IJCNLP 2019

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

Presented by Nikita Nikolaev
BDA&AI Master's Degree student

Introduction

What are we going to talk about?

- BERT and RoBERTa: previous state-of-the-art
- Semantic Textual Similarity (STS) and other tasks
- BERT’s overhead
- **Brand new Sentence-BERT: modifications with siamese and triplet networks**
- Highlights, results, comparisons, etc.

Related Works

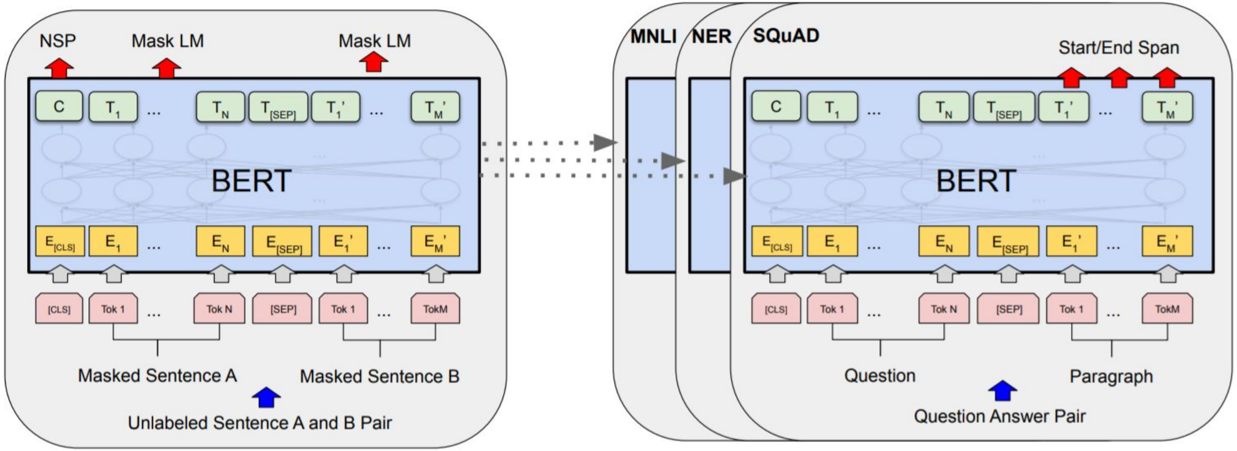
BERT: Bidirectional Encoder Representation from Transformers

Highlights:

- Question answering
- Sentence classification
- Sentence-pair regression

- Input - two sentences, separated by a special [SEP] token

- Multi-head Attention (12 or 24 layers)



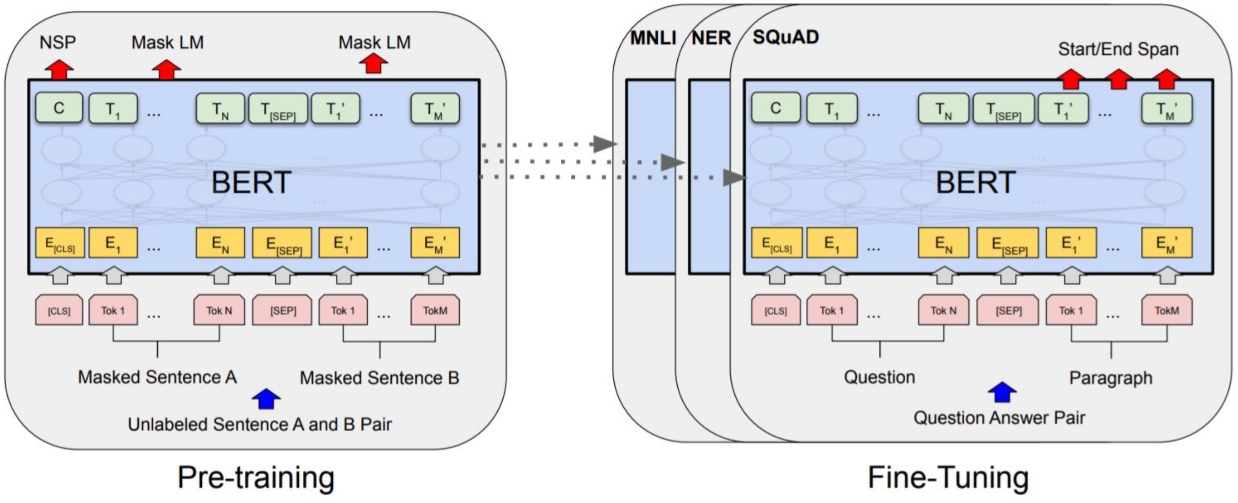
Pre-training
Fine-Tuning

RoBERTa: Robustly **O**ptimized **BERT A**pproach

Highlights:

- Iterates on BERT's pretraining procedure

- Training the model longer
- With bigger batches
- Over more data
- Removing the next sentence prediction objective
- Training on longer sequences
- Dynamically changing the masking pattern applied to the training data

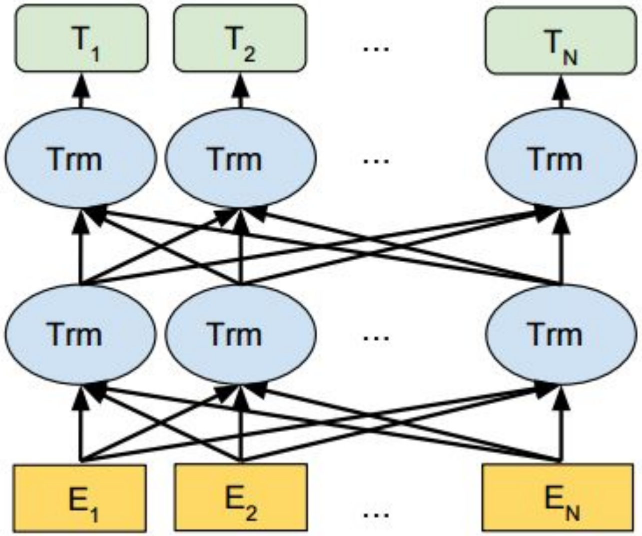


SotA sentence embeddings

- Skip-Thought
- InferSent (outperforms previous)
- Universal Sentence Encoder
- etc.

Comparisons will be available a bit later!

BERT’s disadvantages?



Sentence-BERT: Model Overview

Sentence-BERT

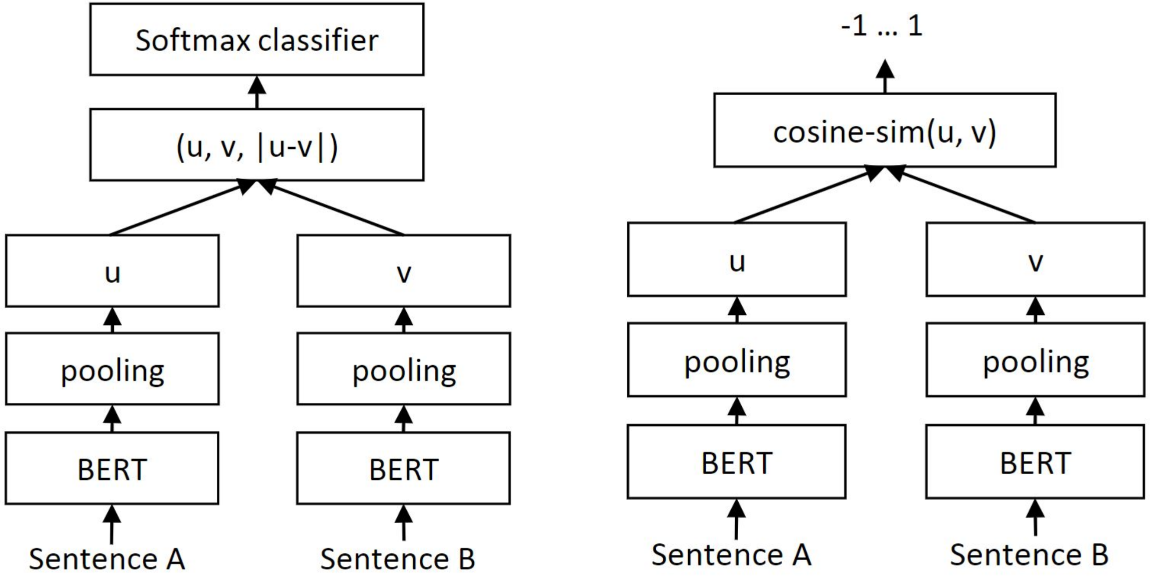
SBERT adds a pooling operation to the output of BERT / RoBERTa to derive a fixed sized sentence embedding.

Strategies:

- Using the output of [CLS] token
- Computing the mean of all output vectors (MEAN-strategy)
- Computing a max-over-time of the output vectors (MAX-strategy)

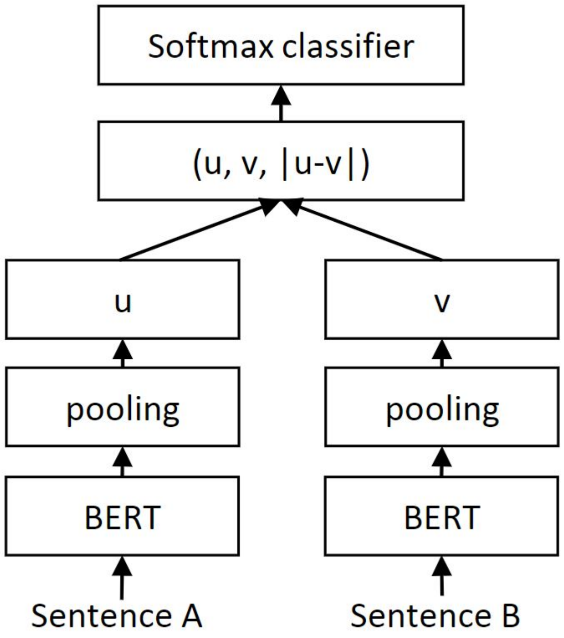
SBERT can be tuned in **less than 20 minutes**, while yielding better results than comparable sentence embedding methods.

Siamese architecture overview



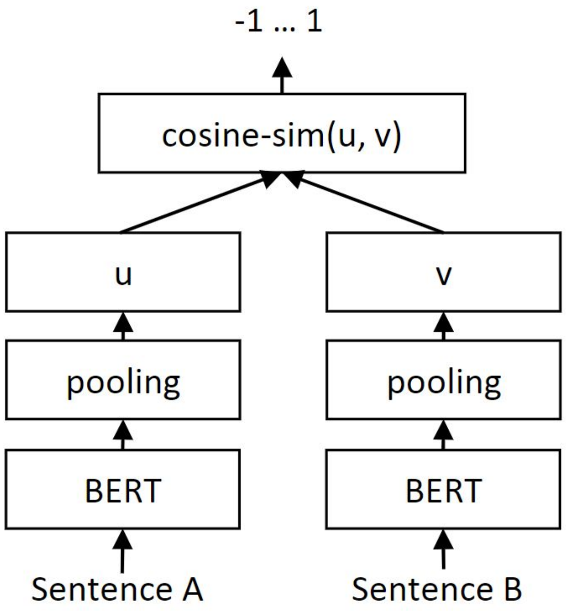
Classification objective

Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).



Regression objective

Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.



Triplet objective

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

Given an anchor sentence **a**, a positive sentence **p**, and a negative sentence **n**, triplet loss tunes the network such that the distance between **a** and **p** is smaller than the distance between **a** and **n**.

Experiments and Results

Experiments and results

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 for various Textual Similarity (STS) tasks. Performance is reported by convention as ρ×100. STS12-STS16: SemEval 2012-2016, STSb: STSbenchmark, SICK-R: SICK relatedness dataset.

Experiments and results

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**

Table 2: Evaluation of SBERT sentence embeddings using the SentEval toolkit. SentEval evaluates sentence embeddings on different sentence classification tasks by training a logistic regression classifier using the sentence embeddings as features. Scores are based on a 10-fold cross-validation.

Thank You For Your Attention

… and remember, attention is all you need!

https://paperswithcode.com/paper/sentence-bert-sentence-embeddings-using