| cases | | | |
|---|---|---|---|
| **doc_1** | **doc_2** | **decision** | **id** |

**doc_1**

| | |
|---|---|
| **authors** | • Xing Wu<br>• Chaochen Gao<br>• Yipeng Su<br>• Jizhong Han<br>• Zhongyuan Wang<br>• Songlin Hu |
| **title** | Smoothed Contrastive Learning for Unsupervised Sentence Embedding |
| **publication_date** | 2022-09-12 00:00:00 |
| **source** | SupportedSources.INTERNET_ARCHIVE |
| **journal** | |
| **volume** | |
| **doi** | |
| **urls** | • https://web.archive.org/web/20220915033902/https://arxiv.org/pdf/2109.04321v2.pdf |
| **id** | id-8668783657730631915 |
| **abstract** | Contrastive learning has been gradually applied to learn high-quality unsupervised sentence embedding. Among the previous un-supervised methods, the latest state-of-the-art method, as far as we know, is unsupervised SimCSE (unsup-SimCSE). Unsup-SimCSE uses the InfoNCE1loss function in the training stage by pulling semantically similar sentences together and pushing apart dis-similar ones.Theoretically, we expect to use larger batches in unsup-SimCSE to get more adequate comparisons among samples and avoid overfitting. However, increasing the batch size does not always lead to improvements, but instead even lead to performance degradation when the batch size exceeds a threshold. Through statistical observation, we find that this is probably due to the introduction of low-confidence negative pairs after in-creasing the batch size. To alleviate this problem, we introduce a simple smoothing strategy upon the InfoNCE loss function, termedGaussian Smoothing InfoNCE (GS-InfoNCE).Specifically, we add random Gaussian noise vectors as negative samples, which act asa smoothing of the negative sample space.Though being simple, the proposed smooth-ing strategy brings substantial improvements to unsup-SimCSE. We evaluate GS-InfoNCEon the standard semantic text similarity (STS)task. GS-InfoNCE outperforms the state-of-the-art unsup-SimCSE by an average Spear-man correlation of 1.38%, 0.72%, 1.17% and0.28% on the base of BERT-base, BERT-large,RoBERTa-base and RoBERTa-large, respectively. |
| **versions** | |

**doc_2**

| | |
|---|---|
| **authors** | • Xing Wu<br>• Chaochen Gao<br>• Yipeng Su<br>• Jizhong Han<br>• Zhongyuan Wang<br>• Songlin Hu |
| **title** | Smoothed Contrastive Learning for Unsupervised Sentence Embedding |
| **publication_date** | 2021-09-09 14:54:24+00:00 |
| **source** | SupportedSources.ARXIV |
| **journal** | None |
| **volume** | |
| **doi** | |
| **urls** | • http://arxiv.org/pdf/2109.04321v2<br>• http://arxiv.org/abs/2109.04321v2<br>• http://arxiv.org/pdf/2109.04321v2 |
| **id** | id2150374846178866255 |
| **abstract** | Contrastive learning has been gradually applied to learn high-quality unsupervised sentence embedding. Among the previous un-supervised methods, the latest state-of-the-art method, as far as we know, is unsupervised SimCSE (unsup-SimCSE). Unsup-SimCSE uses the InfoNCE1loss function in the training stage by pulling semantically similar sentences together and pushing apart dis-similar ones.Theoretically, we expect to use larger batches in unsup-SimCSE to get more adequate comparisons among samples and avoid overfitting. However, increasing the batch size does not always lead to improvements, but instead even lead to performance degradation when the batch size exceeds a threshold. Through statistical observation, we find that this is probably due to the introduction of low-confidence negative pairs after in-creasing the batch size. To alleviate this problem, we introduce a simple smoothing strategy upon the InfoNCE loss function, termedGaussian Smoothing InfoNCE (GS-InfoNCE).Specifically, we add random Gaussian noise vectors as negative samples, which act asa smoothing of the negative sample space.Though being simple, the proposed smooth-ing strategy brings substantial improvements to unsup-SimCSE. We evaluate GS-InfoNCEon the standard semantic text similarity (STS)task. GS-InfoNCE outperforms the state-of-the-art unsup-SimCSE by an average Spear-man correlation of 1.38%, 0.72%, 1.17% and0.28% on the base of BERT-base, BERT-large,RoBERTa-base and RoBERTa-large, respectively. |
| **versions** | |

| decision | id |
|---|---|
| DUPLICATES | 14 |