Unsupervised Sentence Representation via Contrastive Learning with Mixing Negatives

Yanzhao Zhang[1,2], Richong Zhang[1,2*], Samuel Mesh[3], Xudong Liu[1,2], Yongyi Mao[4]

1. SKLSDE, School of Computer Science and Engineering, Beihang University, China
2. Beijing Advanced Institution on Big Data and Brain Computing, Beihang University, China
3. Department of Computer Science, University of Sheffield, UK
4. School of Electrical Engineering and Computer Science, University of Ottawa, Canada

Content

Introduction

- The aim of sentence representation is to train an embedding model maps a sentence to a single vector.
- Recently, contrastive learning(CL) has made great success on unsupervised sentence representation.
- But existing methods only choose random sentence as negative sample, which regard the importance of hard negative.
- In this paper, we first analysis the importance of hard negative, then propose a method MixCSE to construct hard negative.

Traditional Contrastive Learning Framwork

- Encoder f: Mapping a sentence $x_i \in D$ to a feature vector h_i in $\mathbb{R}^d$.
- DataAugment module: Augments the original data or its feature vector to obtain two different views of the data h_i and h_i'.
- Positive sample: Regard h_i as the "anchor feature", then the other view of the same sentence is its positive sample.
- Negative sample: All other sentence will be regard as negative sample.
- The aim of CL is pull positive pairs together and push negative pairs apart:

$$L_{\text{cl}} = -\log \frac{\exp(h_i^T h_i'/\tau)}{\exp(h_i^T h_i'/\tau) + \sum_j^N \exp(h_i^T h_j')/\tau)}$$

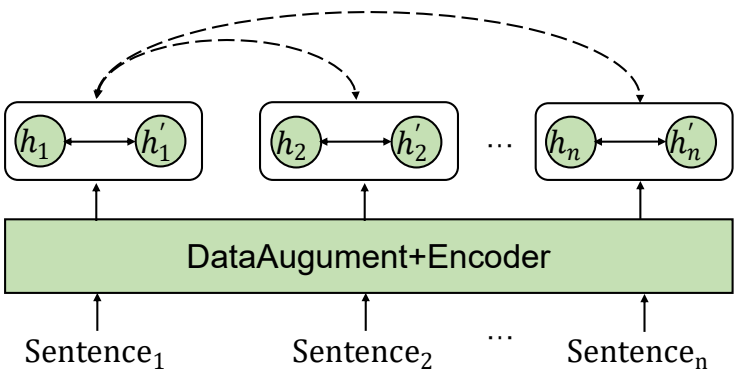


Figure: An illusion of CL.

The importance of Hard Negative

The derivative of L_{cl} with respect to h_i:

$$\frac{\partial L_{\text{cl}}}{\partial h_i} = -\frac{1}{\tau}\left(h_i' + \frac{\exp(h_i^T h_i'/\tau)h_i' + \sum_j^N \exp(h_i^T h_j'/\tau)h_j'}{\exp(h_i^T h_i'/\tau) + \sum_j^M \exp(h_i^T h_j'/\tau)}\right)$$

$$= -\frac{1}{C\tau}\sum_j^N \exp(h_i^T h_j'/\tau)(h_i' - h_j')$$

where $C = \exp(h_i^T h_i') + \sum_j^N \exp(h_i^T h_j'/\tau)$.
We can find that the gradient signal is related to the inner product between h_i and h_j'.
So that a hard negative is important to maintain a strong gradient signal.

Distribution of BERT Embeddings

- Previous work shows that the embeddings obtained from BERT only occupy a sphere cap in the feature space as below.
- h_i can be represented by a set of angles $(\phi_1, \phi_2, \ldots, \phi_{d-1})$, where $\phi_1 \leq \omega$ for some angle $\omega \in (0, \pi)$ and $\phi_2, \ldots, \phi_{d-1}$ are unconstrained. ω is the maxium angle between a point and the Apex direction.

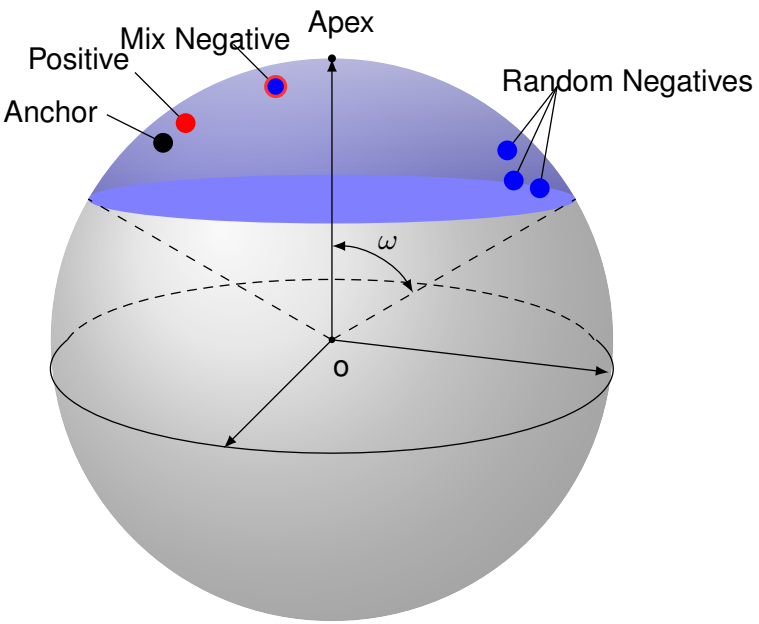


Figure: An illustration of the embedding distribution induced by our method. The mix negative features is closer to the anchor as compared to the random negatives.

Random Negative Sample

Lemma (1)

Suppose that h and h' are two points on the sphere cap $\mathcal{O}_\omega$ and $\measuredangle(h, \mu) = \phi_1$, $\measuredangle(h', \mu) = \phi_1'$, $\measuredangle(\text{proj}(h), \text{proj}(h')) = \beta$ and $\measuredangle(h, h') = \theta$. Then

$$\cos\theta = \cos\phi_1 \cos\phi_1' + sin\phi_1 \sin\phi_1' \cos\beta$$

Lemma (2)

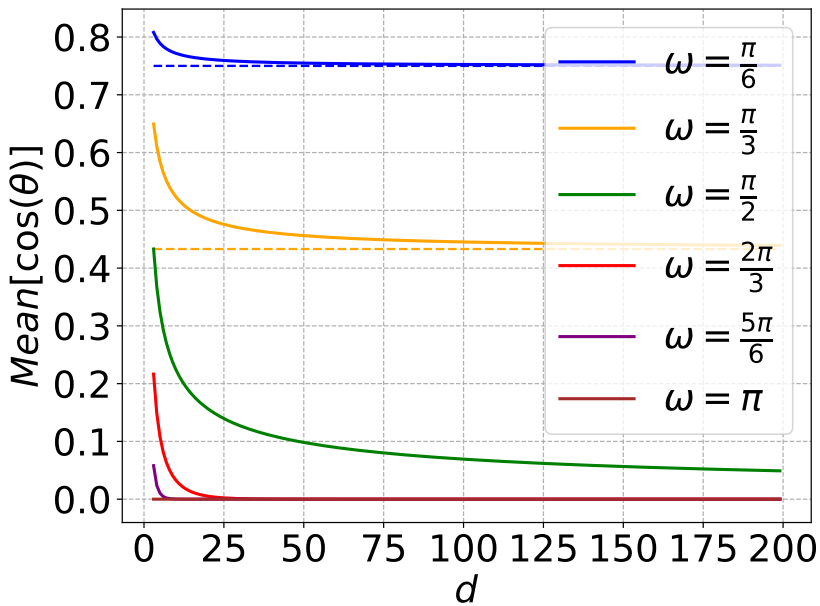
If in Lemma 1, h is fixed and h' is distributed uniformly in the spherical cap $\mathcal{O}_\omega$, the probability density functions of ϕ_1' and β are:

$$P_\phi(\phi_1') = \frac{(\sin\phi_1')^{d-2}}{\int_0^\omega (\sin\phi)^{d-2} d\phi}, \phi_1 \in [0, \omega]$$

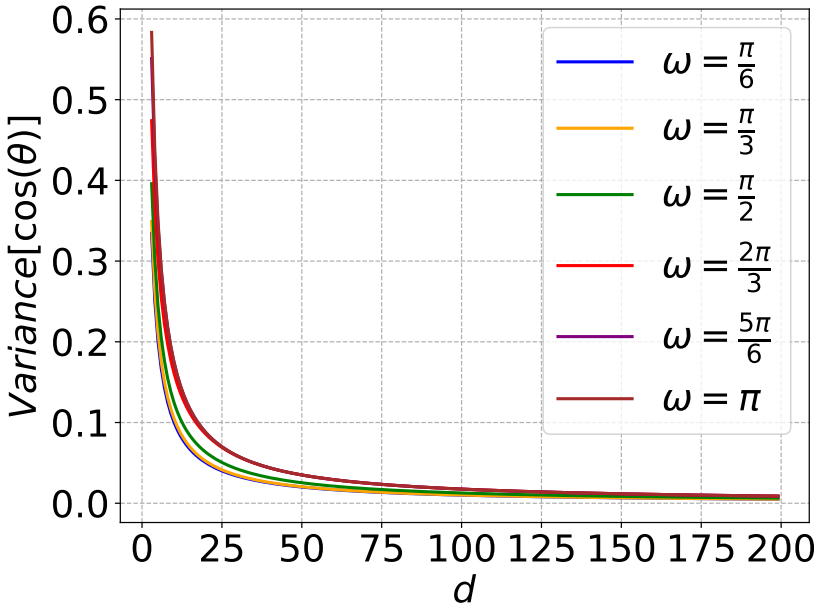
$$P_\beta(\beta) = \frac{(\sin\beta)^{d-2}}{\int_0^\pi (\sin\phi)^{d-2} d\phi}, \beta \in [0, \pi)$$

Random Negative Sample

Using these two lemmas, it is then possible to compute the mean and variance of $\cos\theta$ in the setting of Lemma 2 as shown in below.



(a) Mean of $\cos\theta$



(b) Variance of $\cos\theta$

Conclusion: when d is large enough, there hardly exists any “hard” negative features that are close to the anchor during the training process.

MixCSE

In MixCSE, we construct a negative feature $\widetilde{h'}_{i,j}$ by mixing the positive feature h'_i and a random negative feature h'_j:

$$\widetilde{h'}_{i,j} = \frac{\lambda h'_i + (1-\lambda)h'_j}{||\lambda h'_i + (1-\lambda)h'_j||_2}$$

Then, the CL loss can be rewritten as:

$$L_{\text{mix}} = -\log \frac{\exp\left(\frac{h_i^T h'_i}{\tau}\right)}{\exp\left(\frac{h_i^T h'_i}{\tau}\right) + \sum_j^N \exp\left(\frac{h_i^T h'_j}{\tau}\right) + \exp\left(\frac{h_i^T \text{SG}\left(\widetilde{h'}_{i,j}\right)}{\tau}\right)}.$$

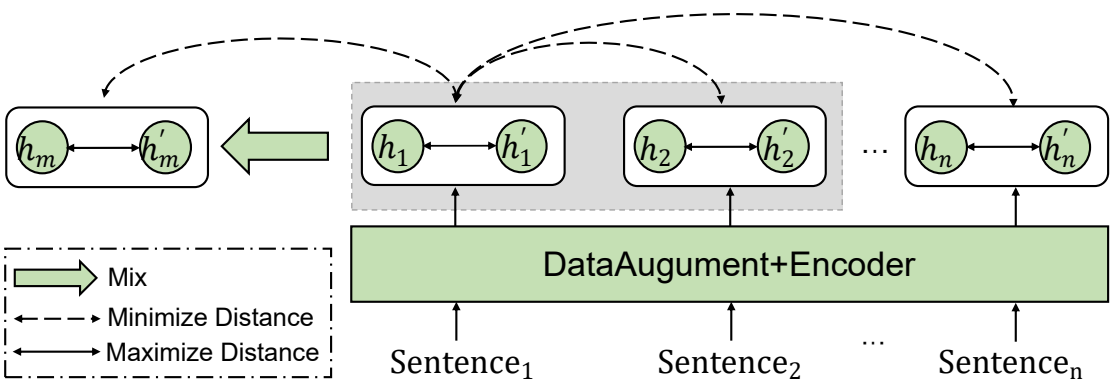


Figure: An illusion of our MixCSE.

MixCSE

The inner product $h_i^T \widetilde{h'}_{i,j}$ of the anchor feature and the mixed negative is

$$h_i^T \widetilde{h'}_{i,j} = \frac{\lambda(h_i^T h_i') + (1-\lambda)(h_i^T h_j')}{||\lambda h_i' + (1-\lambda) h_j'||}$$

When alignment is achieved ($h_i^T h_j' \approx 0$), we have:

$$h_i^T \widetilde{h'}_{i,j} \approx \frac{\lambda}{\sqrt{\lambda^2 + (1-\lambda)^2}}$$

Thus unlike the standard negatives h_j' which gives rise to $h_i^T h_j' \approx 0$, the mixed negatives ensures the inner product value is consistently above zero. Such negative then serve to maintain a **stronger gradient signal.**

MixCSE

Let $L_{\text{mix}}^{\text{no}-\text{sg}}$ represent the contrastive loss with a mixed negative without a stop-gradient. The derivative of $L_{\text{mix}}^{\text{no}-\text{sg}}$ with respect to h_i' is given by

$$\frac{\partial L_{\text{mix}}^{\text{no}-\text{sg}}}{\partial h_i'} = -\frac{1}{C'\tau}((\sum_{j}^{N} \exp(h_i^T h_j'/\tau) + \exp(h_i^T \widetilde{h'}_{i,j}/\tau))h_i^T$$
$$- \frac{\partial \widetilde{h'}_{i,j}}{\partial h_i'} \exp(h_i^T \widetilde{h'}_{i,j}/\tau)h_i^T)$$

We find that if $\widetilde{h'}_{i,j}$ participates in the gradient update, the net effect is that the encoder pushes the positive feature close to the anchor feature. So we stop the gradient of $\widetilde{h'}_{i,j}$.

Experiment

Semantic Textual Similarity (STS): Measuring the similarity between two given sentences.

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg
Avg.Glove	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$	35.20	59.53	49.37	63.39	62.73	48.18	58.60	53.86
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
ConSBERT$_{\text{base}}$	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE-$\text{BERT}_{\text{base}}$	67.17±9.61	79.79±2.72	71.96±3.97	80.21±1.42	77.65±1.24	76.46±1.44	70.57±1.25	74.83±2.32
MixCSE-$\text{BERT}_{\text{base}}$	**71.71±4.04**	**83.14±0.72**	**75.49±1.25**	**83.64±2.32**	**79.00±0.16**	**78.48±0.82**	**72.19±0.46**	**77.66±0.61**
$\text{BERT}_{\text{large}}$	33.06	57.64	47.95	55.83	62.42	49.66	53.87	51.49
$\text{BERT}_{\text{large}}$-flow	65.20	73.39	69.42	74.92	77.63	72.26	62.50	70.76
$\text{BERT}_{\text{large}}$-whitening	64.35	74.60	69.64	74.68	75.94	60.81	72.47	70.35
ConSBERT$_{\text{large}}$	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE-$\text{BERT}_{\text{large}}$	70.21±1.49	83.97±1.18	75.92±0.56	83.9±0.49	78.87±0.75	79.0±1.0	73.89±1.08	77.97±0.7
MixCSE-$\text{BERT}_{\text{large}}$	**72.55±0.49**	**84.32±0.53**	**76.69±0.76**	**84.31±0.10**	**79.67±0.28**	**79.90±0.18**	**74.07±0.13**	**78.80±0.09**

Table: Results on the STS datasets. We implement and reproduce results of SimCSE, and report the average and standard variance of its results. The performances of other comparing models are from their original papers.

Experiment

Transfer Learning(TR): Verifies the quality of the sentence representation on downstream tasks.

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg
Glov.Avg	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
$\text{BERT}_{\text{base}}$	78.66	86.25	94.37	88.66	84.40	92.80	69.54	84.94
$\text{IS-BERT}_{\text{base}}$	81.09	**87.18**	94.96	88.75	85.96	**88.64**	74.24	85.83
$\text{SimCSE-BERT}_{\text{base}}$	71.12±5.94	85.92±0.41	98.56±2.05	88.61±0.18	85.34±0.37	88.4±0.59	73.48±1.26	84.49±0.59
$\text{MixCSE-BERT}_{\text{base}}$	**81.3±1.75**	86.77±0.4	**99.64±0.01**	**89.71±0.17**	85.87±0.49	84.91±0.24	**76.08±0.68**	**86.33±0.26**
$\text{BERT}_{\text{large}}$	60.89	90.15	99.62	86.04	89.95	93.00	69.86	84.22
$\text{SimCSE-BERT}_{\text{large}}$	73.93±2.79	88.87±0.75	99.6±0	89.49±0.16	90.59±0.96	**91.72**±0.95	75.49±0.88	86.8±0.42
$\text{MixCSE-BERT}_{\text{large}}$	**82.95±0.52**	**89.57±0.05**	**99.67±0.01**	**90.14±0.02**	**89.17±0.84**	86.13±0.58	**76.74±0.16**	**87.77±0.11**

Table: Results on the TR datasets. Bold numbers indicate best performance based on the same pretrained model.

Experiment

We analysis the change for positive, negative and mix scores (Fig a) and the change for positive, negative and mix scores (Fig b) during training.

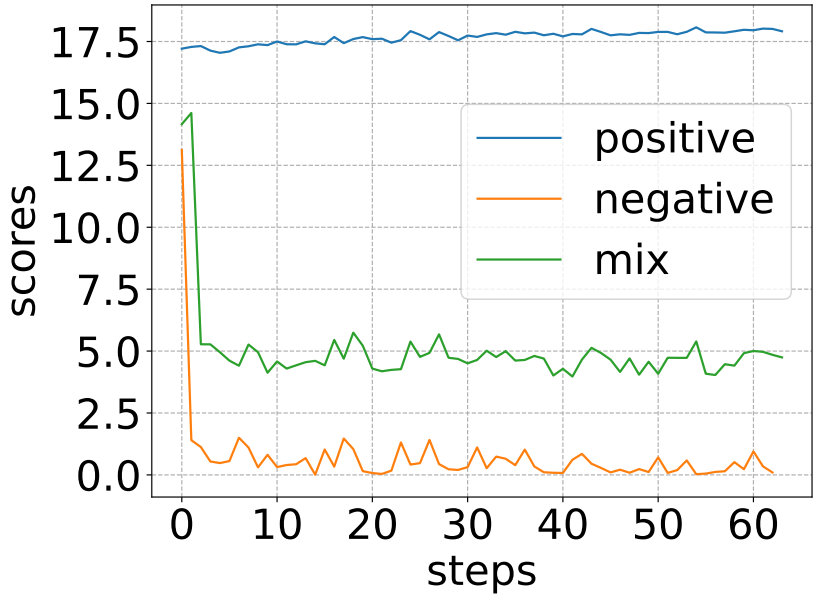


Fig (a)

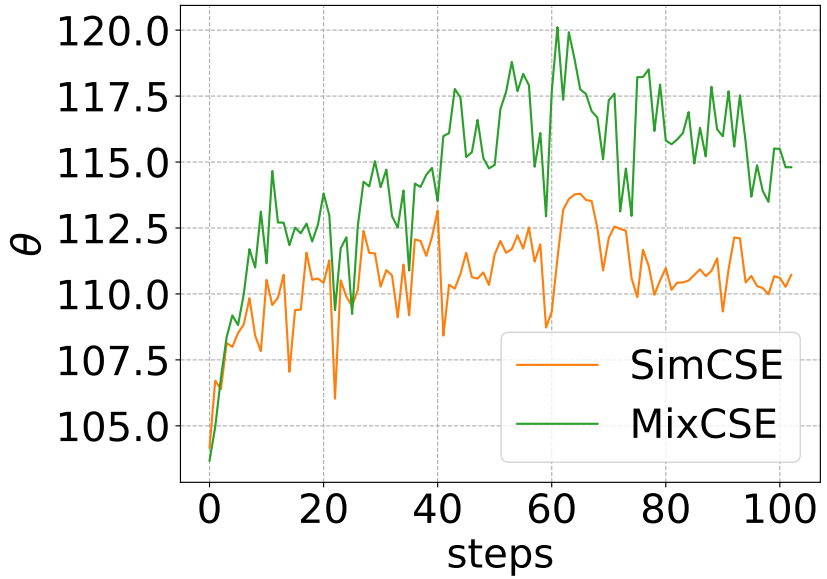


Fig (b)

We can find that mixing negatives can keep a higher similarity scores compared with random sample and MixCSE can get a more uniformity distribution of sentence embedding.

Experiment

Alignment (expected distance between positive features) and Uniformity (expected distances between two random features) are two widely used metric for sentence representation:

$$L_{\text{align}} \triangleq \mathbb{E}_{(x,y)\ P_pos(x,y)}[||f(x) - f(y)||_2^2]$$

$$L_{\text{uniform}} \triangleq \log \mathbb{E}_{(x,y)\ P_data(x,y)}[e^{-2||f(x)-f(y)||_2^2}]$$

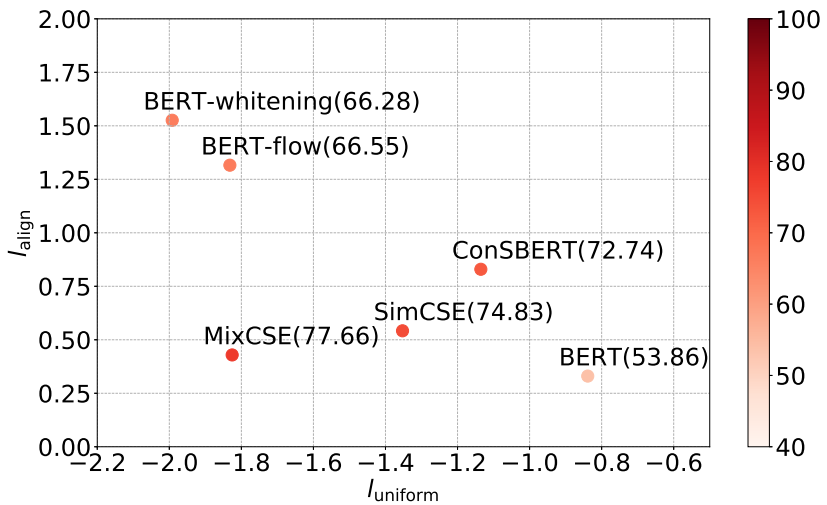


Figure: Alignment and uniformity for different sentence embedding methods measured on the STS-B dev set. Color of points represent average STS performance.

Contribution

1. We prove that hard negatives are important for sentence representation learning and random sampling is not effective for choosing hard negatives even with many repeats of random sampling.
2. We propose MixCSE that constructs hard negatives by mixing the positive features and random negative features for sentence representation.
3. We demonstrate through extensive experiments that MixCSE achieves state-of-the-art results on semantic textual similarity (STS) and transfer tasks (TR).

Thank You for listening