Sparse Contrastive Learning of Sentence Embeddings

Anonymous ACL submission

Abstract

Recently, SimCSE, a simple contrastive learning framework for sentence embeddings, 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. This prompted us to consider whether SimCSE might also have similar harmful parameters. To tackle the problem, 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 hypothesize that parameters with minimal contributions are detrimental, and sparsifying them would result in an improved model performance. Accordingly, a sparsified SimCSE (SparseCSE) is proposed. To systematically explore the ubiquity of detrimental parameters and the removal of them, extensive experiments are conducted on the standard semantic textual similarity (STS) tasks and transfer learning tasks. The results show that the proposed SparseCSE significantly outperform SimCSE. Furthermore, through an in-depth analysis, we establish the validity and stability of our sparsification method, showcasing that the embedding space generated by SparseCSE exhibits an 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 employed 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). With contrastive learning, the semantically similar sentences are brought closer with each other while the dissimilar sentences are pushed apart, thereby a semantically-driven method, namely SimCSE, is established within the space of sentence embeddings.

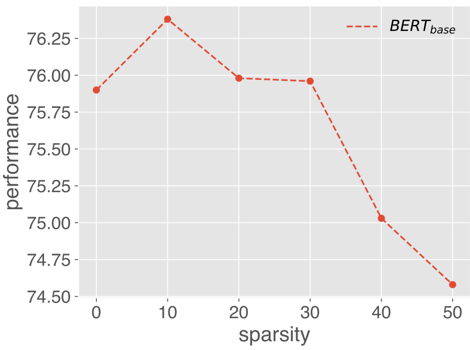

Figure 1: The average performance on STS tasks of SimCSE-BERT$_{\text{base}}$ when pruned at sparsity levels of 10%, 20%, 30%, 40% and 50% respectively. Details of the pruning method can be found in Section 2, while the task specifics and metrics are introduced in Section 3.

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 the 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

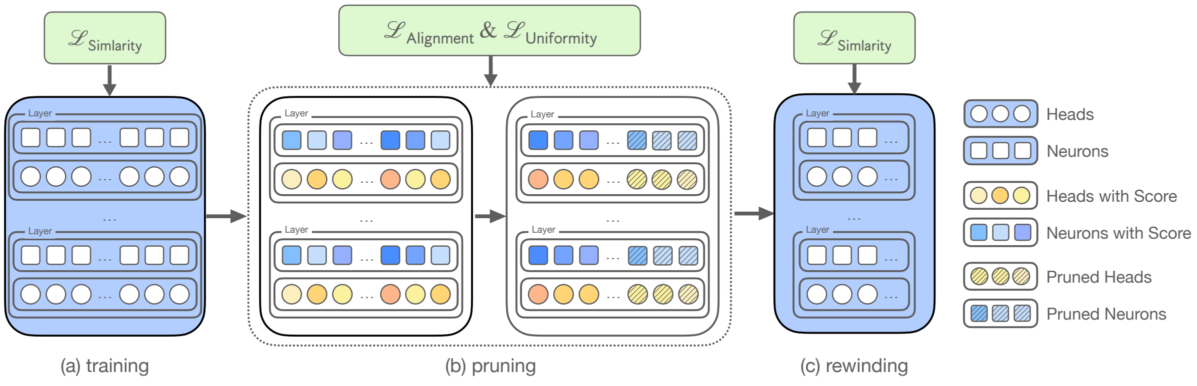

Figure 2: The process of obtaining SparseCSE

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.

Now an important research question arises: How much pruning is needed to best improve the model's performance? Based on a pilot study presented in Figure 1, we observed that model performance on STS tasks does not consistently increase or decrease during pruning. Instead it first exhibits an upward trend when the model is less sparse and then goes down. This suggests that the parameters with the lowest scores are detrimental to model performance, as evidenced by the performance improvement resulting from their pruning. However, an over-sparcification would hurt the performance. Building upon the above observation, 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, we propose a sparsified SimCSE, denoted as SparseCSE. 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 the 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 our formal experiments than in the pilot study, to clearly identify harmful parameters. Finally, the remaining parameters are initialized, and the pruned model is fine-tuned to regain its performance.

We extensively evaluate SparseCSE on seven STS tasks and seven transfer learning tasks. The 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.

2 Our Method

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

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 $\frac{{h_1}^{\intercal} \cdot h_2}{\|h_1\| \cdot \|h_2\|}$.

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, 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

Typical pre-trained language models such as BERT (Devlin et al., 2019) and Roberta (Liu et al., 2019)), are 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 need to be pruned, we associate a set of mask variables with them (Yang et al., 2022a,b) and compare the model's performance before and after the pruning 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 a variable $\xi^{(i)}$ with values in $\{0, 1\}$ is defined for masking each attention head, and it can be represented as:

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

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 $\xi^{(i)}$ equals to 1, it means retaining the attention head, and when it equals to 0 it 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 $\mathbf{W}_2 \in \mathbb{R}^{D_F \times d}$, where D_F denotes the number of neurons in the intermediate layer of FFN. Likewise, we define the variable ν to mask the neurons in the intermediate layer of FFN:

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

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:

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

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:

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

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 $\mathcal{L}_{\text{Score}}$ for further score calculation can be be written as below:

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

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:

	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.

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

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). 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.

3.2 Evaluation

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

	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.

ferent sentence classification tasks and can give an impression on the quality of sentence embeddings.

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 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. The results of transfer task show the same trend prove the ubiquity of the phenomenon found in Table 1.

4 Ablation Studies

4.1 Effects of Rewinding

As shown in the Table 3, the results compare models with and without rewinding. This set of experiments was conducted on the BERT$_{\text{base}}$. Significant differences can be observed, indicating that the rewinding step is essential in this pruning method. Rewinding helps the model restore its original text representation capability.

BERT$_{\text{base}}$	STS.Avg
SparseCSE$_{2\%}$	76.96
SparseCSE$_{2\%}$(w/o RW)	39.55
SparseCSE$_{\text{best}}$	77.49
SparseCSE$_{\text{best}}$(w/o RW)	46.27

Table 3: Effects of the rewinding(RW) step in the pruning methods.

4.2 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 3. 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.

4.3 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 effectiveness, we conducted additional experiments using different λ values (0.25, 0.5, 0.75) as shown in Figure 4. We observed that the coefficient does not have a significant impact on the peak value of

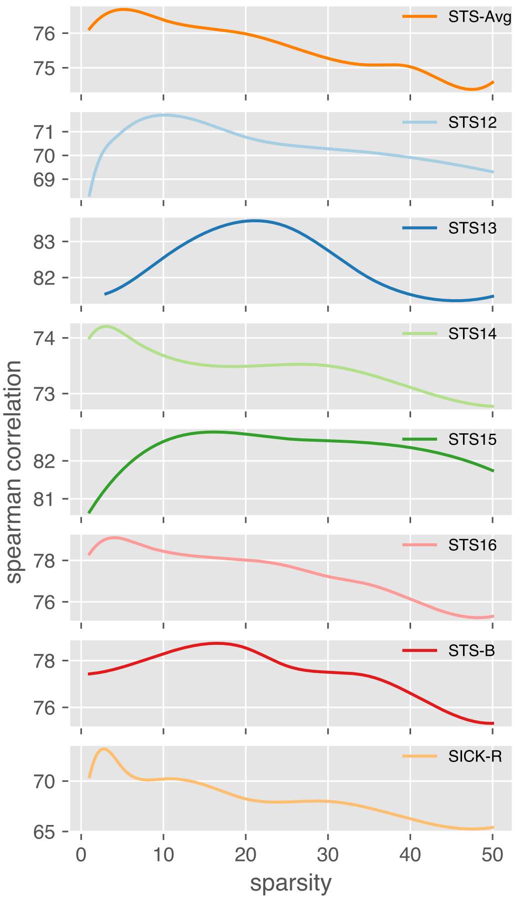

Figure 3: Transitions with varying sparsity on STS tasks.

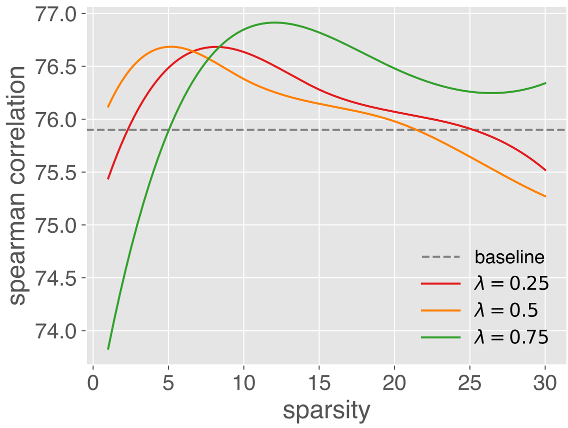

Figure 4: Average STS performance of SparseCSE using $BERT_{base}$ with different λ.

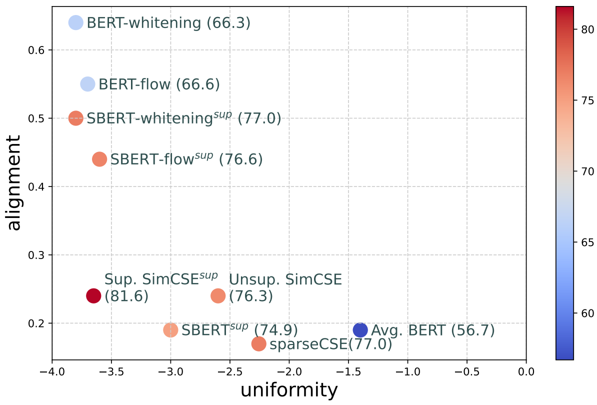

Figure 5: Analysis on alignment and uniformity (the smaller, the better). Points represent average STS performance using $\text{BERT}_{\text{base}}$, with "sup" marked of supervised methods.

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.

4.4 Impact of Pruning MHA and FFN

The main method's pruning strategy advocates for pruning both MHA and FFN. This section breaks down the method, discussing the effects of pruning only MHA and only FFN separately. The results are shown in Table 4, Table 5 and Table 6. It can be observed that pruning only one of these structures impacts the final outcomes across various tasks.

5 Analysis with Other Methods

We compare SparseCSE with other sentence embedding models, including: SimCSE (Gao et al., 2021), BERT(first-last avg.) (Devlin et al., 2019; Su et al., 2021), BERT-flow (Li et al., 2020b), BERT-whitening (Su et al., 2021) and SBERT (Reimers and Gurevych, 2019). BERT (first-last avg.) extracts sentence embeddings by averaging the first and last layers of BERT. BERT-flow applies linear transformations and batch normalization to embeddings from a trained BERT model to improve spatial relationships between sentence embeddings and reduce anisotropy. BERT-whitening similarly adjusts embeddings using a whitening matrix from the covariance matrix. SBERT is a supervised sentence embedding model trained on supervised datasets NLI and STS with the objective of text similarity.

Table 7 presents the sentence embedding performance of various methods on the STS task. Spar-

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
$\text{SparseCSE}_{2\%}$	70.15	82.25	74.16	82.15	78.52	78.71	72.76	76.96
$\text{SparseCSE}_{2\%}(\text{MHA}_{\text{only}})$	71.39	82.92	74.55	82.9	77.94	78.24	70.36	76.9
$\text{SparseCSE}_{2\%}(\text{FFN}_{\text{only}})$	70.98	82.94	74.51	82.01	77.69	78.03	72.09	76.89
$\text{SparseCSE}_{\text{best}}$	71.70	83.41	74.16	82.58	79.10	78.71	72.76	77.49
$\text{SparseCSE}_{\text{best}}(\text{MHA}_{\text{only}})$	69.84	83.49	74.55	82.18	77.58	78.24	70.36	76.61
$\text{SparseCSE}_{\text{best}}(\text{FFN}_{\text{only}})$	70.02	83.09	74.51	82.11	77.61	78.03	72.09	76.78

Table 4: Effects of structures the proposed method prunes. Results on $\text{BERT}_{\text{base}}$.

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
$\text{SparseCSE}_{2\%}$	69.31	83.69	75.72	83.21	79.34	79.41	74.76	77.92
$\text{SparseCSE}_{2\%}(\text{MHA}_{\text{only}})$	68.85	83.76	75.23	82.49	78.55	78.42	74.96	77.47
$\text{SparseCSE}_{2\%}(\text{FFN}_{\text{only}})$	69.57	83.45	75.32	83.42	78.95	78.71	74.31	77.68
$\text{SparseCSE}_{\text{best}}$	70.67	84.60	75.84	83.21	79.60	79.41	75.27	78.32
$\text{SparseCSE}_{\text{best}}(\text{MHA}_{\text{only}})$	69.12	83.92	75.50	81.85	78.99	78.72	73.59	77.38
$\text{SparseCSE}_{\text{best}}(\text{FFN}_{\text{only}})$	70.11	83.11	73.41	83.08	78.40	78.96	75.41	77.50

Table 5: Effects of structures the proposed method prunes. Results on $\text{BERT}_{\text{large}}$.

seCSE shows strong performance across all tasks, outperforming both unsupervised and supervised methods. This advantage is attributed to the superiority of the unsupervised contrastive learning approach inherited from the SimCSE model and the effectiveness of our proposed pruning method.

Figure 5 illustrates the alignment and uniformity scores of these methods along with their performance on the STS task. Benefited from sparsity based on alignment and uniformity properties, sparseCSE demonstrates significant improvements in alignment compared to unsup-SimCSE. As a sparse version of unsup-SimCSE, sparseCSE inherits its advantages in alignment compared to post-training methods like BERT-flow and BERT-whitening, and uniformity compared to BERT(first-last avg.). This highlights that original BERT and post-training adjustments have constraints, while reinforcing sentence representations during training yields superior results. While SBERT was anticipated to outperform unsupervised models but was surpassed by SimCSE, SparseCSE further boosts performance. Notably, we also included supervised SimCSE for comparison with sparseCSE. We found that sparseCSE significantly improves alignment, even when compared to SBERT and supervised SimCSE.

6 Related Work

6.1 Sentence Embedding and SimCSE

Sentence embedding is a key research area in NLP. Unsupervised sentence embedding is especially important due to the scarcity of data for supervised training. Initially, post-training methods (Li et al., 2020b; Su et al., 2021) are used to optimize sentence representation. However, as discussed in section 4.3, Enhancing sentence representation during training can provide better results than post-training metheds. SimCSE's contrastive learning strategy is simple and effective. Following SimCSE, many unsupervised sentence embedding methods (Wu et al., 2022c,b; He et al., 2023; Wang and Dou, 2023) are developed, creating supervised-like tasks from unlabeled data. The proposed pruning method focuses on sentence embedding models using unsupervised contrastive learning. Specifically selecting SimCSE as a representative method for pruning, making this study broadly applicable to similar methods.

6.2 Lottery Ticket Hypothesis

The Lottery Ticket Hypothesis (LTH) (Frankle and Carbin, 2019) suggests that a randomly initialized dense neural network contains a subnetwork that

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
$\text{SparseCSE}_{1\%}$	67.85	81.32	73.09	81.82	81.02	80.29	68.76	76.31
$\text{SparseCSE}_{1\%}(\text{MHA}_{\text{only}})$	67.91	81.91	73.49	82.02	81.13	80.84	69.02	76.62
$\text{SparseCSE}_{1\%}(\text{FFN}_{\text{only}})$	67.83	81.27	73.22	81.70	81.12	80.49	68.68	76.33
$\text{SparseCSE}_{\text{best}}$	68.05	81.82	73.32	82.29	81.02	80.29	68.76	76.48
$\text{SparseCSE}_{\text{best}}(\text{MHA}_{\text{only}})$	68.10	81.42	72.71	82.76	80.42	80.84	69.02	76.47
$\text{SparseCSE}_{\text{best}}(\text{FFN}_{\text{only}})$	67.75	81.51	73.27	82.05	81.08	80.49	68.68	76.40

Table 6: Effects of structures the proposed method prunes. Results on $\text{RoBERTa}_{\text{base}}$.

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
$\text{BERT}_{\text{base}}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
$\text{SBERT}_{\text{base}}{}^{sup}$	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{SBERT}_{\text{base}}\text{-flow}^{sup}$	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{SBERT}_{\text{base}}\text{-whitening}^{sup}$	69.65	77.57	**74.66**	82.27	78.39	**79.52**	**76.91**	77.00
SimCSE-$\text{BERT}_{\text{base}}$	70.37	82.53	73.46	81.58	77.61	76.55	69.22	75.90
$\text{SparseCSE}_{\text{base}}$	**71.70**	**83.41**	74.16	**82.58**	**79.10**	78.71	72.76	**77.49**

Table 7: Sentence embedding performance of $\text{BERT}_{\text{base}}$ on STS tasks (Spearman's correlation). Baselines' results are from Gao et al. 2021. "sup" means supervised methods.

can achieve comparable or better results. Following the hypothesis, many works (Gale et al., 2019a; Desai et al., 2019; Ramanujan et al., 2020; Malach et al., 2020; Brix et al., 2020; Liang et al., 2021; Wu et al., 2022a; Gong et al., 2022; Jaiswal et al., 2023) propose algorithm for getting the winning ticket of various models and find it perform well in many tasks. Among these, structure pruning methods have proven to be effective in pruning transformer models (Prasanna et al., 2020; Hou et al., 2020; Michel et al., 2019; Chen et al., 2020). Inspired by this, we proposed a pruning method for sentence embedding models, resulting in SparseCSE. In Section 3.4, we provide a detailed analysis of the structure pruning methods we used. Furthermore, to address the time-consuming nature of the iterative train-prune-retrain process, many studies (Frankle et al., 2019; Rachwan et al., 2022; Burkholz et al., 2022; You et al., 2022; Shen et al., 2023) have proposed solutions to lower computation costs. Since this paper primarily focuses on optimizing representations for sentence embedding models, efficiency factors will not be discussed in detail. However, it is important to emphasize that there are effective methods to further improve the training efficiency of sparse sentence embedding models.

7 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. The effectiveness of our pruning method is validated, highlighting the crucial role played by alignment and uniformity in optimizing language representation. 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. The effectiveness of our pruning method is validated, highlighting the crucial role played by alignment and uniformity in optimizing language representation. 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.

8 Limitations

We have not extended the method to other sentence embedding models, but discussed its feasibility on SimCSE-derived 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.

Christopher Brix, Parnia Bahar, and Hermann Ney. 2020. Successfully applying the stabilized lottery ticket hypothesis to the transformer architecture. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3909–3915. Association for Computational Linguistics.

Rebekka Burkholz, Nilanjana Laha, Rajarshi Mukherjee, and Alkis Gotovos. 2022. On the existence of universal lottery tickets. 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.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained BERT networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Shrey Desai, Hongyuan Zhan, and Ahmed Aly. 2019. Evaluating lottery tickets under distributional shifts. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 153–162.

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. 2019. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*.

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.

Trevor Gale, Erich Elsen, and Sara Hooker. 2019a. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574.

Trevor Gale, Erich Elsen, and Sara Hooker. 2019b. The state of sparsity in deep neural networks. *CoRR*, abs/1902.09574.

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.

Zhuocheng Gong, Di He, Yelong Shen, Tie-Yan Liu, Weizhu Chen, Dongyan Zhao, Ji-Rong Wen, and Rui Yan. 2022. Finding the dominant winning ticket in pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1459–1472. Association for Computational Linguistics.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Hongliang He, Junlei Zhang, Zhenzhong Lan, and Yue Zhang. 2023. Instance smoothed contrastive learning for unsupervised sentence embedding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12863–12871. AAAI Press.

Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. 2018. Soft filter pruning for accelerating deep convolutional neural networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 2234–2240. ijcai.org.

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*.

Ajay Kumar Jaiswal, Shiwei Liu, Tianlong Chen, Ying Ding, and Zhangyang Wang. 2023. Instant soup: Cheap pruning ensembles in A single pass can draw lottery tickets from large models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 14691–14701. PMLR.

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.

Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2021. Super tickets in pre-trained language models: From model compression to improving generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6524–6538.

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.

Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. 2020. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR.

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.

John Rachwan, Daniel Zügner, Bertrand Charpentier, Simon Geisler, Morgane Ayle, and Stephan Günnemann. 2022. Winning the lottery ahead of time: Efficient early network pruning. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 18293–18309. PMLR.

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. 2020. What's hidden in a randomly weighted neural network? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11890–11899. IEEE.

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.

Alex Renda, Jonathan Frankle, and Michael Carbin. 2020. Comparing rewinding and fine-tuning in neural network pruning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xuan Shen, Zhenglun Kong, Minghai Qin, Peiyan Dong, Geng Yuan, Xin Meng, Hao Tang, Xiaolong Ma, and Yanzhi Wang. 2023. Data level lottery ticket hypothesis for vision transformers. In *Proceedings* of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 1378–1386. ijcai.org.

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.

Hao Wang and Yong Dou. 2023. SNCSE: contrastive learning for unsupervised sentence embedding with soft negative samples. In *Advanced Intelligent Computing Technology and Applications - 19th International Conference, ICIC 2023, Zhengzhou, China, August 10-13, 2023, Proceedings, Part IV*, volume 14089 of *Lecture Notes in Computer Science*, pages 419–431. Springer.

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.

Jiarun Wu, Qingliang Chen, Zeguan Xiao, Yuliang Gu, and Mengsi Sun. 2022a. Pruning adatperfusion with lottery ticket hypothesis. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1632–1646. Association for Computational Linguistics.

Xing Wu, Chaochen Gao, Yipeng Su, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022b. Smoothed contrastive learning for unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4902–4906. International Committee on Computational Linguistics.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022c. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3898–3907. International Committee on Computational Linguistics.

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.

Haoran You, Zhihan Lu, Zijian Zhou, Yonggan Fu, and Yingyan Lin. 2022. Early-bird gcns: Graph-network co-optimization towards more efficient GCN training and inference via drawing early-bird lottery tickets. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 8910–8918. AAAI Press.

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.