SimSCL: A Simple fully-Supervised Contrastive Learning Framework for Text Representation

Youness Moukafih, Abdelghani Ghanem, Karima Abidi, Nada Sbihi, Mounir Ghogho, Kamel Smaïli

SimSCL : A Simple fully-Supervised Contrastive Learning Framework for Text Representation

Youness Moukafih[1,2⋆], Abdelghani Ghanem[1⋆], Karima Abidi[2], Nada Sbihi[1], Mounir Ghogho[1], and Kamel Smaili[2]

[1] TICLab, College of Engineering and Architecture, Université Internationale de Rabat, Morocco
{youness.moukafih,abdelghani.ghanem,mounir.ghogho,nada.sbihi}@uir.ac.ma
[2] LORIA/INRIA-Lorraine 615 rue du Jardin Botanique, BP 101, F-54600 Villers-16s-Nancy, France
{youness.moukafih,karima.abidi,kamel.smaili}@loria.fr

Abstract. During the last few years, deep supervised learning models have been shown to achieve state-of-the-art results for Natural Language Processing tasks. Most of these models are trained by minimizing the commonly used cross-entropy loss. However, the latter may suffer from several shortcomings such as sub-optimal generalization and unstable fine-tuning. Inspired by the recent works on self-supervised contrastive representation learning, we present **SimSCL**, a framework for binary text classification task that relies on two simple concepts: (i) Sampling positive and negative examples given an anchor by considering that sentences belonging to the same class as the anchor as positive examples and samples belonging to a different class as negative examples and (ii) Using a novel FULLY-SUPERVISED contrastive loss that enforces more compact clustering by leveraging label information more effectively. The experimental results show that our framework outperforms the standard cross-entropy loss in several benchmark datasets. Further experiments on Moroccan and Algerian dialects demonstrate that our framework also works well for under-resource languages.

Keywords: Natural Language Processing · Contrastive Learning · Neural Network · Supervised Learning.

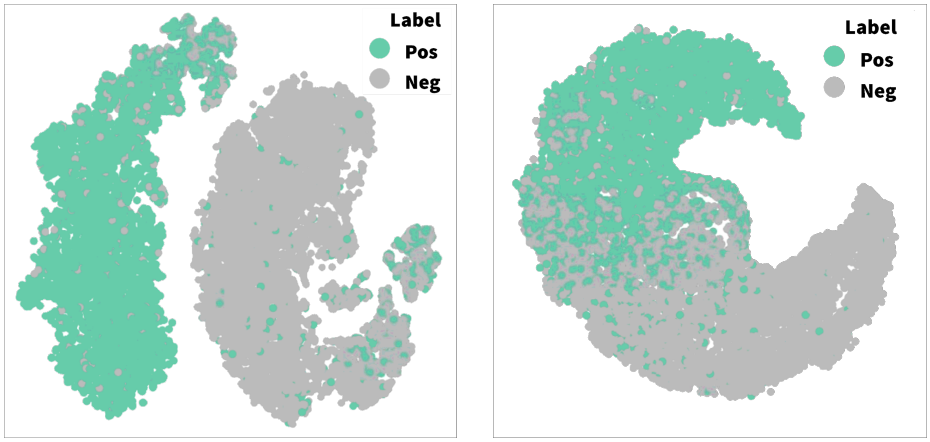
1 Introduction

Over the last few years, deep supervised learning models have achieved tremendous success in a variety of applications across many disciplines varying from Computer Vision (CV) and Automatic Speech Recognition (ASR) to Natural Language Processing (NLP). These models are usually trained by minimizing the commonly-used cross-entropy (CE) objective function. The basic concept of CE is simple and intuitive: each class is assigned a target (usually 1-hot) vector.

⋆ Equal contribution

Despite its popularity, the CE objective loss – the KL-divergence between one-hot vectors of labels and the distribution of the model's output logits – suffers from major robustness issues, which limits its use. In fact, CE suffers from adversarial robustness, as was shown in [1], which demonstrated empirically that training with a CE loss can cause the representations to spread sparsely over the representation space during training. Additionally, introducing noisy data seams to reduce the performance substantially, due to the fact that the cross entropy loss supposes that all the training labels are true, and neglects the fuzziness of noisy labels [2].

To overcome the above-mentioned challenges, many successful alternatives have been proposed to adjust the reference label distribution problems through label smoothing [3,4], Mixup [5], and knowledge distillation [6]. Recently, contrastive learning (CL) algorithms that were developed as estimators of mutual information, has led to major advances in self-supervised representation learning. These methods explicitly aim at training an encoder to learn latent representations of data instances to learn by pulling together representations of augmented views of the same data example (positive pairs), and pushing away representations of augmented views of different data examples (negative examples).



(a) SimSCL (b) Cross-Entropy

Fig. 1: T-SNE plots of the learned sentence representations using SimSCL and Cross-entropy on the SST-2 test set.

Inspired by the recent works on contrastive representation learning strategy, we introduce SimSCL, a simple supervised contrastive learning framework that uses a novel contrastive loss for binary classification task by leveraging label information more effectively. In this work, we consider many positives per an-

chor, unlike previous works on self-supervised contrastive learning which uses only a single positive example per anchor, and many negatives. In other words, the positive points are sampled from the same class as the anchor, instead of being augmented views of the same anchor, as done in self-supervised learning. In figure 2, we show how we select positive examples and negative examples for each class. The use of many positives and many negatives for each anchor in our framework allows the encoder function to better maximize the inter-class and minimize the intra-class similarities (learn effective generalizable features) than the standard framework which relies on the cross-entropy loss using the same model architecture. In figure 1 we can clearly see that our proposed objective function enforces more compact clustering of examples within the same class.

The empirical results show that our proposed framework consistently outperforms the standard cross-entropy loss using the same model architecture on three publicly available benchmark SA datasets, namely, Yelp-2, SST-2, and Amazon-2. Further experiments on Moroccan and Algerian dialects demonstrate that our framework also works well for under-resource languages.

2 Related Work

Our research builds upon previous works in self-supervised representation learning, contrastive learning, and supervised learning. Here, we shed light on the most pertinent papers.

Cross entropy is the de facto choice for the loss function in classification tasks. This prominence is due to many reasons. First, CE has good theoretical grounding in information theory, which makes it useful for theoretical analysis of systems [7]. Second, CE loss has been proven to rival many loss function in large data-sets. However, a number of works have analyzed the shortcomings of the commonly adapted cross-entropy objective function, showing that it leads to poor generalization performance due to poor margins, and sensitivity to noisy labels. Classification models are theoretically evaluated by their ability the separate classes in the representation space. Separability is also of practical use, since large margins can make models robust to small perturbations of the input space and hence, more robust to noise. [1] showed that CE does not maximize the separating margins between classes, and proposed an alternative that solves this problem. This phenomenon can be attributed to the leniency of the penalties of the cross entropy when close to the ground truth label (i.e. CE is eager for the model to be right), and can lead to poor generalization.

Recently, there has been several investigations for the use of contrastive loss for self-supervised learning. Primarily in the computer vision (CV) field, deep Contrastive learning has been use to great effect for learning image representations . For instance, in [8] Hinton and his colleagues propose SimCLR a simple framework, for learning visual representations without specialized architectures

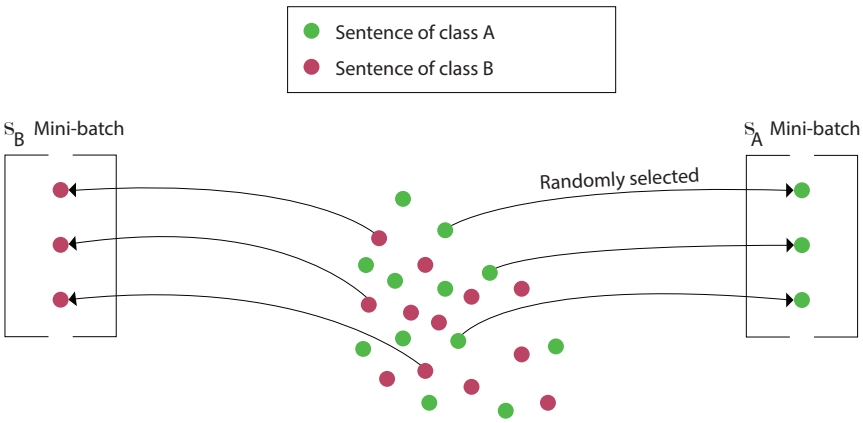


Fig. 2: Overview of the positive and negative examples construction process

or a memory bank, that generates anchor positive pairs by randomly augmenting the same image (e.g. random cropping and Rotation) while anchor-negative pairs are, from augmented views of different images within the same batch and minimizing a contrastive loss shown in Eq 1 that makes augmented views of the same example agree, which were shown to considerably outperform previous methods for self-supervised and semi-supervised learning on various benchmark datasets.

$$\mathcal{L}_i^{self} = -\sum_{i \in I} \log \frac{\exp(z_i . z_{j(i)}/\tau)}{\sum_{a \in A(i)} \exp(z_i . z_a/\tau)} \tag{1}$$

Where $i \in I \equiv \{1...2N\}$ is the index of an arbitrary augmented sample, $j(i)$ is the index of the other augmented sample originating from the same source sample. $A(i) \equiv I \ \{i\}$ is the set of all batch samples except the anchor. $P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$ the set of indices of all positives in the multiviewed batch distinct from i. We can rewrite the set $A(i)$ as $A(i) \equiv P(i) \cup N(i)$, where $N(i) \equiv \{p \in A(i) : \tilde{y}_p \neq \tilde{y}_i\}$.

Most similar to our method is the work done by [9], in that paper the authors proposed two variants of a contrastive supervised loss named SupCon (see $\mathcal{E}$II2*and***Eq**3). SupCon outperforms cross entropy loss and produces state-of-the-art results on ImageNet using ResNet architecture [10] with four different implementations of data augmentation.

$$\mathcal{L}_{out,i}^{sup} = \sum_{i \in I} -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i . z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i . z_a/\tau)} \tag{2}$$

$$\mathcal{L}_{in,i}^{sup} = \sum_{i \in I} -\log \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp(z_i.z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i.z_a/\tau)} \right\} \tag{3}$$

Other researchers have extended these methods to learn representations of graph structured data [11,12,13,14,15]. For instance, in [15] the authors proposed GraphCL a general framework for learning node representations in a self supervised manner using a contrastive loss that aims at maximizing the similarity between the representations of two transformations of the same node's local sub-graph. In the context of NLP, in [16] Mikolov and his colleagues proposed the first contrastive-based framework for learning word-level embeddings by using co- occurring words as positive pairs and k randomly chosen negative samples words form the corpus as negative pairs. More recently, contrastive learning was used for sentence-level representations.[17] proposes a self-supervised contrastive objective by performing both masked language modeling and contrastive learning to learn a universal sentence representations by training a transformer-based encoder to minimize the distance between the embeddings of textual segments randomly sampled from nearby in the same document. More recently, in [18] the authors proposed a novel loss for fine-tuning that includes a supervised contrastive learning term $[(1 - \lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{out}^{sup}]$ novel loss that a supervised contrastive learning objective for fine-tuning transformer-based pre-trained language models that improve performance over a strong RoBERTa-Large baseline on multiple datasets of the GLUE benchmark in the few-shot learning settings.

3 The proposed method

The main goal of the proposed framework is to learn representations by training an encoder network via a novel fully-supervised contrastive loss for classification task. The objective function is meant to capture the similarities between sentences of the same class while distancing the representations of sentences belonging to different classes.

First, we tackle one of the most crucial steps in any contrastive learning framework, namely, creating positive samples. In self-supervised, the common way for creating these positive samples is using various data augmentation strategies such as rotations and cropping in computer vision domain. Nevertheless, directly grafting the way of generating augmented views from the image domain is infeasible, since arbitrarily altering a sentence may change its semantics and thus its sentiment. To address this issue, we consider that sentences of the same sentiment are positive examples of each other. In figure 2, we show how we select positive examples and negative examples for each class.

3.1 Inter-class and Intra-class Distances

We first set up notation and describe the proposed framework for classification tasks that will be essential for the analysis. Let $\mathcal{D} = \{(x_i, y_i)\}_i$ be the dataset, where x_i represent the i^{th} sentence of the dataset and y_i is its label. Let $\mathcal{S}_l^k = \{(x_{k_i}, y_{k_i}) | y_{k_i} = y_{k_j}, \forall i \neq j; 0 \leq i \leq l\}$ denote the set of all sentences belonging to the same class within the corpus with l being the max-length of the sentences. Let $\mathcal{B}_k \sim \mathcal{S}_l^k$ be a mini-batch of randomly sampled examples from the same class. Let $f_w(.)$ denotes the encoder operator where the sub-index w refers to the weights of the encoders to be learnt. Let $H_k = f_w(B_k) \in \mathbb{R}^{N \times d}$ be the highest level l_2 normalized representation of the encoder where N is the batch size and d is the dimension of the embedding vector. The j^{th} row of H_k corresponds to the embedding $h_j^{(l)}$ of sentence k. Mathematically, this can be presented as

$$H_k = \begin{bmatrix} h_1^\top \\ h_2^\top \\ \vdots \\ h_{|\mathcal{B}_k|}^\top \end{bmatrix} \in \mathbb{R}^{N \times d}$$

3.2 Contrastive objective function

Now that we have all the mathematical notions needed, we proceed with the formulation of the objective function of the proposed contrastive learning framework:

$$V_{(C1,C1)} = H_1 H_1^\top = f_w(\mathcal{B}_1) \times f_w(\mathcal{B}_1)^\top \in \mathbb{R}^{N_1 \times N_1} \tag{4}$$

$$V_{(C2,C2)} = H_2 H_2^\top = f_w(\mathcal{B}_2) \times f_w(\mathcal{B}_2)^\top \in \mathbb{R}^{N_2 \times N_2} \tag{5}$$

$$U_{(C1,C2)} = H_1 H_2^\top = f_w(\mathcal{B}_1) \times f_w(\mathcal{B}_2)^\top \in \mathbb{R}^{N_1 \times N_2} \tag{6}$$

$$U_{(C2,C1)} = H_2 H_1^\top = f_w(\mathcal{B}_2) \times f_w(\mathcal{B}_1)^\top \in \mathbb{R}^{N_2 \times N_1} \tag{7}$$

Where, $\mathcal{B}_1 \sim \mathcal{S}_l^1$ and $\mathcal{B}_2 \sim \mathcal{S}_l^2$. $V_{(C1,C1)}$ and $V_{(C2,C2)}$ are the distances between the positive examples, which should be minimized by the objective (inter-class similarity), and $V_{(C1,C2)}$ and $V_{(C2,C1)}$ are the distance between the representation of each batch and the representation of the other batch, which should be maximized (intra-class similarity).

Finally, we concatenate the $V_{(C1,C1)}$ with $V_{(C2,C2)}$ inter-class matrices, and $U_{(C1,C2)}$ with $U_{(C2,C1)}$ intra-class:

$$V = \begin{bmatrix} V_{(C1,C1)} \\ V_{(C2,C2)} \end{bmatrix} \qquad U = \begin{bmatrix} U_{(C1,C2)} \\ U_{(C2,C1)} \end{bmatrix}$$

In order to overcame the mismatching dimension of $V_{(C1,C1)}$ with $V_{(C2,C2)}$ and that of $U_{(C1,C2)}$ with $U_{(C2,C1)}$ we adjust the matrix of the lowest dimension

(column space) to be equal to that of the highest one by adding zeros.

given the previous calculations, we formulate our supervised contrastive objective that we call $\mathcal{L}_{concat}^{sup}$ as follows:

$$\mathcal{L}_{concat}^{sup} = -\frac{1}{N}\sum_{i=1}^{N}\log\left\{\frac{\frac{1}{N_{C(i)}-1}\sum_{j=1,j\neq i}^{N_{C(i)}}\exp(V_{ij}/\tau)}{\frac{1}{N_{C(i)}-1}\sum_{j=1,j\neq i}^{N_{C(i)}}\exp(V_{ij}/\tau)+\frac{1}{N_{\overline{C(i)}}}\sum_{j=1}^{N_{\overline{C(i)}}}\exp(U_{ij}/\tau)}\right\} \tag{8}$$

where, $N = N_{C(i)} + N_{\overline{C(i)}}$, $N_{C(i)}$ is the number of elements of the same class as example i and $N_{(i)}$ the number of the elements of the other class, $\tau \in \mathbb{R}^+$ is a scalar temperature parameter, V_i and U_i are the i^{th} elements of V and U respectively.

Note that, by minimizing the proposed supervised contrastive loss $\mathcal{L}_{concat}^{sup}$, the encoder operator adjust its weights so that representation of sentences with same class label are close to each other (high values for $[\mathrm{V}_i]_{i=1,2...}$), while representation of sentences belonging to different classes are far from each other (low values for $[\mathrm{U}_i]_{i=1,2...}$). Moreover, in contrast to self-supervised learning, our objective function trains the model by exploiting multiple positive examples, resulting in more compact clustering of the embedding space.

It is worth pointing out that our loss is generic, i.e, it can be used for any binary classification problem. In this paper, we only use it for text binary classification task, leaving its exploration for other classification tasks as future work.

4 Main Results

4.1 Datasets

We evaluate the effectiveness of the proposed losses on three popular datasets namely, SST-2, Yelp-2, and Amazon-2, used for benchmarking state-of-the-art sentiment classification learning methods. We also tested our proposed objective function for two other low-resource languages datasets namely, the Moroccan Sentiment Analysis Corpus (MSAC) and the Algerian Sentiment Analysis Corpus (ASAC). MSAC is a multi-domain dataset containing sentences from sport, social and politics domains. ASAC is an Algerian Sentiment Analysis Corpus, it is our own dataset that we collected and annotated taking advantage of the data available on Youtube video comments. The problem was how to extract from Youtube, only the comments concerning the Algerian dialect. As there is no standard method for this problem, we opted for the approach we had proposed

in a previous work [19]. We collected data on YouTube, by selecting several hashtags or keywords used mainly by Algerians. Subsequently, all the data collected was checked and filtered and those which were not Algerian, they were excluded from the corpus. Then, we manually annotated all the comments by assigning each sentence its polarity (we didn't take into consideration neutral sentences). The constructed corpus, written in Arabic and Latin characters, is made up of 3976 positive comments and 4443 negative comments [3]. We summarize each dataset in Table 1.

Table 1: dataset statistics

Dataset	**#Train**	**#Dev**	**#Test**	**#Classes**
SST-2	60k	3.5k	3.5k	2
Yelp-P	600k	50k	38k	2
Amazon-P	3M	600k	400k	2
MSAC	1.6k	0.2k	0.2k	2
ASAC	6.8k	0.8k	0.8	2

4.2 Training Details

Our framework allows various choices of the network architecture without any constraints. However, since the aim of this work is to compare different loss functions on the same model architecture, we opt for simplicity and adopt the commonly used BiLSTM-based encoder.

For SST-2, Yelp-2, and Amazon-2 datsets, $\mathcal{L}_{concat}^{sup}$ was trained for 60 epochs using Adam optimizer with learning rate of 0.001 [20]. We initialize the input layer of the encoder with Glove pre-trained word representations of size 300 [21]. we use an encoder function of 3 hidden layers, a hidden units of 512, and a batch size of 800. We apply dropout of 0.5 on each layer. Note that the CE loss is evaluated by increasing the mini-batch size up to 1000. However, the best results are obtained using a batch size of 500.

For ASAC and MSAC datsets, similarly, the supervised $\mathcal{L}_{concat}^{sup}$ was trained for 15 epochs using Adam optimized with a learning rate of 0.003. However, for these datasets, we use an encoder with 1 hidden layer due to the number of examples that we have in the datasets, a hidden units of 128, and a batch size of 200. We apply dropout of 0.1. Similarly, we the CE is trained for a batch size up to 400, but the best results are obtained using a batch size of 64.

[3] This corpus will be made public

Following common practice, we opt for a linear evaluation of the learned sentence representations. More precisely, we use the learnt representations to train a logistic regression model to solve the text classification task. In practice, the evaluation process was performed using both a linear, and non-linear (ReLU activation) classifier. However, better results were obtained by the latter, achieving an average performance gain of 2% across all datasets.

4.3 Classification Accuracy

Here, we report the obtained results using $\mathcal{L}_{concat}^{sup}$ in different settings on 5 benchmark datasets, and those obtained by the CE and SupCon [9] losses. The results are given in terms of accuracy score measured on the same balanced test set.

Table 2: Linear evaluation of representations with different projection heads g(·) (Accuracy). The representation h (before projection) is 512-dimensional (%).

Dataset \Projection	**Identity**	**Linear**	**Non-linear**
SST-2	93.40	93.60	**94.15**
Yelp-2	95.13	95.31	**95.45**
Amazone-2	93.23	93.61	**94.71**
MSAC	78.48	79.33	**80.10**
ASAC	79.73	80.91	**82.63**

Following common practice, we first study the importance of adding a projection head that maps representations to the space where supervised contrastive loss is applied. Similar to we tested three different MLP architecture: (1) identity mapping; (2) linear projection $z = g(h) = W^{(1)}h \in \mathcal{R}^{512}$; (3) non-linear projection with one additional hidden layer as used by several previous approaches $z = g(h) = W^{(2)}ReLU(W^{(1)}h) \in \mathcal{R}^{512}$. Similar to what have found in previous works, we observe that a non-linear is better than linear and identity functions for projection head (See table 2). Note that, the projection head network is used only in the contrastive training phase, however, we discard it at the fine-tuning and inference phases.

For the evaluation performance, we tested our supervised representation for transfer learning in two settings: (1) the (non-linear) classifier is trained on top of the frozen representation (transfer learning); (2) we train the classifier, where we allow all weights to be adjusted during training (fine-tuned). It is clear from the table 3 that the learnt representations by our loss function are useful for the downstream tasks without adjusting them. In this paper, we provide the results that we obtained with the transfer learning strategy.

Table 3: Comparison of transfer learning and fine-tuning performance (Accuracy).

Dataset	**Transfer learning**	**Fine-tuned**
SST-2	94.05	**94.15**
Yelp-2	**95.53**	95.45
Amazone-2	**94.71**	94.58
MSAC	**80.10**	78.21
ASAC	**82.63**	82.16

Table 4: Performance Results (%)

	Classification Accuracy Results		
Dataset	**CE**	**SupCon** ($\mathcal{L}_{out}^{sup}$)	**SimSCL** ($\mathcal{L}_{concat}^{sup}$)
SST-2	91.28	93.53	**94.15**
Yelp-2	92.12	94.84	**95.45**
Amazon-2	92.94	93.98	**94.71**
MSAC	72.51	78.33	**80.10**
ASAC	78.70	82.11	**82.63**

Table 4 shows the obtained results of biLSTM-based model using our $\mathcal{L}_{concat}^{sup}$ objective function on the previously described datasets; and those obtained by the cross-entropy and SupCon losses. The results are given in terms of accuracy score measured on the same balanced test set. It is clear that in all cases, our framework provides better performance; the gain in performance is significant. Indeed, SimSCL leads to a 4.7% improvement of accuracy on SST-2, 3.6% improvement on Yelp-2, 5% improvement on Amazon, 3.9% improvement on ASAC, and 7.6% improvement on MSAC compared to CE loss. The large performance gap for MSAC dataset demonstrates that cross-entropy struggles with separating classes when dealing with small datasets. Furthermore, the results for MSAC and ASAC prove that our framework is very promising for under-resourced languages, which makes it advantageous over more sophisticated models such as transform-based models (e.g, BERT, RoBERT), which cannot be used for these languages due to the large amount of data needed for pre-training. Moreover, our experiments showed that CE overfits the MSAC dataset very quickly, with a training accuracy of 96% and only 72% accuracy on test. The overfitting problem cannot be explained by the large number of parameters of the biLTSM, since SimSCL-Obj also uses biLTSM (i.e, the same number of parameters as CE). Indeed, the problem can be explained by the fact that CE learns very poor margins between the two classes. Finally, it is worth pointing out that, similar to the cross-entropy loss, our loss function is robust to weights initialization.

5 Conclusion & future work

In this paper, we presented SimSCL, a simple supervised contrastive learning framework for training deep neural network for the binary classification task using a novel loss function. The latter is based on inter-class similarities (to be maximized), and intra-class similarities (to be minimized). We demonstrated, empirically, that SimSCL separates the two classes better than the encoder based on the classical cross-entropy and the SupCon losses. In the future, we plan to extend our method to other domains, such as computer vision, and graph neural network.

References

1. *ang, Tianyu and Xu, Kun and Dong, Yinpeng and Du, Chao and Chen, Ning and Zhu, Jun.* [*"GRethinking softmax cross-entropy loss for adversarial robustness."*]. *arXiv preprint arXiv:1905.10626. 2019.*
2. *Zhang, Tianyi and Wu, Felix and Katiyar, Arzoo and Weinberger, Kilian Q and Artzi, Yoav.* [*"Revisiting few-sample BERT fine-tuning."*]. *arXiv preprint arXiv:2006.05987. 2020.*
3. *Szegedy, Christian and Vanhoucke, Vincent and Ioffe, Sergey and Shlens, Jon and Wojna, Zbigniew.* [*"Rethinking the inception architecture for computer vision."*]. *Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.*
4. *SMüller, Rafael and Kornblith, Simon and Hinton, Geoffrey.* [*"When does label smoothing help?."*]. *arXiv preprint arXiv:1906.02629. 2019.*
5. *SZhang, Hongyi and Cisse, Moustapha and Dauphin, Yann N and Lopez-Paz, David.* [*"mixup: Beyond empirical risk minimization."*]. *arXiv preprint arXiv:1710.09412. 2017.*
6. *Hinton, Geoffrey and Vinyals, Oriol and Dean, Jeff.* [*"Distilling the knowledge in a neural network."*]. *arXiv preprint arXiv:1503.02531. 2015.*
7. *Andreieva, Valeria and Shvai, Nadiia.* [*"Generalization of Cross-Entropy Loss Function for Image Classification."*]. *arXiv preprint arXiv:1503.02537. 2020.*
8. *Chen, Ting and Kornblith, Simon and Norouzi, Mohammad and Hinton, Geoffrey.* [*"A simple framework for contrastive learning of visual representations."*]. *International conference on machine learning. 2020.*
9. *Khosla, Prannay and Teterwak, Piotr and Wang, Chen and Sarna, Aaron and Tian, Yonglong and Isola, Phillip and Maschinot, Aaron and Liu, Ce and Krishnan, Dilip.* [*Supervised contrastive learning."*]. *arXiv preprint arXiv:2004.11362. 2020.*
10. *e, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian.* [*"Deep residual learning for image recognition."*]. *Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.*
11. *Hassani, Kaveh and Khasahmadi, Amir Hosein.* [*"Contrastive multi-view representation learning on graphs."*]. *International Conference on Machine Learning. 2020.*
12. *Zhu, Yanqiao and Xu, Yichen and Yu, Feng and Liu, Qiang and Wu, Shu and Wang, Liang.* [*"Deep graph contrastive representation learning."*]. *arXiv preprint arXiv:2006.04131. 2020.*
13. *Veličković, Petar and Fedus, William and Hamilton, William L and Liò, Pietro and Bengio, Yoshua and Hjelm, R Devon.* [*"Deep graph infomax."*]. *arXiv preprint arXiv:1809.10341. 2018.*

14. *Qiu, Jiezhong and Chen, Qibin and Dong, Yuxiao and Zhang, Jing and Yang, Hongxia and Ding, Ming and Wang, Kuansan and Tang, Jie.* [*"Gcc: Graph contrastive coding for graph neural network pre-training."*]. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.*
15. *Hafidi, Hakim and Ghogho, Mounir and Ciblat, Philippe and Swami, Ananthram.* [*"Graphcl: Contrastive self-supervised learning of graph representations."*]. *arXiv preprint arXiv:2007.08025. 2020.*
16. *Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey.* [*"Efficient estimation of word representations in vector space."*]. *arXiv preprint arXiv:1301.3781. 2013.*
17. *Giorgi, John M and Nitski, Osvald and Bader, Gary D and Wang, Bo.* [*"Declutr: Deep contrastive learning for unsupervised textual representations."*]. *arXiv preprint arXiv:2006.03659. 2020.*
18. *Gunel, Beliz and Du, Jingfei and Conneau, Alexis and Stoyanov, Ves.* [*"Supervised contrastive learning for pre-trained language model fine-tuning."*]. *arXiv preprint arXiv:2011.01403. 2020.*
19. *Abidi, Karima and Menacer, Mohamed Amine and Smaili, Kamel.* [*"CALYOU: A comparable spoken Algerian corpus harvested from youtube."*]. *18th Annual Conference of the International Communication Association (Interspeech). 2017.*
20. *Kingma, Diederik P and Ba, Jimmy.* [*"Adam: A method for stochastic optimization"*]. *arXiv preprint arXiv:1412.6980. 2014.*
21. *Pennington, Jeffrey and Socher, Richard and Manning, Christopher D.* [*"Glove: Global vectors for word representation"*]. *Proceedings of the 2014 conference on empirical methods in natural language processing. 2014.*