# Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks

SAMSUNG SDS

Hyunjin Choi, Judong Kim, Seongho Joe, Youngjune Gwon

Milan, Italy 10 | 15 January 2021

## Motivation

- **BERT –** bidirectional text encoder based on Transformers
  - our baseline architecture to build on **contextualized embedding**
  - **cross-encoder**: two sentences are passed and the target value is predicted
- **Combinatorial explosion**
  - pair regression tasks: combinations of possible sentence pairs
  - e.g., 10,000 sentences → n*(n-1)/2 = about 50 million
- **Need single sentence embedding models**

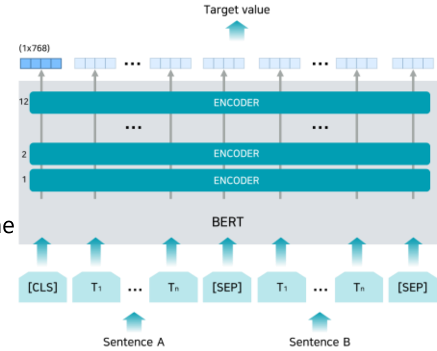
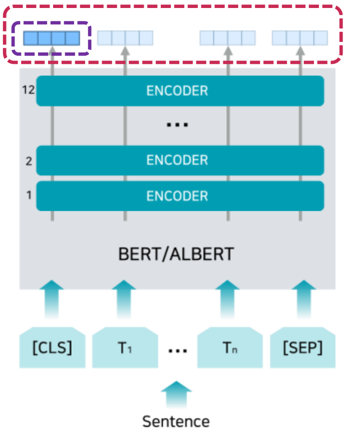
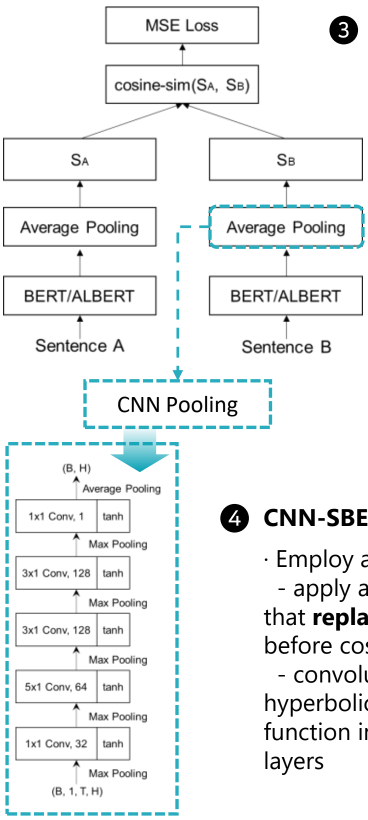



Figure1: BERT for sentence-pair regression tasks

## Models

1. **[CLS] token embedding**
   - [CLS] token: summarizes the information from other tokens
   - the easiest way
2. **Pooled token embedding**
   - Make fixed-length sentence vector by
     (1) averaging the token embedding output
     (2) max pooling




3. **Sentence-BERT/ALBERT (SBERT/SALBERT)**
   - **SBERT**: Reimers & Gurevych
   - **Siamese network architecture**
     - average-pools a pair of the BERT embeddings to fixed-size embeddings
     - using cosine similarity to derive semantically meaningful sentence embeddings
   - **SALBERT**
     - based on **ALBERT**
     - same Siamese network as SBERT
4. **CNN-SBERT/SALBERT (ours)**
   - Employ a **CNN architecture**
     - apply an outer CNN network that **replaces average pooling** before cosine similarity
     - convolutional layers with the hyperbolic tangent activation function interlaced with pooling layers

## Tasks & Datasets

- **Semantic Textual Similarity**
  - Evaluate the **similarity between two sentences** (regression task)
  - Semantic Textual Similarity benchmark (**STSb**): includes 8,628 sentence pairs – train 5,749, dev 1,500, test 1,379
- **Natural Language Inference**
  - determine whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise"
  - Stanford Natural Language Inference (**SNLI**) corpus
  - Multi-Genre Natural Language Inference (**MultiNLI**) corpus

## Experimental Results

| Model | Spearman (Pearson) |
|---|---|
| *Not fine-tuned* | |
| BERT [CLS]-token embedding | 6.43 (1.70) |
| BERT Avg. pooled token embedding | 47.29 (47.91) |
| ALBERT [CLS]-token embedding | 0.86 (4.57) |
| ALBERT Avg. pooled token embedding | 47.84 (46.57) |
| *Fine-tuned on STSb* | |
| BERT [CLS]-token embedding | 12.96 (7.49) |
| BERT Avg. pooled token embedding | 55.76 (54.90) |
| SBERT | 84.66 (84.86) |
| CNN-SBERT | **85.72 (86.15)** |
| ALBERT [CLS]-token embedding | 37.98 (27.89) |
| ALBERT Avg. pooled token embedding | 61.06 (60.41) |
| SALBERT | 74.33 (75.26) |
| CNN-SALBERT | **82.30 (83.08)** |
| *Fine-tuned on NLI (MultiNLI + SNLI)* | |
| BERT [CLS]-token embedding | 32.72 (26.88) |
| BERT Avg. pooled token embedding | 69.57 (68.49) |
| SBERT | **77.22 (74.53)** |
| CNN-SBERT | 76.77 (75.31) |
| ALBERT [CLS]-token embedding | 24.87 (4.11) |
| ALBERT Avg. pooled token embedding | 54.21 (53.58) |
| SALBERT | **74.05 (70.78)** |
| CNN-SALBERT | 73.70 (72.24) |
| *Fine-tuned on NLI (MultiNLI + SNLI) and STSb* | |
| BERT [CLS]-token embedding | 44.77 (38.74) |
| BERT Avg. pooled token embedding | 67.61 (65.30) |
| SBERT | 85.32 (84.51) |
| CNN-SBERT | **85.91 (85.63)** |
| ALBERT [CLS]-token embedding | 40.35 (33.46) |
| ALBERT Avg. pooled token embedding | 60.24 (59.98) |
| SALBERT | 77.59 (77.82) |
| CNN-SALBERT | **83.49 (83.87)** |

Table1: Evaluation on the STSb
by fine-tuning sentence embeddings on STS, NLI, and both

- **Fine-tuning datasets**
  - NLI train sets that are not directly related to STSb still gives a good performance
  - Our best results are obtained by fine-tuning with both NLI and STSb train sets
- **Compare models**
  - [CLS]-token embedding < average pooled token embedding < Siamese network < CNN added Siamese network
  - ALBERT-based models generally achieve lower performance
- **CNN architecture**
  - positive impact on sentence embedding performances
  - improves the ALBERT-based sentence embedding models more than the BERT-based models
  - improvement by CNN to ALBERT models can be as high as 8 points, which is compared to 1 point for the case of BERT models
  - ALBERT exposes more instability compared to BERT and such instability can be alleviated by CNN