SuperConText: Supervised Contrastive Learning Framework for Textual representations

Youness Moukafih, Nada Sbihi, Mounir Ghogho, Kamel Smaïli

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

SuperConText: Supervised Contrastive Learning Framework for Textual representations

YOUNESS MOUKAFIH[1,2], NADA SBIHI[1], MOUNIR GHOGHO[1], (FELLOW, IEEE), KAMEL SMAILI[2]

[1]TICLab, College of Engineering and Architecture, Université Internationale de Rabat, Morocco
[2]Loria, Campus Scientifique, Vandoeuvre Lés-Nancy, France

Corresponding author: First A. Author (e-mail: youness.moukafih@univ-lorraine.fr).

ABSTRACT In the last decade, Deep neural networks (DNNs) have been proven to outperform conventional machine learning models in supervised learning tasks. Most of these models are typically optimized by minimizing the well-known Cross-Entropy objective function. The latter, however, has a number of drawbacks, including poor margins and instability. Taking inspiration from the recent self-supervised Contrastive representation learning approaches, we introduce **Super**vised **Con**trastive learning framework for **Text**ual representations (SuperConText) to address those issues. We pretrain a neural network by minimizing a novel fully-supervised contrastive loss. The goal is to increase both inter-class separability and intra-class compactness of the embeddings in the latent space. Examples belonging to the same class are regarded as positive pairs, while examples belonging to different classes are considered negatives. Further, we propose a simple yet effective method for selecting hard negatives during the training phase. In extensive series of experiments, we study the impact of a number of parameters on the quality of the learned representations (e.g. the batch size). Simulation results show that the proposed solution outperforms several competing approaches on various large-scale text classification benchmarks without requiring specialized architectures, data augmentations, memory banks, or additional unsupervised data. For instance, we achieved top-1 accuracy of 61.94% on the Amazon-F dataset, which is 3.54% above the best result obtained when using the cross-entropy with the same model architecture.

INDEX TERMS Deep learning; Contrastive Learning; Text Classification; Hard Negative Examples

I. INTRODUCTION

Over the past few years, deep neural networks (DNNs) have achieved state-of-the-art results surpassing conventional machine learning algorithms in a variety of applications across many disciplines [1] [3] [4]. The success of deep learning is usually attributed to their ability to automatically learn multiple levels of representations in an end-to-end manner. Most of these models are usually optimized using the well-known Cross-Entropy objective function. Indeed, the concept of cross-entropy is straightforward and intuitive: every class is assigned a vector of a target (usually 1-hot). Despite its popularity, however, the cross-entropy – the KL-divergence between one-hot vectors of labels and the distribution of the model's output logits – suffers from major robustness issues. For example, training a deep neural network by the cross entropy loss is vulnerable to adversarial attacks [2]. Several works have demonstrated, theoretically, that training with cross-entropy loss can cause the representations to spread sparsely over the representation space during training [16]. Additionally, introducing noisy data seems to reduce the performance substantially, due to the fact that the loss considers that all the training labels are true, and neglects the fuzziness of noisy labels [5]. To overcome these issues, many successful alternatives were proposed to address the reference label distribution problems through label smoothing [6], [7], Mixup [9], and knowledge distillation [10]. Recently, contrastive representation learning, which was shown to be related to the estimation of mutual information, has led to major advances in self-supervised learning. Contrastive learning was also shown to achieve state-of-the-art performance on many large-scale benchmark datasets.

Contrastive learning is a particular form of a Siamese neural network, which consists of two or more identical subnetworks, each producing a vector representation of its re-

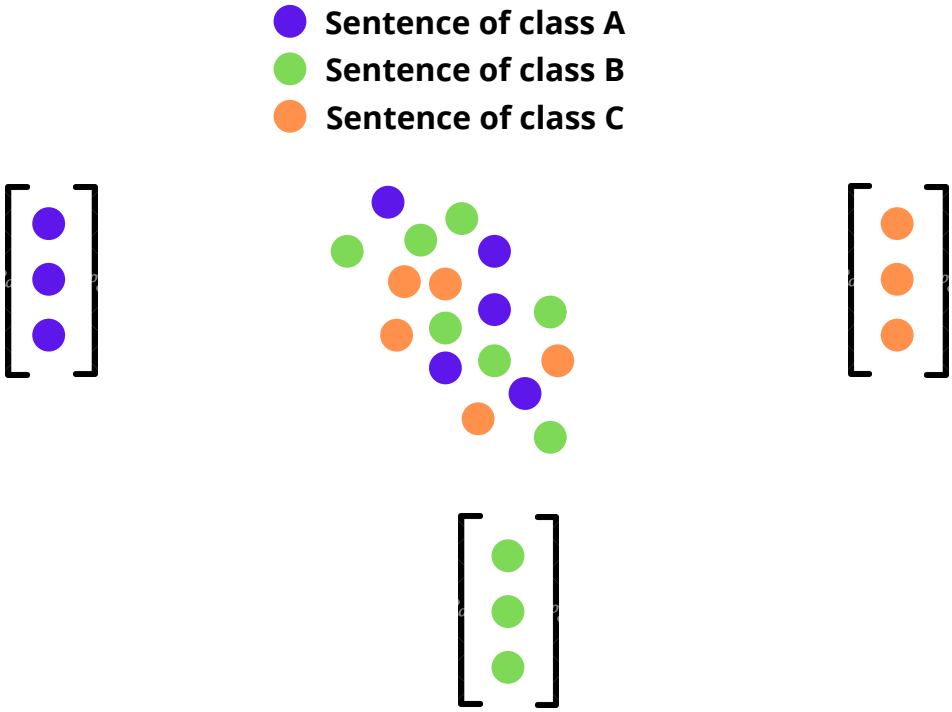
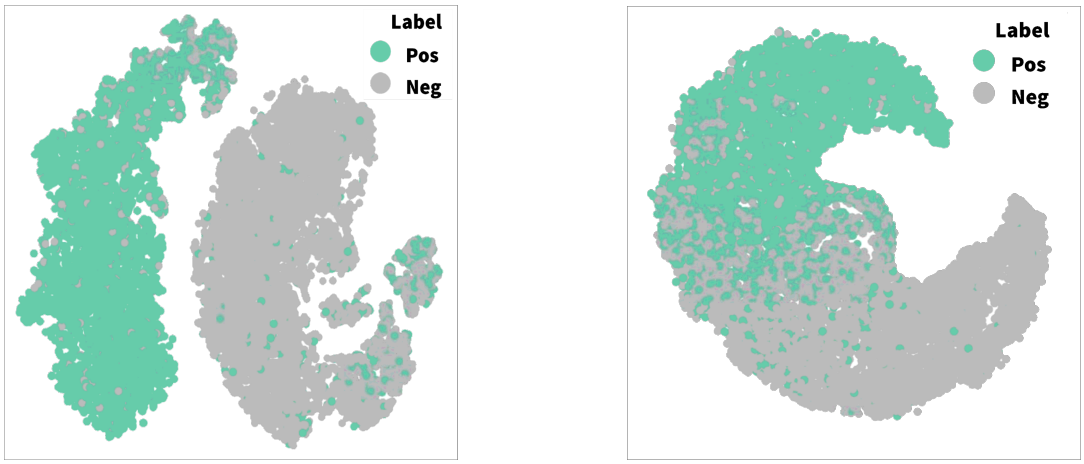


FIGURE 1: Overview of the positive and negative examples construction process.

spective input. Modern Siamese network approaches use data augmentation techniques to generate an augmented view of the same data instance across each sub-network and pull together the learned representations in the embedding space. Maximizing the similarities between each pair of augmented views can result in a trivial solution where all representations are equal to other; this is referred to as collapsing problem. Several approaches have proposed to solve the collapsing problem, one of which is contrastive learning. This latter prevents the undesirable trivial solution by contrasting between positive (similar) examples and many negative (dissimilar) examples. These methods explicitly aim at training a neural network to learn embeddings by pulling together the representations of augmented views, while pushing away the representations of augmented views of different data instances (negative examples), often by using noise-contrastive estimation. The most common strategy is to uniformly sample from the training dataset using examples either from the current batch or from a memory bank. However, it has been observed empirically that contrastive learning methods still suffer from dimensional collapse [38].

In this paper, we propose a supervised contrastive learning framework for text multi-class classification tasks. Inspired by the recent success of joint embedding approaches for learning representations in a self-supervised setting [11] [12], we develop a framework that learns sentence embeddings by maximizing the agreement between the representations of a cluster of instances belonging to the same class using a novel fully-supervised contrastive loss that guides a neural network to better separate the classes. We address the problem of text-multi-class classification applications. We consider many positives per anchor, unlike previous works on self-supervised contrastive learning which use only a single positive example and many negatives. In other words, instead of using augmented views of the same anchor as done in self-supervised contrastive learning (which is not obvious in textual data), we leverage label information to consider many positives and many negatives for each anchor. In figure 1, we show how we select positive/negative examples for each class. The use of many positives and many negatives in our framework allows the encoder function to better maximize the intra-class compactness and the inter-class separability (learns useful and generalizable features) than the standard framework which relies on the cross-entropy loss. Figure 2 shows tSNE plots of the learned representations of the model trained on SST-2 dataset with our loss against those learnt by cross entropy. The increased intra-class compactness and inter-class separability naturally lead to a better text classifier in the fine-tuning stage.

A number of studies have demonstrated that hard negatives (i.e. ones that are hard to distinguish from positives) are important for learning more powerful representations in contrastive learning. Therefore, a number of works have proposed novel sampling strategies. In [13], the authors use hard negative mixing to synthesize new examples from the available hard negatives, while in [14], the authors sample



(a) SuperLoss (b) Cross-Entropy

FIGURE 2: T-SNE plots of the learned sentence embeddings using SuperLoss and the Cross-entropy on SST-2 dataset.

negatives from a ring around each positive (i.e. negatives that are neither too close nor too far from the positive example). In this paper, we propose a novel tunable feature-based sampling strategy for selecting hard negative examples based on the similarities resulting from the pretrained model, and show that our approach improves the performance of the learned representations on downstream classification tasks.

Further, several research studies have shown that the number of negative examples is crucial for learning high-quality representation in self-supervised contrastive learning. Accordingly, recent contrastive learning approaches use large batch sizes, or keep large memory banks. For instance, in [25], the authors proposed Momentum Contrast (MoCo) that uses a queue with features of the last few batches, while in [31], a memory of the whole training data is utilized. It is however shown that increasing the memory/batch size does not always give better results. In this work, we conduct extensive experiments to analyse and investigate whether the conclusions that have been drawn from the most successful approaches for self-supervised contrastive learning are still valid when applied to textual data in a supervised contrastive learning setting.

The experimental results show that our Framework consistently outperforms the commonly-used supervised learning framework based on cross-entropy loss on many publicly available benchmark datasets. For instance, we achieve an accuracy of 61.94% and 95.45% using our framework on Amazon-F and Yelp-P respectively, while the scores obtained by the cross-entropy are 58.40% and 92.12% using the same neural architecture. The experimental results show that our proposed method benefits from large positive examples. However, we find that beyond a certain threshold, increasing the number of positive examples does not improve the quality of the representations. In contrast, the method produces high-quality representations when the number of negative instances is large. In addition, simulations show that the sampling strategy is crucial for learning more generalizeable sentence representations for downstream tasks on several benchmark datasets. Further experiments on Moroccan and Algerian dialects demonstrate that our method also works well for low-resource languages.

We summarize our contributions as follows:

- We propose a novel supervised learning framework for pre-training text representation by leveraging the contrastive learning paradigm.
- We compare the proposed approach with the standard cross-entropy loss-based method using several large-scale text classification datasets.
- We empirically study the effect of the number of positive/negative examples on the performance of the proposed supervised contrastive learning method.
- We propose a novel tunable feature-based sampling strategy for selecting hard negative examples that further improves performance.

II. RELATED WORK

A. CROSS ENTROPY LOSS

Cross Entropy (CE) 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 [15]. Second, CE loss has been proven to rival many loss functions in large datasets [45]. However, it suffers from major robustness issues. Indeed, CE suffers from adversarial robustness, as was shown in [16] 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 seems to degrade performance substantially [19] which is due to the fact that the cross entropy loss supposes that all the training labels are true, and neglects the fuzziness of noisy labels [17], [19]. Classification models are theoretically evaluated by their ability to 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. In [37], [42], the authors 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.

B. SELF-SUPERVISED REPRESENTATION LEARNING

One of the most prominent lines of research that ventures out of unsupervised representation learning is Self-Supervised Learning (SSL). This paradigm uses pretext tasks, which use intrinsic properties of the data in order to evaluate representations [46]. Contrastive SSL is a SSL training technique that tries to discriminate between two types of examples, a) positive examples and b) negative examples, given an anchor, often by using noise-contrastive estimation [49]. In contrastive SSL (e.g., [11], [36], [50]), the loss takes the following form:

$$\mathcal{L}^{self} = -\sum_{i\in\mathcal{I}} \log \frac{\exp(v_i.v_{p(i)}/\tau)}{\sum_{n\in\mathcal{N}(i)} \exp(v_i.v_n/\tau)} \quad (1)$$

where v_i and $v_{p(i)}$ are views of the same data element, $\mathcal{I} \equiv \{1, \cdots, 2N\}$ with N being the number of data elements in the batch, $\mathcal{N}(i) \equiv \mathcal{I}\backslash\{i\}$, and τ is a temperature parameter.

This SSL paradigm has been explored extensively for image representation learning [11], [21] and graph representation learning [22]. These methods sample negative and positive examples based on a certain principle of semantic similarity in the data space. These methods sample negative examples using three main strategies: a) cross-scale based strategies (e.g. Computer vision [51]), b) augmentation based strategies (e.g. graph [34], text [39]), and c) hybrid strategies (cross-scale and augmentation based strategies). Cross-scale-based methods contrast using representations of intermediate layers. That is, given a batch of training examples, the intermediate representations of a data instance are positive examples and the representations of other data instances in that batch are considered to be negative examples. Augmentation based methods contrast using augmented versions of the input data. That is, augmentations of a data instance are considered to be positive examples, while augmentations of different data instances are considered to be negative examples. Data augmentation strategies are not always straightforward, especially in the cases of graphs and text, and are thus still being investigated. Hybrid methods contrast intermediate representations of augmented data [52]. Multiple works have stressed the importance of sampling; [23] showed than sampling harder negative examples is more beneficial for model performance.

However, these methods are known to suffer from large computational costs [53], since the computation of the loss requires multiple forward passes in order to get embeddings for the negative examples. This motivated the development of a new set of methods referred to as non-contrastive. BYOL [24] is a pioneering work in this line of research, which was later followed by many works (e.g SwaV, Barlow Twins, SEER [26]–[28]). These works assert that negative examples regularize the models to prevent them from being *naive*, and they try to replace their role by explicit regularization using certain constraints that prevent the models from learning trivial representations.

C. SUPERVISED CONTRASTIVE REPRESENTATION LEARNING

Recently, many works have extended the self-supervised contrastive learning approach to the fully-supervised setting by leveraging label information for learning representations. In the computer vision field, [29] propose SupCon, an objective function for the task of image classification that bridges the gap between self-supervised learning and fully supervised learning. SupCon was shown to outperform SimCLR [11], Max-Margin [33] and cross-entropy on several benchmark datasets such as ImageNet. The SupCon objective function is defined by:

$$\mathcal{L}^{sup}_{out} = \sum_{i\in\mathcal{I}} -\frac{1}{|\mathcal{P}(i)|} \sum_{p\in\mathcal{P}(i)} \log \frac{\exp(v_i.v_p/\tau)}{\sum_{n\in\mathcal{N}(i)} \exp(v_i.v_n/\tau)} \quad (2)$$

where $\mathcal{I} \equiv \{1, \cdots, N\}$ with N being the batch size, $\mathcal{P}(i)$ is the set of indices of all data elements in the batch that belong to the same class as data element v_i, $\mathcal{N}(i) \equiv \mathcal{I}\backslash\{i\}$,and $|.|$ denotes the cardinality operator.

In [55], the authors proposed a novel objective function for fine-tuning transformer-based language models which consists of a weighted sum of the cross-entropy loss and the above-mentioned supervised contrastive loss: $(1-\lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}^{sup}_{out}$. This approach was shown to outperform a strong RoBERTa-large baseline on the GLUE benchmark dataset in few-shot learning settings. Many works have extended SupCon and proposed new variants. In [56], for instance, the authors evaluated three variants of pixel-wise label-based contrastive loss to pre-train a semantic segmentation model.

D. NEGATIVE MINING IN CONTRASTIVE REPRESENTATION LEARNING

In self-supervised contrastive learning, negative sampling has been shown to be very useful for learning good representations. Several strategies have been proposed to build negatives examples for visual presentations [11], [13], [14], [25]. In most of these works, the aim is to maximize the distance between the representation of a given anchor and those of negative examples that are difficult to discriminate against. In [14], for instance, the authors propose a method which consists of picking two percentiles w_k and w_l ($\in [0, 100]$) and considering h_{n_c} as a negative example for a representation of a query h_q if and only if $h_q^T.h_{n_c}$ is within the w_k-th to the

w_l-th percentile of all $h_n \in Q_q^-$ where Q_q^- denotes a set of negative examples. This enables to easily build hard negative examples (i.e., negatives that are hard to distinguish from the current sample) which are beneficial in learning powerful representations.

III. METHODOLOGY

A. REPRESENTATION LEARNING FRAMEWORK

We propose a supervised contrastive learning framework for textual representations. In our framework, we introduce a novel fully-supervised contrastive loss that we call **SuperLoss**. Our loss is optimized by training an encoder function, a neural network, to maximize the agreement between the normalized representations of a cluster of points with the same class label, while simultaneously pushing away clusters of samples from different classes.

The training takes as input q batches of data, each of which is composed of sentences with the same target, where q is the number of classes. All batches are forward propagated through a neural network to obtain a high-dimensional l_2 normalized embedding. The proposed framework is designed to maximize the agreement of vector representations of points belonging to the same class and contrast them with those of the other classes (see figure 1). To use the pretrained model for classification, we train a linear classifier on top of the frozen learnt representations using cross-entropy loss. As illustrated in figure 3, the framework comprises the following components:

- **Data Sampling**. This is a data loading step in which we randomly sample a batch of sentences from each class. In this paper, we do not use any data augmentation techniques to create positive pairs. Here, we consider sentences of the same class as positive examples of each other, and sentences of other classes as negative examples.

- **Neural Network Encoder**. The neural encoder function is denoted as $h_i = encoder(x_i) \in \mathbb{R}^d$. The output of the encoder provides sentence vector representations. Following the previous works on self-supervised contrastive learning, in all our experiments, the representations are normalized. Our framework allows various choices of the network architecture without any constraints.

- **Projection Head Network**. Following the findings of our previous work on supervised contrastive learning [29], we add a projection head neural network to map the representations to another space before computing the supervised contrastive loss.

- **Constrastive Loss**. The contrastive loss function, which we call SuperLoss, is used to train the neural network encoder and the projection head network.

B. SUPERLOSS FUNCTION

Here, we introduce SuperLoss, a **Super**vised contrastive **Loss**. Its minimization leads to networks which cluster together in the latent space sentences of the same class.

1) Preliminary Mathematical Concepts

Before diving any further into the definition of the objective function and the learning process, we next define certain mathematical notions and notations.

For q matrices $M^{(k)}$, $1 \leq k \leq q$, of the same dimensions, let $\sqcap\limits_{k\in[1,q]} M^{(k)}$ and $\sqsubset\limits_{k\in[1,q]} M^{(k)}$ denote respectively the matrices resulting from their vertical and horizontal concatenations. Let $\text{avg}(M)$ denote the vector obtained by averaging each of the rows of matrix M, i.e. its ith element is the average of the ith row of matrix M.

2) Inter-class and Intra-class Distances

We first define notations 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 available dataset, where x_i represents the i^{th} sentence and y_i is its label. Let $\mathcal{S}_k = \{(x_i, y_i)|y_i = k\}_i$ denote the subset of all sentences of the dataset belonging to class k. Let $\mathcal{B}_k \sim \mathcal{S}_k$ be a mini-batch of randomly sampled examples from $\mathcal{S}_k$. Let $f_w(.)$ denotes the encoder function where the sub-index w refers to the weights of the encoders to be learnt. Let $H_k = f_w(\mathcal{B}_k) \in \mathbb{R}^{N_k \times d}$ be the highest level representation of the encoder where N_k is the batch size and d is the dimension of the embedding vector. The j^{th} row of H_k is the transpose of the embedding vector associated with the j^{th} sentence of $\mathcal{B}_k$, which we denote as h_j, i.e.

$$H_k = \begin{bmatrix} h_1^\top \\ h_2^\top \\ \vdots \\ h_{N_k}^\top \end{bmatrix} \in \mathbb{R}^{N \times d}.$$

As shown in algorithm 1, in each training step our framework starts with sampling at random a batch of sentences from each class. Then we feed all of them to the encoder f_w separately. Finally, our objective function takes as input the normalized representations matrices produced by the encoder.

3) Objective function Formulation

Now that we have all the mathematical notions needed, we proceed with the formulation of the objective function of the proposed contrastive learning framework. First, we calculate the dot product between the representation of each sentence in a class batch with those of all other sentences within the same batch:

$$G_{pos}^k = H_k H_k^\top \in \mathbb{R}^{N_k \times N_k}. \quad (4)$$

Matrix G_{pos}^k contains the similarities between sentences beloning to the same class k. The aim is to maximize these

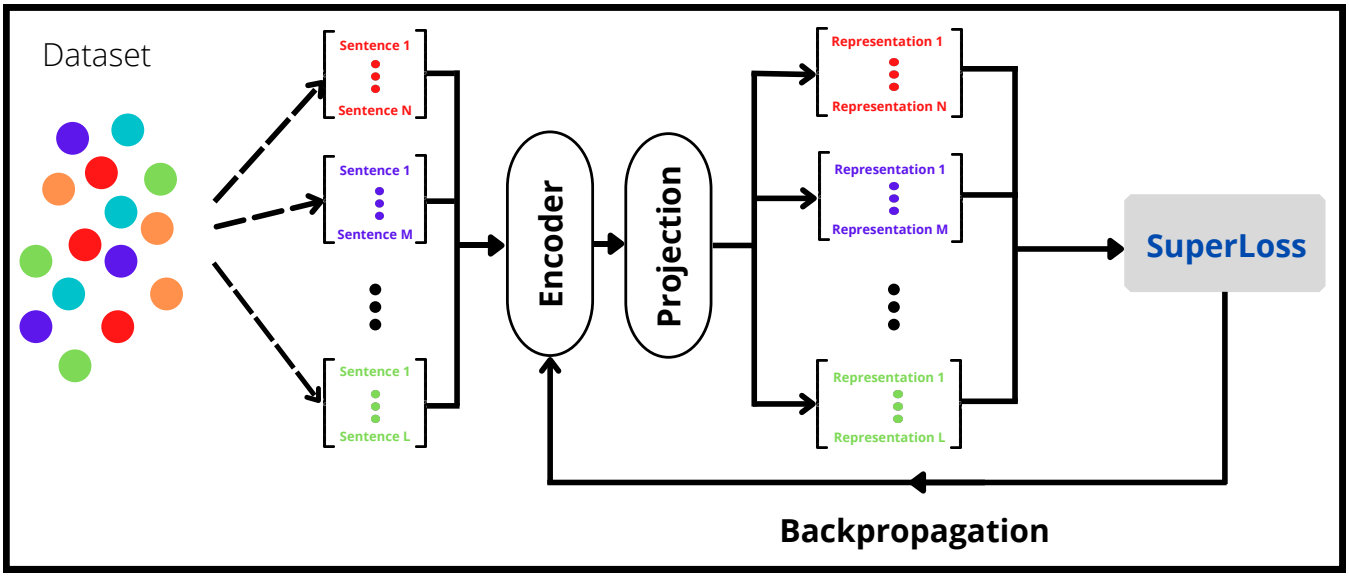


FIGURE 3: The general framework of our proposed approach.

similarities (intra-class similarity).

Then, we calculate the similarities between the representation of sentences beloning to different classes (intra-class similarity):

$$G^k_{neg} := \underset{j\neq k, 1\leq j\leq q}{\sqcap} H_k H_j^\top \in \mathbb{R}^{N_k \times \overline{N_k}} \tag{5}$$

where $\overline{N_k} = \sum_{j\neq k, 1\leq j\leq q} N_j$.

Next, we propose to average each matrix along the column axis after applying the exponential function:

$$v_{pos} = \underset{k\in[1,q]}{\sqcap} \text{avg}\left[\exp\left(G^k_{pos}/\tau\right)\right] \in \mathbb{R}^N \tag{6}$$

$$v_{neg} = \underset{k\in[1,q]}{\sqcap} \text{avg}\left[\exp\left(G^k_{neg}/\tau\right)\right] \in \mathbb{R}^N \tag{7}$$

where $\tau \in \mathbb{R}^+$ is a scalar temperature parameter and $N = \sum_{k=1}^{q} N_k$.
The proposed loss function is defined as follows:

$$\text{SuperLoss} = -\frac{1}{N}\sum_{\ell=1}^{N}\left[\log\left(\frac{v_{pos}[\ell]}{v_{pos}[\ell] + v_{neg}[\ell]}\right)\right] \tag{8}$$

where $v_{pos}[\ell]$ and $v_{neg}[\ell]$ are the ℓ^{th} elements of v_{pos} and v_{neg} respectively.
Here, the encoder's weights will be learnt so as to maximize the elements of v_{pos} (clusters of points belonging to the same class are pulled together in the latent space) and minimize those of v_{neg}, which will result in pushing representations of elements that do not belong to the same class apart from each other.

IV. EXPERIMENTS

In this section, we compare the proposed method with other techniques for sentence representation. First, we describe the datasets used in this paper, then we provide details of the architecture and training process of the proposed method.

A. DATASET AND TRAINING DETAILS

We evaluated the effectiveness of the proposed framework on sentence classification tasks by measuring accuracy on 6 benchmark datasets namely, SST-2, Yelp-P, yelp-F, Amazon-P, Amazon-F, and IMDb. Furthermore, the framework is also tested for representation learning on low-resource language setting datasets namely, MSAC, ASAC. We summarize each dataset based on their main task, domain, number of training examples, and number of classes in Table 1.

B. TRAINING DETAILS

Our framework allows various choices of the network architecture without any constraints. Here, we opt for simplicity and adopt the BiLSTM neural network architecture to compare different objective functions.

For ASAC and MSAC datsets, we used the following settings: we train our framework for 15 epochs using Adam [32] optimized with a learning rate of 0.003. However, for these datasets, we use an encoder with 1 hidden layer only

Algorithm 1: SuperConText Process Description.

Input: n_e: number of epochs , q : number of classes, N_k: batch size for the k^{th} class where $k \in \{1, 2, ..q\}$, $\mathcal{D} = (x_i, y_i)_i$: dataset

Output: f_w: model with trained weights

for $e \in [1, n_e]$ **do**

  For $k \in [1, q]$, $\mathcal{B}_k \underset{N_k}{\sim} \mathcal{D}$, where $|\mathcal{B}_k| = N_k$;

  For $k \in [1, q]$: $H_k := f_w(\mathcal{B}_k) \in \mathbb{R}^{N_k \times d}$;

  For $k \in [1, q]$: $G^k_{pos} := H_k H_k^\top \in \mathbb{R}^{N_k \times N_k}$;

  For $k \in [1, q]$: $G^k_{neg} := \underset{j \neq k, 1 \leq j \leq q}{\sqsubset} H_k H_j^\top \in \mathbb{R}^{N_k \times \overline{N_k}}$; // where $\overline{N_k} = \sum_{j \neq k, 1 \leq j \leq q} N_j$

$$\text{SupeLoss} = -\frac{1}{q} \sum_{k=1}^{q} \frac{1}{N_k} \sum_{i=1}^{N_k} \log \left[\frac{\frac{1}{N_k - 1} \sum_{p=1, p \neq i}^{N_k} \exp\left(G^k_{pos}(i,p)/\tau\right)}{\frac{1}{N_k - 1} \sum_{p=1, p \neq i}^{N_k} \exp\left(G^k_{pos}(i,p)/\tau\right) + \frac{1}{\overline{N_k}} \sum_{n=1}^{\overline{N_k}} \exp\left(G^k_{neg}(i,n)/\tau\right)} \right] \tag{3}$$

  ;

  Back-propagate to Update the Weights;

return f_w

Dataset	#Train	#Dev	#Test	#Classes
SST-2	60k	3.5k	3.5k	2
Yelp-P	600k	50k	50k	2
Yelp-F	600	50k	50k	5
Amazon-P	3M	600k	400k	2
Amazon-F	2.5M	500k	650k	5
IMDB	20k	5k	25k	2
MSAC	1.6k	0.2k	0.2k	2
ASAC	6.8k	0.8k	0.8k	2

TABLE 1: Statistics of datasets used for evaluation.

due to the limited number of examples that we have in these datasets. We use a hidden units of 128 neurons, and a batch size of 200. We apply dropout with probability 0.2. Similarly, the CE is trained for a batch size of up to 400, but the best results are obtained using a batch size of 64. For the remaining datsets, SuperLoss is optimized for 60 epochs using Adam with a learning rate of 0.001. We initialize the input layer of the encoder with Glove pre-trained word representations of size 300 [54]. We use an encoder function of 3 hidden layers of 512 units each, and a batch size of 800. We apply dropout with probability 0.3 on each layer. Note that the CE loss is evaluated by increasing the mini-batch size up to 1000. We run all experiments on 1 GPU server.
Following common protocol, to test our method , 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 multiclass sentence classification task. We report the obtained results of the linear classifiers on top of the learnt representations.

C. CLASSIFICATION ACCURACY

Here, we report the obtained results using SuperLoss on 8 datasets, and those obtained by previous methods which are those based on the CE, Triplet-loss, N-pair-loss, as well as SupCon [29]. The results are given in terms of the accuracy score measured on the same balanced test set.

Following common practice in contrastive learning, we first study the importance of adding a projection head that maps representations to new space where the supervised contrastive loss is applied. Similar to [11], [44], we tested three different MLP architectures: (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 was found in previous works, we observe that a non-linear architecture is better than both the linear and the identity functions for the projection head network (See table 2). Note that the projection head network is used only in the contrastive training phase; it is discarded in the fine-tuning and inference phases.

For the evaluation performance, we evaluated our approach for transfer learning in two different settings: (1) the classifier is trained on top of the frozen representation (transfer learning); (2) we train the classifier (projection head), where we allow all weights to be adjusted during training (fine-tuned). Simulations showed that the learned representations by the proposed objective function are better for the downstream tasks without adjusting them which means that our framework is capable of capturing robust features that better separate the classes. table 3 illustrates the obtained results of both strategies. In this paper, we provide the results that we obtained with the transfer learning strategy.

Table 4 shows the obtained results using our objective function on the previously described datasets, as well as

Dataset \Projection model	Identity	Linear	Non-Linear
SST-2	93.40	93.60	**94.15**
Yelp-P	95.13	95.31	**95.45**
Yelp-F	63.74	64.02	**64.89**
Amazon-P	93.91	94.07	**94.71**
Amazon-F	60.01	60.13	**61.94**
IMDB	84.33	84.45	**86.82**
MSAC	76.16	76.89	**80.10**
ASAC	81.54	81.87	**82.63**

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

Dataset \learning strategy	Transfer-learning	Fine-tuning
SST-2	94.15	**94.42**
Yelp-P	**95.53**	95.45
Yelp-F	**64.89**	64.23
Amazon-P	**94.71**	94.58
Amazon-F	**61.94**	61.11
IMDB	**86.82**	85.14
MSAC	**80.10**	78.21
ASAC	**82.63**	82.16

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

those obtained with other objective functions. The results are given in terms of the accuracy score measured on the same balanced test set. It is seen that in 87% of the cases, our framework achieves better performance; the gain in performance is significant. Indeed, SuperLoss leads to, for instance, a 2.87% improvement of accuracy on SST-2, 3.33% improvement on Yelp-P, 3.34% improvement on Amazon-F, 3.93% improvement on ASAC,and 7.59% improvement on MSAC compared to CE loss, respectively. The large performance gap for MSAC dataset demonstrates that cross-entropy struggles with separating the classes when dealing with small datasets. Furthermore, the results for MSAC and ASAC prove that our framework is very promising for under-resourced languages. Moreover, our experiments showed that CE overfits the MSAC dataset very quickly, with a training accuracy of 96.33% and only 72% accuracy on test. The overfitting problem cannot be explained by the large number of parameters of the model, since our objective function also uses the same model architecture (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.

V. ABLATION STUDY

We investigate here the effects of different parameters on performance. All experiments have been conducted using Yelp-F dataset. We run each experiment with 10 different seeds, and report the average test accuracy.

Loss	Cross-Entropy	Triplet-loss	N-pair-loss	SupCon	SuperLoss
SST-2	91.28	90.68	92.39	93.53	**94.15**
Yelp-P	92.12	92.48	92.87	94.84	**95.45**
Yelp-F	63.64	62.51	63.92	**65.10**	64.89
Amazon-P	92.94	93.11	93.66	93.98	**94.71**
Amazon-F	58.40	57.32	58.69	60.84	**61.94**
IMDB	84.60	85.93	86.08	86.13	**86.82**
MSAC	72.51	70.09	74.50	78.33	**80.10**
ASAC	78.70	78.76	79.73	82.11	**82.63**

TABLE 4: Performance Results (%)

A. IMPACT OF THE MODEL ARCHITECTURE

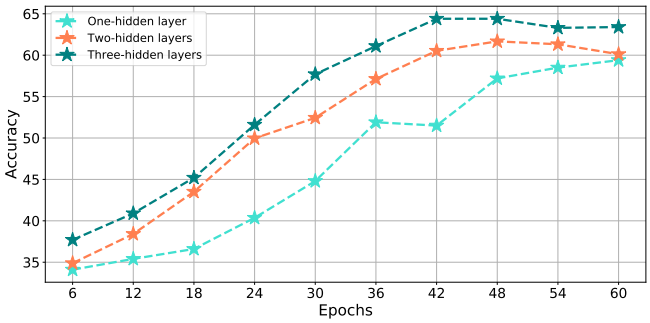
We consider several BiLSTM encoder-based architectures with growing capacity, particularly the number of hidden layers and the number of hidden units of the model. In Figure 4a depicts how changing the model's architecture affects the quality of the learnt representations on the downstream task. We found that increasing the number of hidden layers works better for the proposed framework. In this work, we used three layers due to the GPU memory constraint. Our experiments show that SuperLoss surpasses the cross-entropy loss using the same model architecture for all configurations. Figure 4b shows the obtained accuracy score for different numbers of hidden units ($\{100, 200, 300, 512, 768\}$). It is clear that by increasing the dimension of the hidden layers the model works better, though the gain beyond 512 is small.

B. TRAINING WITH LARGE BATCH SIZE

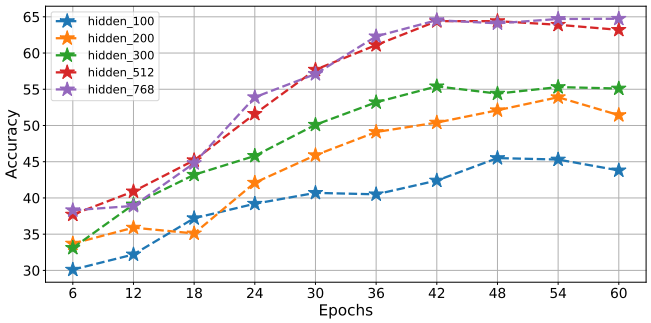
Here, we show empirically the impact of batch size on the quality of the models' representations trained with the same number of epochs (60 epochs). Figure 4c shows the accuracy of a linear classifier trained upon the learned 512-dimensional representations while varying the batch size. Similar to self-supervised contrastive learning, we found that training the model on larger batch sizes have a significant (high-quality representation) advantage over the smaller ones. Note that with the CE loss, the highest scores are obtained for a batch size of 500; larger batch sizes decreased the accuracy of the downstream classification task. In contrast, for our objective function, larger batch sizes provides more negative examples, thus improving the results. In this ablation study, we evaluated the model's representations with batch sizes of {500, 750, 1000}. However, we believe that by increasing the batch size further, the model can learn higher quality features that can be useful on downstream task.

C. IMPACT OF THE TEMPERATURE

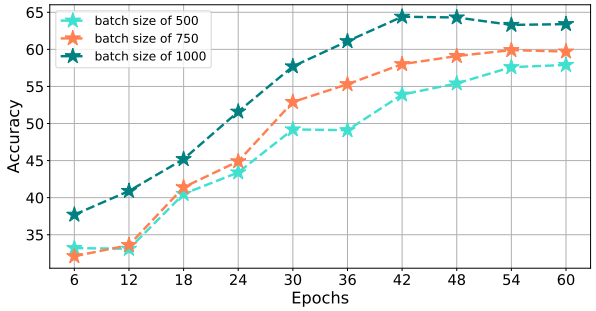
Figure 4d shows the impact of scalar temperature parameter on Top-1 accuracy performance of our framework. Empirical observations show that smaller temperature benefits training more than higher ones (lower temperature increases the influence of examples that are harder to separate). However, very low temperatures are harder to train due to numerical instability. Thus, an appropriate temperature can help the model learn from hard negatives. The empirical behavior of the effect of the temperature parameter is in line with the observations made in previous work related to self-supervised/fully-supervised contrastive learning.



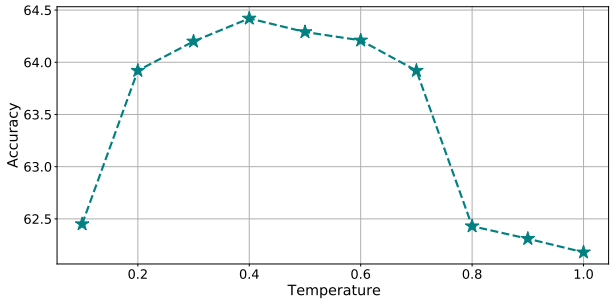
(a) Linear evaluation for models trained with different choices of number of hidden layers and epochs.



(b) Linear evaluation for models trained with different choices of number of units in each layer and epochs.



(c) Linear evaluation for models trained with different choices of batch size.



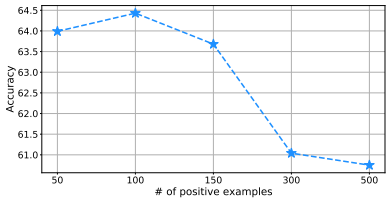
(d) Linear evaluation for models trained with different temperature value τ.

D. EFFECT OF THE NUMBER OF POSITIVE/NEGATIVES EXAMPLES

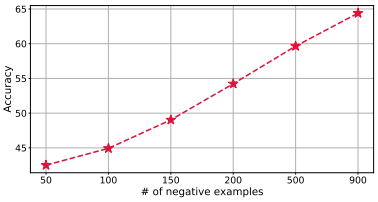
In recent contrastive learning approaches, the number of negative examples has been shown to be a key component for learning high-quality representations. The majority of these methods sample negatives from a very large batches or a memory bank to increase the number of negative examples beyond the batch size and have reported significant performance gains with increasing batch sizes. In this research, we study the impact of the number of the negative/positive examples. To the best of our knowledge, this work is the first to consider the effect of the number of positive and negative examples on the fully-supervised contrastive learning setting. Similar to self-supervised contrastive learning, we found that increasing the number of negative examples is beneficial for learning representations in our framework. Our experiments show that a high number of negatives helps our loss function to encourage the encoder to find features that can better separate the representations of different classes in the latent space. In this paper, we use 900 (see figure 5c) negative examples for each class due to the GPU memory constraint. However, we believe that increasing further the number of negative examples will produce better results. In figure 5a, we report the top-1 accuracy performance of the downstream classification task for different values of the number of positive examples. In each simulation, we fixed the number of positive instances for a given class and the negative examples are uniformly sampled from the remaining classes (e.g 300 positive points and 175 negative points for each remaining classes). We observe that training our supervised objective function with a high number of positive examples leads to good representations. However, simulations show that at beyond a certain threshold, increasing the number of positives decreases the accuracy of the downstream task.

VI. NEGATIVE SAMPLING STRATEGY

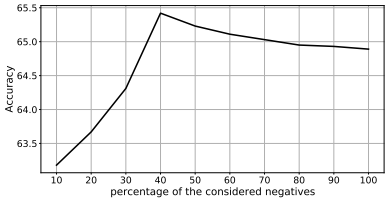
Contrastive learning is recently proposed to learn feature embeddings in a self-supervised manner. The latter relies on the positive and negative instances. As revealed by recent studies, negative examples are crucial in learning robust representations. Accordingly, different strategies have been proposed to sample negatives that are hard to distinguish for a given anchor in the latent space [11], [13], [14], [25]. In this paper, we propose a simple yet effective strategy for selecting hard negative examples for supervised contrastive learning for text representations. In the proposed framework, in each iteration, we maximize the distance of the average similarity of a given anchor with all instances from the remaining classes which means that all negative examples are considered as hard negative. To overcome this, we first train the model for a number of epochs (20 epochs on YELP-F dataset) using all negative instances within the batches, then we modify the training strategy by maximizing the distance of anchors with those pairs that have a similarity higher than a fine-tuned threshold. By doing so, our loss will guide the encoder function to produce representations by considering



(a) The effect of the number of positive examples.



(b) The effect of the number of negative examples



(c) Accuracy scores with different value of the threshold hyper-parameter

Dataset / Loss	SuperLoss	SuperLoss*
SST-2	94.15	**94.62**
Yelp-P	95.45	**95.89**
Yelp-F	94.71	**94.93**
Amazon-p	61.94	**62.76**
Amazon-F	61.94	**62.88**
IMDB	86.82	**86.94**
MSAC	80.10	**81.32**
ASAC	82.63	**83.45**

TABLE 5: Performance Results (%)

only truly hard negatives. In our strategy, we simply select these examples by first ordering the similarities for a given an anchor, then we select the most similar instances (hard negatives). Following this new strategy, the simulations show that, indeed, the distances between anchors and the negative examples become higher compared to those obtained using the previous learning strategy. We also noticed that for a given anchor, the selected hard negative examples are those from the closest class (for class 'Very Positive' the majority of the negative examples are from the class 'Positive' which is the most similar class to that of the anchor). We fine-tune the similarity threshold using the validation set by selecting the top most similar examples to the anchor. Figure 5c shows the different top-1 accuracy obtained as a function of similarity threshold. We evaluate our negative sampling strategy on several benchmark datasets. Experimental results show that the strategy is beneficial.

In table 5, we report the SuperLoss* which refers to the results obtained using the proposed negative sampling strategy. In bold, we report the best performance. As it can be seen in the table, $SuperLoss^*$ outperforms $SuperLoss$ in most cases which means that the proposed strategy leads the encoder to learn better representations (relevant features for distinguishing the classes in the latent space).

VII. CONCLUSION

We proposed SuperConText, a new framework for learning text representations using a novel supervised contrastive loss. SuperConText encourages an encoder function to learn representations by maximizing the average agreement between the representation of an anchor and those of N positive pairs, determined as elements belonging to the same class as the anchor, while distancing the anchor's representation with those of negative examples. Simulations show that the proposed framework outperforms several methods based on other objective functions on various benchmark datasets. We have conducted a number of experiments to understand the effects of both negative and positive examples on the quality of learned representations. We further introduced a simple yet effective negative sampling strategy to enhance the quality of the representations. The experimental results show that the proposed strategy improves performance in most cases.

REFERENCES

[1] He, 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**, 770–778.
[2] Carlini, Nicholas and Wagner, David. Towards evaluating the robustness of neural networks. 2017 ieee symposium on security and privacy (sp) **2017**, 39–57.
[3] Yu, Dong and Deng, Lin. Automatic speech recognition. Springer **2016**, 1.
[4] Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 **2014**.
[5] 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**.
[6] 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**, ,2818–2826.
[7] Müller, Rafael and Kornblith, Simon and Hinton, Geoffrey. When does label smoothing help?. arXiv preprint arXiv:1906.02629 **2019**.
[8] Müller, Rafael and Kornblith, Simon and Hinton, Geoffrey. When does label smoothing help?. arXiv preprint arXiv:1906.02629 **2019**.
[9] Zhang, Hongyi and Cisse, Moustapha and Dauphin, Yann N and Lopez-Paz, David. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 **2017**.
[10] Hinton, Geoffrey and Vinyals, Oriol and Dean, Jeff. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2015**.
[11] 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**, , 1597–1607.
[12] Weng, Bowen and Xiong, Huaqing and Zhao, Lin and Liang, Yingbin and Zhang, Wei. Momentum Q-learning with finite-sample convergence guarantee. arXiv preprint arXiv:2007.15418 **arXiv preprint arXiv:2007.15418**.

[13] Kalantidis, Yannis and Sariyildiz, Mert Bulent and Pion, Noe and Weinzaepfel, Philippe and Larlus, Diane. Hard negative mixing for contrastive learning. Advances in Neural Information Processing Systems **2020**, 33, 21798–21809.

[14] Wu, Mike and Mosse, Milan and Zhuang, Chengxu and Yamins, Daniel and Goodman, Noah. Conditional negative sampling for contrastive learning of visual representation. rXiv preprint arXiv:2010.02037 **2020**.

[15] Andreieva, Valeria and Shvai, Nadiya. Generalization of cross-entropy loss function for image classification. Proceedings of the IEEE conference on computer vision and pattern recognition **2020**.

[16] Pang, Tianyu and Xu, Kun and Dong, Yinpeng and Du, Chao and Chen, Ning and Zhu, Jun. Rethinking softmax cross-entropy loss for adversarial robustness. arXiv preprint arXiv:1905.10626 **2019**.

[17] Sukhbaatar, Sainbayar and Bruna, Joan and Paluri, Manohar and Bourdev, Lubomir and Fergus, Rob. Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080 **2014**.

[18] Zhang, Zhilu and Sabuncu, Mert. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **2018**.

[19] Zhang, Zhilu and Sabuncu, Mert. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systemsn **2018**, 31.

[20] Gutmann, Michael and Hyvärinen, Aapoan. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Proceedings of the thirteenth international conference on artificial intelligence and statistics **2010**, 297–304.

[21] Dwibedi, Debidatta and Aytar, Yusuf and Tompson, Jonathan and Sermanet, Pierre and Zisserman, Andrew. With a little help from my friends: Nearest-neighbor contrastive learning of visual representation. Proceedings of the IEEE/CVF International Conference on Computer Vision **2021**, 9588–9597.

[22] Hafidi, Hakim and Ghogho, Mounir and Ciblat, Philippe and Swami, Ananthram. Negative sampling strategies for contrastive self-supervised learning of graph representations. Signal Processing **2016**, 190, 108310.

[23] Robinson, Joshua and Chuang, Ching-Yao and Sra, Suvrit and Jegelka, Stefanie. Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 **2020**.

[24] Grill, Jean-Bastien and Strub, Florian and Altché, Florent and Tallec, Corentin and Richemond, Pierre and Buchatskaya, Elena and Doersch, Carl and Avila Pires, Bernardo and Guo, Zhaohan and Gheshlaghi Azar, Mohammad and others. Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **2020**, 33, 21271–21284.

[25] He, Kaiming and Fan, Haoqi and Wu, Yuxin and Xie, Saining and Girshick, Ross. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition **2020**, 9729–97384.

[26] Caron, Mathilde and Misra, Ishan and Mairal, Julien and Goyal, Priya and Bojanowski, Piotr and Joulin, Armand. Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **2020**, 33, 9912–9924.

[27] Goyal, Priya and Caron, Mathilde and Lefaudeux, Benjamin and Xu, Min and Wang, Pengchao and Pai, Vivek and Singh, Mannat and Liptchinsky, Vitaliy and Misra, Ishan and Joulin, Armand and othersn. Self-supervised pretraining of visual features in the wild. arXiv preprint arXiv:2103.01988 **2021**.

[28] Zbontar, Jure and Jing, Li and Misra, Ishan and LeCun, Yann and Deny, Stéphane. Barlow twins: Self-supervised learning via redundancy reduction. International Conference on Machine Learning **2021**, 12310–12320.

[29] 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. Advances in Neural Information Processing Systems **2020**, 33, 18661–18673.

[30] Elsayed, Gamaleldin and Krishnan, Dilip and Mobahi, Hossein and Regan, Kevin and Bengio, Samy. Large margin deep networks for classification. in neural information processing systems **2018**.

[31] Unsupervised feature learning via non-parametric instance discrimination. Proceedings of the IEEE conference on computer vision and pattern recognition **2018**, 3733–3742.

[32] Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 **2014**, 3733–3742.

[33] You, Yuning and Chen, Tianlong and Sui, Yongduo and Chen, Ting and Wang, Zhangyang and Shen, Yang, Meng. Large-margin softmax loss for convolutional neural networks. ICML **2016**, 2, 7.

[34] Liu, Weiyang and Wen, Yandong and Yu, Zhiding and Yang, Meng. Graph contrastive learning with augmentations. Advances in Neural Information Processing Systems **2020**.

[35] Oord, Aaron van den and Li, Yazhe and Vinyals, Orio. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 **2018**

[36] Tian, Yonglong and Krishnan, Dilip and Isola, Phillip. Contrastive multiview coding. European conference on computer vision **2020**

[37] Elsayed, Gamaleldin and Krishnan, Dilip and Mobahi, Hossein and Regan, Kevin and Bengio, Samy. Large margin deep networks for classification. Advances in neural information processing systems **2018**

[38] Jing, Li and Vincent, Pascal and LeCun, Yann and Tian, Yuandong. Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint arXiv:2110.09348 **2021**.

[39] Yan, Yuanmeng and Li, Rumei and Wang, Sirui and Zhang, Fuzheng and Wu, Wei and Xu, Weiran. Consert: A contrastive framework for self-supervised sentence representation transfer. arXiv preprint arXiv:2105.11741 **2021**.

[40] Xie, Yaochen and Xu, Zhao and Zhang, Jingtun and Wang, Zhengyang and Ji, Shuiwang. Self-supervised learning of graph neural networks: A unified review. IEEE Transactions on Pattern Analysis and Machine Intelligence **2022**.

[41] schannen, Michael and Djolonga, Josip and Rubenstein, Paul K and Gelly, Sylvain and Lucic, Mario. On mutual information maximization for representation learning. arXiv preprint arXiv:1907.13625 **2019**.

[42] Nar, Kamil and Ocal, Orhan and Sastry, S Shankar and Ramchandran, Kannan. Cross-entropy loss and low-rank features have responsibility for adversarial examples. arXiv preprint arXiv:1901.08360 **2019**.

[43] 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**.

[44] Moukafih, Youness and Ghanem, Abdelghani and Abidi, Karima and Sbihi, Nada and Ghogho, Mounir and Smaïli, Kamel. SimSCL: A Simple fully-Supervised Contrastive Learning Framework for Text Representation. Australasian Joint Conference on Artificial Intelligence **2022**.

[45] Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li. Imagenet: A large-scale hierarchical image databas. 2009 IEEE conference on computer vision and pattern recognitio **2009**

[46] DWu, Chao-Yuan and Manmatha, R and Smola, Alexander J and Krahenbuhl, Philipp. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 **2018**

[47] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. Sampling matters in deep embedding learning. Proceedings of the IEEE international conference on computer vision **2017**

[48] Yuan, Yuhui and Yang, Kuiyuan and Zhang, Chao. Sampling matters in deep embedding learning. Proceedings of the IEEE international conference on computer vision **2017**

[49] Gutmann, Michael and Hyvärinen, Aapo. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. Proceedings of the thirteenth international conference on artificial intelligence and statistics **2010**

[50] Henaff, Olivier. Data-efficient image recognition with contrastive predictive coding. International conference on machine learning **2020**

[51] Kaku, Aakash and Upadhya, Sahana and Razavian, Narges. Intermediate Layers Matter in Momentum Contrastive Self Supervised Learning. Advances in Neural Information Processing Systems **2021**

[52] Haidar, Md Akmal and Rezagholizadeh, Mehdi and Ghaddar, Abbas and Bibi, Khalil and Langlais, Philippe and Poupart, Pascal. CILDA: Contrastive Data Augmentation using Intermediate Layer Knowledge Distillation. arXiv preprint arXiv:2204.07674 **2022**

[53] Cao, Yun-Hao and Wu, Jianxin. Rethinking self-supervised learning: Small is beautiful. arXiv preprint arXiv:2103.13559 **2021**

[54] 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 (EMNLP) **2014**

[55] 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**

[56] Zhao, Xiangyun and Vemulapalli, Raviteja and Mansfield, Philip Andrew and Gong, Boqing and Green, Bradley and Shapira, Lior and Wu, Ying. Contrastive learning for label efficient semantic segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision **2021**

YOUNESS MOUKAFIH has received the master's degree in big data analytics and smart systems from Sidi Mohammed Ben Abdellah University, Fez, Morocco, in 2018. He is currently pursuing the joint Ph.D. degree with University of Lorraine and International University of Rabat. His research interests include machine learning, natural language processing, computer vision and graph representation learning.

NADA SBIHI is an Assistant Professor at ESIN engineer school and member of TICLab at International Univesity of Rabat. She obtained her engineering degree from the National Institute of Posts and Telecommunications (INPT) in 2006, in computer science, networks and systems. After three years of experience as a Business Intelligence engineer at Sofrecom, she continued her studies at Pierre and Marie Currie University (Paris 6), where she obtained a master's degree and then a doctorate. His doctoral studies were carried out at INRIA under the supervision of Dr James ROBERTS. Before joining the UIR faculty, she worked at TICLab as a post-doctoral fellow under the supervision of Professor Mounir GHOGHO. His research themes are : Big Data and artificial intelligence and their applications in various fields: intelligent transport, social networks, content-oriented networks, urban pollution and health.

MOUNIR GHOGHO (Fellow, IEEE) has received the M.Sc. degree in 1993 and the PhD degree in 1997 from the National Polytechnic Institute of Toulouse, France. He was an EPSRC Research Fellow with the University of Strathclyde (Scotland), from Sept 1997 to Nov 2001. In Dec 2001, he joined the school of Electronic and Electrical Engineering at the University of Leeds (England), where he was promoted to full Professor in 2008. While still affiliated with the University of Leeds, in 2010 he joined the International University of Rabat (Morocco) where he is currently Dean of the College of Doctoral Studies and Director of TICLab (ICT Research Laboratory). He is a Fellow of IEEE, a recipient of the 2013 IBM Faculty Award, and a recipient of the 2000 UK Royal Academy of Engineering Research Fellowship. He is the co-founder and co-director of the CNRS-associated International Research lab DataNet, in the field of Big Data and artificial intelligence. His research interests are in Machine Learning, Signal Processing and Wireless Communication, on which he has published over 300 papers in journals and conferences. He has coordinated around 20 research projects and supervised over 30 PhD students in the UK and Morocco. In the past, he served as an associate editor of many journals including the IEEE Signal Processing Magazine and the IEEE Transactions on Signal Processing.

KAMEL SMAILI is Professor of Computer Science at University of Lorraine. He is an expert in natural language processing. He carries out his research activity at LORIA, a UL laboratory. He obtained an engineering degree in computer science from the University (USTHB), Algiers, Algeria in 1986, a doctorate from the University Henri Poincaré in 1991 and an HDR from the University of Nancy 2 in 2001. Between 1999 and 2001, he was seconded to the CNRS. In 2014, he created the SMarT research team (https://smart.LORIA.fr/) at LORIA. The main activity of his team is processing natural language using statistical and neural paradigms and more generally the development of deep machine learning methods for NLP. He has supervised 19 doctoral students, 17 of whom have defended and two are at the end of their thesis. Every year, he reviews ANR and foreign project files (NeuroInsight 2022). Since 2015, he co-directs an associated international laboratory (LIA CNRS) named DataNet. In this laboratory, are involved two laboratories of the University of Lorraine and several Moroccan universities including the UIR. He has published more than 150 articles in international journals and conferences[1]. He coordinated several international projects, the last one ended in November 2019, the Chist-Era AMIS (Access Multilingual Information opinions) project. He leads a new project TRADEF (ASTRID ANR) that started in January 2023. He has organized and chaired international conferences (ICNLSSP 2017 and ICALP 2019) and he is a member of several international conference program committees such as: ICASSP, Interspeech, LREC, etc. He has been invited to give lectures in Japan, France, Spain, Poland, Algeria, Tunisia and Morocco. He has been a member of more than 70 thesis and HDR juries in France, Germany, Spain, India, Ireland, Algeria, Morocco and Tunisia.

[1] https://members.LORIA.fr/KSmaili/publications/