Sentence Embedding

Wei Wang

2021.03

Overview

- Sentence Embedding: learning semantically meaningful representations for each sentence.
- Unsupervised Sentence Embedding: unsupervised learning goals.

  (reconstruct self or surrounding sentences)

  SDAE, FastSent, QT, IS-BERT
- Supervised Sentence Embedding: labeled data.

  InferSent, USE, SBERT

Overview

- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks(EMNLP 2019)(SBERT)
- DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)
- An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)(IS-BERT)

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

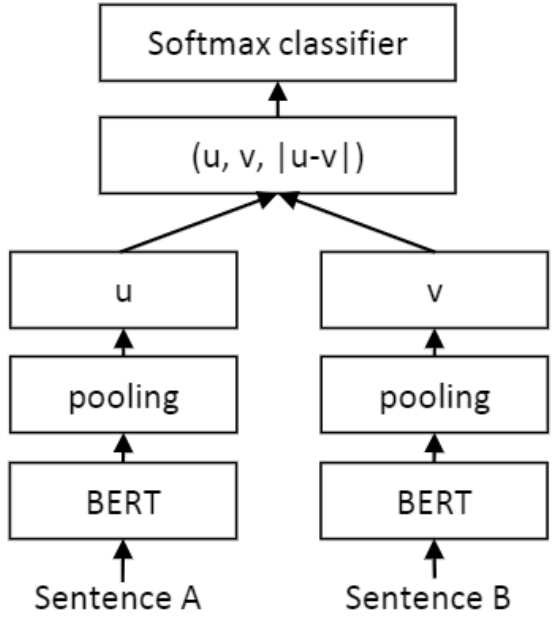
Introduction

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks(EMNLP 2019)

- BERT set new state-of-the-art performance on various sentence classification and sentence-pair regression tasks.
- BERT uses a cross-encoder: Two sentences are passed to the transformer network and the target value is predicted.
- However, this setup is unsuitable for various pair regression tasks due to too many possible combinations.
- Sentence-BERT (SBERT), a modification of the BERT network using siamese and triplet networks that is able to derive semantically meaningful sentence embeddings, which can be used for large-scale semantic similarity comparison.

Method

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks(EMNLP 2019)



- SBERT adds a pooling operation on top of the encoder to derive a fixed sized sentence embedding.

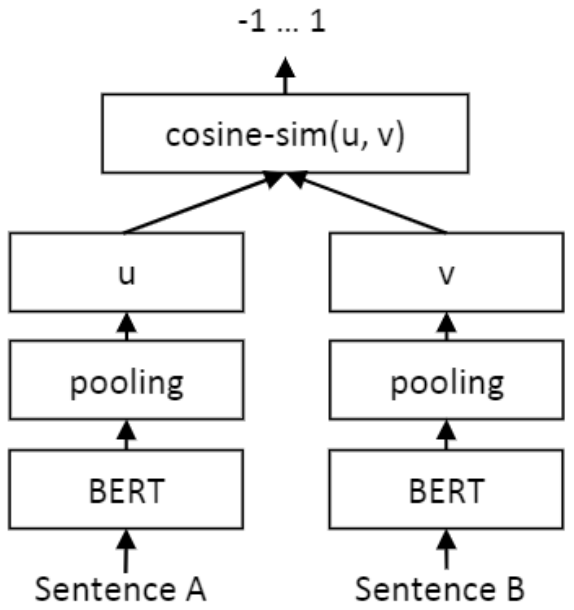
  CLS, MEAN, MAX

- Three objective functions.

Classification $o = \text{softmax}(W_t(u, v, |u - v|))$

Regression cosine-sim(u, v)

Triplet $max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$



- Trained on SNLI and Multi-Genre NLI

Experiment-Semantic Textual Similarity (Unsupervised-STS)

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks(EMNLP 2019)

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

Table 1: Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels

- Using the output of BERT leads to rather poor performances.
- The proposed method outperforms both InferSent and Universal Sentence Encoder substantially.
- We only observe minor difference between SBERT and SRoBERTa for generating sentence embeddings.

Experiment-Semantic Textual Similarity (Supervised-STS)

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

- The author uses the training set to fine-tune SBERT using the regression objective function. At prediction time, they compute the cosine-similarity between the sentence embeddings.
- This two-step approach had an especially large impact for the BERT cross-encoder.
- There is not a significant difference between BERT and RoBERTa.

Model	Spearman
Not trained for STS	
Avg. GloVe embeddings	58.02
Avg. BERT embeddings	46.35
InferSent - GloVe	68.03
Universal Sentence Encoder	74.92
SBERT-NLI-base	77.03
SBERT-NLI-large	79.23
Trained on STS benchmark dataset	
BERT-STSb-base	84.30 $\pm$ 0.76
SBERT-STSb-base	84.67 $\pm$ 0.19
SRoBERTa-STSb-base	**84.92** $\pm$ 0.34
BERT-STSb-large	**85.64** $\pm$ 0.81
SBERT-STSb-large	84.45 $\pm$ 0.43
SRoBERTa-STSb-large	85.02 $\pm$ 0.76
Trained on NLI data + STS benchmark data	
BERT-NLI-STSb-base	**88.33** $\pm$ 0.19
SBERT-NLI-STSb-base	85.35 $\pm$ 0.17
SRoBERTa-NLI-STSb-base	84.79 $\pm$ 0.38
BERT-NLI-STSb-large	**88.77** $\pm$ 0.46
SBERT-NLI-STSb-large	86.10 $\pm$ 0.13
SRoBERTa-NLI-STSb-large	86.15 $\pm$ 0.35

Experiment-SentEval

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks(EMNLP 2019)

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Avg. GloVe embeddings	77.25	78.30	91.17	87.85	80.18	83.0	72.87	81.52
Avg. fast-text embeddings	77.96	79.23	91.68	87.81	82.15	83.6	74.49	82.42
Avg. BERT embeddings	78.66	86.25	94.37	88.66	84.40	92.8	69.45	84.94
BERT CLS-vector	78.68	84.85	94.21	88.23	84.13	91.4	71.13	84.66
InferSent - GloVe	81.57	86.54	92.50	**90.38**	84.18	88.2	75.77	85.59
Universal Sentence Encoder	80.09	85.19	93.98	86.70	86.38	**93.2**	70.14	85.10
SBERT-NLI-base	83.64	89.43	94.39	89.86	88.96	89.6	**76.00**	87.41
SBERT-NLI-large	**84.88**	**90.07**	**94.52**	90.33	**90.66**	87.4	75.94	**87.69**

- SBERT is able to achieve the best performance in 5 out of 7 tasks.
- Even though transfer learning is not the purpose of SBERT, it outperforms other state-of-the-art sentence embeddings methods on this task.
- Average BERT embeddings / CLS-token output from BERT return sentence embeddings that are infeasible to be used with cosine similarity or with Euclidean distance.

Experiment-Ablation Study

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks(EMNLP 2019)

- When trained with the classification objective function on NLI data, the pooling strategy has a rather minor impact. The impact of the concatenation mode is much larger.
- The most important component is the elementwise difference |u − v|.
- When trained with the regression objective function, we observe that the pooling strategy has a large impact. There, the MAX strategy perform significantly worse than MEAN or CLS-token strategy.

	NLI	STSb
Pooling Strategy		
MEAN	**80.78**	**87.44**
MAX	79.07	69.92
CLS	79.80	86.62
Concatenation		
(u, v)	66.04	-
$(\|u - v\|)$	69.78	-
$(u * v)$	70.54	-
$(\|u - v\|, u * v)$	78.37	-
$(u, v, u * v)$	77.44	-
$(u, v, \|u - v\|)$	**80.78**	-
$(u, v, \|u - v\|, u * v)$	80.44	-

Table 6: SBERT trained on NLI data with the classification objective function, on the STS benchmark (STSb) with the regression objective function. Configurations are evaluated on the development set of the STSb using cosine-similarity and Spearman's rank correlation. For the concatenation methods, we only report scores with MEAN pooling strategy.

Summary

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks(EMNLP 2019)

- SBERT achieve a significant improvement over state-of-the-art sentence embeddings methods.
- Replacing BERT with RoBERTa did not yield a significant improvement in our experiments.
- SBERT is computationally efficient.

- Compared with InferSent, mainly replacing BiLSTM encoder with BERT.
- NLI is a high-level understanding task that involves reasoning about the semantic relationships within sentences.
- Better supervised sentence embedding?
  - better encoder than BERT
  - better labeled data than NLI

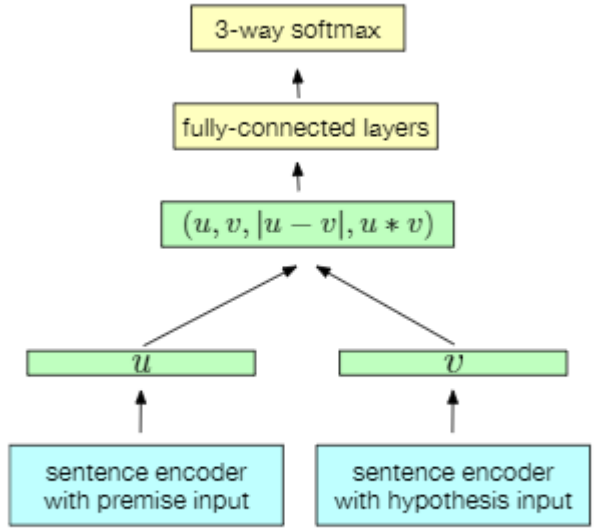


Figure 1: **Generic NLI training scheme.**

DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations

John M. Giorgi
University of Toronto
Department of Computer Science
The Donnelly Centre
Vector Institute
Toronto, ON M5G 1M1, Canada
john.giorgi@utoronto.ca

Osvald Nitski
University of Toronto
Department of Mechanical and Industrial Engineering
Peter Munk Cardiac Center, University Health Network
Toronto, ON M5S 3G8, Canada
osvald.nitski@mail.utoronto.ca

Gary D. Bader
University of Toronto
Department of Molecular Genetics
Department of Computer Science
The Donnelly Centre
Toronto, ON M5S 3E1, Canada
gary.bader@utoronto.ca

Bo Wang
University of Toronto
Peter Munk Cardiac Center, University Health Network
CIFAR AI Chair, Vector Institute
Toronto, ON M5G 1M1, Canada
bowang@vectorinstitute.ai

Introduction

Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)

- Recent work has demonstrated strong transfer task performance using pretrained sentence-level embeddings.
- However, the highest performing solutions require at least some labelled data, limiting their usefulness to languages and domains where labelled data exists for the chosen pretraining tasks.
- We propose a self-supervised, contrastive objective that can be used alongside MLM to pretrain a transformer.

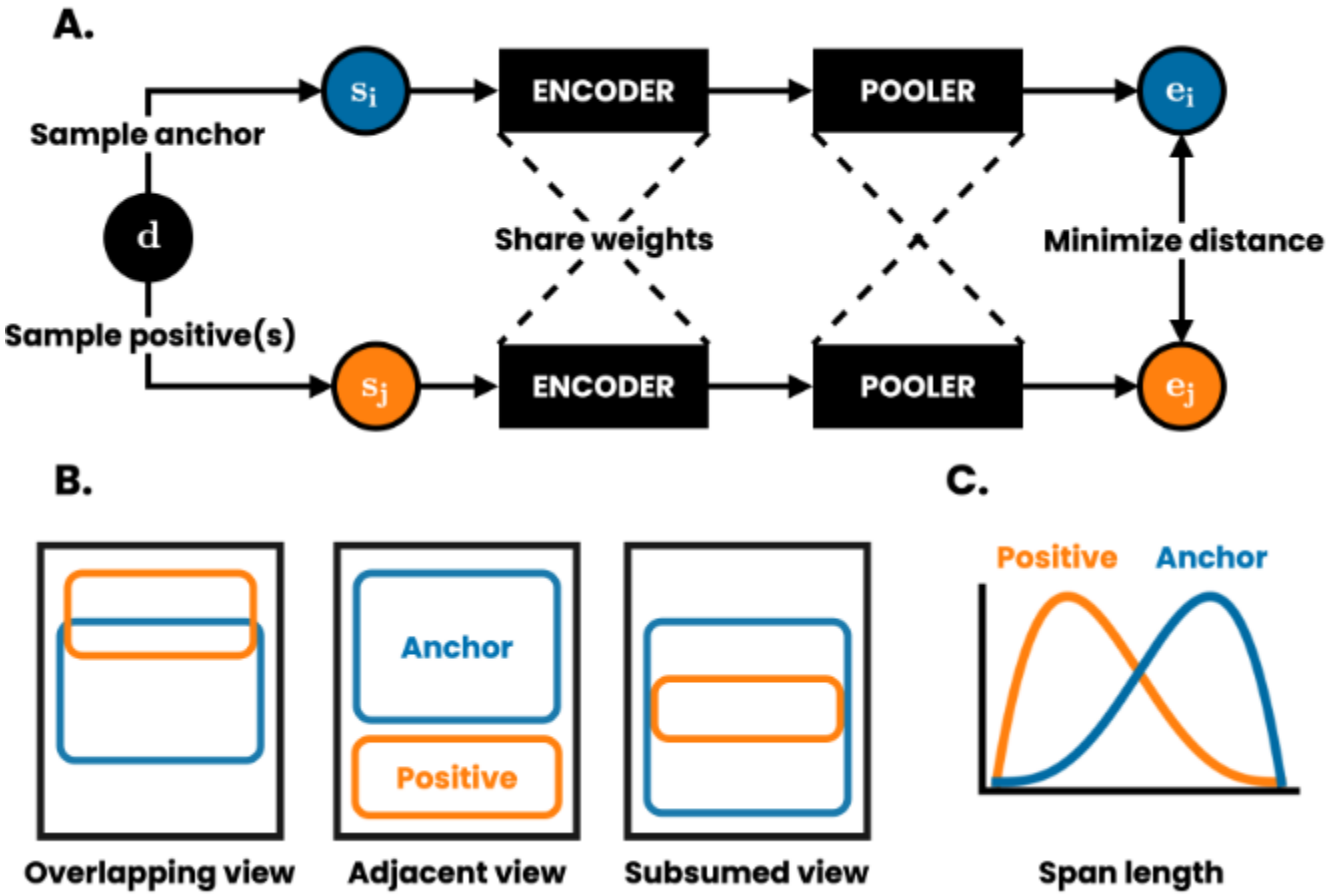
Method

Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)

- For each document d in a minibatch of size N, we sample A anchor spans per document and P positive spans per anchor.
- Positive spans can overlap with, be adjacent to or be subsumed by the sampled anchor span.
- The length of anchors and positives are randomly sampled from beta distributions.

$$\mathcal{L}_{\text{contrastive}} = \sum_{i=1}^{AN} \ell(i, i+AN) + \ell(i+AN, i)$$

$$\ell(i,j) = -\log \frac{\exp(\text{sim}(\boldsymbol{e}_i, \boldsymbol{e}_j)/\tau)}{\sum_{k=1}^{2AN} \mathbb{1}_{[i \neq k]} \cdot \exp(\text{sim}(\boldsymbol{e}_i, \boldsymbol{e}_k)/\tau)}$$



$$\mathcal{L} = \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{MLM}}$$

Method

Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)

The sampling procedure produces three types of positives: positives that partially overlap with the anchor, positives adjacent to the anchor, and positives subsumed by the anchor,

and two types of negatives: easy negatives sampled from a different document than the anchor, and hard negatives sampled from the same document as the anchor.

$$\ell_{\text{anchor}} = \lfloor p_{\text{anchor}} \times (\ell_{\max} - \ell_{\min}) + \ell_{\min} \rfloor$$
$$s_i^{\text{start}} \sim \{0 \dots n - \ell_{\text{anchor}}\}$$
$$s_i^{\text{end}} = s_i^{\text{start}} + \ell_{\text{anchor}}$$
$$\boldsymbol{s}_i = \boldsymbol{x}^d_{s_i^{\text{start}}:s_i^{\text{end}}}$$

$$\ell_{\text{positive}} = \lfloor p_{\text{positive}} \times (\ell_{\max} - \ell_{\min}) + \ell_{\min} \rfloor$$
$$s_{i+pAN}^{\text{start}} \sim \{s_i^{\text{start}} - \ell_{\text{positive}} \dots s_i^{\text{end}}\}$$
$$s_{i+pAN}^{\text{end}} = s_{i+pAN}^{\text{start}} + \ell_{\text{positive}}$$
$$\boldsymbol{s}_{i+pAN} = \boldsymbol{x}^d_{s_{i+pAN}^{\text{start}}:s_{i+pAN}^{\text{end}}}$$

Experiment

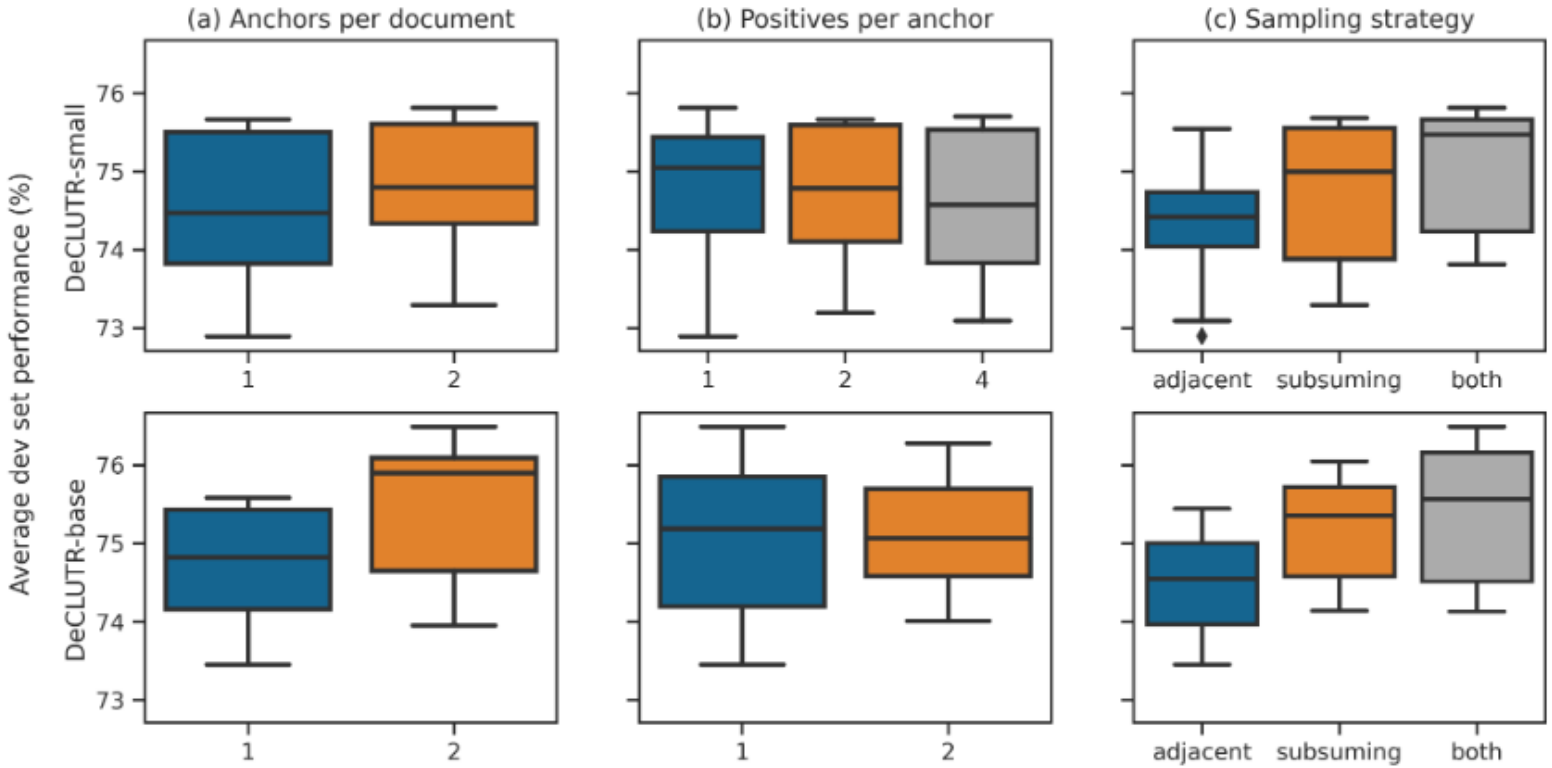
Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)

Model	Parameters	Embed. dim.	SentEval			
			Downstream	Probing	Avg.	Δ
Bag-of-words (BoW) weak baselines						
GloVe	–	300	65.50	62.42	63.96	-12.02
fastText	–	300	68.57	63.16	65.87	-10.11
Supervised and semi-supervised						
InferSent	38M	4096	76.46	72.58	74.52	-1.46
USE	147M	512	**79.13**	66.70	72.91	-3.06
Sentence Transformers	125M	768	77.59	63.22	70.40	-5.57
Unsupervised						
Transformer-small	82M	768	72.69	**74.27**	73.48	-2.50
Transformer-base	125M	768	72.22	73.38	72.80	-3.18
DeCLUTR-small (ours)	82M	768	76.80	73.84	75.32	-0.66
DeCLUTR-base (ours)	125M	768	78.16	73.80	**75.98**	–

- Both DeCLUTR-small and DeCLUTR-base significantly boost downstream task performance while maintaining high probing task performance.

Experiment

Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)



- sampling multiple anchors per document has a large positive impact on the quality of the learned embeddings.
- a positive sampling srategy that allows positives to be adjacent to and subsumed by the anchor outperforms a strategy which only allows adjacent or subsuming views.

Experiment

Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)

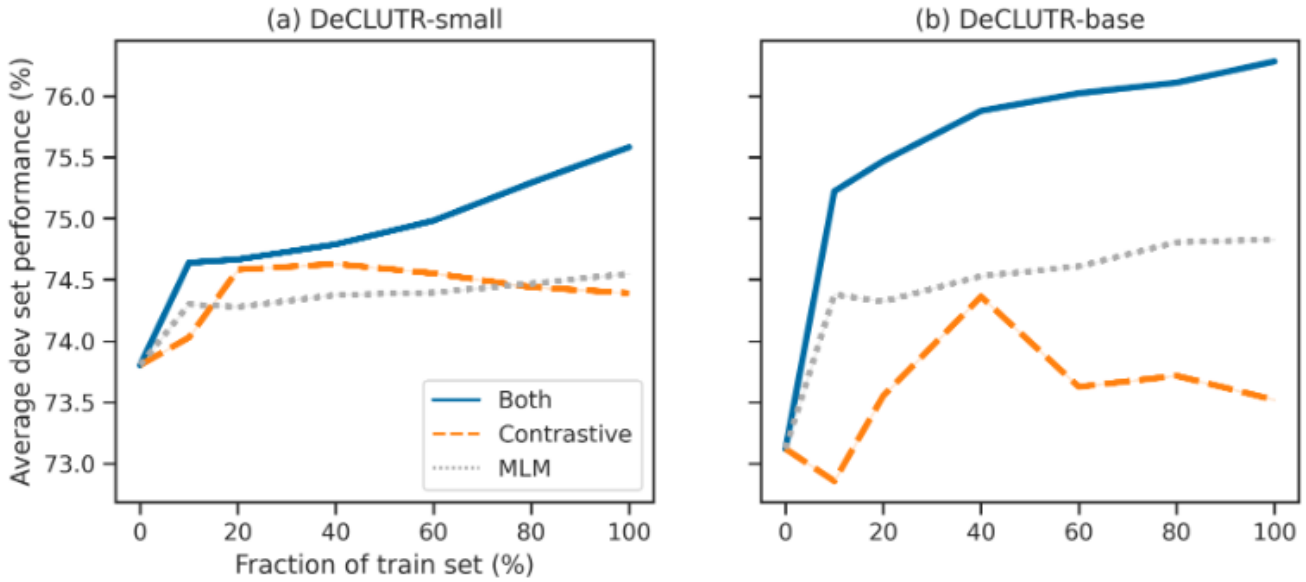


Figure 3: Effect of training objective, train set size and model capacity on SentEval performance. DeCLUTR-small has 6 layers and ∼82M parameters; DeCLUTR-base has 12 layers and ∼125M parameters. Average scores are reported from the development set of the SentEval benchmark. 100% corresponds to 1 epoch of training with all 495,243 documents from our OpenWebText subset.

- pretraining the model with both the MLM and contrastive objectives improves performance over training with either objective alone.

Summary

Deep Contrastive Learning for Unsupervised Textual Representations(Arxiv 2020)

- The author proposed a self-supervised objective for learning universal sentence representations. The objective is conceptually simple, easy to implement, and applicable to any text encoder.
- Results on the SentEval benchmark demonstrated the effectiveness of the proposed method.

- Need document example which contain many spans.
- Not learning representation of a complete sentence.
- Better unsupervised sentence embedding?
  - Better positives and better negatives.

An Unsupervised Sentence Embedding Method by Mutual Information Maximization

Yan Zhang[1*†], **Ruidan He**[2*], **Zuozhu Liu**[3], **Kwan Hui Lim**[1], **Lidong Bing**[2]
[1]Singapore University of Technology and Design
[2]DAMO Academy, Alibaba Group
[3]ZJU-UIUC Institue

Introduction

An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)

- BERT is inefficient for sentence-pair tasks such as clustering or semantic search as it needs to evaluate combinatorially many sentence pairs which is very time-consuming.
- SBERT is trained on corpus with high-quality labeled sentence pairs, which limits its application to tasks where labeled data is extremely scarce.
- The author proposes a novel unsupervised sentence embedding model with light-weight feature extractor on top of BERT for sentence encoding, and train it with a novel self-supervised learning objective.

Method

An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)

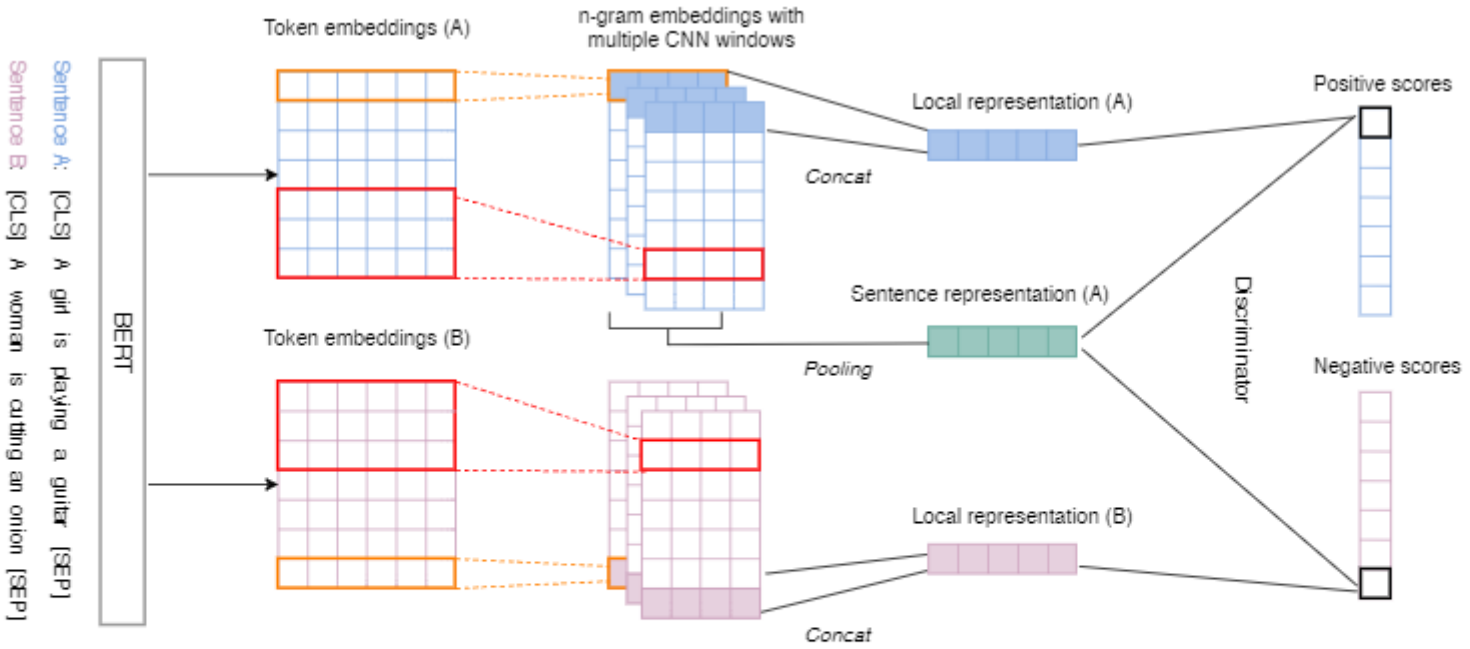


Figure 1: Model Architecture. Two sentences are encoded by BERT and multiple CNNs with different window sizes to get concatenated local n-gram token embeddings. A discriminator T takes all pairs of {sentence representation, token representation} as input and decides whether they are from the same sentence. In this example, we treat sentence "A" as the positive sample and "B" as negative, then n-gram embeddings of "A" will be summarized to a global sentence embedding via pooling. The discriminator produces scores for all token representations from both "A" and "B" to maximize the MI estimator in Eq.2.

Method

An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)

- The learning objective is to maximize the mutual information (MI) between the global sentence representation E(x) and each of its local token representation $F^i(x)$.
- the Jensen-Shannon estimator is defined as

$$\begin{aligned}\widehat{\mathcal{I}}_{\omega}^{JSD}(\mathcal{F}_{\theta}^{(i)}(\mathbf{x});\mathcal{E}_{\theta}(\mathbf{x})) :=& \\ E_{\mathbb{P}}[-sp(-T_{\omega}(\mathcal{F}_{\theta}^{(i)}(\mathbf{x}),\mathcal{E}_{\theta}(\mathbf{x})))] & \\ - E_{\mathbb{P}\times\tilde{\mathbb{P}}}[sp(T_{\omega}(\mathcal{F}_{\theta}^{(i)}(\mathbf{x}'),\mathcal{E}_{\theta}(\mathbf{x})))],&\end{aligned} \tag{2}$$

- The end-goal learning objective over the whole dataset $\mathcal{X}$ is defined as

$$\omega^*,\theta^* = \underset{\omega,\theta}{\operatorname{argmax}} \frac{1}{|\mathcal{X}|}\Big(\sum_{\mathbf{x}\in\mathcal{X}}\sum_{i=1}^{l_x}\widehat{\mathcal{I}}_{\omega}^{JSD}(\mathcal{F}_{\theta}^{(i)}(\mathbf{x});\mathcal{E}_{\theta}(\mathbf{x}))\Big), \tag{3}$$

Experiment-Unsupervised Evaluations-Unsupervised STS

An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Using unlabeled data (unsupervised methods)								
Unigram-TFIDF†	-	-	58.00	-	-	-	52.00	-
SDAE†	-	-	12.00	-	-	-	46.00	-
ParagraphVec DBOW†	-	-	43.00	-	-	-	42.00	-
ParagraphVec DM†	-	-	44.00	-	-	-	44.00	-
SkipThought†	-	-	27.00	-	-	-	57.00	-
FastSent†	-	-	**63.00**	-	-	-	61.00	-
Avg. GloVe embeddings‡	55.14	**70.66**	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings‡	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector‡	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
Ours: IS-BERT-NLI	**56.77**	69.24	61.21	**75.23**	**70.16**	**69.21**	**64.25**	**66.58**
Using labeled NLI data (supervised methods)								
InferSent - GloVe‡	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
USE‡	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
SBERT-NLI‡	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89

- BERT out-of-the-box gives surely poor results on STS tasks. all supervised methods outperform other unsupervised baselines. As expected, IS-BERT-NLI is in general inferior to these two supervised baselines.

Experiment-Supervised Evaluations-SentEval

An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Using unlabeled data (unsupervised methods)								
Unigram-TFIDF†	73.7	79.2	90.3	82.4	-	85.0	73.6	-
SDAE†	74.6	78.0	90.8	86.9	-	78.4	73.7	-
ParagraphVec DBOW†	60.2	66.9	76.3	70.7	-	59.4	72.9	-
SkipThought†	76.5	80.1	93.6	87.1	82.0	92.2	73.0	83.50
FastSent†	70.8	78.4	88.7	80.6	-	76.8	72.2	-
Avg. GloVe embeddings‡	77.25	78.30	91.17	87.85	80.18	83.0	72.87	81.52
Avg. BERT embeddings‡	78.66	86.25	94.37	88.66	84.40	**92.8**	69.54	84.94
BERT CLS-vector‡	78.68	84.85	94.21	88.23	84.13	91.4	71.13	84.66
Ours: IS-BERT-task	**81.09**	**87.18**	**94.96**	**88.75**	**85.96**	88.64	**74.24**	**85.91**
Using labeled NLI data (supervised methods)								
InferSent - GloVe‡	81.57	86.54	92.50	90.38	84.18	88.2	75.77	85.59
USE‡	80.09	85.19	93.98	86.70	86.38	93.2	70.14	85.10
SBERT-NLI‡	83.64	89.43	94.39	89.86	88.96	89.6	76.00	87.41

- IS-BERT-task is able to outperform other unsupervised baselines on 6 out of 7 tasks, and it is on par with InferSent and USE which are strong supervised baselines trained on NLI task

Experiment-Supervised Evaluations-Supervised STS

An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)

- BERT and SBERT performs similarly on this task. IS-BERT-STSb (ssl+ft) outperforms both baselines.
- When directly fine-tuning IS-BERT on the labeled data, it performs much worse than SBERT.
- However, when comparing IS-BERT-STSb(ft) with IS-BERT-STSb(ssl+ft), adding self-supervised learning before fine-tuning leads to more than 10% performance improvements.

Model	ρ
BERT-STSb	84.30
SBERT-STSb	84.67
Ours: IS-BERT-STSb (ft)	74.76
Ours: IS-BERT-STSb (ssl + ft)	**85.04**

Table 4: Spearman's rank correlation ρ on the STSb test set. Results of baselines are extracted from (Reimers and Gurevych, 2019)

Summary

An Unsupervised Sentence Embedding Method by Mutual Information Maximization(EMNLP 2020)

- The author proposed IS-BERT for unsupervised sentence representation learning with a novel MI maximization objective.
- IS-BERT outperforms all unsupervised sentence embedding baselines on various tasks and is competitive with supervised sentence embedding methods in certain scenarios.
- IS-BERT achieves substantially better results in this scenario as it has the flexibility to be trained on the task-specific corpus without label restriction.

- A new way of constructing positive and negative cases: global-to-local.
- Better unsupervised sentence embedding?
  - Better positives and better negatives.

Thanks