:	Couple kissing outside on street.
Query 4: A man is standing on the streets taking photographs.	
SimCSE:	People run a marathon on a city street with a crowd watching.
MCSE:	A guy wearing a white shirt is taking a picture.
Query 5: Two boys are playing in pool filled with sparkling blue water.	
SimCSE:	A little girl is swimming under the crystal blue water.
MCSE:	Two children are swimming in a pool.
Query 6: An old man sitting on a bench staring at the ocean.	
SimCSE:	A man sitting on a bench by the ocean.
MCSE:	An old man sits on a bench overlooking the water.

Table 8: Retrieved examples from Flickr30k test set.