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

(2021)

Beliz Gunel, Jingfei Du, Alexis Conneau, Ves Stoyanov

Summary

Contributions

The standard process for natural language classification consist of two steps: a first pre-training stage where a large language model is pre-trained on an auxiliary task using unlabeled data, followed by a fine-tuning stage using the cross-entropy loss. However, in the second stage, the loss of choice, which is the cross-entropy loss has sever limitations such as its instability. In order to overcome these limitations and push for a good generalization that captures the similarity and the dissimilarity between examples of different classes, which in turn is more label efficient loss, the authors proposed a new formulation of the contrastive loss in the supervised setting, which is used as an auxiliary objective without any specialized architectures, data augmentation or memory banks, and resulting in better performances and robustness.

Method

In order to overcome the many stability and generalization issues encountered during the fine-tuning stage, the authors proposed to leverage the commonalities between the examples of each class and contrast them with examples from other classes, which induced the model to focus on the important dimension of the multi-dimensional hidden representations and resulting in a better stability, robustness and better results with a limited amount of labels.

For a multi-class classification problem with C classes and a batch of N examples $\{x_i, y_i\}_{i=1,...N}$. Let the encoder be denoted as $\Phi(\cdot) \in \mathbf{R}^d$ with l_2 normalized hidden output representations before the softmax projection which gives the outputs $\hat{y}_i$, N_{y_i} be the total number of examples in the batch that have the same label as y_i, and $\tau > 0$ as an adjustable scalar temperature parameter and λ as a calar weighting hyperparameter tuned for each downstream task and setting. The total loss used for fine-tuning the model is as follows:

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

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

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

where the representations used for the SCL (Supervised Contrastive Learning) loss are the hidden representations corresponding to the token [CLS] for single sentence and sentence-pair tasks.

To summarize, the new loss consists of an additional auxiliary term (SCL) that acts over the hidden representations directly, and pushed the model to encoder input sentences of the same labels similarly in the embeddings space, and away from the inputs of different labels as illustrated bellow.

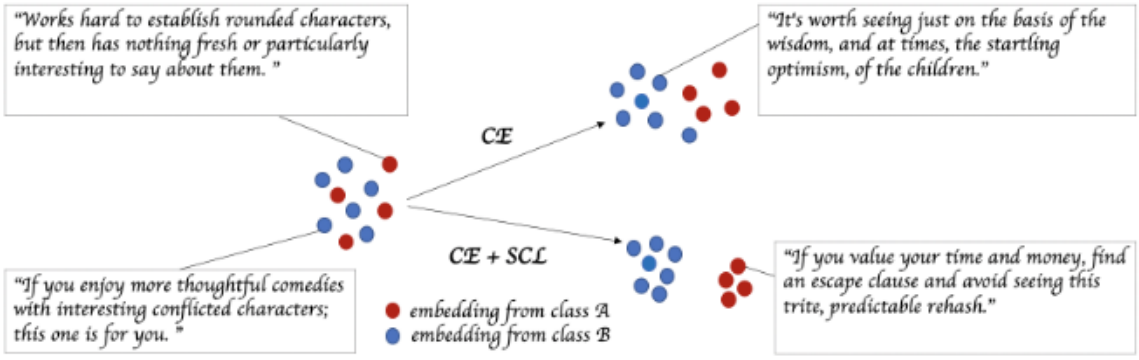

Figure 1: Our proposed objective includes a cross-entropy term (CE) and a supervised contrastive learning (SCL) term, and it is formulated to push examples from the same class close and examples from different classes further apart. We show examples from the SST-2 sentiment analysis dataset from the 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.

Results

examples as tSNE plots, where we have 20, 100 labeled examples and full dataset respectively for fine-tuning in Figure 3 in the Appendix.

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

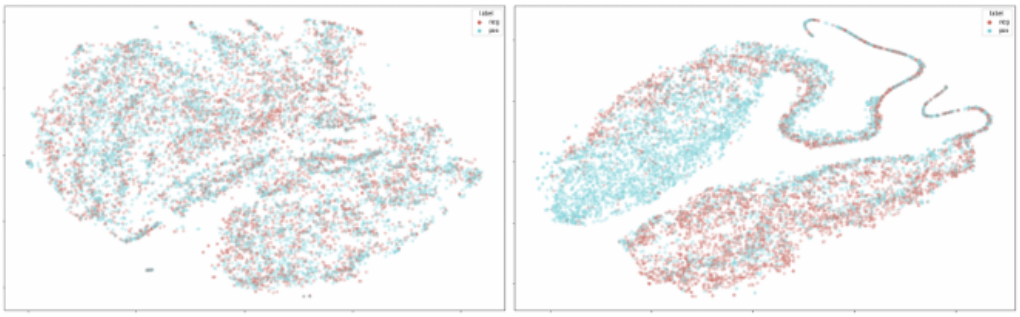

Figure 2: tSNE plots of the learned CLS embeddings on the SST-2 test set in the few-shot learning setting of having 20 labeled examples to fine-tune on – comparing RoBERTa-Large fine-tuned with CE only (left) and with our proposed objective CE+SCL (right) for the SST-2 sentiment analysis task. Blue: positive examples; red: negative examples.

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.

Model	Loss	SST-2	CoLA	MRPC	RTE	QNLI	MNLI	Avg
$\text{RoBERTa}_{\text{Large}}$	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
$\text{RoBERTa}_{\text{Large}}$	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
$\text{RoBERTa}_{\text{Base}}$	CE	16	94.1±0.5	83.3±0.7	88.2±0.8	84±0.6	15.9
$\text{RoBERTa}_{\text{Base}}$	CE + SCL	16	94.9±0.6	83.7±0.9	92.5±0.4	85.3±0.5	15.08
$\text{RoBERTa}_{\text{Base}}$	CE	64	94.2±0.4	83.3±0.5	89.2±0.5	84±0.4	8.43
$\text{RoBERTa}_{\text{Base}}$	CE + SCL	64	94.7±0.2	83.8±0.6	92.6±0.5	85.7±0.7	7.44
$\text{RoBERTa}_{\text{Base}}$	CE	256	94.1±0.4	84±0.5	90±0.7	84.4±0.6	**2.46**
$\text{RoBERTa}_{\text{Base}}$	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).

Model	Loss	N	Amazon-2	Yelp-2
$\text{RoBERTa}_{\text{Large}}$	CE	40	87.4±6.4	90.8±2.2
$\text{RoBERTa}_{\text{Large}}$	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.