Sparse Contrastive Learning of Sentence Embeddings

Anonymous ACL submission

Abstract

Recently, SimCSE has shown the feasibility of contrastive learning in training sentence embeddings and illustrates its expressiveness in spanning an aligned and uniform embedding space. However, prior studies have shown that dense models could contain harmful parameters that affect the model performance, and it is no wonder that SimCSE can also be invented with such parameters. Driven by this, parameter sparsification is applied, where alignment and uniformity scores are used to measure the contribution of each parameter to the overall quality of sentence embeddings. Drawing from a preliminary study, we consider parameters with minimal contributions to be detrimental, as their sparsification results in improved model performance. To discuss the ubiquity of detrimental parameters and remove them, more experiments on the standard semantic textual similarity (STS) tasks and transfer learning tasks are conducted, and the results show that the proposed sparsified SimCSE (SparseCSE) has excellent performance in comparison with SimCSE. Furthermore, through in-depth analysis, we establish the validity and stability of our sparsification method, showcasing that the embedding space generated by SparseCSE exhibits improved alignment compared to that produced by SimCSE. Importantly, the uniformity remains uncompromised.

1 Introduction

The task of learning universal sentence embeddings using large-scale pre-trained models has been extensively explored in prior research (Logeswaran and Lee, 2018; Reimers and Gurevych, 2019; Li et al., 2020a; Zhang et al., 2020a; Gao et al., 2021; Liu et al., 2021; Yan et al., 2021; Feng et al., 2022). More recently, contrastive learning has been proposed as a method to enhance the quality of sentence embeddings (Qiu et al., 2022; Zhang et al., 2020a; Gao et al., 2021; Liu et al., 2021; Yan et al., 2021). By employing contrastive learning, semantically similar sentences are brought closer together while dissimilar sentences are pushed apart, thereby a semantically-driven structure is established within the space of sentence embeddings.

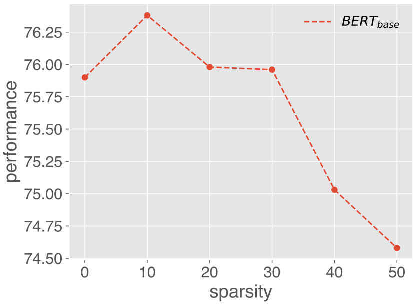

Figure 1: The average performance of SimCSE-BERT$_{\text{base}}$ on STS tasks when pruned at sparsity levels of 10%, 20%, 30%, 40% and 50% respectively. Details of the pruning method can be found in Section 2, while the task specifics and metrics are introduced in Section 3.

Unsupervised SimCSE (unsup-SimCSE) is a notable framework for contrastive sentence embeddings (Gao et al., 2021). It utilizes dropout as a simple data augmentation technique to create positive pairs and employs a cross-entropy objective based on cosine similarity for contrastive learning. Inspired by recent research on parameter sparsification (Xia et al., 2022; Prasanna et al., 2020; Hou et al., 2020; Michel et al., 2019), particularly the works on the lottery ticket hypothesis (LTH) (Frankle and Carbin, 2019; Bai et al., 2022; Frankle et al., 2020; Yang et al., 2022b) showing its effectiveness in improving model performance through pruning, we hypothesize that certain parameters in SimCSE might hinder the representation of universal sentence embeddings. By removing these parameters, we anticipate an improvement in the model's performance.

To accurately estimate the contribution of each parameter, it is essential to consider properties that characterize contrastive representation learning. In the literature (Wang and Isola, 2020), two such properties have been proposed: alignment and uniformity. Alignment measures the proximity of features derived from positive pairs, indicating how well the model captures semantic similarity. On the other hand, uniformity pertains to the distribution of features across the hypersphere, ensuring that the representations are spread out evenly. These properties offer valuable insights into understanding and evaluating contrastive representation learning. Utilizing alignment and uniformity as guiding principles, we propose an innovative approach, named alignment and uniformity score, to quantify parameter contribution during the preparation phase for pruning.

Based on a pilot study presented in Figure 1, we observed that model performance does not consistently decrease during pruning, instead it exhibits an upward trend when the model is less sparse. This suggests that the parameters with the lowest scores are detrimental to model performance, as evidenced by the performance improvement resulting from their pruning. Building upon this, we conducted a series of more extensive and detailed experiments to explore the ubiquity of detrimental parameters and assess the stability of our proposed pruning method.

Specifically, our approach consists of three stages: training, parameter sparsification, and rewinding. First, we train an unsupervised SimCSE model using a pre-trained language model (LM). Then, we estimate alignment and uniformity scores for each parameter based on the trained model's feedback. Parameters with low scores are pruned and varying sparsity is attempted in formal experiments than in pilot study to clearly identify harmful parameters. Finally, the remaining parameters are initialized, and the pruned model is fine-tuned to regain its performance. Our model is thus named SparseCSE.

We extensively evaluate SparseCSE on seven STS tasks and seven transfer learning tasks. Results show that SparseCSE outperforms SimCSE, demonstrating its superior performance. Our pruning method is also shown to effectively identify the optimal sparsity for pruning, further enhancing performance. Further analysis reveals the stability of our pruning method across multiple tasks. Comparison with other works highlights the similarity of SparseCSE to SimCSE in uniformity and its competitive performance in alignment.

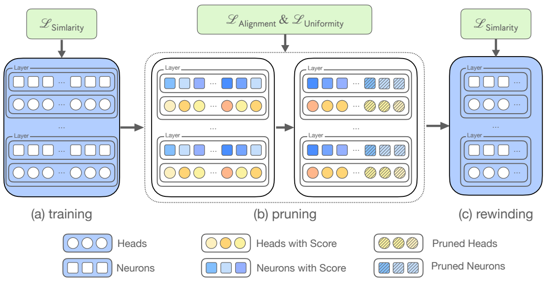

Figure 2: The process of obtaining SparseCSE

2 Our Method

Similar to the lottery ticket approach (Frankle and Carbin, 2019), our method is illustrated in Figure 2, following a training, pruning, and rewinding paradigm.

2.1 Training and Rewinding

To effectively train a model that captures universal sentence embeddings, we adopt a contrastive framework similar to the previous work (Gao et al., 2021). This framework is also utilized during the rewinding stage. In this framework, we employ dropout to create positive representation pairs (h_i, h_i^+) for each sentence x_i in a collection of sentences $x_i{}_{i=1}^{m}$. The training objective for this contrastive framework, using a mini-batch of N pairs, can be expressed as follows:

$$\mathcal{L}_{\text{similarity}}^{(i)} = -\log \frac{e^{\mathbf{sim}(h_i, h_i^+)/\tau}}{\sum_{j=1}^{n} e^{\mathbf{sim}(h_i, h_j^+)/\tau}},$$

where τ is a temperature hyperparameter and $\mathbf{sim}(h_1, h_2)$ represents the cosine similarity $\frac{{h_1}^{\mathsf{T}} \cdot h_2}{\|h_1\| \cdot \|h_2\|}$.

During training, an initial pretrained language model (LM) is utilized, and all parameters are involved in this phase. However, during rewinding, only the remaining parameters after pruning are applied to the LM, with their values initialized to their early-stage pre-training values. The objective of rewinding is to enable the pruned model to restore its performance prior to pruning.

2.2 Pruning

BERT (Devlin et al., 2019) (or Roberta (Liu et al., 2019)) is composed of multiple stacked encoder layers known as transformers. Each transformer encoder consists of a multi-head self-attention block (MHA) and a feed-forward network block (FFN). In line with prior research (Prasanna et al., 2020;

Hou et al., 2020; Michel et al., 2019), our pruning approach primarily focuses on sparsifying the attention heads in the MHA blocks and the intermediate neurons in the FFN blocks. To determine which parameters to prune, we associate a set of mask variables with them (Yang et al., 2022a,b) and compare the model's performance before and after the operation.

For a MHA block with N_H independent heads, the i-th head is parameterized by $\mathbf{W}_Q^{(i)}$, $\mathbf{W}_K^{(i)}$, $\mathbf{W}_V^{(i)} \in \mathbb{R}^{d\times d_A}$, and $\mathbf{W}_O^{(i)} \in \mathbb{R}^{d_A\times d}$, all parallel heads are further summed to produce the final output. Then variable $\xi^{(i)}$ with values in $\{0,1\}$ is defined for masking each attention head, and it can be represented as:

$$\mathrm{MHA}(\mathbf{X}) = \sum_{i=1}^{N_H} \xi^{(i)} \mathrm{Attn}^{(i)}_{\mathbf{W}_Q^{(i)},\mathbf{W}_K^{(i)},\mathbf{W}_V^{(i)},\mathbf{W}_O^{(i)}}(\mathbf{X}),$$

where the input $\mathbf{X} \in \mathbb{R}^{l\times d}$ represents a l-length sequence of d-dimensional vectors and $\xi^{(i)}$ is designed as a switching value, when $\xi^{(i)}$ equal to 1, it means keeping the attention head retained, and when it equal to 0 means removing that attention head from the MHA.

On the other hand, a FFN block includes two fully-connected layers parameterized by $\mathbf{W}_1 \in \mathbb{R}^{d\times D_F}$ and $\mathbf{W}_2 \in \mathbb{R}^{D_F\times d}$, denoting D_F as the number of neurons in the intermediate layer of FFN. Likewise, we define variable ν to mask the neurons in the intermediate layer of FFN:

$$\mathrm{FFN}(\mathbf{A}) = \sum_{i=1}^{D_F} \nu^{(i)} \mathrm{GELU}_{\mathbf{W}_1,\mathbf{W}_2}(\mathbf{A}),$$

where the input $\mathbf{A} \in \mathbb{R}^{l\times d}$ defines a d-dimensional vectors with l-length sequence.

2.3 Alignment and Uniformity Score

In order to determine the parameters that have a greater impact on the distribution of universal sentence embeddings, we introduce a joint objective based on the alignment and uniformity properties (Wang and Isola, 2020).

Here is the formulation of the alignment loss:

$$\mathcal{L}_{\textsf{Alignment}} \triangleq \log \mathop{\mathbb{E}}_{\mathbf{x_i},\mathbf{x_i}^+\sim\mathcal{N}_{pos}} \|\mathbf{h_i} - \mathbf{h_i}^+\|^2,$$

where $h_i, {h_i}^+$ are representations of $x_i, {x_i}^+$, which are a pair of positive sentences in a batch of N_{pos} sentences. It indicates that the sentences with similar semantics are expected to be closer in the embedding space.

And, here is the formulation of the uniformity loss:

$$\mathcal{L}_{\textsf{Uniformity}} \triangleq \log \mathop{\mathbb{E}}_{\mathbf{x_i},\mathbf{x_j}\sim\mathcal{N}} \mathrm{e}^{-2\|\mathbf{h_i}-\mathbf{h_j}\|^2},$$

where h_i, h_j are representations of x_i, x_j, which are different sentences in a batch of N sentences. It indicates that sentence embeddings with different semantics are supposed to distribute on the hypersphere by larger distances.

To balance the alignment and uniformity, we introduce a coefficient λ to quantify the tradeoff. The joint loss $\mathcal{L}_{\textsf{Score}}$ for further score calculation can be be written as below:

$$\mathcal{L}_{\textsf{Score}} = \lambda \cdot \mathcal{L}_{\textsf{Alignment}} + (1-\lambda)\cdot \mathcal{L}_{\textsf{Uniformity}},$$

Finally, according to the literature (Molchanov et al., 2017), the scores of the attention heads in MHA and the intermediate neurons in FFN can be depicted as:

$$\mathbb{I}_{\textsf{head}}^{(i)} = \mathbb{E}_{\mathcal{D}} \left| \frac{\partial \mathcal{L}_{\textsf{Score}}}{\partial \xi^{(i)}} \right|, \quad \mathbb{I}_{\textsf{neuron}}^{(i)} = \mathbb{E}_{\mathcal{D}} \left| \frac{\partial \mathcal{L}_{\textsf{Score}}}{\partial \nu^{(i)}} \right|,$$

where $\mathcal{D}$ is a data distribution, $\mathbb{E}$ represents expectation.

After estimating the scores, we rank the attention heads and intermediate neurons respectively with the scores, and prune the parameters with low scores according to the constraint of the given sparsity.

3 Experiments

3.1 Baselines & Implementation

We start by training unsup-SimCSE models using popular language models (BERT$_{\text{base}}$, BERT$_{\text{large}}$, Roberta$_{\text{base}}$) as our baselines. Both training and rewinding process of sparseCSE follow the training details of SimCSE (Gao et al., 2021). Pruning process is produced with varying sparsity levels from 1% to 50% and different value of coefficient λ. More details are shown in Appendix A.

3.2 Evaluation

Following SimCSE (Gao et al., 2021), we evaluate sentence embeddings on 7 semantic textual similarity (STS) tasks, which include STS

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
SimCSE-BERT$_{\text{base}}$	70.37	82.53	73.46	81.58	77.61	76.55	69.22	75.9
SparseCSE$_{2\%}$	$70.15^{-0.22}$	$82.25^{-0.28}$	$74.16^{+0.70}$	$82.15^{+0.57}$	$78.52^{+0.91}$	$78.71^{+2.16}$	$72.76^{+3.54}$	$76.96^{+1.06}$
SparseCSE$_{\text{best}}$	$71.70^{+1.33}_{10\%}$	$83.41^{+0.88}_{25\%}$	$74.16^{+0.70}_{2\%}$	$82.58^{+1.00}_{25\%}$	$79.10^{+1.49}_{4\%}$	$78.71^{+2.16}_{2\%}$	$72.76^{+3.54}_{2\%}$	$77.49^{+1.59}$
SimCSE-BERT$_{\text{large}}$	69.93	84.04	75.15	82.99	78.32	79.12	74.16	77.67
SparseCSE$_{2\%}$	$69.31^{-0.62}$	$83.69^{-0.35}$	$75.72^{+0.57}$	$83.21^{+0.22}$	$79.34^{+1.02}$	$79.41^{+0.29}$	$74.76^{+0.60}$	$77.92^{+0.25}$
SparseCSE$_{\text{best}}$	$70.67^{+0.74}_{1\%}$	$84.60^{+0.56}_{8\%}$	$75.84^{+0.69}_{8\%}$	$83.21^{+0.22}_{1\%}$	$79.60^{+1.28}_{8\%}$	$79.41^{+0.29}_{1\%}$	$75.27^{+1.11}_{3\%}$	$78.32^{+0.64}$
SimCSE-Roberta$_{\text{base}}$	67.45	81.28	72.74	81.31	80.87	80.12	68.37	76.02
SparseCSE$_{1\%}$	$67.85^{+0.40}$	$81.32^{+0.04}$	$73.09^{+0.35}$	$81.82^{+0.51}$	$81.02^{+0.15}$	$80.29^{+0.17}$	$68.76^{+0.39}$	$76.31^{+0.29}$
SparseCSE$_{\text{best}}$	$68.05^{+0.60}_{4\%}$	$81.82^{+0.54}_{4\%}$	$73.32^{+0.58}_{4\%}$	$82.29^{+0.98}_{20\%}$	$81.02^{+0.15}_{2\%}$	$80.29^{+0.17}_{1\%}$	$68.76^{+0.39}_{1\%}$	$76.48^{+0.46}$

Table 1: Performance of sparseCSE on STS tasks. Each backbone has three rows: the baseline, the result with optimal sparsity based on average score, and the result with optimal sparsity based on each task. The optimal sparsity values are shown in the bottom right corner. The improvements over the baseline are highlighted in red in the upper right corner.

2012–2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS Benchmark (Cer et al., 2017) and SICK-Relatedness (Marelli et al., 2014). STS tasks can reveal the ability of clustering semantically similar sentences, which is one of the main goals for sentence embeddings. Furthermore, we also introduce 7 transfer learning tasks into evaluation as a supplementary prove, and the details of the tasks are shown in Appendix B.

3.3 Main Results

Table 1 shows the results on STS tasks. The best results based on each task are all improved, and the model on BERT$_{\text{base}}$ improves the average result from 75.9% to 77.49%. We also determine an optimal sparsity corresponding to the best average score of all tasks. We observe that pruning the models with this specific sparsity level leads to improvements in almost every task. The results of transfer task are shown in Table 2 in Appendix B, where the same trend prove the ubiquity of the phenomenon found in Table 1.

Considering the results comprehensively, we observe that the pruned models tend to exhibit optimal performance at lower sparsity levels. In order to carry out an in-depth analysis of this phenomenon, a detailed discussion on varying sparsity and trade-off of alignment and uniformity are produced in Appendix C and D.

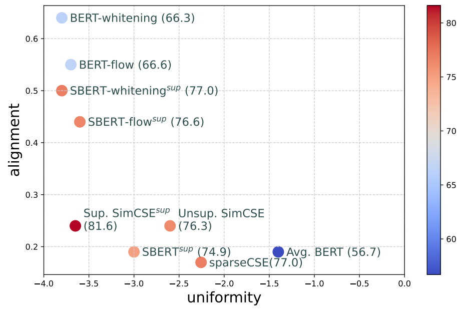

Figure 3: Analysis on alignment and uniformity (the smaller, the better). Points represent average STS performance using BERT$_{\text{base}}$, with "sup" marked of supervised methods.

Evaluation on Alignment and Uniformity Figure 3 illustrates the uniformity and alignment scores of various methods along with their performance on the STS task. The methods include BERT (Devlin et al., 2019), SimCSE (Gao et al., 2021), SBERT (Reimers and Gurevych, 2019), BERT-flow (Li et al., 2020b), BERT-whitening (Su et al., 2021) and sparseCSE. As a sparse version of Unsupervised SimCSE, sparseCSE inherits its advantages in alignment compared to post-processing methods (like BERT-flow and BERT-whitening) and uniformity compared to pre-trained embeddings (like BERT). Benefited from sparsity based on both alignment and uniformity properties, sparseCSE demonstrates significant improvements in alignment compared to the state-of-the-art models including the original model and some supervised models (like SBERT and supervised SimCSE), while achieving comparable uniformity scores.

4 Conclusions

In conclusion, this paper introduces a parameter sparsification technique based on alignment and uniformity scores, resulting in the development of SparseCSE, which exhibits impressive performance. Through extensive evaluation on STS tasks, transfer learning tasks, and comparison in terms of alignment and uniformity, SparseCSE demonstrates its competitive edge in sentence embedding.

5 Limitations

We can't extend the application of our pruning methods to a wider range of sentence embedding models beyond SimCSE. Consequently, our focus in the main results revolves around assessing the performance of SimCSE and SparseCSE, without delving into the applicability of other models.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 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, Denver, Colorado. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, 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, San Diego, California. Association for Computational Linguistics.

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, Montréal, Canada. Association for Computational Linguistics.

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, Atlanta, Georgia, USA. Association for Computational Linguistics.

Reinald Kim Amplayo, Arthur Brazinskas, Yoshi Suhara, Xiaolan Wang, and Bing Liu. 2022. Beyond opinion mining: Summarizing opinions of customer reviews. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3447–3450. ACM.

Yue Bai, Huan Wang, Zhiqiang Tao, Kunpeng Li, and Yun Fu. 2022. Dual lottery ticket hypothesis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo 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, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv: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 *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3259–3269. PMLR.

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 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic BERT with adaptive width and depth. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9119–9130. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020b. 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, Online. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Fangyu Liu, Ivan Vulic, 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 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1442–1459. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 216–223. European Language Resources Association (ELRA).

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. 2017. Pruning convolutional neural networks for resource efficient inference. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278. ACL.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 115–124. The Association for Computer Linguistics.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics.

Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 813–823. ACM.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

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

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, pages 200–207. ACM.

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 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Lang. Resour. Evaluation*, 39(2-3):165–210.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. 2022. Structured pruning learns compact and accurate models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1513–1528. Association for Computational Linguistics.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5065–5075. Association for Computational Linguistics.

Yi Yang, Chen Zhang, and Dawei Song. 2022a. Sparse teachers can be dense with knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3904–3915. Association for Computational Linguistics.

Yi Yang, Chen Zhang, Benyou Wang, and Dawei Song. 2022b. Doge tickets: Uncovering domain-general language models by playing lottery tickets. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 144–156. Springer.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020a. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1601–1610. Association for Computational Linguistics.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020b. 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, Online. Association for Computational Linguistics.

A Implementation Details

We follow the training details of SimCSE (Gao et al., 2021) for both training and rewinding process of sparseCSE, including hyperparameter settings and a dataset of one million randomly selected sentences from English Wikipedia.

We prune the baseline models on the dataset STS Benchmark (Cer et al., 2017). The dataset was originally used to evaluate the alignment and uniformity of sentence embeddings in SimCSE (Gao et al., 2021), and we ascertain that it can significantly contribute to the computation of pruning scores and serve as a guiding factor in the pruning process. The objective is to enhance the model with valuable information from alignment and uniformity. It is noteworthy that opting for a pruning process, as opposed to training, is a judicious decision. This is particularly relevant due to the limitation of the small dataset for calculating alignment and uniformity objectives, making model training impractical. During the pruning process, we explore different sparsity levels from a predefined set (1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%), and use a λ value of 0.5 for the main experiment. Additionally, we examine the impact of different λ values (0.25 and 0.75) in further analysis.

B Transfer Tasks for Evaluation

The transfer learning tasks contain MR (Pang and Lee, 2005), CR (Amplayo et al., 2022), SUBJ (Pang and Lee, 2004), MPQA (Wiebe et al., 2005), SST-2 (Socher et al., 2013), TREC (Voorhees and Tice, 2000) and MRPC (Dolan and Brockett, 2005), which are different sentence classification tasks and can give an impression on the quality of sentence embeddings.

The results on transfer learning tasks are shown in table 2. And the average improvement on BERT$_{\text{base}}$, BERT$_{\text{large}}$ and Roberta$_{\text{base}}$ achieves 1.79%, 0.99% and 0.78%, respectively. For instance, when applying 2% sparsity to the BERT$_{\text{base}}$ model, we achieve the best average improvement of 1.53 on transfer tasks shown in Table 2. All tasks benefit from this pruning sparsity, with improvements of 2.04, 1.94, 0.46, 0.53, 1.20, 2.00, and 2.55.

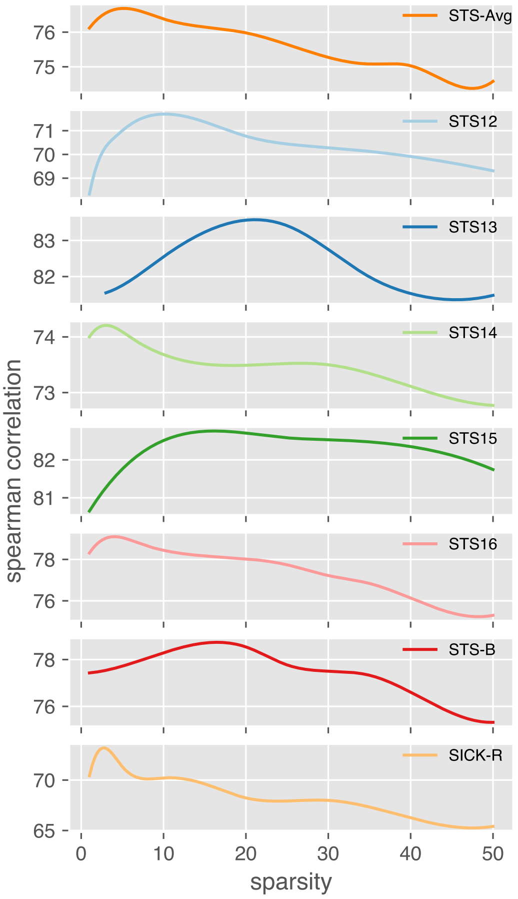

Figure 4: Transitions with varying sparsity on STS tasks.

C Searching within Varying Sparsity

The transition of the BERT$_{\text{base}}$ model's performance, as measured by the average score across the seven STS tasks, as well as the discrete scores of these tasks, is illustrated in Figure 4. It is evident from the figure that for each task, the model's performance initially improves and then declines as the sparsity level increases, showing a peak. In every task, this peak appears steadily around a fixed sparsity corresponding to the optimal sparsity value in the main results. This indicates that the best performance observed in the main results for each task is not an isolated occurrence but rather a continuous trend.

D Tradeoff of Alignment and Uniformity

In our approach, the alignment loss and uniformity loss work together to guide parameter scoring, with the coefficient λ regulating their relative influence. To further investigate the contributions of alignment and uniformity strategies to model ef-

	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	Avg
SimCSE-BERT$_{\text{base}}$	78.84	84.21	93.83	88.87	83.75	86.40	72.99	84.13
SparseCSE$_{2\%}$	$80.88^{+2.04}$	$86.15^{+1.94}$	$94.29^{+0.46}$	$89.40^{+0.53}$	$84.95^{+1.20}$	$88.40^{+2.00}$	$75.54^{+2.55}$	$85.66^{+1.53}$
SparseCSE$_{\text{best}}$	$80.90^{+2.06}_{3\%}$	$86.15^{+1.94}_{2\%}$	$94.58^{+0.75}_{7\%}$	$89.43^{+0.56}_{4\%}$	$85.83^{+2.08}_{3\%}$	$88.40^{+2.00}_{2\%}$	$76.12^{+3.13}_{8\%}$	$85.92^{+1.79}$
SimCSE-BERT$_{\text{large}}$	84.02	88.11	94.8	89.59	89.9	90.20	75.48	87.44
SparseCSE$_{2\%}$	$84.26^{+0.24}$	$89.43^{+1.32}$	$95.27^{+0.47}$	$89.83^{+0.24}$	$89.57^{-0.33}$	$92.40^{+2.20}$	$76.46^{+0.98}$	$88.17^{+0.73}$
SparseCSE$_{\text{best}}$	$84.65^{+0.63}_{3\%}$	$89.43^{+1.32}_{2\%}$	$95.27^{+0.47}_{2\%}$	$90.07^{+0.48}_{9\%}$	$89.57^{-0.33}_{2\%}$	$93.80^{+3.60}_{6\%}$	$76.52^{+1.04}_{3\%}$	$88.44^{+0.99}$
SimCSE-Roberta$_{\text{base}}$	81.39	86.94	93.20	87.11	87.10	84.20	74.09	84.86
SparseCSE$_{1\%}$	$82.18^{+0.79}$	$88.05^{+1.11}$	$93.53^{+0.33}$	$87.59^{+0.48}$	$87.48^{+0.38}$	$84.00^{-0.20}$	$74.78^{+0.69}$	$85.37^{+0.51}$
SparseCSE$_{\text{best}}$	$82.18^{+0.79}_{1\%}$	$88.21^{+1.27}_{3\%}$	$93.53^{+0.33}_{1\%}$	$87.59^{+0.48}_{1\%}$	$87.48^{+0.38}_{1\%}$	$86.00^{+1.80}_{7\%}$	$74.78^{+0.69}_{1\%}$	$85.64^{+0.78}$

Table 2: The result of transfer learning tasks. Data annotation method is the same as the previous table.

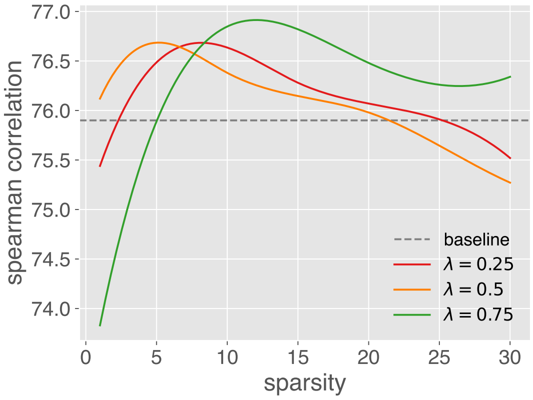

Figure 5: Average STS performance of SparseCSE using $BERT_{base}$ with different λ.

fectiveness, we conducted additional experiments using different λ values (0.25, 0.5, 0.75) as shown in Figure 5. We observed that the coefficient does not have a significant impact on the peak value of each task. However, it does influence the pattern of how model performance varies with sparsity. When $\lambda = 0.5$, the pruned model's performance exhibits a rapid increase and decrease at lower sparsity levels, resulting in a distinct peak. On the other hand, with $\lambda = 0.25$, the performance trend shows a relatively flatter increase and decrease, with the peak occurring at slightly higher sparsity levels. These findings suggest that alignment and uniformity play similar roles in guiding contrastive representation learning, but they have different effects on parameter filtering.