Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning

Anonymous authors
Paper under double-blind review

Abstract

State-of-the-art natural language understanding classification models follow two-stages: pre-training a large language model on an auxiliary task, and then fine-tuning the model on a task-specific labeled dataset using cross-entropy loss. Cross-entropy loss has several shortcomings that can lead to sub-optimal generalization and instability. Driven by the intuition that good generalization requires capturing the similarity between examples in one class and contrasting them with examples in other classes, we propose a supervised contrastive learning (SCL) objective for the fine-tuning stage. Combined with cross-entropy, the SCL loss we propose obtains significant improvements over a strong RoBERTa-Large baseline on multiple datasets of the GLUE benchmark in the few-shot learning settings, and it does not require any specialized architecture, data augmentation of any kind, memory banks, or additional unsupervised data. The new objective leads to models that are more robust to different levels of noise in the training data, and can generalize better to related tasks with limited labeled data.

1 Introduction

State-of-the-art for most existing natural language processing (NLP) classification tasks is achieved by systems that are first pre-trained on auxiliary language modeling tasks and then fine-tuned on the task of interest with cross-entropy loss (Radford et al., 2019; Howard & Ruder, 2018; Liu et al., 2019; Devlin et al., 2019). Although commonly used, cross-entropy loss – the KL-divergence between one-hot vectors of labels and the distribution of model's output logits – has several shortcomings. Cross entropy loss leads to poor generalization performance (Liu et al., 2016; Cao et al., 2019), and it lacks robustness to noisy labels (Zhang & Sabuncu, 2018; Sukhbaatar et al., 2015) or adversarial examples (Elsayed et al., 2018; Nar et al., 2019). Effective alternatives have been proposed to change the reference label distributions through label smoothing (Szegedy et al., 2016; Müller et al., 2019), Mixup (Zhang et al., 2018), CutMix (Yun et al., 2019), knowledge distillation (Hinton et al., 2015) or self-training (Yalniz et al., 2019; Xie et al., 2020).

Fine-tuning using cross entropy loss in NLP also tends to be unstable (Zhang et al., 2020; Dodge et al., 2020), especially when supervised data is limited, a scenario in which pre-training is particularly helpful. To tackle the issue of unstable fine-tuning, recent work proposes local smoothness-inducing regularizers (Jiang et al., 2020) and regularization methods inspired by the trust region theory (Aghajanyan et al., 2020) to prevent representation collapse that lead to poor generalization performance. Empirical analysis suggests that fine-tuning for longer, reinitializing top few layers (Zhang et al., 2020), and using debiased Adam optimizer during fine-tuning (Mosbach et al., 2020) can make the fine-tuning procedure more stable.

We are inspired by the learning strategy that humans deploy when given a few examples – try to find the commonalities between the examples of each class and contrast them with examples from other classes. We hypothesize that a similarity-based loss will be able to hone in on the important dimensions of the multidimensional hidden representations and lead to better few-shot learning results and be more stable while fine-tuning pre-trained models. We propose a novel objective for fine-tuning that includes a supervised contrastive learning term that pushes examples from the same class close and examples of different classes further apart. The new term is similar to the contrastive objective used for self-supervised representation learning in various domains such as image, speech, and video. (Sohn, 2016; Oord et al., 2018; Wu et al., 2018; Bachman et al., 2019; Hénaff et al., 2019;

Baevski et al., 2020; Conneau et al., 2020; Tian et al., 2020; Hjelm et al., 2019; Han et al., 2019; He et al., 2020; Misra & Maaten, 2020; Chen et al., 2020a;b). Unlike these methods, however, we use a contrastive objective for supervised learning of the final task, instead of contrasting different augmented views of examples.

Adding supervised contrastive learning (SCL) term to the fine-tuning objective significantly improves performance on several natural language understanding tasks from the GLUE benchmark (Wang et al., 2019) over the state-of-the-art models using cross entropy loss for few-shot learning settings (20, 100, 1000 labeled examples). Models trained with SCL are not only robust to the noise in the training data, but also generalize better to related tasks with limited labeled data. Our approach does not require any specialized architectures (Bachman et al., 2019; Hénaff et al., 2019), memory banks (Wu et al., 2018; Tian et al., 2020; Misra & Maaten, 2020), data augmentation of any kind, or additional unsupervised data. To the best of our knowledge, our work is the first to successfully integrate a supervised contrastive learning objective for fine-tuning pre-trained language models.

- We propose a novel objective for fine-tuning of pre-trained language models that includes a supervised contrastive learning term, as described in Section 2.
- We obtain strong improvements on few-shot learning settings (20, 100, 1000 labeled examples) as shown in Table 2, leading up to 10.7 points improvement for 20 labeled examples.
- We demonstrate that our proposed objective is more robust across augmented training datasets with varying noise levels as shown in Table 3, leading to 7 points average improvement on MNLI across augmented training sets.
- We show that the task-models fine-tuned with our proposed objective has better generalization ability to a related task with limited labeled data as shown in Table 7, leading to 2.9 points improvement on Amazon-2 along with significant reduction in variance across few-shot training samples, when transferred from the source SST-2 task model.

2 Approach

We propose a novel objective that includes a supervised contrastive learning term for fine-tuning pre-trained language models. The loss is meant to capture similarities between examples of the same class and contrast them with examples from other classes.

For a multi-class classification problem with C classes, we work with a batch of training examples of size N, $\{x_i, y_i\}_{i=1,...N}$. $\Phi(\cdot) \in \mathbf{R}^d$ denotes the l_2 normalized embedding of the final encoder hidden layer before the softmax projection; N_{y_i} is the total number of examples in the batch that have the same label as y_i; $\tau > 0$ is an adjustable scalar temperature parameter that controls the separation of classes; and λ is a scalar weighting hyperparameter that we tune for each downstream task. The loss is given by the following formulas:

$$\mathcal{L} = (1 - \lambda) \cdot \mathcal{L}_{CE} + \lambda \mathcal{L}_{SCL} \tag{1}$$

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \cdot log\hat{y}_{i,c} \tag{2}$$

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \log \frac{\exp\left(\Phi(x_i) \cdot \Phi(x_j)/\tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp\left(\Phi(x_i) \cdot \Phi(x_k)/\tau\right)} \tag{3}$$

The overall loss is a weighted average of CE and the SCL loss, as given in equation (1). The canonical definition of CE that we use is given in equation (2). The novel SCL loss is given in equation (3).

This loss can be applied using a variety of encoders $\Phi(\cdot) \in \mathbf{R}^d$ – for example a ResNet for a computer vision application or a pre-trained large language model such as BERT for an NLP application. In this work, we focus on fine-tuning pre-trained language models for single sentence and sentence-pair classification. For single sentence classification, each example x_i consists of sequence of tokens prepended with the special $[CLS]$ token $x_i = [[CLS], t_1, t_2, \ldots, t_L, [EOS]]$.

The length of sequence L is constrained such that $L < L_{\max}$. Similarly, for sentence-pair classification tasks, each example x_i is a concatenation of two sequences of tokens $[t_1, t_2, \dots t_L]$ and $[s_1, s_2, \dots, s_M]$ corresponding to the sentences with special tokens delimiting them: $x_i = [[CLS], t_1, t_2, \dots, t_L, [SEP], s_1, s_2, \dots, s_M, [EOS]]$. The length of concatenated sequences is constrained such that $L + M < L_{\max}$. In both cases, $\Phi(x_i) \in \mathbf{R}^d$ uses the embedding of $[CLS]$ token as the representation for example x_i. These settings follow standard practices for fine-tuning pre-trained language models for classification (Devlin et al., 2019; Liu et al., 2019).

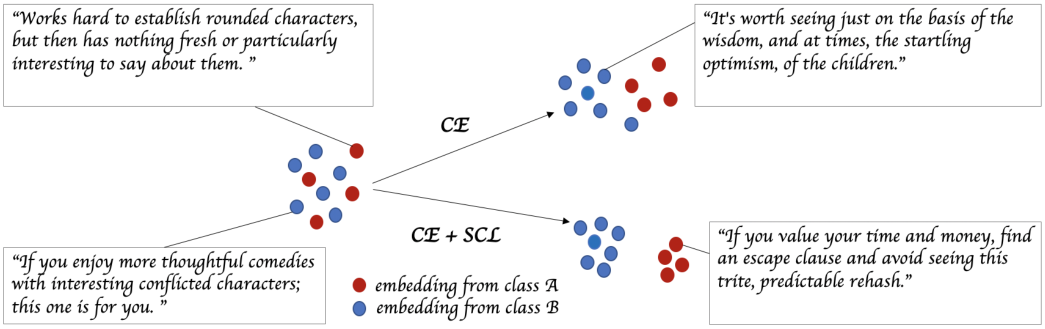
"Works hard to establish rounded characters, but then has nothing fresh or particularly interesting to say about them. "
"It's worth seeing just on the basis of the wisdom, and at times, the startling optimism, of the children."
CE
CE + SCL
"If you enjoy more thoughtful comedies with interesting conflicted characters; this one is for you. "
● embedding from class A
● embedding from class B
"If you value your time and money, find an escape clause and avoid seeing this trite, predictable rehash."

Figure 1: Our proposed objective includes a cross-entropy term (CE) and supervised contrastive learning (SCL) term, and it is formulated to push examples from the same class close and examples of different classes further apart. We show examples from SST-2 sentiment analysis dataset from the GLUE benchmark, where class A (shown in red) is negative movie reviews and class B (shown in blue) is positive movie reviews. Although we show a binary classification case for simplicity, the loss is generally applicable to any multi-class classification setting.

Empirical observations show that both l_2 normalization of the encoded embedding representations and an adjustable scalar temperature parameter τ improve performance. Lower temperature increases the influence of examples that are harder to separate, effectively creating harder negatives. Using hard negatives has been previously shown to improve performance in the context of margin-based loss formulations such as triplet loss (Schroff et al., 2015). The empirical behavior of the adjustable temperature parameter is consistent with the observations of previous work related to supervised contrastive learning. (Chen et al., 2020a; Khosla et al., 2020).

Relationship to Self-Supervised Contrastive Learning Self-supervised contrastive learning has shown success in learning powerful representations, particularly in the computer vision domain. (Chen et al., 2020a; He et al., 2020; Tian et al., 2020; Mnih & Kavukcuoglu, 2013; Gutmann & Hyvärinen, 2012; Kolesnikov et al., 2019) Self-supervised learning methods do not require any labeled data; instead they sample a mini batch from unsupervised data and create *positive* and *negative* examples from these samples using strong data augmentation techniques such as AutoAugment (Cubuk et al., 2019) or RandAugment (Cubuk et al., 2020) for computer vision. *Positive* examples are constructed by applying data augmentation to the same example (cropping, flipping, etc. for an image), and *negative* examples are simply all the other examples in the sampled mini batch. Intuitively, self-supervised contrastive objectives are learning representations that are invariant to different views of *positive* pairs; while maximizing the distance between *negative* pairs. The distance metric used is often the inner product or the Euclidean distance between vector representations of the examples.

For a batch of size N, self-supervised contrastive loss is defined as:

$$\mathcal{L}_{self} = \sum_{i=1}^{2N} -\log \frac{\exp\left(\Phi(x'_{2i-1}) \cdot \Phi(x'_{2i})/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{i \neq k} \exp\left(\Phi(x'_i) \cdot \Phi(x'_k)/\tau\right)} \tag{4}$$

where $\Phi(\cdot) \in \mathbf{R}^d$ is the l_2 normalization embedding from the encoder before the final classification softmax layer; $\tau > 0$ is a scalar temperature parameter. A is defined as a data augmentation block that generates two randomly generated augmented examples, x'_{2i} and x'_{2i-1} from the original example x_i: $A(\{x_i, y_i\}_{i=1,\dots N}) = \{x'_i, y'_i\}_{i=1,\dots 2N}$. As an example, A can be RandAugment for a computer vision application; or it could be a back-translation model for an NLP application.

3 Related Work

Traditional Machine Learning and Theoretical Understanding Several works have analyzed the shortcomings of the widely adopted cross-entropy loss, demonstrating that it leads to poor generalization performance due to poor margins (Liu et al., 2016; Cao et al., 2019), and lack of robustness to noisy labels (Zhang & Sabuncu, 2018; Sukhbaatar et al., 2015) or adversarial examples (Elsayed et al., 2018; Nar et al., 2019). On the other hand, there has been a body of work that has explored the performance difference for classifiers trained with discriminative (i.e., optimizing for $p(y|x)$, where y is the label and x is the input) losses such as cross-entropy loss and generative losses (i.e. optimizing for $p(x|y)$). Ng & Jordan (2001) show that classifiers trained with generative losses can outperform their counterparts trained with discriminative losses in the context of Logistic Regression and Naive Bayes. Raina et al. (2003) show that a hybrid discriminative and generative objective outperforms both solely discriminative and generative approaches. In the context of contrastive learning, Saunshi et al. (2019) propose a theoretical framework for analyzing contrastive learning algorithms through hypothesizing that semantically similar points are sampled from the same latent class, which allows showing formal guarantees on the quality of learned representations.

Contrastive Learning There has been several investigations for the use of contrastive loss formulations for self-supervised, semi-supervised, and supervised learning methods, primarily in the computer vision domain. Chen et al. (2020a) propose a framework for contrastive learning to learn visual representations without specialized architectures or a memory bank and show state-of-the-art results on ImageNet ILSVRC-2012 (Russakovsky et al., 2015), outperforming previous methods for self-supervised, semi-supervised and transfer learning. Similarly, Khosla et al. (2020) propose a supervised contrastive loss that outperforms cross entropy loss and gets state-of-the-art results on ImageNet on both ResNet-50 and ResNet-200 (He et al., 2016) with AutoAugment (Cubuk et al., 2019) data augmentation. They also show increased robustness on the ImageNet-C dataset (Hendrycks & Dietterich, 2019), and demonstrate that supervised contrastive loss is less sensitive to hyperparameter settings such as optimizers or data augmentations compared to cross-entropy loss. Liu & Abbeel (2020) propose a hybrid discriminative-generative training of energy-based models where they approximate the generative term with a contrastive loss using large batch sizes and show improved classification accuracy of WideResNet-28-10 (Zagoruyko & Komodakis, 2016) on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) datasets, outperforming state-of-the-art discriminative and generative classifiers. They also demonstrate improved performance for WideResNet-28-10 on robustness, out-of-distribution detection, and calibration, compared to other state-of-the-art generative and hybrid models. Finally, Fang & Xie (2020) propose pre-training language models using a self-supervised contrastive learning objective at the sentence level using back-translation as the augmentation method, followed by fine-tuning by predicting whether two augmented sentences originate from the same sentence – showing improvements over fine-tuning BERT on a subset of GLUE tasks.

Stability and Robustness of Fine-tuning Language Models There has been several works on analyzing robustness of fine-tuning large pre-trained language models, since they tend to overfit to the labeled task data and fail to generalize to unseen data when there is limited labeled data for the downstream task. To improve the generalization performance, Jiang et al. (2020) propose a local smoothness-inducing regularizer to manage the complexity of the model and a Bregman proximal point optimization method, an instance of trust-region methods, to prevent aggressive updating of the model during fine-tuning. They show state-of-the-art performance on GLUE, SNLI (Bowman et al., 2015), SciTail (Khot et al., 2018), and ANLI (Nie et al., 2020) natural language understanding benchmarks. Similarly, Aghajanyan et al. (2020) propose a regularized fine-tuning procedure inspired by trust-region theory that replaces adversarial objectives with parametric noise sampled from normal or uniform distribution in order to prevent representation collapse during fine-tuning for better generalization performance, without hurting the performance. They show improved performance on a range of natural language understanding and generation tasks including DailyMail/CNN (Hermann et al., 2015), Gigaword (Napoles et al., 2012), Reddit TIFU (Kim et al., 2019), and the GLUE benchmark. There has also been some empirical analysis that suggests fine-tuning for more epochs, reinitializing top few layers (Zhang et al., 2020) instead of only the classification head, and using debiased Adam optimizer instead of BERTAdam (Devlin et al., 2019) during fine-tuning (Mosbach et al., 2020) make the fine-tuning procedure more stable across different runs.

4 Experimental Setup

4.1 Datasets and Training Details

We use datasets from the GLUE natural language understanding benchmark (Wang et al., 2019) for evaluation. We include both single sentence classification tasks and sentence-pair classification tasks to test whether our hypothesis is generally applicable across tasks. We summarize each dataset based on their main task, domain, number of training examples and number of classes in Table 1.

In our few-shot learning experiments, we sample half of the original validation set of GLUE benchmark and use it as our test set, and sample $\sim$500 examples for our validation set from the original GLUE validation set, both taking the label distribution of the original validation set into account. For each task, we want the validation set to be small enough to avoid easy overfitting on the validation set, and big enough to avoid high-variance when early-stopping at various epochs for the few-shot learning experiments. For full dataset experiments, such as the ones shown in Table 5, Table 6, Table 8, and Table 9, we sample a validation set from the original training set of the GLUE benchmark based on the size of the original validation set of GLUE, and report our test results on the original validation set of GLUE.

We run each experiment with 10 different seeds, and report the average test accuracy, standard deviation, along with p-values with respect to the baseline. We pick the best hyperparameter combination based on the average validation accuracy across 10 seeds. For few-shot learning experiments such as the ones shown in Table 2, Table 3, and Table 10, we sample 10 different training set samples based on the total number of examples N specified from the original training set of the GLUE benchmark, taking the label distribution of the original training set into account. We report the average and the standard deviation of the test accuracies of the top 3 models based on their validation accuracies out of 10 random training set samples. Best hyperparameter combination is picked based on the average validation accuracy of the top 3 models. The reason why we focus on the top 3 models for this setting is that we would like to reduce the variance across training set samples.

Dataset	**Task**	**Domain**	**#Train**	**#Classes**
SST-2	sentiment analysis	movie reviews	67k	2
CoLA	grammatical correctness	linguistic publications	8.5k	2
MRPC	paraphrase	news	3.7k	2
RTE	textual entailment	news/Wikipedia	2.5k	2
QNLI	question answering/textual entailment	Wikipedia	105k	2
MNLI	textual entailment	multi-domain	393k	3

Table 1: GLUE Benchmark datasets used for evaluation.

We use fairseq Ott et al. (2019) library and the open-source RoBERTa-Large model for all of our experiments. During all the fine-tuning runs, we use Adam optimizer with a learning rate of 1e-5, batch size of 16 (unless specified otherwise), and dropout rate of 0.1. For each experiment that includes the SCL term, we conduct a grid-based hyperparameter sweep for $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ and $\tau \in \{0.1, 0.3, 0.5, 0.7\}$. We observe that models with best test accuracies across all experimental settings overwhelmingly use the hyperparameter combination $\tau = 0.3$ and $\lambda = 0.9$.

4.2 Constructing Augmented Noisy Training Datasets

Machine learning researchers or practitioners often do not know how noisy their datasets are, as input examples might be corrupted or ground truth labeling might not be perfect. Therefore, it is preferable to use robust training objectives that can get more information out of datasets of different noise levels, even where there is limited amount of labeled data. We simulate augmented training datasets of different noise levels using a back-translation model (Edunov et al., 2018), where we increase the temperature parameter to create more noisy examples. Back-translation refers to the procedure of translating an example in language A into language B and then translating it back to language A, and it is a commonly used data augmentation procedure for NLP applications, as the new examples obtained through back-translation provide targeted inductive bias to the model while preserving the meaning of the original example. Specifically, we use WMT'18 English-German and German-English translation models, use random sampling to get more diverse examples, and employ and augmentation ratio of 1:3 for supervised examples:augmented examples. We observe that employing random sampling with a tunable temperature parameter is critical to get diverse paraphrases for the supervised examples, consistent with previous work (Edunov et al., 2018; Xie et al., 2019), since commonly used beam search results in very regular sentences that do not provide diversity to the existing data distribution. We keep the validation and test sets same with the experiments shown in Table 2.

5 Analysis and results

5.1 GLUE Benchmark Few-shot Learning Results

We proposed adding the SCL term inspired by the learning strategy of humans when they are given few examples. In Table 2, we report our few-shot learning results on SST-2, QNLI, and MNLI from the GLUE benchmark with 20, 100, 1000 labeled training examples. Details of the experimental setup are explained in Section 4. We use a very strong baseline of fine-tuning RoBERTa-Large with cross-entropy loss. We observe that the SCL term improves performance over the baseline significantly across all datasets and data regimes, leading to 10.7 points improvement on QNLI, 3.4 points improvement on MNLI, and 2.2 points improvement on SST-2, where we have 20 labeled examples for training. This shows that our proposed objective is effective both for binary single sentence classification such as sentiment analysis and grammatical correctness; and sentence pair classification tasks such as textual entailment and paraphrasing – when we are given few labeled examples. We see that as we increase the number of labeled examples, performance improvement over the baseline decreases, leading to 1.9 points improvement on MNLI for 100 examples and 0.6 points improvement on QNLI for 1000 examples. We also would like to acknowledge that improvements over the baseline when N=1000 on both SST-2 and MNLI are not statistically significant. In addition, we conduct a control study where we investigate the importance of l_2 normalization and temperature scaling where we replace SCL loss with CE loss but keep the normalization and scaling, as shown in Table 10 in the Appendix as the method CE+CE.

In Figure 2, we show tSNE plots of the learned representations of CLS embeddings on SST-2 test set when trained with 20 labeled examples, comparing CE with and without the SCL term. We can see that SCL term enforces more compact clustering of examples with the same label; while the distribution of embeddings learned with CE is close to random. We include a more detailed comparison of tSNE plots for CE and CE+SCL, where we have 20, 100 labeled examples and full dataset respectively for training in Figure 3 in the Appendix.

Model	**Loss**	**N**	**SST-2**	**QNLI**	**MNLI**
RoBERTa$_{\text{Large}}$	CE	20	85.9±2.1	65.0±2.0	39.3±2.5
RoBERTa$_{\text{Large}}$	CE + SCL	20	**88.1**±**3.3**	**75.7**±**4.8**	**42.7**±**4.6**
	p-value		5e-10	1e-46	1e-8
RoBERTa$_{\text{Large}}$	CE	100	91.1±1.3	81.9±0.4	59.2±2.1
RoBERTa$_{\text{Large}}$	CE + SCL	100	**92.8**±**1.3**	**82.5**±**0.4**	**61.1**±**3.0**
	p-value		3e-17	1e-20	2e-4
RoBERTa$_{\text{Large}}$	CE	1000	94.0±0.6	89.2±0.6	81.4±0.2
RoBERTa$_{\text{Large}}$	CE + SCL	1000	**94.1**±**0.5**	**89.8**±**0.4**	**81.5**±**0.2**
	p-value		0.6	1e-12	0.5

Table 2: Few-shot learning test results on the GLUE benchmark where we have N=20,100,1000 labeled examples for training. Reported results are the mean and the standard deviation of the test accuracies of the top 3 models based on validation accuracy out of 10 random training set samples, along with p-values for each experiment.

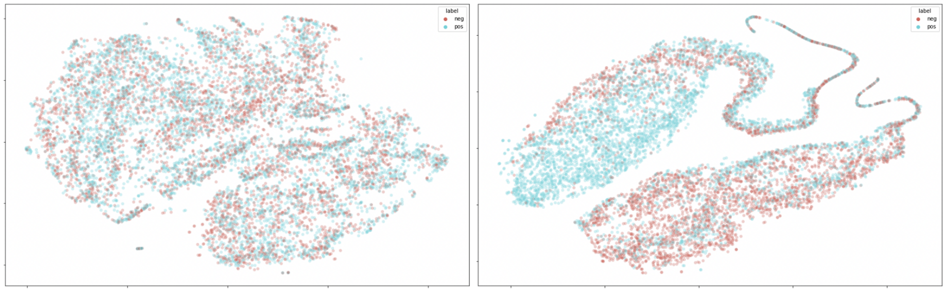

Figure 2: tSNE plots of learned CLS embedding on SST-2 test set where we have 20 labeled examples, comparing CE with and without SCL term. Blue: positive examples; red: negative examples.

5.2 Robustness Across Augmented Noisy Training Datasets

In Table 3, we report our results on augmented training sets with varying levels of noise. We have 100 labeled examples for training for each task, and we augment their training sets with noisy examples using a back-translation model, as described in detail in Section 4.2. Note that we use the back-translation model to simulate training datasets of varying noise levels and not as a method to boost model performance. Experimental setup follows what is described in Section 4 for few-shot learning experiments. T is the temperature for the back-translation model used to augment the training sets, and higher temperature corresponds to more noise in the augmented training set.

We observe consistent improvements over the RoBERTa-Large baseline with our proposed objective across all datasets across all noise levels, with 0.4 points improvement on SST-2, 2.5 points improvement on QNLI, and 7 points improvement on MNLI on average across augmented training sets. The improvement is particularly significant for inference tasks (QNLI, MNLI) when the noise levels are higher (higher temperature), leading to 7.7 points improvement on MNLI when T=0.7, and 4.2 points improvement on QNLI when T=0.9. We show some samples of augmented examples used in this robustness experiment in Table 4. For T=0.3, examples mostly stay the same with minor changes in their phrasing, while for T=0.9, some grammatical mistakes and factual errors are introduced.

Dataset	**Loss**	**Original**	**T=0.3**	**T=0.5**	**T=0.7**	**T=0.9**	**Average**
SST-2	CE	91.1±1.3	92.0±1.3	91.4±1.0	**91.7±1.3**	90.0±0.5	91.3±1.2
SST-2	CE + SCL	**92.8±1.3**	**92.6±0.9**	**91.5±1.0**	91.2±0.6	**91.5±1.0**	**91.7±1.0**
QNLI	CE	81.9±0.4	81.1±2.3	80.0±2.9	78.9±3.7	75.9±4.0	79.0±3.5
QNLI	CE + SCL	**82.5±0.4**	**82.7±1.9**	**81.9±2.5**	**81.3±0.6**	**80.1±2.5**	**81.5±2.0**
MNLI	CE	59.2±2.1	54.0±1.1	55.3±2.4	54.6±2.2	47.0±1.8	52.7±3.9
MNLI	CE + SCL	**61.1±3.0**	**61.2±2.3**	**62.1±0.9**	**62.3±1.1**	**53.0±2.1**	**59.7±4.3**

Table 3: Results on the GLUE benchmark for robustness across noisy augmented training sets. Average shows the average performance across augmented training sets.

5.3 GLUE Benchmark Full Dataset Results

In Table 5, we report results using our proposed objective on six downstream tasks from the GLUE benchmark. We use a very strong baseline of fine-tuning RoBERTa-Large with cross-entropy loss, which is currently the standard practice for the state-of-the-art NLP classification models. Details of the experimental setup are explained in Section 4.

We observe that adding the SCL term to the objective improves the performance over the RoBERTa-Large baseline that lead to 3.1 points improvement on MRPC, 3.5 points improvement on QNLI, and an average improvement of 1.2 points across all 6 datasets. We conduct these experiments to investigate the effect of the SCL term in high-data regimes, as we observe that it's effective in few-shot learning settings. We acknowledge that only MRPC and QNLI results are statistically significant, and we report the results on the other datasets as a finding for the sake of completeness.

Dataset	Type	Sentence
SST-2	Original	As possibly the best actor working in movies today.
SST-2	Augmented (T=0.3)	As perhaps the best actor who now stars in films.
SST-2	Original	The young stars are too cute; the story and ensuing complications are too manipulative.
SST-2	Augmented (T=0.9)	The babies are too cute, the image and complications that follow too manipulative.
QNLI	Original	Brain tissue is naturally soft, but can be stiffened with what liquid?
QNLI	Augmented (T=0.3)	Brain tissue is omitted naturally, but with what fluid it can be stiffened?
QNLI	Original	In March 1968, CBS and Sony formed CBS/Sony Records, a Japanese business joint venture.
QNLI	Augmented (T=0.9)	CBS was founded by CBS and Sony Records in March 1962, a Japanese company.
MNLI	Original	However, the link did not transfer the user to a comment box particular to the rule at issue.
MNLI	Augmented (T=0.3)	However, the link did not send the user to a comment field specifically for the rule.
MNLI	Original	Tenants could not enter the apartment complex due to a dangerous chemical spill.
MNLI	Augmented (T=0.9)	Tenants were banned from entering the medical property because of a blood positive substance.

Table 4: Sample of augmented examples with different noise levels for the robustness experiment shown in Table 3. Higher temperature (T) corresponds to more noise in the augmented training set.

We hypothesize larger batch sizes lead to better performance, but we leave that for future work as that requires additional engineering effort. We show evidence for this hypothesis in our ablation studies that we show in Table 6, where we conduct the full dataset experiments for CE+SCL with the same experimental setup described here for Table 5 on SST-2, CoLA, QNLI, and MNLI for batch sizes 16, 64, and 256 using RoBERTa-Base. We observe that as we increase the batch size, performance improves significantly across all datasets. Specifically, we observe 0.3 points improvement on SST-2, 0.8 points improvement on CoLA, 0.4 points improvement on QNLI, and 1.3 points improvement on MNLI, when we increase the batch size from 16 to 256 for CE+SCL. We also investigate the effect of SCL term in the overall training speed, and we measure that with average updates per second metric, shown in Table 6. For batch size 16, the batch size we use throughout the paper across all experimental settings, effect of SCL is negligible – decreasing average updates per second from 15.9 to 15.08. As we increase the batch size, effect of SCL to training speed becomes more significant – decreasing average updates per second from 2.46 to 1.54 for batch size 256. In addition, we conduct a control study where we investigate the importance of l_2 normalization and temperature scaling where we replace SCL loss with CE loss but keep the normalization and scaling (denoted as CE+CE) both for full dataset results in Table 8, and for batch size ablation in Table 9 in the Appendix.

Model	Loss	SST-2	CoLA	MRPC	RTE	QNLI	MNLI	Avg
RoBERTa$_{\text{Large}}$	CE	96.0±0.4	86.0±0.5	86.4±2.4	85.5±1.8	90.4±0.8	88.4±1	88.8
RoBERTa$_{\text{Large}}$	CE + SCL	**96.3±0.4**	**86.1±0.8**	**89.5±0.9**	**85.7±0.5**	**93.9±0.7**	**88.6±0.7**	**90**
	p-value	0.07	0.63	0.01	0.06	0.01	0.16	

Table 5: Test results on the validation set of GLUE benchmark. We compare fine-tuning RoBERTa-Large with CE with and without SCL. Best hyperparameter configuration picked based on average validation accuracy. We report average accuracy across 10 seeds for the model with best hyperparameter configuration, its standard deviation, and p-values.

Model	Loss	Bsz	SST-2	CoLA	QNLI	MNLI	Avg ups/sec
RoBERTa$_{\text{Base}}$	CE	16	94.1±0.5	83.3±0.7	88.2±0.8	84±0.6	15.9
RoBERTa$_{\text{Base}}$	CE + SCL	16	94.9±0.6	83.7±0.9	92.5±0.4	85.3±0.5	15.08
RoBERTa$_{\text{Base}}$	CE	64	94.2±0.4	83.3±0.5	89.2±0.5	84±0.4	8.43
RoBERTa$_{\text{Base}}$	CE + SCL	64	94.7±0.2	83.8±0.6	92.6±0.5	85.7±0.7	7.44
RoBERTa$_{\text{Base}}$	CE	256	94.1±0.4	84±0.5	90±0.7	84.4±0.6	**2.46**
RoBERTa$_{\text{Base}}$	CE + SCL	256	**95.2±0.3**	**84.5±0.5**	**92.9±0.3**	**86.6±0.6**	1.54

Table 6: Ablation study on performance and training speed shown as average updates per second (Avg ups/sec) for fine-tuning RoBERTa-Base with respect to the batch size (Bsz).

5.4 Generalization Ability of Task Models

In this experiment, we first fine-tune RoBERTa-Large on SST-2 using its full training set and get a task model with and without SCL term. Then, we transfer this task model to two related single sentence sentiment analysis binary classification tasks for the movie reviews domain – Amazon-2 and Yelp-2 (Zhang et al., 2015). For both, we sample 20 labeled examples for each class, and follow the few-shot learning experimental setup described in Section 4. In Table 7, we demonstrate that using the SCL term for both source (SST-2) and target domains (Amazon-2, Yelp-2) lead to better generalization ability, with 2.9 points improvement on Amazon-2 and 0.4 points improvement on Yelp-2 along with significant reduction in variance across training set samples.

Model	**Loss**	**N**	**Amazon-2**	**Yelp-2**
RoBERTa$_{\text{Large}}$	CE	40	87.4±6.4	90.8±2.2
RoBERTa$_{\text{Large}}$	CE + SCL	40	**90.3±0.6**	**91.2±0.4**

Table 7: Generalization of the SST-2 task model (fine-tuned using the full training set) to related tasks (Amazon-2, Yelp-2) where there are 20 labeled examples for each class.

6 Conclusion

We propose a supervised contrastive learning objective for fine-tuning pre-trained language models and demonstrate significant improvements over a strong RoBERTa-Large baseline on multiple datasets of the GLUE benchmark in the few-shot learning settings. We also show that our proposed objective leads to models that are more robust to different levels of noise in the training data and can generalize better to related tasks with limited labeled task data. Currently, data augmentation methods in NLP and their effects on the downstream tasks are neither as effective nor as well understood as their counterparts in the computer vision domain. In future work, we plan to study principled and automated data augmentation techniques for NLP that would allow extending our supervised contrastive learning objective to both semi-supervised and self-supervised learning settings.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. *ArXiv*, abs/2008.03156, 2020.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

Kaidi Cao, Colin Wei, Adrien Gaidon, N. Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020b.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

E. Cubuk, Barret Zoph, Dandelion Mané, V. Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, 2019.

E. D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3008–3017, 2020.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*, abs/2002.06305, 2020.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *EMNLP*, 2018.

Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *NeurIPS*, 2018.

Hongchao Fang and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *ArXiv*, abs/2005.12766, 2020.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13:307–361, 2012.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 1483–1492, 2019.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020.

Olivier J. Hénaff, A. Srinivas, J. Fauw, Ali Razavi, C. Doersch, S. Eslami, and A. Oord. Data-efficient image recognition with contrastive predictive coding. *ArXiv*, abs/1905.09272, 2019.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

K. Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, W. Kay, Mustafa Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NeurIPS*, 2015.

Geoffrey E. Hinton, Oriol Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

R. Devon Hjelm, A. Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 328–339, 2018.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *ACL*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

Tushar Khot, A. Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *AAAI*, 2018.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks. In *NAACL-HLT*, 2019.

A. Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929, 2019.

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Hao Liu and P. Abbeel. Hybrid discriminative-generative training via contrastive learning. *ArXiv*, abs/2007.09070, 2020.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, 2016.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

I. Misra and L. V. D. Maaten. Self-supervised learning of pretext-invariant representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706–6716, 2020.

A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *NeurIPS*, 2013.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *ArXiv*, abs/2006.04884, 2020.

R. Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *NeurIPS*, 2019.

Courtney Napoles, Matthew R. Gormley, and Benjamin Van Durme. Annotated gigaword. In *AKBC-WEKEX@NAACL-HLT*, 2012.

K. Nar, O. Ocal, S. Sastry, and K. Ramchandran. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *ArXiv*, abs/1901.08360, 2019.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NeurIPS*, 2001.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, J. Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. 2020.

A. Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Rajat Raina, Yirong Shen, Andrew Y. Ng, and Andrew McCallum. Classification with hybrid generative/discriminative models. In *NeurIPS*, 2003.

Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/saunshi19a.html`.

Florian Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir D. Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. In *ICLR*, 2015.

Christian Szegedy, V. Vanhoucke, S. Ioffe, Jon Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019. URL `https://openreview.net/forum?id=rJ4km2R5t7`.

Zhirong Wu, Yuanjun Xiong, S. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv: Learning*, 2019.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6022–6031, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016.

Hongyi Zhang, M. Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning. *ArXiv*, abs/2006.05987, 2020.

X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, 2015.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.

A Appendix

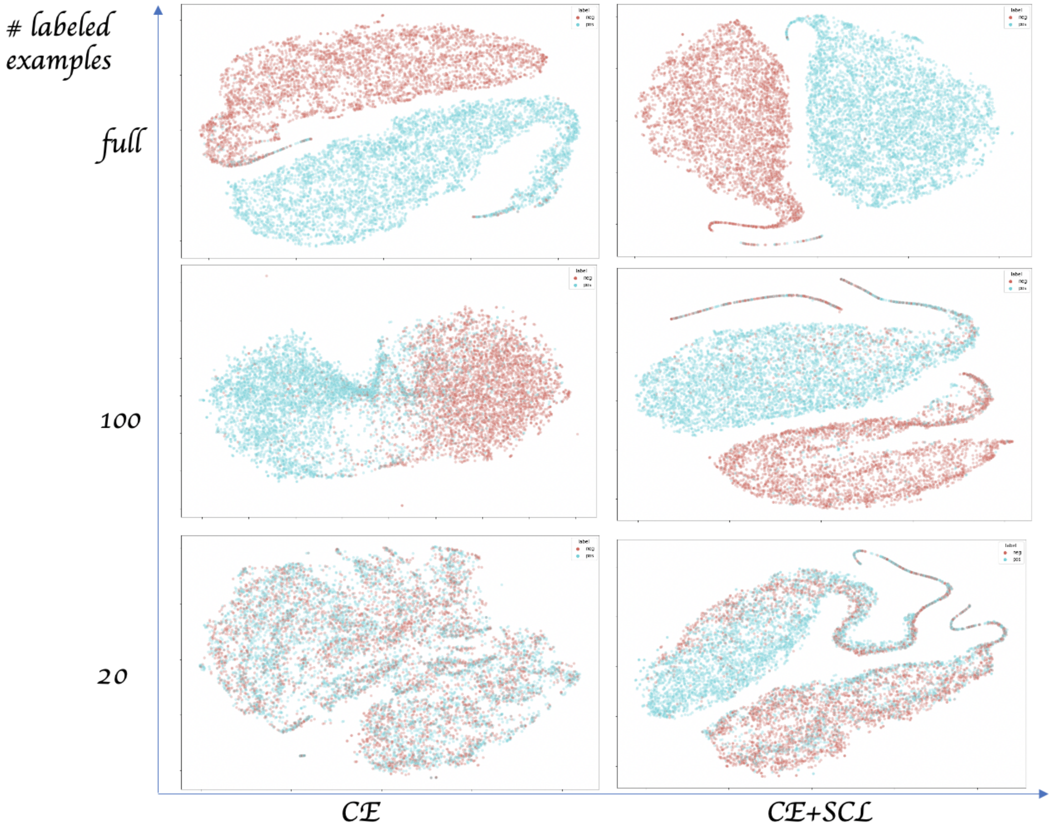

Figure 3: tSNE plots of learned CLS embedding on SST-2 test set where we have 20, 100 labeled examples, and full dataset respectively, comparing CE with and without SCL term. Blue: positive examples; red: negative examples.

Model	Loss	SST-2	CoLA	MRPC	RTE	QNLI	MNLI	Avg
RoBERTa$_{\text{Large}}$	CE	96.0±0.4	86.0±0.5	86.4±2.4	85.5±1.8	90.4±0.8	88.4±1	88.8
RoBERTa$_{\text{Large}}$	CE + SCL p-value	**96.3±0.4** 0.07	86.1±0.8 0.63	**89.5±0.9** 0.01	**85.7±0.5** 0.06	**93.9±0.7** 0.01	88.6±0.7 0.16	**90**
RoBERTa$_{\text{Large}}$	CE + CE p-value	96±0.4 0.39	86.3±0.4 0.13	89±1 0.01	84.9±1 0.1	93.9±0.8 0.01	**89±1** 0.12	89.9
RoBERTa$_{\text{Large}}$	Khosla et al. (2020) p-value	96±0.3 0.4	**86.7±1** 0.42	89.3±1.2 0.01	85.2±1 0.22	92.4±0.7 0.01	88.8±0.9 0.13	89.7

Table 8: Test results on the validation set of GLUE benchmark. We compare fine-tuning RoBERTa-Large with CE with and without SCL, CE+CE and the two-stage method of Khosla et al. (2020). Best hyperparameter configuration is picked based on the average validation accuracy. We report average accuracy across 10 seeds for the model with the best hyperparameter configuration, its standard deviation, and p-values. CE+CE refers to the case where we replace SCL loss with the CE loss but keep l2 normalization and temperature scaling.

Model	**Loss**	**Bsz**	**SST-2**	**CoLA**	**QNLI**	**MNLI**	**Avg ups/sec**
RoBERTa$_{\text{Base}}$	CE	16	94.1±0.5	83.3±0.7	88.2±0.8	84±0.6	15.9
RoBERTa$_{\text{Base}}$	CE + SCL	16	94.9±0.6	83.7±0.9	92.5±0.4	85.3±0.5	15.08
RoBERTa$_{\text{Base}}$	CE + CE	16	94.8±0.7	83.6±0.4	91.6±0.5	85±0.3	15.25
RoBERTa$_{\text{Base}}$	CE	64	94.2±0.4	83.3±0.5	89.2±0.5	84±0.4	8.43
RoBERTa$_{\text{Base}}$	CE + SCL	64	94.7±0.2	83.8±0.6	92.6±0.5	85.7±0.7	7.44
RoBERTa$_{\text{Base}}$	CE + CE	64	94.6±0.7	83.5±0.6	92.1±0.8	85±0.8	7.64
RoBERTa$_{\text{Base}}$	CE	256	94.1±0.4	84±0.5	90±0.7	84.4±0.6	**2.46**
RoBERTa$_{\text{Base}}$	CE + SCL	256	**95.2±0.3**	**84.5±0.5**	**92.9±0.3**	**86.6±0.6**	1.54
RoBERTa$_{\text{Base}}$	CE + CE	256	94.3±0.5	83.5±0.3	91.9±0.4	84.6±0.8	1.77

Table 9: Ablation study on performance and training speed shown as average updates per second (Avg ups/sec) for fine-tuning RoBERTa-Base with respect to the batch size (Bsz). CE+CE refers to the case where we replace SCL loss with the CE loss but keep l2 normalization and temperature scaling.

Model	**Loss**	**N**	**SST-2**	**QNLI**	**MNLI**
RoBERTa$_{\text{Large}}$	CE	20	85.9±2.1	65.0±2.0	39.3±2.5
RoBERTa$_{\text{Large}}$	CE + SCL	20	**88.1±3.3**	**75.7±4.8**	**42.7±4.6**
	p-value		5e-10	1e-46	1e-8
RoBERTa$_{\text{Large}}$	CE + CE	20	86.5±2.2	75.1±3.5	40.8±3.7
	p-value		0.03	4e-68	3e-4
RoBERTa$_{\text{Large}}$	CE	100	91.1±1.3	81.9±0.4	59.2±2.1
RoBERTa$_{\text{Large}}$	CE + SCL	100	**92.8±1.3**	**82.5±0.4**	**61.1±3.0**
	p-value		3e-17	1e-20	2e-4
RoBERTa$_{\text{Large}}$	CE + CE	100	91.7±0.5	81.7±0.5	56±4.0
	p-value		1e-4	3e-4	2e-8
RoBERTa$_{\text{Large}}$	CE	1000	94.0±0.6	89.2±0.6	81.4±0.2
RoBERTa$_{\text{Large}}$	CE + SCL	1000	**94.1±0.5**	**89.8±0.4**	**81.5±0.2**
	p-value		0.6	1e-12	0.5
RoBERTa$_{\text{Large}}$	CE + CE	1000	94±0.7	89.3±1	81.2±0.2
	p-value		0.78	0.06	0.12

Table 10: Few-shot learning test results on the GLUE benchmark where we have N=20,100,1000 labeled examples for training. Reported results are the mean and the standard deviation of the test accuracies of the top 3 models based on validation accuracy out of 10 random training set samples, along with p-values for each experiment. CE+CE refers to the case where we replace SCL loss with the CE loss but keep l2 normalization and temperature scaling.