

*mi*CSE: MUTUAL INFORMATION CONTRASTIVE LEARNING FOR LOW-SHOT SENTENCE EMBEDDINGS

Anonymous authors

Paper under double-blind review

ABSTRACT

This paper presents *mi*CSE, a mutual information-based Contrastive learning framework that significantly advances the state-of-the-art in few-shot sentence embedding. The proposed approach imposes alignment between the attention pattern of different views during contrastive learning. Learning sentence embeddings with *mi*CSE entails enforcing the syntactic consistency across augmented views for every single sentence, making contrastive self-supervised learning more sample efficient. As a result, the proposed approach shows strong performance in the few-shot learning domain. While it achieves superior results compared to state-of-the-art methods on multiple benchmarks in few-shot learning, it is comparable in the full-shot scenario. The proposed approach is conceptually simple, easy to implement and optimize, yet empirically powerful. This study opens up avenues for efficient self-supervised learning methods that are more robust than current contrastive methods for sentence embedding.

1 INTRODUCTION

Measuring sentence similarity has been a challenging endeavor due to the ambiguity and variability of linguistic expressions. The community’s strong interest in the topic can be attributed to its applicability in numerous language processing applications, such as sentiment analysis, information retrieval, and semantic search Pilehvar & Navigli (2015); Iyyer et al. (2015). It has been recently shown that the Transformer-based language models already perform surprisingly well Reimers & Gurevych (2019). Simultaneously, to unfold their full potential, language models pre-trained on large generic corpora require fine-tuning on the downstream task and corpora Devlin et al. (2018); Pfeiffer et al. (2020); Mosbach et al. (2021). In terms of sentence embeddings, contrastive learning schemes have already been adopted successfully van den Oord et al. (2018); Liu et al. (2021); Gao et al. (2021); Carlsson et al. (2021). The idea of contrastive learning is that positive and negative pairs are generated given a batch of samples. Whereas the positive pairs are obtained via augmentation, negative pairs are often created by random collation of sentences. Following the construction of pairs, contrastive learning forces the network to learn feature representations by pushing apart different samples (negative pairs) or pulling together similar ones (positive pairs). While some methods seek to optimize for selecting “hard” negative for negative pair generation Zhou et al. (2022), others investigated better augmentation techniques for positive pair creation. In this regard, many methods have been proposed to create augmentations to boost representation learning. A standard approach for the augmentation aim at input *data level* (a.k.a *discrete* augmentation), which comprises word level operations such as swapping, insertion, deletion and substitution Xie et al. (2017); Coulombe (2018); Wei & Zou (2019). In contrast to that, *continuous* augmentation operate at the *representation level*, comprising approaches like interpolation or “mixup” on the embedding space Chen et al. (2020); Cheng et al. (2020); Guo et al. (2019). Most recently, augmentation was also proposed in a more continuous fashion operating in a *parameter level* via simple techniques such as drop-out Gao et al. (2021); Liu et al. (2021); Klein & Nabi (2022) or random span masking Liu et al. (2021). The intuition is that “drop-out” acts as minimal data augmentation, providing an expressive *semantic variation*. However, it will likely affect *syntactic alignment* across views. Since positive pairs are constructed from identical sentences, utilizing drop-out noise, we hypothesize that the syntactic dependency over the views should be preserved. Building on this idea, we maximize the *syntactic dependence* by enforcing distributional similarity over the attention values across the augmentation views. To this end, we employ maximization of the mutual information (MI) on the attention tensors of the

positive pairs. However, since attention tensors can be very high-dimensional, computing mutual information (MI) can quickly become a significant burden if not intractable. Common approaches for MI estimation are non-parametric, e.g., relying on binning, non-parametric kernel density estimators - see Suzuki et al. (2008); Kwak & Choi (2002); Kraskov et al. (2004). These estimators often do not scale well and cannot be computed efficiently on GPUs Gao et al. (2014). In contrast, the parametric modeling of attention seems to be a more reasonable choice for statistical modeling of Transformer attention distribution (e.g., Gaussian distribution Bahuleyan et al. (2018), Dirichlet distribution Deng et al. (2018), Weibull and Log-Normal distribution Fan et al. (2020)). This paper proposes a simple parametric solution to alleviate the computational burden of MI computation, which can be deployed efficiently. Specifically, we adopt the Log-Normal distribution for modeling the attention pattern. On the one hand, empirical evidence confirms this model to be a good fit. On the other hand, it facilitates the optimization objective to be defined in closed form. In this case, mutual information can be provably reformulated as a function of correlation, allowing native GPU implementation. As discussed above, the proposed approach builds upon the contrastive learning paradigm known to suffer from model collapse. This issue becomes even more problematic when enforcing MI on the attention level, as it tightens the positive pairs via regularizing the attention. Therefore the selection of negative pairs becomes more critical in our setup. To this end, we utilize momentum contrastive learning to generate harder negatives He et al. (2020). A “tighter” binding on positive pairs and repulsion on “harder” negative pairs empowers the proposed contrastive objective, yielding more powerful representations.

Combining ideas from momentum contrastive learning and attention regularization, we propose *miCSE*, a conceptually simple yet empirically powerful method for sentence embedding, with the goal of integrating semantic and syntactic information of a sentence in an information-theoretic and Transformer-specific manner. We conjecture the relation between attention maps and a form of syntax to be the main driver behind the success of our approach. To validate that, we performed a controlled empirical observation on this matter which suggests the lack of syntax-related properties of the sentences in a previous work (i.e., SimCSE Gao et al. (2021)) compared to *miCSE* (see Fig. 1). We speculate that our proposed method injects syntactic information into the model as an inductive bias, facilitating representation learning with fewer samples. The adopted syntactic information inductive biases provide a structural prior as an implicit form of supervision during training Wilcox et al. (2020), which promotes few-shot learning capabilities in neural language models. To validate this, we introduced a low-shot setup for training sentence embeddings. In this benchmark, we finetune the language model *only* with a small number of training samples. Note that this is a very challenging setup. The inherent difficulty can be attributed to the need to mitigate the domain shift in the low-shot self-supervised learning scheme. We emphasize the importance of this task, as in many real-world applications, only small datasets are often available. Examples of such cases are NLP for low-resource languages or expert-produced texts (e.g., medical records by doctors), personalized LM for social media analysis (e.g., personalized hate speech recognition on Twitter), etc. In the low-shot sentence embedding benchmark, our proposed method significantly improves over the state-of-the-art. We believe this is the first work that explores how to combine semantic and syntactic information through attention regularization, and empirically demonstrates this benefit for low-shot sentence embeddings.

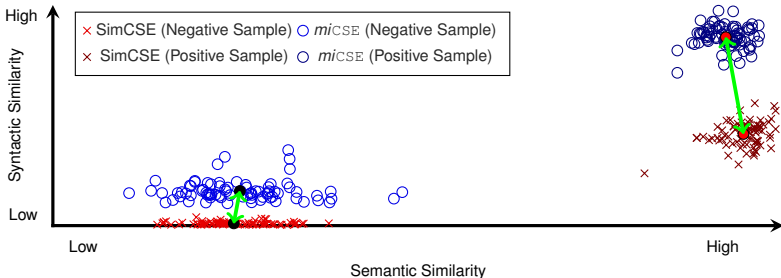


FIGURE 1. Sentence embeddings of positive and negative contrastive pairs in terms of semantic and syntax, comparing SimCSE and *miCSE*. Semantic similarity is measured in terms of cosine similarity, syntactic similarity measured with mutual information on attention-level. (●) and (●) denote centroids of positive and negative centroids, (↔) their distance.

Related works: Recently, VaSCL Zhang et al. (2022), ConSERT Yan et al. (2021) and PCL Wu et al. (2022) proposed contrastive representation learning with diverse augmentation strategies on positive pair. However, we proposed a principled approach for enforcing *alignment* in positive pairs at contrastive learning. Similar to us, ESIMCSE Wu et al. (2021) and MoCoSE Cao et al. (2022) proposed to exploit a momentum contrastive learning model with negative sample queue for sentence embedding to boost *uniformity* of the representations. However, unlike us, they do not enforce any further tightening objective on the positive pairs nor consider few-shot learning. Very recently, authors in InforMin-CL Chen et al. (2022) proposed information minimization-based contrastive learning. Specifically, the authors propose to minimize the information entropy between positive embeddings generated by drop-out augmentation. Our model differs from this paper and the method in Bachman et al. (2019); Yang et al. (2021); Zhang et al. (2020); Sordoni et al. (2021); Wu et al. (2020), which focuses on using mutual information for self-supervised learning. A key difference compared to these methods is that they estimate MI directly on the representation space. In contrast, our method computes the MI on attention. Other related works include Chuang et al. (2022); Liu et al. (2022).

The contributions of the proposed work are threefold: **First**, we propose to inject syntactic information into language models by adding an attention-level objective. **Second**, we introduce Attention Mutual Information (AMI), a simple and efficient objective for sample efficient self-supervised contrastive learning. **Third**, we introduce low-shot learning for sentence embedding. We show that our method performs comparably to the state-of-the-art in the full-shot scenario and significantly better in few-shot learning.

2 METHOD

The proposed approach aims to exploit the syntactic structure of the sentences in a contrastive learning scheme. Compared to conventional contrastive learning that solely operates at the level of *semantic* similarity in the embedding space, the proposed approach injects *syntactic* information into the model. This is achieved by regularizing the attention space of the model during training. We let \mathcal{D} denote a dataset consisting of string sequences (sentences) from corpus \mathcal{X} with $\mathcal{D} = \{x_1, x_2, \dots, x_{|\mathcal{X}|}\}$, where we assume x_i to be a tokenized sequence of length n with $x_i \in \mathbb{N}^n$. For mapping the input data to the embedding space, we use a bi-encoder f_θ parametrized by θ . Bi-encoders entail the computation of embeddings for similarity comparison, whereby each sentence in a pair is encoded separately. Hence, the instantiation of a bi-encoder on augmented input data induces multiple views. For the following, we let $v \in \{1, 2\}$ denote the index of the view, where each view corresponds to a different augmentation. Consequently, encoding a data batch \mathcal{D}_b yields embedding matrices $E_v \in \mathbb{R}^{|\mathcal{D}_b| \times U}$, where U denotes the dimensionality of the embeddings. Employing a Transformer, encoding the input data yields the embedding matrices and the associated attention tensors W_v . Then learning representation of the proposed approach entails the optimization of a joint loss:

$$\min_{\theta} \mathcal{L}_C(E_1, E_2) + \mathcal{L}_D(W_1, W_2) \quad (1)$$

with $(E_1, W_1), (E_2, W_2) = f_\theta(\mathcal{D}_b)$. Here, \mathcal{L}_C is responsible for the semantic alignment, corresponding to the standard InfoNCE van den Oord et al. (2018) loss that seeks to pull positive pairs close together while pushing away negative pairs in the embedding space. In contrast, \mathcal{L}_D is responsible for the syntactic alignment, operating on the attention space. However, in comparison to \mathcal{L}_D is employed only on the attention tensors of the positive pairs.

2.1 EMBEDDING-LEVEL MOMENTUM-CONTRASTIVE LEARNING (INFOANCE)

The InfoNCE-loss seeks to pull positive pairs together in the embedding space while pushing negative pairs apart. Specifically, InfoNCE on embeddings pushes for the similarity of each sample and its corresponding augmented embedding. Negatives pairs are constructed in two ways, reflected by the two terms in the denominator of Eq. 2. First, in-batch negative pairs are constructed by pairing each sentence with another random sentence (sharing no semantic similarity), pushing for dissimilarity. Second, using embeddings obtained from a momentum encoder known as MoCo He et al. (2020); Cao et al. (2022). The momentum encoder is a replication of the encoder f_θ , whose parameters are updated more slowly. Specifically, while the parameters of f_θ encoder are updated via back-propagation, the parameters of the momentum encoder are updated using an exponential

Algorithm 1 Mutual Information estimation

Input: Batch \mathcal{D}_b , encoder f_θ , multinomial sampler p_{mult}
Output: Average mutual information $\frac{1}{|R|\cdot|\mathcal{D}_b|} \sum_{i,r}^{D_b,R} MI_i^r$

$(E_1, W_1), (E_2, W_2) \leftarrow f_\theta(\mathcal{D}_b)$ ▷ Transformer encoding creating views
for $i \leftarrow 1 \dots |\mathcal{D}_b|$ **do**
 $w_i, {}^+w_i \leftarrow \text{EXTRACT}(W_1, W_2, i)$ ▷ Extract attention tensor for each sample
 $\{({}^+)w_i^1, \dots, ({}^+)w_i^R\} \leftarrow \pi({}^+)w_i$ ▷ Slicing the attention tensors
 $s \leftarrow$ number of text tokens in x_i
 for $r \leftarrow 1 \dots |R|$ **do**
 $J_r \leftarrow p_{mult}(1/s^2, \dots, 1/s^2, \mathbf{0})$ ▷ Sampling indices of valid attentions
 $MI_i^r \leftarrow \text{AMI}(\bigcup_{j \in J_r} w_i^r[j], \bigcup_{j \in J_r} {}^+w_i^r[j])$
 end for
end for
procedure $\text{AMI}(w_1, w_2)$
 $z_1, z_2 \leftarrow \log(w_1), \log(w_2)$ ▷ Log-Normal to Normal transform
 $\rho \leftarrow \cos(z_1 - \bar{z}_1, z_2 - \bar{z}_2)$ ▷ Compute correlation coefficient on centered attentions
 Return $-\frac{1}{2}(1 - \rho^2)$ ▷ Mutual information for tensor slice
end procedure

moving average from the former. The negative embeddings are produced from samples from previous batches, which are stored in queue \mathcal{Q} and are forward-passed through the momentum encoder.

Then the InfoNCE van den Oord et al. (2018) loss is defined as:

$$\mathcal{L}_C = - \sum_i^{|\mathcal{D}_b|} \log \frac{\exp(\text{sim}(e_i, {}^+e_i)/\tau)}{\sum_{j:i \neq j}^{|\mathcal{D}_b|} \exp(\text{sim}(e_i, e_j)/\tau) + \sum_j^{|\mathcal{Q}|} \exp(\text{sim}(e_i, q_j)/\tau)}, \quad (2)$$

where $e_i \in E_1$ and ${}^+e_i \in E_2$ denote the embeddings of different augmentations of x_i , $\text{sim}(\cdot)$ is the cosine similarity metric, q_j denote representations obtained from momentum encoder, and $\tau \in \mathbb{R}$ is temperature scalar.

2.2 ATTENTION-LEVEL MUTUAL INFORMATION (AMI)

Preliminaries and notations: We first briefly review the attention mechanism and explain the notation used in the rest of this section. A Transformer stack consists of a stack of L layers, with input data cascading up the layer stack. Each layer comprises a self-attention module and a feed-forward network in its simplest form. Passing sentences through the encoder stack entails simultaneous computation of attention weights. These attention weights indicate the relative importance of every token. To this end, key-value pairs are computed for each token of the input sequence Within each self-attention module. This entails the computation of three different matrices: key matrix K , value matrix V , and query matrix Q . The values of the attention weights W are obtained according to $W = \text{softmax}(f(Q, K)) \in \mathbb{R}^{n \times n}$, where $f(\cdot)$ is a scaled dot-product. Output features are then generated as obtained according to WV . In order to attend to different sub-spaces Vaswani et al. (2017) simultaneously, the attention mechanism is replicated H times, which is referred to as multi-head attention. During training the encoder, the self-attention tensors W values are subject to a random deterministic process, with randomness arising due to drop-out. Hence, the proposed approach seeks to optimize syntactic alignment by maximizing mutual information between the attention tensors W_v of the augmentation views. To regularize the joint attention space, we propose a pipeline consisting of four steps:

1) Attention Tensor Slicing: Given that augmentation has different effects on the attention distribution depending on the depth (layer) and the position (head) in the Transformer stack, we propose to slice the attention tensor. Chunking the attention has multiple advantages. On the one hand, this allows for preserving the locality of distribution change. This is important as it can be empirically observed that distribution divergence between views decreases with increasing depth in the encoding stack. On the other hand, restricting the space permits the use of a simple distributional model such as bivariate distribution. For the sake of economy in notation, we will restrict the attention tensor of a

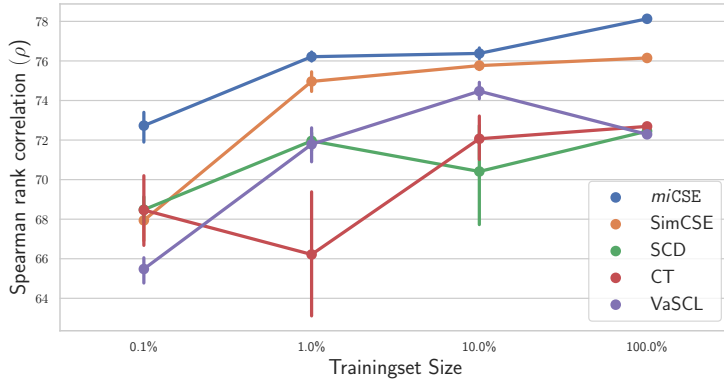


FIGURE 2. Few-shot performance of different algorithms. *miCSE* (—), *SimCSE* Gao et al. (2021) (—), *SCD* Klein & Nabi (2022) (—), *CT* Carlsson et al. (2021) (—), *VaSCL* Zhang et al. (2022) (—). Performance is shown in Spearman’s correlation average of the STS benchmark. Training set size: 0.1%, 1.0%, 10.0%, 100.0% of the data.

single encoded sample for the following. To this end, a slicing function $\pi : \mathbb{R}^{L \times H \times n \times n} \rightarrow \mathbb{R}^{R \times n \times n}$ cuts the attention tensor for sample i into R (indexed) elements: $\pi(\mathbf{w}_i) = \{\mathbf{w}_i^1, \dots, \mathbf{w}_i^R\}$.

2) Attention Sampling: Different sentences in the batch are typically in token sequences of different lengths. To accommodate the different lengths and facilitate efficient training, sequences are typically padded with [PAD]-token for length equality. Although this allows for efficient batch encoding on GPU, attentions arising from [PAD]-tokens have to be discarded when looking at statistical relationships. In order to accommodate for the different lengths of tokenized sequences, perform a sampling step for attention values within each grid cell \mathbf{w}_i^r . To this end, we leverage multinomial distribution $P_{mult}(p_1, \dots, p_{n^2})$, where s correspond to the number of non-padding tokens with $1 \leq s \leq n$. Specifically, we sample from the pool of s^2 attention values, each with a probability of $\frac{1}{s^2}$, with the remaining elements associated with probability 0. As a result, we obtain a set J_r consisting of m indices of the attention tensors for each slice $r \in R$:

$$J_r = \{j_1, \dots, j_m\} \sim P_{mult}\left(\underbrace{1/s^2, \dots, 1/s^2}_{1, \dots, s^2}, \underbrace{0, \dots, 0}_{(n-s)^2, \dots, n^2}\right) \quad (3)$$

It should be noted that for the same slice r across the views, the same index set is used for sampling: $\tilde{\mathbf{w}}_i^r = \bigcup_{j \in J_r} \mathbf{w}_i^r[j]$ and ${}^+\tilde{\mathbf{w}}_i^r = \bigcup_{j \in J_r} {}^+\mathbf{w}_i^r[j]$.

3) Attention Mutual Information Estimation: We propose using mutual information to measure the similarity of attention patterns for different views. Specifically, we follow Fan et al. (2020) and adopt the Log-Normal distribution for modeling the attention distribution, which is prudent for several reasons. First, Empirical observation confirms attention asymmetry. Second, leveraging a non-symmetric distribution accommodates that the attention tensor W decomposes into K and Q , which enables attention to be non-symmetric. Third, adopting the log-normal models facilitates the optimization objective to be defined in closed form and hence easy to optimize, particularly on GPUs. Mutual information for two normally distributed tuple vectors (z_1, z_2) can be written as a function of correlation I.M. & A.M. (1957) given by:

$$I(z_1, z_2) = -\frac{1}{2} \log(1 - \rho^2), \quad (4)$$

where ρ corresponds to the correlation coefficient computed from z_1 and z_2 . Hence, we compute the mutual information with for each slice r and sample x_i as $MI_i^r = I(\log(\tilde{\mathbf{w}}_i^r), \log({}^+\tilde{\mathbf{w}}_i^r))$. Application of $\log(\cdot)$ function is employed to accommodate for the Log-Normal to Normal random variable transformation. For details on the implementation, see Alg. 1.

4) Mutual Information Aggregation: In order to compute the loss component for attention regularization, we need to aggregate the distributional similarities for the entire tensor. Aggregation is obtained by averaging the individual similarities obtained for each slice $r \in R$ and each sample x_i

Semantic Textual Similarity (STS) Benchmark								
Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
BERT	21.54	32.11	21.28	37.89	44.24	20.29	42.42	31.40
BERT [◇] (first-last avg)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
GloVe [♣] (avg.)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT-flow [◇]	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT-whitening [◇]	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS [♡]	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
SG-OPT [♠]	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
C ^T	67.43	79.18	69.05	76.92	74.62	73.24	68.38	72.69
SCD [†]	66.94	78.03	69.89	78.73	76.23	76.30	73.18	74.19
Mirror-BERT [†]	69.10	81.10	73.00	81.90	75.70	78.00	69.10	75.40
SimCSE	68.69	82.05	72.91	81.15	79.39	77.93	70.93	76.15
ConSERT [†] _{large}	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
VaSCL [†]	69.08	81.95	74.64	82.64	80.57	80.23	71.23	77.19
MoCoSE [†]	71.58	81.40	74.47	83.45	78.99	78.68	72.44	77.27
InforMin-CL [†]	70.22	83.48	75.51	81.72	79.88	79.27	71.03	77.30
PCL [†]	72.74	83.36	76.05	83.07	79.26	79.72	72.75	78.14
<i>mi</i> CSE	71.71	83.09	75.46	83.13	80.22	79.70	73.62	78.13

TABLE 1. Sentence embedding performance on STS tasks measured as Spearman’s correlation using BERT_{base}, except for VaSCL that uses RoBERTa. Unless states otherwise, [CLS]-embedding was used. ♣: results from Reimers & Gurevych (2019); ♡: results from Zhang et al. (2020); ♠ results from Kim et al. (2021); ◇ results from Gao et al. (2021); † by the respective authors; other results are by ourselves. (●) denotes top-3 results, (●) denotes the proposed approach, **bold** denotes best result.

in the batch. With $\lambda \in \mathbb{R}$ some weighting scalar, the attention alignment loss term is:

$$\mathcal{L}_D(W_1, W_2) = -\frac{\lambda}{|R| \cdot |D_b|} \sum_i^{\mathcal{D}_b} \sum_r^R MI_i^r \quad (5)$$

3 EXPERIMENTS

In this section, we describe the experimental setting used for the evaluation, present our main results, and discuss different aspects of our method by providing several empirical analyses.

3.1 EXPERIMENTAL SETUP

Model and Hyperparameters: Training is started from a pre-trained transformer LM. Specifically, we employ the Hugging Face Wolf et al. (2020) implementation of BERT_{base}. For each approach evaluated, we follow the same hyperparameters proposed by the authors. In the InfoNCE loss we set $\tau = 0.05$. In order to determine the hyperparameter λ a coarse grid search $\{1.0, 0.1, \dots, 1.0e-5\}$ was conducted to assess the magnitude. Upon determination, a fine grid search was conducted once with 10 steps. We set $\lambda = 2.5e-3$ for training 100% of the data in a single episode with a batchsize of 50 at a learning rate of $3.0e-5$ and 250 warm-up steps. The number of optimization steps is largely kept constant for training the different dataset sizes. For the training set of size 10^6 (= 100%) we train for 1 epoch, for the size of 10^5 (= 10%), we train for 10 epochs, etc. The momentum encoder is associated with a sample queue of size $|Q| = 384$. The momentum encoder parameters are updated with a factor of 0.995, except for the MLP pooling layer, which is kept identical to the online network. Additionally, we increase the drop-out for the momentum encoder network from the default rate (0.1) to 0.3.

Data and Evaluation: Following Gao et al. (2021), we train the model in an unsupervised fashion on sentences from Wikipedia. In order to train the model in a few-shot learning scenario, we create random sample sets of different sizes $\{10^6, 10^5, 10^4, 5.0 \cdot 10^3, 10^3\}$. We repeated the training set creation for each size for 5 times with different random seeds.

Mutual Information Estimation: Following the observations in Voita et al. (2019), we restrict the computation of the mutual information to the upper part of the layer stack. Specifically, we select the layers between 8 and 12 (= last layer in $BERT_{base}$). To accommodate input sequences of varying lengths and make computation more efficient, we pool together pairs of adjacent heads (without overlap) while preserving the head separation. From each of the $(4 \times \frac{H}{2})$ chunks of pooled attentions, we random sample 150 joint-attention pairs for each embedding of the bi-encoder.

3.2 EXPERIMENTAL RESULTS

Unsupervised Sentence Embedding: We compare *mi*CSE to previous state-of-the-art sentence embedding methods on STS tasks. For comparisons, we favored comparable architectures (bi-encoder) that facilitate seamless integration of the proposed approach and methods of comparable backbone.

For semantic text similarity, we evaluated on 7 STS tasks: Agirre et al. (2012; 2013; 2014; 2015; 2016), STS Benchmark Cer et al. (2017) and SICK-Relatedness Marelli et al. (2014). These datasets come in sentence pairs with correlation labels in the range of 0 and 5, indicating the semantic relatedness of the pairs. Specifically, we employ the SentEval toolkit Conneau & Kiela (2018) for evaluation. It should be noted that all our STS experiments are conducted in a *fully unsupervised* setup, not involving any STS training data. The benchmark measures the relatedness of two sentences based on the cosine similarity of their embeddings. The evaluation criterion is Spearman’s rank correlation (ρ). For comparability, we follow the evaluation protocol of Gao et al. (2021), employing Spearman’s rank correlation and aggregation on all the topic subsets. Results for the sentence similarity experiment are presented in Tab. 1. As can be seen, the proposed approach is slightly lower in terms of average performance than state-of-the-art algorithms such as PCL. However, our proposed method has the most consistent performance across all benchmarks, with performance always in the top-3. A more in-depth analysis shows the best performance on the SICK-R benchmark, where it outperforms the second best approach SCD by (+0.44) and PCL by (+0.87). We highlight the comparison to the closest method SimCSE, where the proposed approach has an average gain of (+3.94). This improvement is due to the two additional components (i.e., AMI and MoCo) we add to this baseline method.

Low-shot Sentence Embedding: In this experiment, the performance of several SOTA sentence embedding approaches is benchmarked elaboratively. Similar to Sec. 3.2, we evaluate 7 STS tasks, STS Benchmark, and SICK-Relatedness with Spearman’s ρ rank correlation as the evaluation metric. However, in contrast to the previous section, models are trained on different subsets of the data, namely {100%, 10%, 1%, 0.1%} of the Wikipedia dataset used in Gao et al. (2021). Results for the low-shot sentence similarity experiment can be presented in Fig. 2. As can be seen, the proposed approach gains by increasing the training set size and consistently outperforms all the baselines in all

<i>Semantic Textual Similarity</i>				
Model	0.1%	1%	10%	100%
CT Carlsson et al. (2021)	68.46 ± 2.33	66.21 ± 4.06	72.06 ± 1.46	72.69
AMI+CT	71.12 ± 1.11	72.20 ± 0.49	73.20 ± 0.78	73.55
Mirror-BERT Liu et al. (2021)	40.13 ± 5.08	42.17 ± 1.69	42.47 ± 3.66	43.32
AMI+Mirror-BERT	43.99 ± 1.26	45.26 ± 2.60	44.72 ± 1.36	47.48
Mirror (avg.) Liu et al. (2021)	71.48 ± 1.19	71.80 ± 1.18	70.38 ± 1.18	69.81
AMI+Mirror-BERT (avg.)	71.49 ± 0.95	72.54 ± 0.49	70.68 ± 1.19	71.34
SimCSE Gao et al. (2021)	67.94 ± 1.16	74.96 ± 0.65	75.76 ± 0.24	76.15
AMI+SimCSE	73.85 ± 0.49	76.21 ± 0.28	76.31 ± 0.46	76.88
MoCo+SimCSE	69.54 ± 1.61	75.73 ± 0.91	76.73 ± 0.29	76.81
<i>mi</i> CSE	73.68 ± 0.89	76.40 ± 0.48	76.38 ± 0.35	78.13

TABLE 2. Sentence embedding few-shot learning performance on STS tasks measured as Spearman’s correlation using $BERT_{base}$. Unless states otherwise, [CLS]-embedding was used, number corresponds to the average performance across all benchmarks, **bold** denotes best performance, (○) denotes integration of the proposed approach. Results computed by ourselves.

training subsets. Interestingly, our proposed method reaches the performance of SimCSE trained on the entire dataset with only 0.5% of the data. We believe it shows the impact of exploiting syntax information for data augmentation during training. It should be noted that the performance gain is most significant when conducted on a single token rather than token averaging. We attribute this to token averaging to a certain degree equivalent to attention regularization. On the *extremely* low data regime, the proposed approach shows very strong performance up (+11) compared to SimCSE - see Fig. 3a. It suggests the potential resilience of our self-supervised fine-tuning to very small batch training.

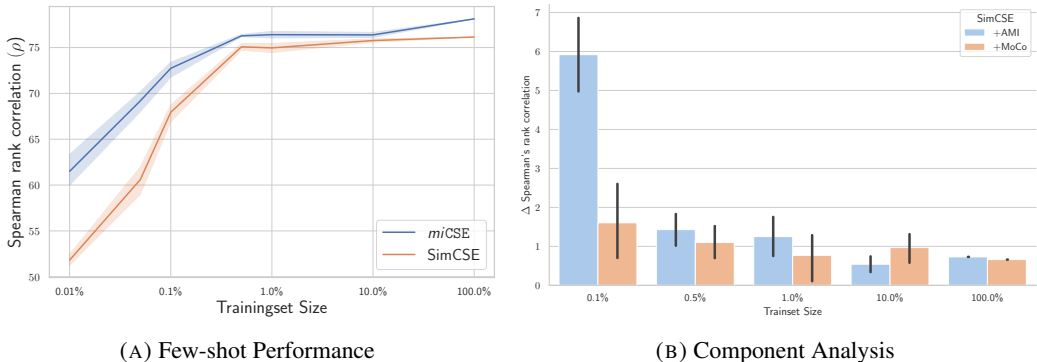


FIGURE 3. Few-shot performance analysis of models trained with different ratios of dataset size. Performance is shown in Spearman’s correlation average of the STS benchmark. **Left:** Few-shot performance of SimCSE Gao et al. (2021) (—) and the proposed approach *miCSE* (—). **Right:** Few-shot ablation study with y-axis showing change (Δ) in Spearman’s rank correlation ρ , showing the effect of adding individual components w.r.t. the SimCSE baseline.

3.3 EXPERIMENTAL ANALYSIS OF COMPONENTS

Untitled document Saved 255 words Given that AMI acts as a regularizer on Transformer attention, we evaluate the applicability in conjunction with other contrastive learning methods. We evaluate the following approaches CT Carlsson et al. (2021), Mirror-BERT Liu et al. (2021), and SimCSE Gao et al. (2021). Evaluation is conducted on 7 STS tasks, STS Benchmark, and SICK-Relatedness with Spearman’s ρ rank correlation as the evaluation metric. Results for the low-shot sentence similarity experiment are presented in Tab. 2. As can be seen, our proposed AMI can boost the performances of all approaches in all data portions.

Additionally, it shows the most significant boost in performance in combination with SimCSE. In addition, we observe that the impact of AMI grows with declining training set size. In combination with SimCSE, AMI leads to a performance gain of up to (+5.91) at 0.1% of the data. We also observe that adding AMI to all the approaches significantly reduces the variance for all methods. This can probably be attributed to the regularization effect of the proposed AMI component. In addition, we conducted an ablation study to assess the effect of AMI and MoCo w.r.t. the baseline SimCSE. As can be seen in Fig. 3b, both AMI and MoCo improve the baseline at different data ratios. Again, AMI provides a particularly strong performance boost in the low-data regime.

In contrast, the impact of MoCo diminishes with decreasing training set size. We emphasize that our approach gets the best of both worlds by integrating these two components. In practice, this can be directly exploited for different few-shot setups by adjusting the relative importance of AMI via tuning hyper-parameter λ .

3.4 EXPERIMENTAL ANALYSIS ON SYNTAX VS. SEMANTIC

In light of the lack of a rigorous benchmark for analyzing syntax in sentence embedding, we performed two qualitative analyses visualized in Fig. 1 and Fig 4. Our empirical observations are:

Observation (i) *There is higher semantic and syntactic similarity between positive pairs compared to the negative pairs:* Our contrastive learning approach assumes that positive pairs exhibit more

syntactic similarity than negative pairs (i.e., syntactic inductive bias). To validate this hypothesis, we plot the semantic similarity against syntactic similarity - for both positive and negative pairs. Specifically, we analyzed the embeddings and attention values of the trained model with SimCSE and the proposed approach. Input to the models was randomly sampled sentences from Wikipedia. Interestingly enough, although training the proposed model involves maximization of MI over the attention w.r.t. positive pairs, we also observe the reflection of syntactic information in the negative pairs. As can be seen in Fig. 1, the negative pairs end up in the low left corner, whereas the positive pairs are in the upper right corner.

Observation (ii): Negative pairs with similar syntax shows higher attention similarity, compared to pairs with dissimilar syntax: For a more in-depth analysis of this, we further sub-divided the negative pairs into two groups: *a*) negative pairs with similar dependency trees, *b*) negative pairs with dissimilar dependency trees. For simplicity, we adopted a binary similarity scheme - “similar” implies an identical dependency tree, whereas “dissimilar” corresponds to a non-identical dependency tree. To highlight the inter-group syntax similarity, samples of each group were normalized w.r.t. the centroid of the opposite group. As can be seen in Fig 4 (by the increased distance between the cluster centers), the proposed approach encodes a notion of syntactic similarity. Note that this margin appeared solely due to enforcing the AMI on attention for the positive pairs, leading to the emergence of a notion of “syntax” on negative pairs.

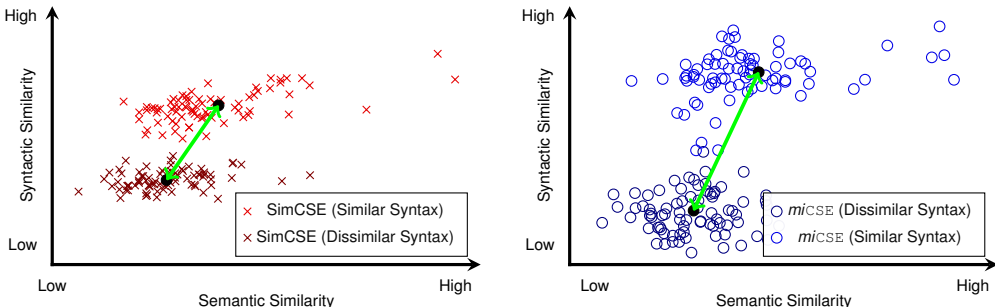


FIGURE 4. Comparison of negative contrastive pairs sentence embeddings in terms of semantic and syntax. **Left:** SimCSE **Right:** *miCSE*. Semantic similarity measured in terms of cosine similarity, syntactic similarity measured with mutual information on attentions. Negative pairs sub-divided into pairs with similar/dissimilar dependency trees. (●) denote cluster centroids, (↔) distance between centroids. Ranges are aligned.

Discussion on the relation of *Syntax* and *Attention*: The proposed approach aligns the attention patterns for drop-out augmented input pairs. We posit that conducting such a regularization enforces constraints w.r.t. the syntax tree of the sentence embeddings. This is motivated by recent literature findings, which suggest that the Transformers attention captures the syntactic grammatical relationships of the sentences Ravishankar et al. (2021); Clark et al. (2019); Raganato et al. (2018); Voita et al. (2019). Additionally, recent research explicitly targets the extraction of topologies from attention maps for diverse tasks on syntactic and grammatical structure Kushnareva et al. (2021); Cherniavskii et al. (2022); Perez & Reinauer (2022). Although no “one-to-one” mapping connects syntactic structures and attention patterns, the attention tensor, at the bare minimum, encodes a “holistic notion” of the grammatical structure of sentences.

4 CONCLUSION

We proposed a method to inject structural similarity into language models for self-supervised representation learning for sentence embeddings. The proposed approach integrates the inductive bias at the level of Transformer attention by enforcing mutual information on positive pairs obtained by drop-out augmentation. Leveraging attention regularization makes the proposed approach much more sample efficient. Consequently, it outperforms methods with a significant margin in low-shot learning scenarios while having state-of-the-art performance in full-shot to comparable approaches. Future work will investigate the extension of the proposed approach to discrete augmentation.

REFERENCES

- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 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 (SemEval 2012)*, pp. 385–393, 2012. URL <https://www.aclweb.org/anthology/S12-1051>.
- Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *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*, pp. 32–43, 2013. URL <https://www.aclweb.org/anthology/S13-1004>.
- Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 81–91, 2014. doi: 10.3115/v1/S14-2010. URL <https://www.aclweb.org/anthology/S14-2010>.
- 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. 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)*, pp. 252–263, 2015. doi: 10.18653/v1/S15-2045. URL <https://www.aclweb.org/anthology/S15-2045>.
- Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 497–511. Association for Computational Linguistics, 2016. doi: 10.18653/v1/S16-1081. URL <https://www.aclweb.org/anthology/S16-1081>.
- Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.
- Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. Variational attention for sequence-to-sequence models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1672–1682, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL <https://aclanthology.org/C18-1142>.
- Rui Cao, Yihao Wang, Yuxin Liang, Ling Gao, Jie Zheng, Jie Ren, and Zheng Wang. Exploring the impact of negative samples of contrastive learning: A case study of sentence embedding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3138–3152, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.248. URL <https://aclanthology.org/2022.findings-acl.248>.
- Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. Semantic re-tuning with contrastive tension. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Ov_sMNau-PF.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, 2017. doi: 10.18653/v1/S17-2001. URL <https://www.aclweb.org/anthology/S17-2001>.
- Jiaao Chen, Zichao Yang, and Diyi Yang. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2147–2157, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.194. URL <https://aclanthology.org/2020.acl-main.194>.

- Shaobin Chen, Jie Zhou, Yuling Sun, and He Liang. An information minimization contrastive learning model for unsupervised sentence embeddings learning. In *International Conference of Computational Linguistics (COLING)*, 2022.
- Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5961–5970, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.529. URL <https://aclanthology.org/2020.acl-main.529>.
- Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Acceptability judgements via examining the topology of attention maps, 2022. URL <https://arxiv.org/abs/2205.09630>.
- Yung-Sung Chuang, Rumén Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljagic, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert’s attention. *arXiv preprint arXiv:1906.04341*, 2019.
- Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. 2018. URL <https://www.aclweb.org/anthology/L18-1269>.
- Claude Coulobme. Text data augmentation made simple by leveraging nlp cloud apis. *ArXiv*, abs/1812.04718, 2018.
- Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. Latent alignment and variational attention. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/b691334ccf10d4ab144d672f7783c8a3-Paper.pdf>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Xinjie Fan, Shujian Zhang, Bo Chen, and Mingyuan Zhou. Bayesian attention modules. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16362–16376. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/bcff3f632fd16ff099a49c2f0932b47a-Paper.pdf>.
- Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. *CoRR*, abs/1411.2003, 2014. URL <http://arxiv.org/abs/1411.2003>.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL <https://aclanthology.org/2021.emnlp-main.552>.
- Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. *ArXiv*, abs/1905.08941, 2019.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.
- Gel’fand I.M. and Yaglom A.M. Calculation of amount of information about a random function contained in another such function. In *Amer. Math. Soc. Transl. Ser.: Series 2*, volume 12, pp. 199–246, 1957.

- Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1681–1691, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1162. URL <https://aclanthology.org/P15-1162>.
- Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2528–2540, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.197. URL <https://aclanthology.org/2021.acl-long.197>.
- Tassilo Klein and Moin Nabi. Scd: Self-contrastive decorrelation for sentence embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), May 2022.
- Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL <https://link.aps.org/doi/10.1103/PhysRevE.69.066138>.
- Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Baranikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 635–649, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.50. URL <https://aclanthology.org/2021.emnlp-main.50>.
- N. Kwak and Chong-Ho Choi. Input feature selection by mutual information based on parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1667–1671, 2002. doi: 10.1109/TPAMI.2002.1114861.
- Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1442–1459, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.emnlp-main.109>.
- Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. Trans-encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. In *International Conference on Learning Representations, 2022*. URL <https://openreview.net/forum?id=AmUhwTOHgm>.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. pp. 216–223, 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.
- Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations, 2021*. URL <https://openreview.net/forum?id=nzplWnVAyah>.
- Ilan Perez and Raphael Reinauer. The topological bert: Transforming attention into topology for natural language processing, 2022. URL <https://arxiv.org/abs/2206.15195>.
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pp. 46–54, Online, 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.7>.

- Mohammad Taher Pilehvar and Roberto Navigli. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128, 2015. ISSN 0004-3702. doi: <https://doi.org/10.1016/j.artint.2015.07.005>. URL <https://www.sciencedirect.com/science/article/pii/S000437021500106X>.
- Alessandro Raganato, Jörg Tiedemann, et al. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics, 2018.
- Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3031–3045, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.264. URL <https://aclanthology.org/2021.eacl-main.264>.
- Nils Reimers and Iryna Gurevych. 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)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL <https://aclanthology.org/D19-1410>.
- Alessandro Sordani, Nouha Dziri, Hannes Schulz, Geoff Gordon, Philip Bachman, and Remi Tachet Des Combes. Decomposed mutual information estimation for contrastive representation learning. In *International Conference on Machine Learning*, pp. 9859–9869. PMLR, 2021.
- Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In Yvan Saeys, Huan Liu, Iñaki Inza, Louis Wehenkel, and Yves Van de Pee (eds.), *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008*, volume 4 of *Proceedings of Machine Learning Research*, pp. 5–20, Antwerp, Belgium, 15 Sep 2008. PMLR. URL <https://proceedings.mlr.press/v4/suzuki08a.html>.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL <http://arxiv.org/abs/1807.03748>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.
- Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pp. 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1670. URL <https://aclanthology.org/D19-1670>.
- Ethan Wilcox, Peng Qian, Richard Futrell, Ryosuke Kohita, Roger Levy, and Miguel Ballesteros. Structural supervision improves few-shot learning and syntactic generalization in neural language models. *arXiv preprint arXiv:2010.05725*, 2020.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language

- processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020.
- Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. Pcl: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings, 2022. URL <https://arxiv.org/abs/2201.12093>.
- Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding. *arXiv preprint arXiv:2109.04380*, 2021.
- Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. Data noising as smoothing in neural network language models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=H1VYHY9gg>.
- Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5065–5075, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.393. URL <https://aclanthology.org/2021.acl-long.393>.
- Chuangang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual representation learning. *arXiv preprint arXiv:2104.12565*, 2021.
- Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew Arnold. Virtual augmentation supported contrastive learning of sentence representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 864–876, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.70. URL <https://aclanthology.org/2022.findings-acl.70>.
- Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. An unsupervised sentence embedding method by mutual information maximization. pp. 1601–1610, 2020. doi: 10.18653/v1/2020.emnlp-main.124. URL <https://www.aclweb.org/anthology/2020.emnlp-main.124>.
- Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. Debaised contrastive learning of unsupervised sentence representations. *arXiv preprint arXiv:2205.00656*, 2022.

A APPENDIX

In the following sections, we add additional details omitted in the main paper due to space restrictions. First, we illustrate the cosine similarity distribution according to human judgment (ground-truth) in Sec. B. Next, in Sec. C, we visualize the 2D histogram of joint distributions between views. In Sec. D, we present detailed results of the few-shot performance of *miCSE* in contrastive and non-contrastive setup. Finally, the exact relation between mutual information and correlation is presented in Sec. E.

B COSINE-SIMILARITY DISTRIBUTION

To directly show the strengths of our approaches on STS tasks, we illustrate the cosine similarity on embeddings distributions of STS-B pairs in combination with human ratings in Fig. 5. The STS dataset comes in sentence pairs together with correlation labels in the range of 0 and 5, indicating the semantic relatedness of the pairs. Here, the x-axis is the sample similarity of sentences according to human judgment (ground-truth), and the y-axis represents cosine similarity between pairs using

embeddings. Color coding corresponds to ground-truth similarity. Compared to the baseline model (SimCSE), *miCSE* better distinguishes sentence pairs with different levels of similarities, as can be seen from the stronger correlation between embedding distance and human rating. This property leads to better performance on STS tasks. In addition, we observe that *miCSE* generally shows a more scattered distribution while preserving a lower variance on semantically similar sentence pairs. This observation further validates that *miCSE* can potentially achieve a better alignment-uniformity balance.

C VISUALIZATION OF JOINT DISTRIBUTION

To analyze the impact of the proposed approach compared to the baseline SimCSE at the attention level, we visualized the joint distribution of the attention values created by the two views created by the bi-encoder. The joint distribution and mutual information are closely related. More specifically, given two random variables X and Y , the associated mutual information can be expressed in terms of the joint distribution as:

$$I(X, Y) = \sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (6)$$

where $p(x, y)$ denotes the joint-distribution and $p(x), p(y)$ the marginals. Assuming random variables that are normally distributed, the joint distribution of random variables is distinctly shaped depending on the correlation coefficient ρ . See Sec. E details on the relationship between entropy and the correlation coefficient. In the extreme case of totally unrelated marginals $\rho = 0$, the joint distribution assumes a circular shape having the lowest possible mutual information. On the other end of the spectrum, in the case of perfect correlation, the joint distribution assumes collinearity (45° diagonal), with mutual information assuming maximal value. In order to avoid visual clutter, we sliced the attention tensor into 12 slices, pooling together every 3 adjacent heads and every 4 adjacent layers. Slicing the tensor at a higher resolution leads to visually very similar results. The axes of the joint distribution (2d histogram) correspond to the marginals' distribution. As *miCSE* maximizes the mutual information, one can observe a reduction in the scatter of the joint distribution compared to SimCSE.

D DETAILED COMPARISON WITH SIMCSE

Our proposed method is built on top of contrastive learning. Thus it intrinsically relies on the existence of the negative pairs. To complement the performance comparison of contrastive learning in Fig. 3a, we designed an experiment to analyze the extent to which attention regularization alone (AMI) can compensate for the lack of negative pairs. To that end, we conducted training with positive pairs only. See Tab. 3 and Fig. 7 for results. The integration of mutual attention information boosts the performance by up to (+15) across all training set sizes. It suggests the potential application of our proposed attention regularization for non-contrastive learning.

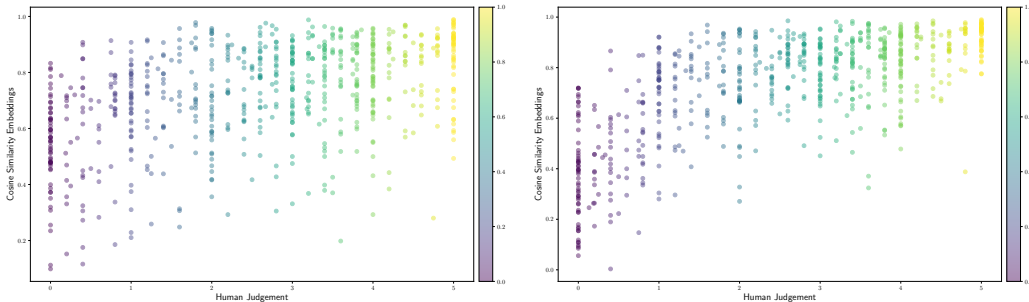


FIGURE 5. Scatter plots of cosine similarities between sentence pairs in STS. Pairs are shown based on ground-truth human scores (higher means more similar) along the x-axis, and the y-axis is the cosine similarity. Color coding corresponds to ground-truth similarity. **Left:** SimCSE, **Right:** *miCSE* (best viewed in color)

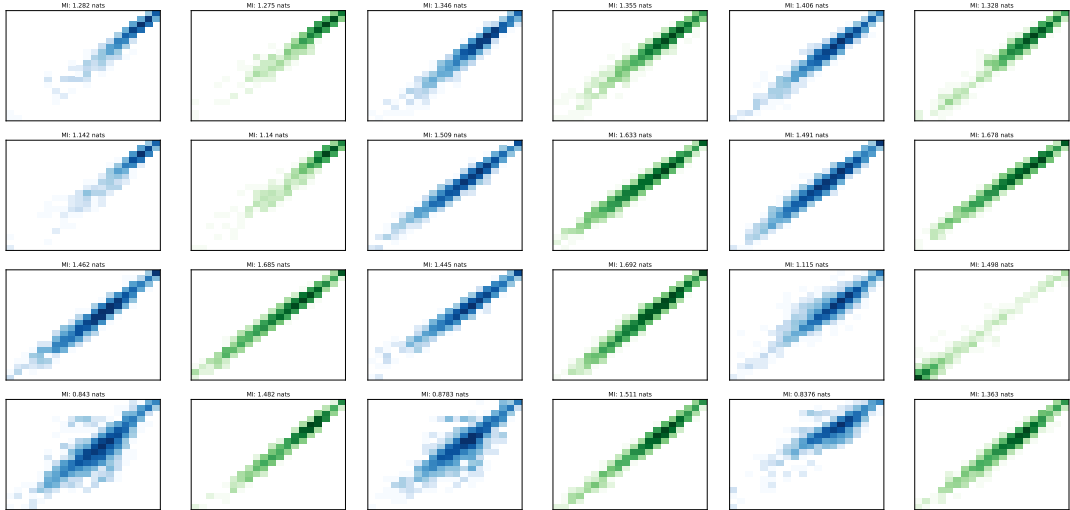


FIGURE 6. Joint distribution between two augmentation induced views. Images depict 12 attention slices per methods, obtained by slicing the attention tensor for the input sentence “the best thing you can do is to know your stuff.” Increasing depth in layer stack from left to right, top to bottom. (●) SimCSE, (●): *mi*CSE(best viewed in color)

Semantic Textual Similarity				
Model	0.1%	1%	10%	100%
SimCSE (with negatives)	66.69 ± 1.03	74.08 ± 0.81	75.01 ± 0.23	76.15
* <i>mi</i> CSE(with negatives)	73.85 ± 0.49	76.21 ± 0.28	76.31 ± 0.46	78.13
SimCSE (w/o negatives)	43.02 ± 4.48	41.30 ± 1.63	42.56 ± 6.87	40.18
* <i>mi</i> CSE(w/o negatives)	57.00 ± 1.32	56.41 ± 3.38	53.38 ± 4.70	54.34

TABLE 3. Sentence embedding few-shot learning performance on STS tasks measured as Spearman’s correlation. Top: performance in contrastive setup with in-batch negatives. Bottom: performance with positive samples only. The number corresponds to the average performance across all benchmarks.

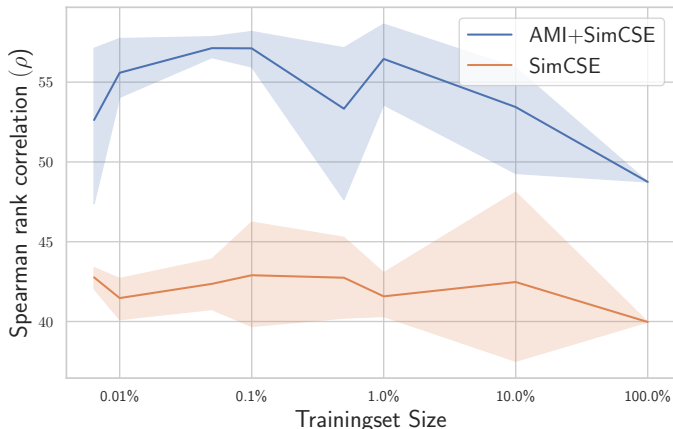


FIGURE 7. Few-shot performance of SimCSE Gao et al. (2021) (—) and the proposed approach AMI in combination with SimCSE (—). Performance is shown in Spearman’s correlation average of the STS benchmark at different ratios of dataset sizes used for training. Training in non-contrastive setting with positive-only pairs.

E BIVARIATE NORMAL MUTUAL INFORMATION

General Log-Normal Properties: Similar to the normal distribution, the log-normal distribution $\log \mathcal{N}(w|\mu_w, \sigma_w^2)$ has two parameters μ_w and σ_w capturing mean and variance. It follows that applying the log transformation on a random variable w , we yield random variable $z = \log(w)$, which is normally distributed: $z \sim \mathcal{N}(\mu_z, \sigma_z^2)$.

Mutual Information: Given a vectors of tuples (X_1, X_2) containing i.i.d. points sampled the joint bivariate normal distribution of $p(A, B) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \in \mathbb{R}^2, \Sigma \in \mathbb{R}^{2 \times 2}$. It can be shown that there exists an exact relationship between mutual information and the correlation coefficient ρ I.M. & A.M. (1957) derived from X_1 and X_2 . To that end, we expand the notation:

$$\boldsymbol{\mu} = (\mu_1 \quad \mu_2), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad (7)$$

The marginal and the join entropy terms for Gaussian distributed variables can be written as:

$$H(X_i) = \frac{1}{2} \log(2\pi e \sigma_i^2) = \frac{1}{2} + \frac{1}{2} \log(2\pi) + \log(\sigma_i), \quad i \in \{1, 2\} \quad (8)$$

$$H(X_1, X_2) = \frac{1}{2} \log [(2\pi e)^2 |\Sigma|] = 1 + \log(2\pi) + \log(\sigma_1\sigma_2) + \frac{1}{2}(1 - \rho^2). \quad (9)$$

Given that Mutual Information can be written in terms of entropy as:

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \quad (10)$$

Then it follows by inserting Eq. 8,9 in Eq. 10:

$$I(X_1, X_2) = -\frac{1}{2}(1 - \rho^2) \quad (11)$$