Advancing Semantic Textual Similarity Modeling: A Regression Framework with Translated ReLU and Smooth K2 Loss

Anonymous ACL submission

Abstract

Since the introduction of BERT and RoBERTa, research on Semantic Textual Similarity (STS) has made groundbreaking progress. Particularly, the adoption of contrastive learning has substantially elevated state-of-the-art performance across various STS benchmarks. However, contrastive learning categorizes text pairs as either semantically similar or dissimilar, failing to leverage fine-grained annotated information and necessitating large batch sizes to prevent model collapse. These constraints pose challenges for researchers engaged in STS tasks that require nuanced similarity levels or those with limited computational resources, compelling them to explore alternatives like Sentence-BERT. Nonetheless, Sentence-BERT tackles STS tasks from a classification perspective, overlooking the progressive nature of semantic relationships, which results in suboptimal performance. To bridge this gap, this paper presents an innovative regression framework and proposes two simple yet effective loss functions: Translated ReLU and Smooth K2 Loss. Experimental analyses demonstrate that our method achieves convincing performance across seven established STS benchmarks, especially when supplemented with task-specific training data. [1]

1 Introduction

Semantic Textual Similarity (STS) constitutes a fundamental task in natural language processing, wielding significant influence across a multitude of applications including text clustering, information retrieval, and recommendation systems. Despite the remarkable precision achieved by interactive architectures within these tasks, their inability to support offline computation limits their viability in large-scale text analysis scenarios. In response to this, the seminal work of Sentence-BERT (Reimers and Gurevych, 2019) introduces a dual-tower architecture to encode the sentences within a pair separately, thereby facilitating the derivation of independent embeddings. This approach showcases superior efficacy and has rapidly gained widespread acceptance, now serving as a cornerstone for various downstream tasks. Consequently, further improvements to Sentence-BERT hold high research and practical value.

Nevertheless, the advent of contrastive learning methods, exemplified by SimCSE (Gao et al., 2021), has demonstrated more pronounced enhancements across renowned English STS benchmarks like STS12-16 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS-B (Cer et al., 2017), and SICK-R (Marelli et al., 2014). This has shifted the research focus in recent years towards integrating contrastive learning techniques with pre-trained language models (PLMs) like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). An intuitive comparison is that using the NLI dataset (Bowman et al., 2015; Williams et al., 2018) as a training corpus, SimCSE-RoBERTa$_{\text{base}}$ attains an average Spearman's correlation score of 82.52 across these STS tasks, hugely surpassing the 74.21 achieved by Sentence-RoBERTa$_{\text{base}}$.

Such discernible performance disparity has inadvertently overshadowed the advantages of Sentence-BERT, especially in terms of data utilization efficiency and computational resource demands. Contrastive learning, by its self-supervised nature, predominantly recognizes text pairs as either similar or dissimilar. This binary categorization restricts contrastive learning methods to using triple form data composed of an anchor sentence, a positive instance, and a hard negative instance for training in supervised settings (Gao et al., 2021). Many practical scenarios, however, tend to provide more finely-grained labeled data (e.g., highly relevant, moderately relevant, relevant, not relevant) (Liu et al., 2023), where contrastive learning ap-

[1] Our code and checkpoints are available at https://anonymous.4open.science/r/STS-Regression.

proaches can usually only exploit text pairs whose similarity indicators are at the endpoints.

Additionally, since contrastive learning enhances model discriminability by treating other samples within the same batch as negative instances, it requires large batch sizes, thereby consuming substantial computational resources. For example, SimCSE's supervised learning settings include a batch size of 512 and 3 epochs. To accommodate this configuration on consumer-grade GPUs, SimCSE constrains the maximum input length to 32 (Gao et al., 2021). In contrast, Sentence-BERT and our proposed methodology necessitate a mere batch size of 16 and 1 epoch to reach convergence. Additionally, our default maximum input length is 256, significantly longer than SimCSE's.

The aforementioned drawbacks highlight the difficulty in completely replacing Sentence-BERT with contrastive learning methods. Hence, some cutting-edge works (Zhang et al., 2023b) continue to employ Sentence-BERT for sentence embedding derivation. Nonetheless, given that STS tasks typically categorize text pairs by degrees of semantic similarity, and Sentence-BERT approaches these tasks from a classification standpoint, neglecting the progressive relationships between categories, there exists a clear opportunity for improvement. As a illustration, consider an STS task with five categories, labeled consecutively from 1 to 5. Traditional classification strategies would yield identical loss for a sample scored at 2, irrespective of its prediction as 3 or 4, an approach evidently suboptimal.

To rectify such deficiency, this paper proposes a novel framework that converts multi-category STS tasks into regression problems, thus effectively capturing the progressive relationships between categories. For a given dataset, we first map its original labels to a sequential array of integers, ensuring that samples with higher similarity scores are assigned correspondingly greater integers. Then, we set the number of nodes in the output layer to one, thereby enabling the model to produce a continuous prediction value. Finally, the model parameters are updated according to the difference between predicted and actual scores.

Distinct from standard regression problems, the ground truth within our transformed multi-category STS tasks manifest as a series of discrete points along the numerical axis. Therefore, instead of precisely matching the target points, the floating-point predictive values just need to be sufficiently close to get correctly classified. To accommodate this process, we introduce a zero-gradient buffer zone to widely utilized L1 Loss and MSE Loss, unveiling two innovative loss functions: Translated ReLU and Smooth K2 Loss.

Comprehensive evaluations across seven STS benchmarks substantiate that our regression framework surpasses traditional classification strategies in handling multi-category STS tasks. Additionally, we find that further updating our model's checkpoint with the STS-B and SICK-R training sets allows our method to achieve superior Spearman correlation relative to contrastive learning methods, reaching state-of-the-art performance. These findings reinforce the effectiveness of our proposed solution and the importance of utilizing task-specific data, an aspect often neglected in contrastive learning paradigms.

The main contributions of this study are outlined as follows:

- Building upon the foundation of Sentence-BERT, we develop a regression framework adept at modeling the progressive relationships between categories in multi-class STS tasks. This not only enhances performance but also, due to regression's intrinsic properties, simplifies the prediction process for K-category problems to require only a single output node, significantly minimizing the model's output layer parameter count.

- We propose two innovated loss functions, Translated ReLU and Smooth K2 Loss, specifically tailored to address classification problems involving progressive relationships between categories.

- Through empirical evidence, we demonstrate that, when combined with task-specific data, our Siamese network approach can attain better results than contrastive learning schemes.

2 Related Work

In this chapter, we primarily review two types of STS task solutions directly related to our work:

Siamese Neural Network Architectures: These approaches (Reimers and Gurevych, 2019; Conneau et al., 2017; Thakur et al., 2021), proposed relatively earlier in the field, have been widely applied across various domains owing to their effectiveness on annotated data. Although their performance on the seven STS benchmarks

(STS 12-16, STS-B, SICK-R) is generally inferior to contemporary contrastive learning methods, this discrepancy largely arises from the absence of task-specific training data. Thus, models have the flexibility to opt for alternative sources, such as wiki datasets (Gao et al., 2021) or NLI datasets (Bowman et al., 2015; Williams et al., 2018), which adapt readily to triplet format. Given our goal of tackling multi-category STS tasks, our model architecture remains rooted in the Siamese network. However, in contrast to preceding efforts, we introduce an innovative regression framework designed to explicitly capture the progressive relationships between categories.

Contrastive Learning Methods: Contrastive learning has become the dominant paradigm for addressing STS tasks, characterized by a vast amount of research (Jiang et al., 2022; Zhang et al., 2023a). However, contrastive learning loss functions, epitomized by InfoNCE Loss (Oord et al., 2018), concentrate solely on the binary semantic categorization of texts and cannot directly utilize fine-grained labeled corpus. Furthermore, the necessity for large batch sizes to ensure negative sample diversity and prevent model collapse imposes considerable computational demands. For instance, supervised SimCSE's training requires 58GB of GPU memory (Jiang et al., 2023), whereas our proposed method, even with a maximum sequence length eight times that of SimCSE, demands merely 42GB.

3 Methodology

This chapter delineates our methodological framework, beginning with a detailed exposition of the designed network architecture and its operational workflow in Section 3.1. Then, in Sections 3.2 and 3.3, we present the two novel loss functions proposed in this study.

3.1 Network Architecture

As illustrated in Figure 1, we utilize a Siamese neural network with shared parameters for encoding input sentences via BERT to obtain corresponding word embedding matrices. Subsequently, sentence embeddings, denoted as u and v for paired sentences A and B, are derived through average pooling. These embeddings, both vectors of the hidden dimension, are then concatenated alongside their element-wise difference $|u - v|$ and passed through a fully connected layer with parameters sized at $3 \times$ hidden_dimension to produce the model's predicted continuous similarity score.

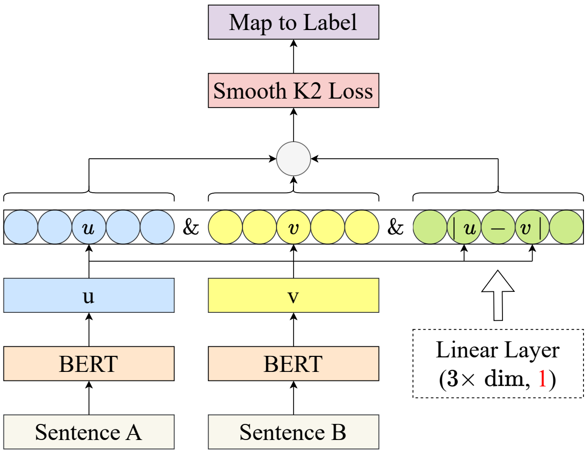

Figure 1: Our Regression Framework. Here, the two BERT models share same parameters, with "dim" representing the embedding dimensions of u and v.

Our methodology diverges from the original dual-tower structures employed by Sentence-BERT and InferSent (Conneau et al., 2017) in three critical aspects:

1. We model STS tasks, characterized by a progressive relationship between categories, as regression problems. This is achieved by mapping labels from the original dataset to a sequence of incrementing integers reflective of their similarity relations, thus conveying to the model that categories are not independent but progressively related.

2. Building on this, we streamline the output node count in our final fully connected layer to one, enabling the model to directly yield a similarity score rather than a categorical probability distribution. Through this adjustment, for STS tasks containing K categories, we effectively reduce the parameter size of the output layer from $3 \times$ hidden_dimension $\times K$ to $3 \times$ hidden_dimension $\times$ 1. In light of the expanding dimensions of hidden layers in contemporary PLMs, this optimization can save substantial computational resources.

3. Contrasting with InferSent and Sentence-BERT's classification-based approach, which assigns target classes for sentence pairs based on the highest probability, our regression framework categorizes based on the closeness between the predicted and actual values.

To better understand this process, consider an STS task with four categories, labeled as "very relevant," "moderately relevant," "slightly relevant," and "not relevant." After clarifying the progressive relationship between these categories, we would map them to four consecutive integers 0, 1, 2, 3, respectively, ranging from "not relevant" to "very relevant." This mapping strategy is highly flexible, allowing for task-specific adjustments in numerical nodes and intervals. Subsequently, we encode the paired sentences separately and calculate their semantic similarity, resulting in a floating-point prediction value. By rounding this value, it can be converted into a discrete label. For instance, a prediction of 2.875 for a sample pair would be classified as "very relevant," as it approximates closely to the boundary point 3. Similarly, if a sample's prediction is 1.333, it would be approximated to 1 and thus classified as "slightly relevant" because 1.333 is closer to 1 among the four boundary points 0, 1, 2, 3.

Extending from the above examples, it can be seen that if we map the original labels to natural numbers spaced by d, as long as the difference between the model's prediction and the ground truth is less than $\frac{d}{2}$, the sample will be correctly classified. However, conventional regression loss functions, represented by L1 Loss and MSE Loss, always enforce the difference between the model's prediction and the true value to be zero—a requirement that is unnecessary for our scenario. Thus, we introduce a zero-gradient buffer zone into both functions, resulting in the creation of Translated ReLU and Smooth K2 Loss.

3.2 Translated ReLU

We first present Translated ReLU, mathematically formulated in Equation 1. Herein, d represents the interval between mapped category labels, with $d = 1$ for a sequence of consecutive natural numbers.

$$x \rightarrow \text{abs(prediction - label)} \geq 0$$
$$f(x) = \begin{cases} 0 & x < x_0 \leq \frac{d}{2} \\ k(x - x_0) & x_0 \leq x \end{cases} \tag{1}$$
$$f(x) = \max\big(0, k(x - x_0)\big)$$

As previously discussed, when the difference between the model's predicted value and the ground truth is less than $\frac{d}{2}$, it signifies a correct classification of the sample. Traditional regression loss functions, however, mandate absolute congruence between predictions and true values, applying a penalty for any deviation. This stringent requirement to some extent diverts the model's focus from difficult samples that have not yet been correctly classified and ignores the inherent variability within classes.

To circumvent this limitation, we introduce an adjustable threshold hyperparameter x_0, and set the loss function to zero for values within $[0, x_0]$. This modification posits that a divergence less than x_0 between prediction and ground truth is deemed sufficiently precise, thus exempt from penalty or gradient update. For disparities exceeding x_0, Translated ReLU imposes a linear penalty. To maintain accurate classification, x_0 must not exceed $\frac{d}{2}$, with the interval between x_0 and $\frac{d}{2}$ acting as a margin akin to that in Hinge Loss. This margin can enhance model robustness by penalizing correctly predicted samples that lack adequate confidence. Additionally, a parameter k is specified to control the slope of the function.

The graphical depiction of Translated ReLU is exhibited on the left side of Figure 2, with parameters set to $k = 2$ and $x_0 = 0.25$. This configuration resembles the ReLU activation function, albeit with a rightward translation. Our study employs Translated ReLU as a loss function and will compare its effects with those of L1 Loss in ensuing sections to demonstrate the significance of zero-gradient buffer zone for augmenting model performance.

3.3 Smooth K2 Loss

Translated ReLU is characterized by its simplicity and efficacy. Nonetheless, we acknowledge its limitation pertaining to the abrupt lack of smoothness at the demarcation point $x = x_0$, alongside a constant gradient that fails to accommodate varying strengths of updates based on the distance between predictions and actual values. To address these concerns, we introduce another loss function termed Smooth K2 Loss to provide a smoother transition and a gradient that dynamically adjusts in accordance with the magnitude of discrepancy from the ground truth. The formulation and the derivative of Smooth K2 Loss are specified as follows:

$$x \rightarrow \text{abs(prediction - label)} \geq 0$$
$$f(x) = \begin{cases} 0 & x < x_0 \leq \frac{d}{2} \\ k(x^2 - 2x_0x + x_0^2) & x_0 \leq x \end{cases} \tag{2}$$
$$\frac{\partial f(x)}{\partial x} = \begin{cases} 0 & x < x_0 \leq \frac{d}{2} \\ 2k(x - x_0) & x_0 \leq x \end{cases}$$

Echoing the structure of Translated ReLU, Smooth K2 Loss also incorporates a zero-gradient buffer zone, but exhibits a quadratic function for $x \geq x_0$, as illustrated on the right side of Figure 2. Given the differential mathematical underpinnings

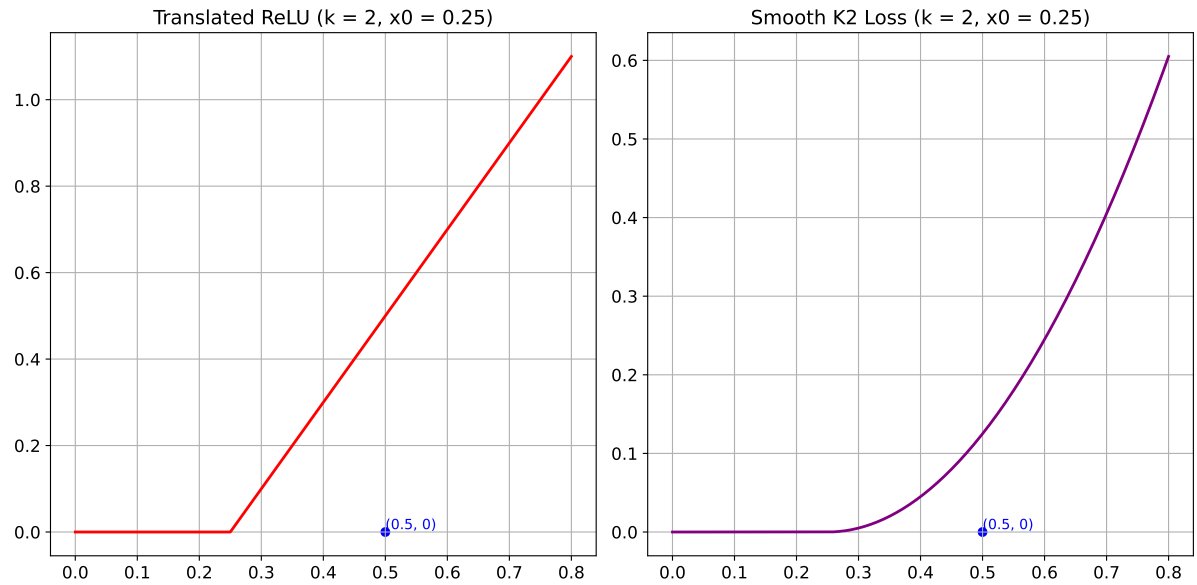

Figure 2: Comparison of Translated ReLU and Smooth K2 Loss, both with $k = 2, x_0 = 0.25$.

of these two loss functions, Smooth K2 Loss is recommended for scenarios with high-quality data and strong credibility. In contrast, when dealing with datasets that contain considerable noise, Translated ReLU may be a more suitable choice.

Additionally, prior to the application of Translated ReLU and Smooth K2 Loss, it is advisable to consider reassigning prediction values that transcend the defined category range to the nearest boundary. For instance, in a classification task where the category labels can be sequentially converted to 0, 1, 2 and 3, if the model predicts a value of 3.57 for a sample with an actual label of 3, this might be deemed acceptable and potentially obviate the need for a loss adjustment. This rationale stems from the observation that, despite the prediction's deviation exceeding $\frac{d}{2} = 0.5$, the absence of subsequent boundary points beyond 3 warrants a relaxation of this criterion.

4 Experiment

This chapter provides empirical validation of our regression framework and two innovative loss functions: Translated ReLU and Smooth K2 Loss. We commence by comparing the performance across different modeling strategies for multi-category STS tasks and various loss functions (Section 4.1). Subsequently, we demonstrate that, when supplemented with task-specific training data, our Siamese neural network architecture outperforms prevailing contrastive learning methods (Section 4.2). Following this, we examine the influence of varying hyperparameter settings on model performance (Section 4.3). Finally, we present ablation studies of our proposed methodology (Section 4.4).

4.1 STS Performance without Task-specific Training Data

Our experimental setup closely mirrors that of Sentence-BERT, leveraging fine-tuning on BERT or RoBERTa with a composite corpus derived from SNLI and MNLI datasets. These NLI datasets categorize sentence pairs into three distinct classes: contradiction, neutral, and entailment. Sentence-BERT maps these to 0, 2, 1, respectively, and employs a classification strategy for training (Reimers and Gurevych, 2019). In contrast, our method sequentially maps contradiction, neutral, and entailment to 0, 1 and 2. This mapping reflects the natural order of semantic similarity, from least to most similar, thereby enabling our regression framework to more effectively capture the progressive relationships between categories.

For computational efficiency, we uniformly set the batch size to 16 and limit training to a single epoch, with model checkpoints preserved based on performance metrics on the STS-B development set. The specific hyperparameter settings for Translated ReLU and Smooth K2 Loss are cataloged in Table 1. During the evaluation phase, we assess the model's average Spearman correlation across seven STS tasks via the SentEval (Conneau and Kiela, 2018) toolkit. The results of the aforementioned experiments are summarized in Table 2, from which we distill insights along three pivotal aspects:

PLM	Loss	k	x_0
BERT$_{\text{base}}$	Translated ReLU	2.5	0.25
BERT$_{\text{base}}$	Smooth K2 Loss	2	0.25
RoBERTa$_{\text{base}}$	Translated ReLU	1	0.25
RoBERTa$_{\text{base}}$	Smooth K2 Loss	3	0.25

Table 1: Hyperparameter configurations for employing Translated ReLU and Smooth K2 Loss across various model combinations.

1. **Classification Strategy vs. Regression Strategy:** Our regression framework, particularly when utilizing Smooth K2 Loss, yields an average Spearman correlation of 76.03 for BERT$_{\text{base}}$ and 76.04 for RoBERTa$_{\text{base}}$. These figures significantly outstrip those attained through Sentence-BERT and the classification method with Cross-Entropy Loss, highlighting the regression-based modeling's superiority in both reducing the output layer's parameter size and enhancing semantic discrimination for multi-category STS tasks.

2. **Efficacy of the Zero-Gradient Buffer Zone:** The adoption of Translated ReLU improves performance for both BERT and RoBERTa beyond what is achieved with L1 Loss. Likewise, employing Smooth K2 Loss surpasses MSE Loss on both PLMs. These comparisons underline the benefit of integrating a zero-gradient buffer zone in balancing model's focus across diverse samples within regression-modeled multi-category classification tasks.

3. **Adaptive Gradients Aligned with Prediction Errors:** Models trained with Smooth K2 Loss outshine those utilizing Translated ReLU, and models employing MSE Loss exceed those with L1 Loss. This evidences the advantages of dispensing differentiated gradients in line with prediction-ground truth deviations, especially when leveraging high-quality datasets like NLI.

Collectively, these findings substantiate the merit of adopting a regression framework for multi-category STS tasks and enhancing traditional regression loss functions with a zero-gradient buffer zone to optimize model performance.

4.2 STS Performance with Task-specific Training Data

Although the Siamese neural network, augmented by our regression framework and innovative loss functions, has exhibited significant performance enhancements, a disparity persists relative to prevailing contrastive learning methods. To bridge this gap, we exploit another critical advantage of the Siamese architecture: its capacity to fully utilize task-specific training data.

Among the seven STS benchmarks (STS12-16, STS-B, and SICK-R), STS-B and SICK-R come with their own training datasets. The STS-B training set comprises 5,749 sentence pairs with similarity scores ranging from 0 to 5, whereas the SICK-R training set includes 4,500 pairs, scored from 1 to 5. To standardize these scores, we apply a transformation $5 \times \frac{\text{label}(z)-1}{4}$ to each sample z in the SICK-R dataset. Subsequently, we concatenate these two sets and round all sample labels to integers, resulting in a task-specific training dataset containing 10,249 sentences pairs. While the sample quantity provided by this newly introduced dataset is approximately only one percent of the NLI corpus, combining them has been sufficient for us to surpass leading contrastive learning approaches.

Continuing from the checkpoint established in Section 4.1, we further fine-tune our model using this compact, task-specific dataset with Smooth K2 Loss. Adhering to our protocol, checkpoints are preserved based on STS-B development set performance. The updated results across the seven STS benchmarks are summarized in Table 3, illustrating an improvement in our method's average Spearman correlation for BERT$_{\text{base}}$ and RoBERTa$_{\text{base}}$ from 76.03 and 76.04 to 82.93 and 83.23, respectively. These outcomes exceed those achieved by leading contrastive learning methods, such as SimCSE, PromptBERT, Jina Embeddings 2 (Günther et al., 2023), and Nomic Embed (Nussbaum et al., 2024), and set new SOTA performance.

Contrastive learning methods, by contrast, are generally unable to leverage the detailed, multi-level annotated information provided by STS datasets. The prevalent contrastive learning loss function, InfoNCE Loss, serves as an illustrative case for this limitation. For any input sentence x_i, InfoNCE Loss computes the similarity between its encoding $f(x_i)$ and that of its positive instance $f(x_i^+)$ in the numerator, while the denominator aggregates similarity calculations between $f(x_i)$ and encodings of other samples within the same batch, aiming to draw similar samples closer and push dissimilar ones apart. The standard formulation of InfoNCE Loss, where N represents the batch size

Models	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B	SICK-R	Avg.
Implementation on $BERT_{base}$								
Sentence-BERT_base ♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
BERT_base + Cross Entropy	70.01	71.18	70.10	78.37	72.92	74.88	**73.58**	73.01
BERT_base + L1 Loss	69.76	69.56	68.13	76.33	70.96	73.61	70.28	71.23
BERT_base + Translated ReLU	**72.51**	75.46	72.34	78.46	72.64	76.54	72.02	74.28
BERT_base + MSE Loss	72.38	76.47	74.35	78.71	72.95	77.91	70.67	74.78
BERT_base + Smooth K2 Loss	72.39	**78.33**	**75.28**	**80.26**	**74.52**	**78.78**	72.65	**76.03**
Implementation on $RoBERTa_{base}$								
Sentence-$\text{RoBERTa}_\text{base}$ ♣	71.54	72.49	70.80	78.74	73.69	77.77	**74.46**	74.21
$\text{RoBERTa}_\text{base}$ + Cross Entropy	71.15	74.29	72.66	79.44	74.12	76.56	73.02	74.46
$\text{RoBERTa}_\text{base}$ + L1 Loss	68.12	62.27	64.20	72.80	67.28	72.44	66.82	67.70
$\text{RoBERTa}_\text{base}$ + Translated ReLU	71.13	76.07	72.18	78.13	73.94	77.59	70.94	74.28
$\text{RoBERTa}_\text{base}$ + MSE Loss	**72.67**	77.09	72.93	79.52	74.12	**77.88**	69.85	74.87
$\text{RoBERTa}_\text{base}$ + Smooth K2 Loss	72.53	**78.28**	**73.88**	**80.88**	**75.35**	77.44	73.94	**76.04**

Table 2: Spearman correlation for models across seven STS tasks **without** using task-specific training data. This table is partitioned to facilitate **a single variable comparison**. ♣: results from (Reimers and Gurevych, 2019).

Models	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B	SICK-R	Avg.
Contrastive Pre-training Model								
Jina Embeddings v2 base	74.28	84.18	78.81	87.55	85.35	84.85	78.98	82.00
Nomic Embed Text v1	65.19	81.67	74.00	83.58	81.87	76.43	75.41	76.88
Implementation on $BERT_{base}$								
SimCSE BERT_base ♠	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
$\text{PromptBERT}_\text{base}$ ♡	**75.48**	85.59	80.57	85.99	**81.08**	**84.56**	80.52	81.97
Ours + STS-B SICK-R train	73.68	**88.42**	**86.10**	**86.56**	79.63	84.12	**82.01**	**82.93**
Implementation on $RoBERTa_{base}$								
SimCSE $\text{RoBERTa}_\text{base}$ ♠	76.53	85.21	80.95	86.03	82.57	85.83	80.50	82.52
$\text{PromptRoBERTa}_\text{base}$ ♡	**76.75**	85.93	82.28	86.69	**82.80**	**86.14**	80.04	82.95
Ours + STS-B SICK-R train	73.83	**89.00**	**84.16**	**87.95**	81.94	84.64	**81.07**	**83.23**

Table 3: Spearman correlation for models across seven STS tasks. ♠: results from (Gao et al., 2021). ♡: results from (Jiang et al., 2022).

and τ a temperature hyperparameter, is as follows:

$$\ell_i = -\log \frac{e^{\cos(f(x_i), f(x_i^+))/\tau}}{\sum_{j=1}^{N} e^{\cos(f(x_i), f(x_j)^+)/\tau}} \tag{3}$$

While this mechanism effectively refines the semantic space distribution of PLMs, it is constrained to utilizing only text pairs with the highest similarity ratings. Since InfoNCE Loss merely includes numerator and denominator components, it distinguishes only whether two texts are similar or not. Given the denominator is composed of other samples within the same batch, the only part that can be filled with labeled data is the numerator.

In contexts where more detailed, domain-specific data is available, the shortcomings of contrastive learning in not being able to effectively harness multi-level label information, only performing coarse semantic distinctions, becomes more evident. A potential pathway is to combine our regression framework with contrastive learning. By supplementing a contrastively trained model with our Siamese neural network architecture, it may be possible to capture finer semantic nuances. This avenue of exploration holds promise for future work, potentially enhancing the applicability and efficacy of our approach.

4.3 Performance under Different Hyperparameter Settings

In this study, we introduce two innovative loss functions, Translated ReLU and Smooth K2 Loss, each characterized by two critical hyperparameters: k and x_0. The parameter k primarily controls the gradient of the loss function, while x_0 sets the tolerance threshold for model predictions. To discern the influence of these hyperparameters on model performance, we conduct a series of experiments during both the initial training phase with NLI datasets and the subsequent fine-tuning phase with task-specific training data. The outcomes of these investigations are consolidated in Tables 4. Rather than executing an exhaustive grid search, initial values are selected based on our preliminary insights, followed by incremental adjustments. This implies that there may still be room for further improvement in our model's performance.

PLM	Loss	k	x_0	Performance
Without task-specific training data				
$\text{BERT}_{\text{base}}$	Translated ReLU	1.5	0.25	74.21
$\text{BERT}_{\text{base}}$	Translated ReLU	2	0.25	74.21
$\text{BERT}_{\text{base}}$	Translated ReLU	2.5	0.25	**74.28**
$\text{BERT}_{\text{base}}$	Smooth K2 Loss	3	0.25	75.75
$\text{BERT}_{\text{base}}$	Smooth K2 Loss	2.5	0.25	75.89
$\text{BERT}_{\text{base}}$	Smooth K2 Loss	2	0.25	**76.03**
$\text{RoBERTa}_{\text{base}}$	Translated ReLU	2	0.25	74.00
$\text{RoBERTa}_{\text{base}}$	Translated ReLU	1.5	0.25	74.11
$\text{RoBERTa}_{\text{base}}$	Translated ReLU	1	0.25	**74.28**
$\text{RoBERTa}_{\text{base}}$	Smooth K2 Loss	2.5	0.25	75.89
$\text{RoBERTa}_{\text{base}}$	Smooth K2 Loss	3	0.2	75.90
$\text{RoBERTa}_{\text{base}}$	Smooth K2 Loss	3	0.25	**76.04**
With task-specific training data				
$\text{BERT}_{\text{base}}$	Smooth K2 Loss	4	0.2	82.89
$\text{BERT}_{\text{base}}$	Smooth K2 Loss	3.5	0.25	82.89
$\text{BERT}_{\text{base}}$	Smooth K2 Loss	4	0.3	82.90
$\text{BERT}_{\text{base}}$	Smooth K2 Loss	4	0.25	**82.93**
$\text{RoBERTa}_{\text{base}}$	Smooth K2 Loss	4	0.3	82.86
$\text{RoBERTa}_{\text{base}}$	Smooth K2 Loss	3.5	0.25	82.90
$\text{RoBERTa}_{\text{base}}$	Smooth K2 Loss	3.5	0.3	83.18
$\text{RoBERTa}_{\text{base}}$	Smooth K2 Loss	3	0.25	**83.23**

Table 4: Impact of different hyperparameter settings (k, x_0) on model performance.

The experimental results from Table 4 reveal minor fluctuations in model performance across diverse hyperparameter configurations, which affirms the resilience and robustness of our proposed methodology. This stability highlights the inherent adaptability of our regression framework as well as loss functions, suggesting their applicability across a wide range of modeling scenarios without necessitating extensive hyperparameter optimization.

4.4 Ablation Studies

In Section 4.1, we initially demonstrate the effectiveness of our regression framework by comparing the performance differences of models utilizing both classification-based and regression-based strategies for STS tasks. Then, we elucidate the significance of zero-gradient buffer zones by comparing the performance of models when selecting Translated ReLU or L1 Loss, and Smooth K2 Loss or MSE Loss as the loss function. These comparisons directly align with the three core innovations of this paper and fulfill the role of ablation experiments.

Here, we enhance our ablation study with an evaluation of our network architecture as depicted in Figure 1. Specifically, we aim to determine the necessity of concatenating u, v, and their element-wise difference $|u - v|$ in the final linear layer of the model. For this purpose, we employ both BERT and RoBERTa models under the same experimental conditions outlined in Section 4.1, with results detailed in Table 5. The findings indicate that the concatenation method $(u, v, |u - v|)$ is the most effective for both PLMs, thus further validating the rationality of our proposed scheme.

PLM	Concatenation	Spearman
$\text{BERT}_{\text{base}}$	(u, v)	53.30
$\text{BERT}_{\text{base}}$	$(\|u - v\|)$	54.84
$\text{BERT}_{\text{base}}$	$(u, v, \|u - v\|)$	**76.03**
$\text{RoBERTa}_{\text{base}}$	(u, v)	60.99
$\text{RoBERTa}_{\text{base}}$	$(\|u - v\|)$	59.10
$\text{RoBERTa}_{\text{base}}$	$(u, v, \|u - v\|)$	**76.04**

Table 5: Average Spearman's correlation scores obtained by models on seven STS tasks with different concatenation methods in the last linear layer of the Siamese neural network architecture.

5 Conclusion

In this paper, we propose an innovative regression framework accompanied by two simple yet efficacious loss functions: Translated ReLU and Smooth K2 Loss, to address multi-category STS tasks. Compared to traditional classification strategies, our regression framework achieves superior performance while reducing the parameter count of the model's output layer. Further empirical evidence demonstrates that when supplemented with task-specific training data, our approach can surpass prevailing contrastive learning methods, achieving state-of-the-art performance on seven STS benchmarks.

Limitations

Due to the lack of suitable baselines and limited computational resources, the experiments in this paper are primarily centered on the discriminative PLMs such as BERT and RoBERTa, rather than recently advanced generative models (e.g., LLaMA (Touvron et al., 2023)). However, it is important to note that, compared to generative PLMs, BERT possesses a much smaller parameter count, which leads to higher inference efficiency. This attribute is particularly valuable in large-scale information retrieval scenarios.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645*.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. RankCSE: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13785–13802.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Bowen Zhang, Kehua Chang, and Chunping Li. 2023a. Cot-bert: Enhancing unsupervised sentence representation through chain-of-thought. *arXiv preprint arXiv:2309.11143*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.