Sentence Representations via Gaussian Embedding

Shohei Yoda **Hayato Tsukagoshi** **Ryohei Sasano** **Koichi Takeda**
Graduate School of Informatics, Nagoya University
yoda.shohei.a1@s.mail.nagoya-u.ac.jp,
tsukagoshi.hayato.r2@s.mail.nagoya-u.ac.jp,
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

Recent progress in sentence embedding, which represents a sentence's meaning as a point in a vector space, has achieved high performance on several tasks such as the semantic textual similarity (STS) task. However, a sentence representation cannot adequately express the diverse information that sentences contain: for example, such representations cannot naturally handle asymmetric relationships between sentences. This paper proposes GaussCSE, a Gaussian-distribution-based contrastive learning framework for sentence embedding that can handle asymmetric inter-sentential relations, as well as a similarity measure for identifying entailment relations. Our experiments show that GaussCSE achieves performance comparable to that of previous methods on natural language inference (NLI) tasks, and that it can estimate the direction of entailment relations, which is difficult with point representations.

1 Introduction

Sentence embeddings are representations to describe a sentence's meaning and are widely used in natural language tasks such as document classification (Liu et al., 2021), sentence retrieval (Wu et al., 2022), and question answering (Liu et al., 2020). In recent years, machine-learning-based sentence embedding methods with pre-trained language models have become mainstream, and various methods for learning sentence embeddings have been proposed (Reimers and Gurevych, 2019; Gao et al., 2021). However, as these methods represent a sentence as a point in a vector space and primarily use symmetric measures such as the cosine similarity to measure the similarity between sentences, they cannot capture asymmetric relationships between two sentences, such as entailment and hierarchical relations.

In this paper, we propose GaussCSE, a Gaussian-distribution-based contrastive sentence embedding to handle such asymmetric relationships between sentences by extending Gaussian embedding for words (Luke and Andrew, 2015). Figure 1 shows examples of sentence representations obtained by a previous method and by GaussCSE. Whereas the previous method represents a sentence as a point, GaussCSE represents a sentence as a region in the embedding space, and when two sentences have an entailment relation, the embedding of the entailing sentence contains the embedding of the entailed one. In these examples, S1 entails S2, but with previous methods, it is difficult to determine the entailment relation only from their embeddings. In contrast, by taking into account the variances of the distributions, GaussCSE can capture the asymmetric relationship where S1 entails S2 but S2 does not entail S1, as well as the fact that S3 is not in the entailment relationship with either S1 or S2.

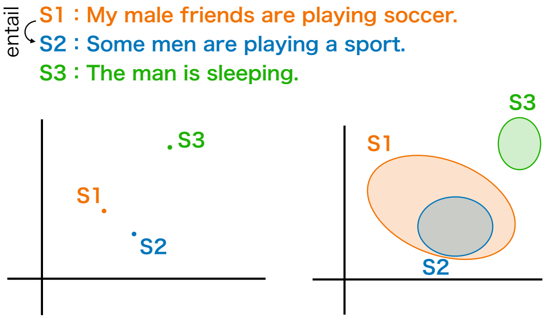
entail
S1 : My male friends are playing soccer.
S2 : Some men are playing a sport.
S3 : The man is sleeping.
S3
S1
S2
S3
S1
S2

Figure 1: Sentence representations in embedding spaces of a previous method (left) and GaussCSE (right).

To validate the usefulness of GaussCSE, we performed comparative experiments on two tasks: the natural language inference (NLI) task, and the task of predicting the entailment direction. The results demonstrate that GaussCSE can accurately predict the entailment direction while maintaining good performance on the NLI task.[1]

[1] We released our code and fine-tuned models at https://github.com/yoda122/GaussCSE.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 418–425
March 17-22, 2024 ©2024 Association for Computational Linguistics

2 Sentence Representations via Gaussian Embedding

GaussCSE is a method to obtain Gaussian embeddings of sentences by fine-tuning a pre-trained language model through contrastive learning. In this section, we first review a representative study of Gaussian embeddings and then review SimCSE, a method that acquires sentence embeddings via contrastive learning. We also review embedding methods that focus on asymmetric relations, which is closely related to our research. We then describe GaussCSE, which extends Gaussian embeddings and SimCSE.

2.1 Gaussian Embedding

One representative study on Gaussian embeddings sought to embed a word as a Gaussian distribution $\mathcal{N}$ (Luke and Andrew, 2015). In this method, the embedding N_i of a word w_i is represented as $\mathcal{N}(x; \mu_i, \Sigma_i)$ by using the mean vector μ_i in n-dimensional space and the variance-covariance matrix Σ_i.

The similarity between two words is measured using the Kullback-Leibler (KL) divergence, as defined by the following equation:

$$D_{\mathrm{KL}}(N_i||N_j) = \int_{x \in \mathbb{R}^n} \mathcal{N}(x; \mu_i, \Sigma_i) \log \frac{\mathcal{N}(x; \mu_i, \Sigma_i)}{\mathcal{N}(x; \mu_j, \Sigma_j)}. \quad (1)$$

The KL divergence is an asymmetric measure whose value changes when the arguments are reversed, which makes it suitable for capturing asymmetric relationships between embeddings, such as entailment relations.

2.2 Supervised SimCSE

In recent years, there has been a significant amount of research on methods for acquiring vector-based sentence embeddings (e.g., Zhang et al., 2020; Li et al., 2020; Tsukagoshi et al., 2021; Jiang et al., 2022; Chuang et al., 2022; Klein and Nabi, 2022). One of the most representative methods is supervised SimCSE (Gao et al., 2021), which trains sentence embedding models through contrastive learning on NLI datasets.

NLI datasets contain collections of sentence pairs, where each pair comprises a premise and a hypothesis and is labeled with “entailment,” “neutral,” or “contradiction.” Specifically, supervised SimCSE uses sentence pairs labeled with “entailment” as positive examples and those labeled with “contradiction” as hard negative examples. This approach achieves high performance on semantic textual similarity (STS) tasks, which evaluate how well sentence embedding models capture the semantic similarities between the sentences in a pair.

2.3 Sentence Embeddings for Asymmetric Relations

Similar to our approach, there are several studies that focus on the asymmetric relationships between sentences. Sen2Pro (Shen et al., 2023) represents sentences as probability distributions by sampling embeddings multiple times from pre-trained language models to reflect model and data uncertainty. RSE (Wang and Li, 2023) enriches sentence embeddings by incorporating relationships between sentences, such as entailment and paraphrasing, allowing for a more comprehensive representation of information. Unlike these methods, we propose a fine-tuning method utilizing contrastive learning for generating probabilistic distributed representations of sentences.

2.4 GaussCSE

To handle asymmetric relationships between sentences, we fine-tune pre-trained language models for representing sentences as Gaussian distributions via contrastive learning. We call this approach GaussCSE. First, a sentence s_k is fed to BERT, and the sentence's vector representation v_k is obtained from the embedding of the `[CLS]` token. When using RoBERTa, where the `[CLS]` token does not exist, the beginning-of-sentence token <s> is used as an alternative. Then, v_k is fed to two distinct linear layers, thus obtaining a mean vector μ_k and a variance vector σ_k, which is a diagonal component of a variance-covariance matrix. Note that, for computational efficiency, we adopt the same approach as in the previous study (Luke and Andrew, 2015); that is, we represent the variance by using only the diagonal elements of the variance-covariance matrix. Subsequently, we use μ_k and σ_k to obtain a Gaussian distribution N_k as a sentence representation.

We then define a similarity measure by the following equation to measure the asymmetric similarity of sentence s_i with respect to sentence s_j:

$$\mathrm{sim}(s_i||s_j) = \frac{1}{1 + D_{\mathrm{KL}}(N_i||N_j)}. \quad (2)$$

Because the KL divergence's range is $[0, \infty)$, the range of $\mathrm{sim}(s_i||s_j)$ is $(0, 1]$. When the variance of

N_i is greater than the variance of N_j, $D_{\mathrm{KL}}(N_i||N_j)$ tends to be larger than $D_{\mathrm{KL}}(N_j||N_i)$, which means that sim$(s_j||s_i)$ tends to be larger than sim$(s_i||s_j)$. Note that sim$(s_j||s_i)$ can be computed with the same computational complexity as cosine similarity, owing to representing the variance using only the diagonal elements of the variance-covariance matrix.[2]

When learning entailment relations, as with word representation by Gaussian embedding, GaussCSE performs learning such that the embedding of a sentence that entails another sentence has greater variance than the embedding of the sentence that is entailed. To achieve this, we use sentence pairs in an entailment relationship and increase the variance for premise (*pre*) sentences while decreasing it for hypothesis (*hyp*) sentences in NLI datasets. This is accomplished by training the model to increase sim$(hyp||pre)$ relative to sim$(pre||hyp)$ in accordance with the characteristics of the KL divergence as described above. Conversely, we decrease sim$(hyp||pre)$ when the premise does not entail the hypothesis, thus indicating that the sentences are not semantically related. As the KL divergence is more sensitive to differences in the mean than differences in the variance, this operation is expected to increase the distance between the two sentences' distributions.

Following the supervised SimCSE approach, we use contrastive learning with NLI datasets to train the model. During training, we aim to increase the similarity between positive examples and decrease the similarity between negative examples. We use the following three sets for positive and negative examples.

Entailment set The set of premise and hypothesis pairs labeled with "entailment." These semantically similar sentences are brought closer to each other.

Contradiction set The set of premise and hypothesis pairs labeled with "contradiction." These sentences with no entailment are used as negative examples and are spread apart from each other.

Reversed set The set of sentence pairs obtained by reversing each pair in the "entailment set." These sentences, whose entailment relation is reversed, are used as negative examples to increase the variance of premise sentences and decrease the variance of hypothesis sentences.

We compute sim$(hyp||pre)$ for both positive and negative examples. Specifically, the similarities of positive and negative examples in the three sets are computed by using n triplets of sentences (s_i, s_i^+, s_i^-), where s_i is premise, s_i^+ and s_i^- are entailment and contradiction hypotheses. The loss function for contrastive learning is defined as follows:

$$\begin{aligned}
V_E &= \Sigma_{j=1}^{n} e^{\mathrm{sim}(s_j^+||s_i)/\tau},\\
V_C &= \Sigma_{j=1}^{n} e^{\mathrm{sim}(s_j^-||s_i)/\tau},\\
V_R &= \Sigma_{j=1}^{n} e^{\mathrm{sim}(s_j||s_i^+)/\tau},\\
\mathcal{L} &= \sum_{i=1}^{n} -\log \frac{e^{\mathrm{sim}(s_i^+||s_i)/\tau}}{V_E + V_C + V_R},
\end{aligned} \tag{3}$$

where n is a batch size and τ is a temperature hyperparameter.

By performing learning with such a loss function, the model is expected to learn close mean vectors for sentences that are semantically similar. For entailment pairs, it is expected that the variance of the entailing sentence will become large and that of the entailed sentence will become small.

3 Experiments

We validated the effectiveness of GaussCSE through experiments on two tasks: NLI and prediction of the entailment direction.

3.1 NLI Task

We evaluated GaussCSE by comparing it with previous methods for recognizing textual entailment. NLI tasks usually perform three-way classification, but we performed two-way classification by collapsing the "neutral" and "contradiction" cases as "non-entailment," following revious studies on sentence embeddings. When the value of sim$(hyp||pre)$ was greater than a threshold, the relation was classified as "entailment"; otherwise, it was classified as "non-entailment."

We used the Stanford NLI (SNLI) (Bowman et al., 2015), Multi-Genre NLI (MNLI) (Williams et al., 2018), and SICK (Marelli et al., 2014) datasets for evaluation.[3] We adopted the accuracy as the evaluation metric and we used the threshold that achieved the highest accuracy on the development set to calculate the accuracy.

[2] More details are provided in Appendix A.

[3] The details of each dataset are in Appendix B

3.2 Entailment Direction Prediction Task

To validate that GaussCSE can capture asymmetric relationships, we performed the task of predicting which sentence entailed the other when given two sentences A and B in an entailment relation. We used the similarity to determine the entailment direction, where A is determined to be the entailing sentence if $\mathrm{sim}(B||A)$ was larger than $\mathrm{sim}(A||B)$. For this task, we used only sentence pairs labeled "entailment" in the datasets, and we adopted the accuracy as the evaluation metric. Note that SICK has instances with the bilateral entailment label. As there is no unique entailment direction between a pair of such sentences, we excluded such sentence pairs from the dataset in this experiment.

3.3 Experimental Setup

We used BERT-base, BERT-large, RoBERTa-base, and RoBERTa-large in transformers[4] as pre-trained language models, and report the results for BERT-base and RoBERTa-large in Section 3.4.[5] Following Gao et al. (2021), we combined the SNLI and MNLI datasets to form the training dataset. We conducted a statistical test for differences in accuracies when using the same pre-trained language model and dataset. Specifically, we tested the differences in accuracies obtained by the different loss functions with McNemar's test at a significance level of 0.05. Each experiment was conducted with five different random seeds, and the average was used as the final score. Details of other configurations are provided in the Appendix E.

We conducted experiments with four different loss functions, each with different training data: the entailment set alone (ent), the entailment and contradiction sets (ent+con), the entailment and reversed sets (ent+rev), and all sets (ent+con+rev).

3.4 Results

NLI task Table 1 lists the experimental results of the NLI task. The performance of supervised SimCSE[6] trained on BERT-base is given as a baseline. Among the four settings, those using both the entailment and contradiction sets (ent+con and ent+con+rev) performed relatively well, achieving comparable performance to that of SimCSE. Because the reversed set comprised semantically similar sentence pairs, treating such similar sentence pairs as negative examples did not contribute to performance in the NLI task.

Model	Loss function	SNLI	MNLI	SICK	Avg.
SimCSE (BERT-base)		74.96	78.18	86.11	79.75
BERT -base	ent	72.44	67.92	67.70	69.35
	ent+con	**77.63**	**77.71**	80.38	78.57
	ent+rev	69.32	66.04	67.93	67.76
	ent+con+rev	76.64	76.85	**83.15**	**78.88**
RoBERTa -large	ent	72.54	68.67	69.96	70.39
	ent+con	**78.05**	**79.96**	81.05	79.68
	ent+rev	69.17	66.47	67.84	67.82
	ent+con+rev	76.68	79.07	**84.17**	**79.97**

Table 1: Experimental results of the NLI task.

Model	Loss function	SNLI	MNLI	SICK	Avg.
Length-baseline		92.63	82.64	69.14	81.47
BERT -base	ent	64.84	61.11	60.10	62.01
	ent+con	64.55	56.84	69.67	63.68
	ent+rev	**97.60**	**92.64**	**87.80**	**92.68**
	ent+con+rev	**97.38**	**91.92**	**86.22**	91.84
RoBERTa -large	ent	66.91	60.88	61.56	63.11
	ent+con	64.57	55.31	71.38	63.75
	ent+rev	**97.89**	**93.97**	**88.71**	**93.52**
	ent+con+rev	**97.42**	93.61	86.57	92.53

Table 2: Experimental results of the entailment direction prediction task.

Entailment Direction Prediction Task Table 2 lists the experimental results of entailment direction prediction. The performance of a baseline method which determines longer sentence as entailing one (length-baseline) is also given. We can see that the leveraging of the reversed set significantly improved the accuracy, and outperformed the baseline method. This indicates that GaussCSE succeeds in acquiring embeddings that can recognize the direction of the entailment by using the reverse set as negative examples.

Regarding the differences in accuracy among the datasets, accuracies of over 97% and over 93% were achieved on the SNLI and MNLI datasets, respectively, whereas the accuracy on the SICK dataset was relatively low, 89% at the highest. These results were presumably due to the datasets' characteristics regarding the different lengths of sentence pairs.[7] However, the fact that GaussCSE achieved 89% accuracy by leveraging the reversed set even on the SICK dataset indicates that it took the semantic content of sentences into account in capturing entailment relationships.

Considering the overall experimental results of the two tasks, we can conclude that by leveraging

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

[5] All the experimental results are in Appendix C and D.

[6] https://github.com/princeton-nlp/SimCSE

[7] Sentence length ratios of these datasets are provided in Appendix F.

both contradiction and reverse sets as negative examples, GaussCSE could achieve high accuracy in predicting the direction of entailment relations while retaining the performance of the NLI task.

4 Conclusion

In this paper, we have presented GaussCSE, a Gaussian-distribution-based contrastive sentence embedding to handle asymmetric relationships between sentences. GaussCSE fine-tunes pre-trained language models via contrastive learning with asymmetric similarity. Through experiments on the NLI task and entailment direction prediction, we have demonstrated that GaussCSE achieves comparative performance to previous methods on NLI task and also accurately estimate the direction of entailment relations, which is difficult with conventional sentence representations.

In this study, we used a Gaussian distribution to represent the spread of the meaning of a sentence in the embedding space, we would like to conduct a comparison with the use of other types of embedding, such as Hyperbolic Embeddings (Nickel and Kiela, 2017) or Box Embeddings (Dasgupta et al., 2022) in future work.

Limitations

Our proposed method involves supervised learning to acquire Gaussian-based sentence representations, but the optimal choices of the probability distribution and domain representation are not yet known. Additionally, for low-resource languages on which large-scale NLI datasets may not be available for use as supervised training data, alternative training approaches will need to be explored. To address these challenges, future investigations could consider alternative embedding methods such as box embeddings going beyond Gaussian-based approaches, as well as experiments using multilingual models. Furthermore, it would be beneficial to explore unsupervised learning techniques that are less dependent on language resources.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 21H04901.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4207–4218.

Shib Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Li, and Andrew McCallum. 2022. Word2Box: Capturing Set-Theoretic Semantics of Words using Box Embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2263–2276.

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.

Loshchilov Ilya and Hutter Frank. 2019. Decoupled Weight Decay Regularization. In *Proceedings of 7th International Conference on Learning Representations (ICLR)*.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Prompt-BERT: Improving BERT Sentence Embeddings with Prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8826–8837.

Tassilo Klein and Moin Nabi. 2022. SCD: Self-Contrastive Decorrelation of Sentence Embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 394–400.

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.

Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. RikiNet: Reading Wikipedia pages for natural question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6762–6771.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. 2021. Effective Convolutional Attention Network for Multi-label Clinical Document Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5941–5953.

Vilnis Luke and McCallum Andrew. 2015. Word Representations via Gaussian Embedding. In *Proceedings of 3rd International Conference on Learning Representations (ICLR)*.

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

Maximilian Nickel and Douwe Kiela. 2017. Poincaré Embeddings for Learning Hierarchical Representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 6341–6350.

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.

Lingfeng Shen, Haiyun Jiang, Lemao Liu, and Shuming Shi. 2023. Sen2Pro: A probabilistic perspective to sentence embedding from pre-trained language model. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 315–333.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2021. DefSent: Sentence Embeddings using Definition Sentences. 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 411–418.

Bin Wang and Haizhou Li. 2023. Relational sentence embedding for flexible semantic matching. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 238–252.

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 (NAACL)*, pages 1112–1122.

Bohong Wu, Zhuosheng Zhang, Jinyuan Wang, and Hai Zhao. 2022. Sentence-aware Contrastive Learning for Open-Domain Passage Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1062–1074.

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.

A Computation Complexity of KL divergence

The KL divergence between Gaussian distributions can be computed analytically using the following formula:

$$D_{KL}(N_i \| N_j) = \frac{1}{2}[\log \frac{|\Sigma_j|}{|\Sigma_i|} + tr(\Sigma_j^{-1}\Sigma_i) + (\mu_i - \mu_j)^T \Sigma_j^{-1}(\mu_i - \mu_j) - d],$$

where d denotes the dimension of N_1 and N_2. Since we set all elements except the diagonal components of the covariance matrix to zero, Σ^{-1} becomes the reciprocal of each component in Σ and $|\Sigma|$ can be computed as the product of its diagonal components. The calculations for each term can be done in $O(d)$, resulting in an overall computational complexity of $O(d)$, which is the same with the computational complexity of cosine similarity.

B Details of NLI Datasets

SNLI, MNLI and SICK datasets comprise pairs of premise and hypothesis sentences. SNLI contains approximately 570,000 sentence pairs, where the premise sentences were obtained by crawling image descriptions, and the hypothesis sentences were manually generated and annotated by human annotators. MNLI contains approximately 430,000 sentence pairs, and its construction method was similar to that of SNLI. The key difference is that MNLI includes premise sentences from both written and spoken speech in a wider range of styles, degrees of formality, and topics as compared to SNLI. SICK contains approximately 10,000 sentence pairs. Like SNLI, the premise sentences in SICK were constructed from sources such as image descriptions; however, a portion of the premise sentences was automatically replaced by using specific rules to generate the hypothesis sentences.

C Full Results of the NLI Task

Table 3 shows experimental results of the NLI task for all pre-trained models. In addition to accuracy (Acc.), we adopted area under the precision-recall curve (AUPRC) as the evaluation metrics for this NLI task. To calculate the AUPRC, we varied the threshold for determining whether two sentences

Model	Loss function	SNLI		MNLI		SICK	
		Acc.	AUPRC	Acc.	AUPRC	Acc.	AUPRC
SimCSE (BERT-base)		74.96	66.76	78.18	75.88	86.11	81.41
BERT -base	ent	72.44	60.65	67.92	56.96	67.70	68.26
	ent+con	**77.63**	**70.95**	**77.71**	**74.21**	80.38	**82.12**
	ent+rev	69.32	54.21	66.04	53.87	67.93	63.60
	ent+con+rev	76.64	67.07	76.85	71.34	**83.15**	79.45
BERT -large	ent	73.51	62.79	69.88	61.96	70.85	72.56
	ent+con	**77.79**	**71.11**	**78.31**	**75.23**	**81.24**	**83.73**
	ent+rev	69.46	54.67	66.23	55.28	68.13	64.73
	ent+con+rev	77.02	68.02	**77.86**	73.65	**83.73**	80.99
RoBERTa -base	ent	72.10	59.98	68.77	58.39	67.50	67.02
	ent+con	**77.60**	**70.58**	**78.76**	**75.90**	**81.21**	**83.26**
	ent+rev	69.35	54.21	66.19	54.50	66.54	61.90
	ent+con+rev	76.37	66.39	**77.74**	73.01	**82.95**	80.46
RoBERTa -large	ent	72.54	60.74	68.67	60.21	69.96	72.01
	ent+con	**78.05**	**71.41**	**79.96**	**78.12**	81.05	**84.91**
	ent+rev	69.17	54.54	66.47	55.96	67.84	68.05
	ent+con+rev	76.68	67.14	79.07	75.58	**84.17**	82.41

Table 3: Experimental results of the NLI task for all combination of a pre-trained model and loss function.

were in an entailment relation from 0 to 1 in steps of 0.001.

D Full Results of the Entailment Direction Prediction Task

Table 4 shows experimental results of the entailment direction prediction task for all combinations of pre-trained models and loss functions.

E Detail of Experimental Setup

The fine-tuning epoch size is 3, the temperature hyperparameter is 0.05, and the optimizer is AdamW (Ilya and Frank, 2019). The embedding dimensions were 768 for BERT-base and RoBERTa-base and 1024 for BERT-large and RoBERTa-large. These settings are the same as SimCSE (Gao et al., 2021). Fine-tuning for BERT-base and RoBERTa-base took about 40 minutes on a single NVIDIA A100. Fine-tuning for BERT-large and RoBERTa-large took about 2 hours on the same GPU. We carry out grid-search of batch size $\in \{16, 32, 64, 128\}$ and learning rate $\in \{1\text{e}{-}5, 3\text{e}{-}5, 5\text{e}{-}5\}$ on the SNLI development set, then used the best-performing combination in the in-training evaluation described below. The learning rate is 0 at the beginning and increases linearly to a set value in the final step. Table 5 summarizes the detailed grid-search results. The values in the table represent the AUC values of the precision-recall curve for the NLI task for each batch size and learning rate, where each value was multiplied by 100.

In each experiment, the AUC of the precision-recall curve for the NLI task on the SNLI development set was calculated every 100 training steps, and the model with the best performance was used for the final evaluation on the test set.

Model	Loss function	SNLI	MNLI	SICK	Avg.
Length-baseline		92.63	82.64	69.14	81.47
BERT -base	ent	64.84	61.11	60.10	62.01
	ent+con	64.55	56.84	69.67	63.68
	ent+rev	**97.60**	**92.64**	**87.80**	**92.68**
	ent+con+rev	**97.38**	**91.92**	**86.22**	91.84
BERT -large	ent	62.06	60.09	62.09	61.41
	ent+con	62.43	54.87	69.01	62.10
	ent+rev	**97.66**	**92.76**	**88.03**	**92.81**
	ent+con+rev	**97.55**	**93.11**	**85.94**	92.20
RoBERTa -base	ent	65.84	60.41	59.69	61.98
	ent+con	65.66	55.24	69.97	63.62
	ent+rev	**97.74**	**93.15**	**87.90**	92.93
	ent+con+rev	**97.44**	**93.10**	**88.43**	**92.99**
RoBERTa -large	ent	66.91	60.88	61.56	63.11
	ent+con	64.57	55.31	71.38	63.75
	ent+rev	**97.89**	**93.97**	**88.71**	**93.52**
	ent+con+rev	**97.42**	93.61	86.57	92.53

Table 4: Experimental results of the entailment direction prediction task for all combinations of pre-trained models and loss functions.

F Ratio of Length of Sentence Pairs

Figure 2 shows histograms of the ratios of the length of the premise sentence to that of the hypothesis sentence for each sentence pair in each dataset.

Model	Batch size	Learning rate		
		1e-5	3e-5	5e-5
BERT-base	16	63.05	65.72	**66.21**
	32	62.02	64.69	64.84
	64	60.44	62.93	64.20
	128	58.99	61.26	62.66
BERT-large	16	64.66	**65.65**	61.09
	32	63.73	65.56	63.42
	64	62.24	65.01	62.46
	128	60.72	63.41	64.68
RoBERTa-base	16	64.66	65.78	**66.31**
	32	63.06	65.09	65.68
	64	61.59	64.18	64.95
	128	60.48	62.54	63.84
RoBERTa-large	16	66.22	**67.17**	61.69
	32	65.96	67.10	60.64
	64	64.26	66.01	66.88
	128	63.07	64.91	65.72

Table 5: Grid-search results.

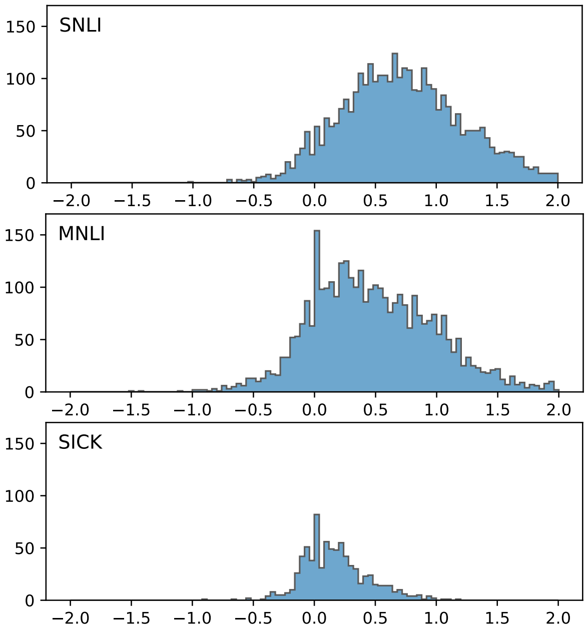

Figure 2: Histograms representing the distributions of the logarithmic values of the length ratios of the premise sentences and their corresponding hypothesis sentences in the SNLI, MNLI, and SICK datasets. The horizontal axis represents the logarithm of the length ratio, and the vertical axis represents the number of sentence pairs.