SimCSE: Simple Contrastive Learning of Sentence Embeddings

- https://arxiv.org/pdf/2104.08821.pdf
- https://github.com/princeton-nlp/SimCSE
- 주로 텍스트 데이터에 적용가능한 간단한 contrastive learning 방법 제시
- 해당 도메인 데이터를 쓰지 않고 NLI 데이터로 학습한 SentenceBERT 성능을 능가함 (unsupervised)
- NLI 데이터셋을 사용하여 추가적인 성능 향상 방법도 제시 (supervised)

2021.07.29, 정욱재

Backgrounds

- Sentence BERT (SBERT): https://arxiv.org/pdf/1908.10084.pdf

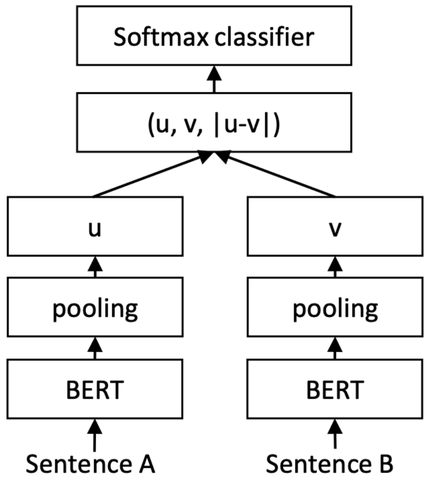


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

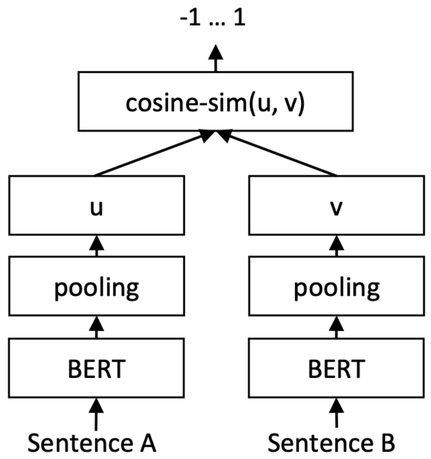


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

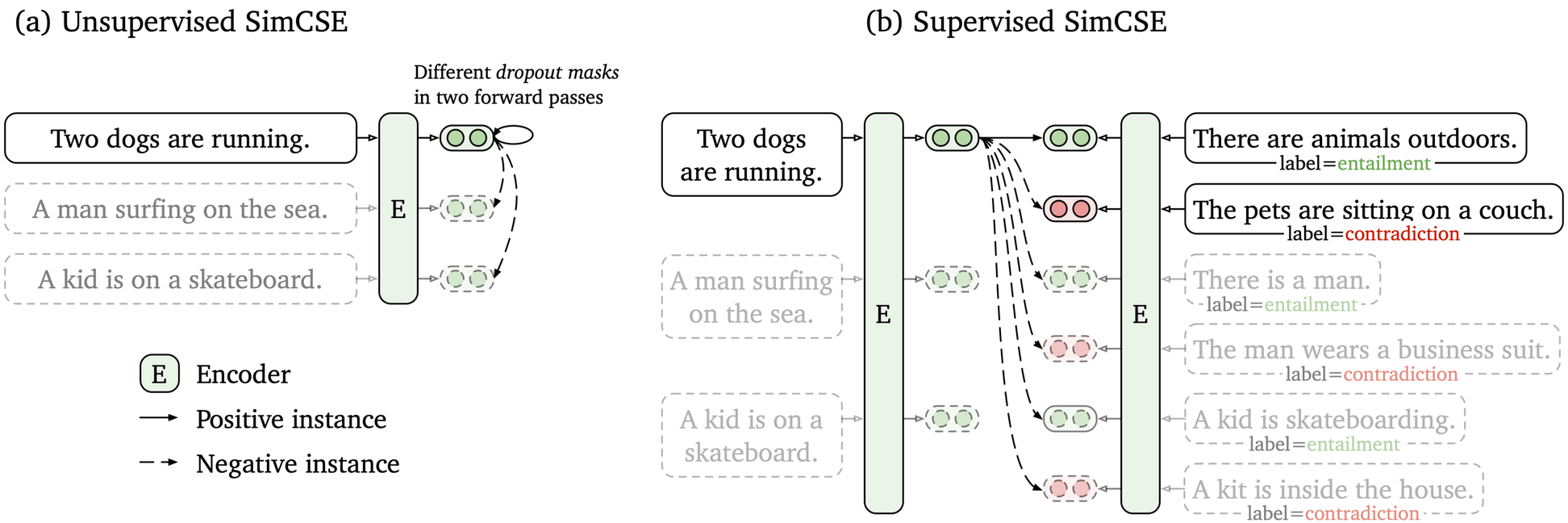
Background

- Wang and Isola (2020) (https://arxiv.org/abs/2005.10242) 에서 제시하길 Contrastive Learning은 alignment와 uniformity가 중요하다.
- Alignment: 주어진 positive pair들에 대해, embedding간의 expected distance
- Uniformity: 임베딩 자체가 얼마나 균등하게 분포되어 있나
- 두 값 모두 lower is better

$$\ell_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2.$$

$$\ell_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2},$$

SimCSE


(a) Unsupervised SimCSE
Different *dropout masks* in two forward passes
Two dogs are running.
A man surfing on the sea.
A kid is on a skateboard.
E
E Encoder
Positive instance
Negative instance
(b) Supervised SimCSE
Two dogs are running.
A man surfing on the sea.
A kid is on a skateboard.
E
E
There are animals outdoors.
label=entailment
The pets are sitting on a couch.
label=contradiction
There is a man.
label=entailment
The man wears a business suit.
label=contradiction
A kid is skateboarding.
label=entailment
A kit is inside the house.
label=contradiction

SimCSE

- Unsupervised SimCSE
- Dropout noise를 data augmentation으로 봄

Data augmentation			STS-B
None			**79.1**
Crop	*10%*	*20%*	*30%*
	75.4	70.1	63.7
Word deletion	*10%*	*20%*	*30%*
	74.7	71.2	70.2
Delete one word			74.8
w/o dropout			71.4
MLM 15%			66.8
Crop 10% + MLM 15%			70.8

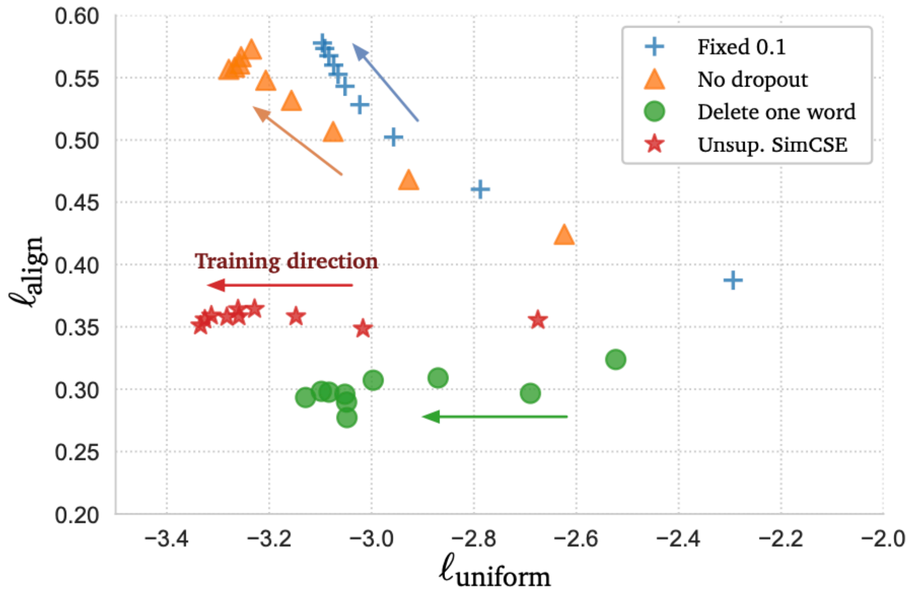


Figure 2: ℓ_{align}-ℓ_{uniform} plot for unsupervised SimCSE, "no dropout", "fixed 0.1" (same dropout mask for x_i and x_i^+ with $p = 0.1$), and "delete one word". We visualize checkpoints every 10 training steps and the arrows indicate the training direction. For both ℓ_{align} and ℓ_{uniform}, *lower numbers are better*.

SimCSE

- Anisotropy, isotropy problem
- Word embedding도 그렇고, sentence embedding도 그렇고 일반적으로 상당히 높은 similarity를 보임 -> 임베딩이 anisotropy하게 분포되어 있다.
- Word Embedding의 Singular value를 살펴봤을 때 몇몇개의 dominant한 값을 제외하고는 0에 가까운 값을 보임
- 이 논문에서 Contrastive learning이 실제로 이 값들을 flatten해서 성능을 더 끌어올릴 수 있을까?

Results

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
GloVe embeddings (avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-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
∗ SimCSE-$\text{BERT}_{\text{base}}$	**66.68**	**81.43**	**71.38**	**78.43**	**78.47**	**75.49**	**69.92**	**74.54**
$\text{RoBERTa}_{\text{base}}$ (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\text{base}}$-whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	**68.68**	**82.62**	**73.56**	**81.49**	**80.82**	**80.48**	**67.87**	**76.50**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	**69.87**	**82.97**	**74.25**	**83.01**	**79.52**	**81.23**	**71.47**	**77.47**
Supervised models								
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
$\text{SBERT}_{\text{base}}$♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{SBERT}_{\text{base}}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{SBERT}_{\text{base}}$-whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
∗ SimCSE-$\text{BERT}_{\text{base}}$	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
$\text{SRoBERTa}_{\text{base}}$♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
$\text{SRoBERTa}_{\text{base}}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

Table 6: Sentence embedding performance on STS tasks (Spearman's correlation, "all" setting). We highlight the highest numbers among models with the same pre-trained encoder. ♣: results from Reimers and Gurevych (2019); ♡: results from Zhang et al. (2020); all other results are reproduced or reevaluated by ourselves. For BERT-flow (Li et al., 2020) and whitening (Su et al., 2021), we only report the "NLI" setting (see Table D.3).

Results

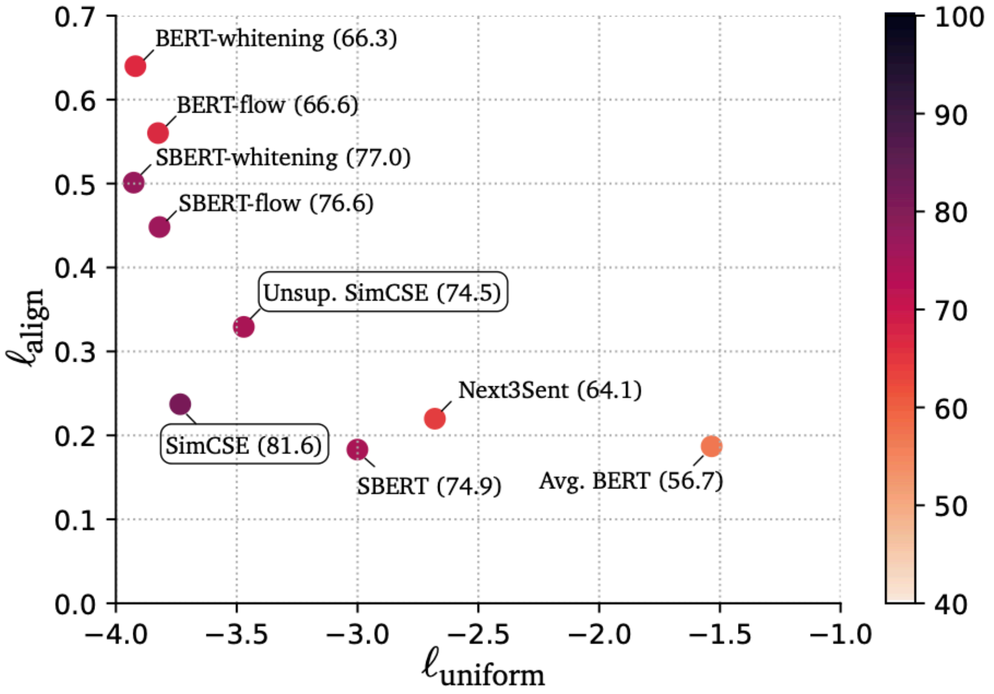


Figure 3: ℓ_{align}-ℓ_{uniform} plot of models based on $\text{BERT}_{\texttt{base}}$. Color of points and numbers in brackets represent average STS performance (Spearman's correlation). *Next3Sent*: "next 3 sentences" from Table 3.

Results

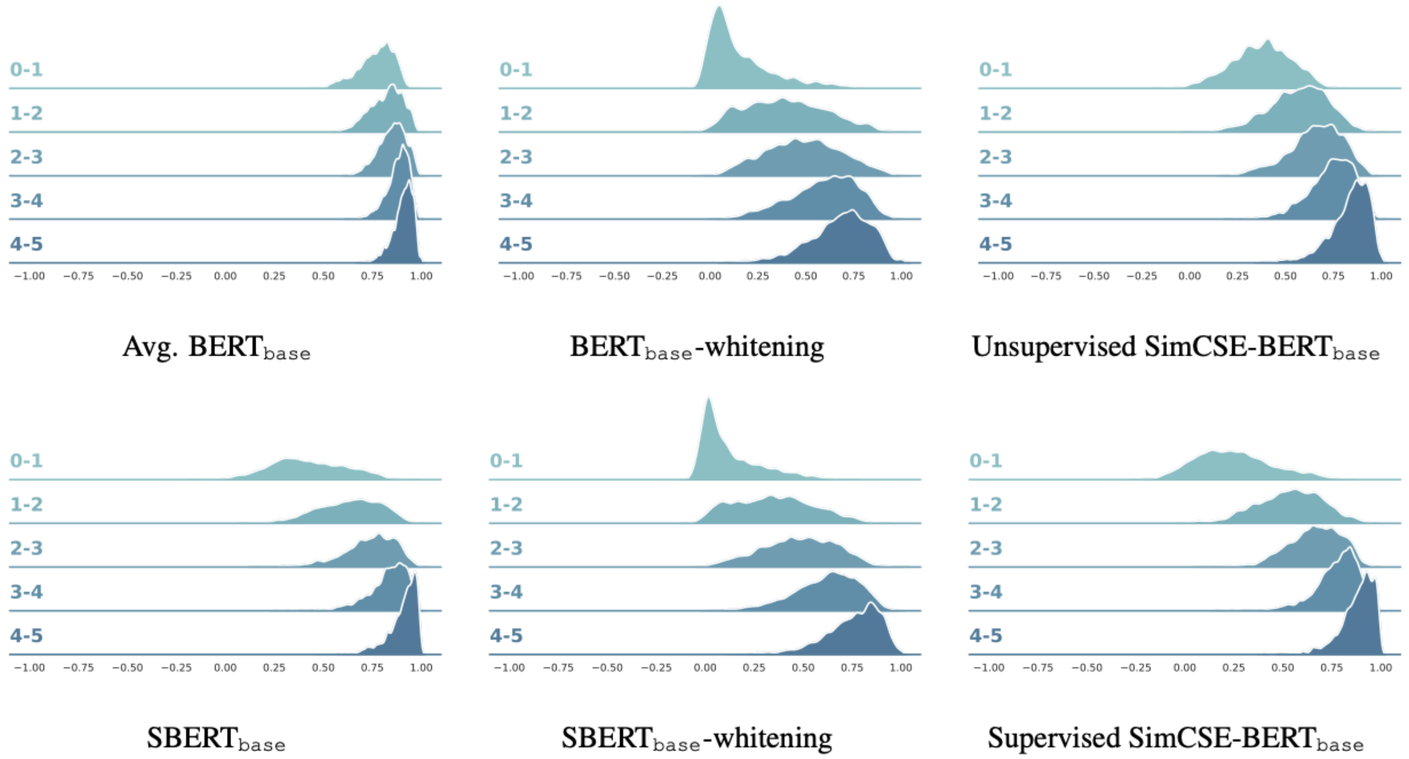
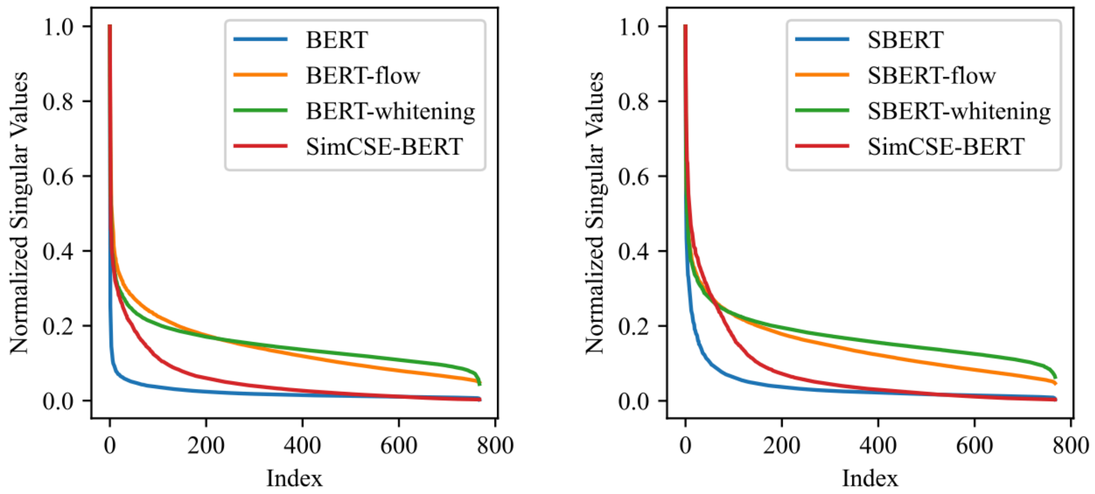


Figure 4: Density plots of cosine similarities between sentence pairs in full STS-B. Pairs are divided into 5 groups based on ground truth ratings (higher means more similar) along the y-axis, and x-axis is the cosine similarity.

Results



Unsupervised models Supervised models

Figure E.1: Singular value distributions of sentence embedding matrix from sentences in STS-B. We normalize the singular values so that the largest one is 1.