ESimCSE: Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, Songlin Hu

Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Kuaishou Technology, Beijing, China

Published in 2022 Coling

발제자: 윤예준

https://arxiv.org/pdf/2109.04380.pdf

목차

- Introduction
- Method
- Experiments & Results

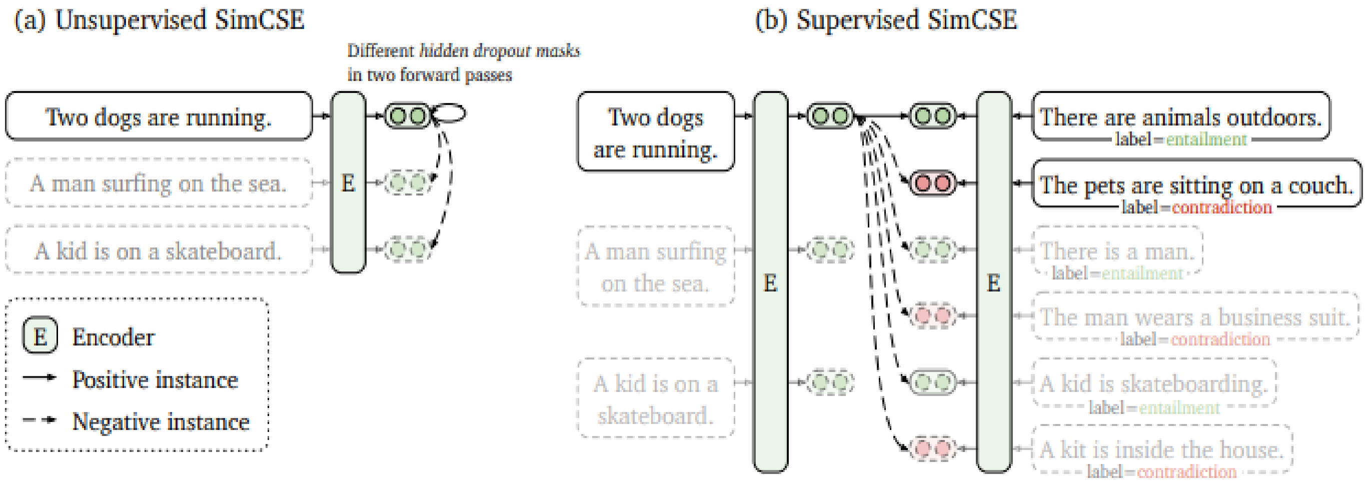
ESimCSE Outline

- **SimCSE의 샘플 구축 방법에서 발생한 문제에 대한 개선 방법 제안**

- Negative pair 수를 늘리기 위해 Momentum Contrast method 사용 제안

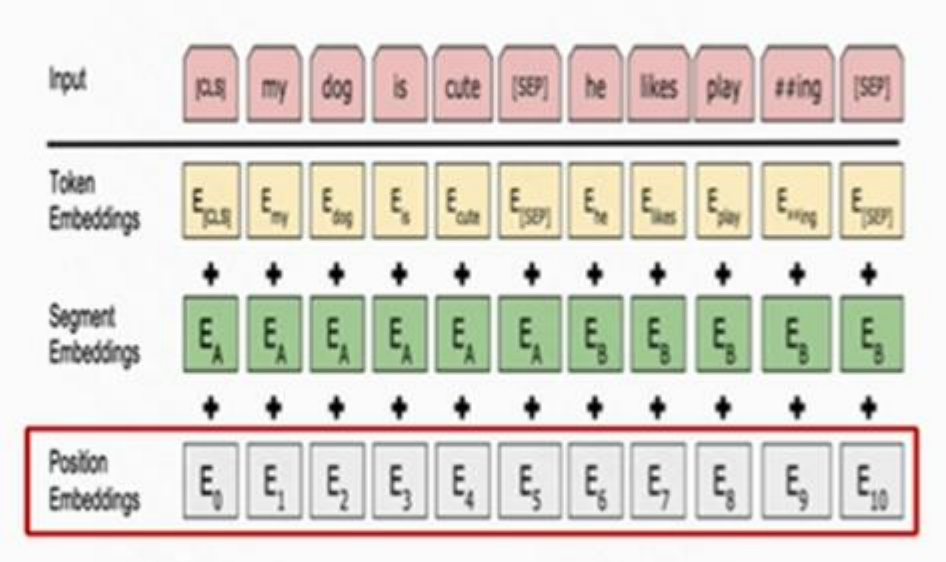
1. Introduction

SimCSE



$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z'_i})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z'_j})/\tau}}, \qquad (4)$$

SimCSE 샘플 구축에 대한 문제점



Length Diff	Avg. Similarity Diff
> 3	16.34
≤ 3	**18.18** (+1.84)

Table 1: The average similarity difference between the model (SimCSE-BERT) predictions and the normalized ground truths.

Word Repetition

- 문장 길이 편향을 완화하기 위한 방법

Method	Text	Similarity
original sentence	I like this apple because it looks so fresh and it should be delicious.	1.0
random insertion	I **don't** like this apple because **but** it looks so **not** fresh and it should be **dog** delicious.	0.69
random deletion	I like this ~~apple~~ because it looks so ~~fresh~~ and it should be ~~delicious~~.	0.32
word repetition	I like **like** this apple because it looks so **so** fresh and **and** it should be delicious.	0.99
word repetition	I **I** like this apple **apple** because it looks **looks** so fresh **fresh** and it should be delicious **delicious**.	0.98

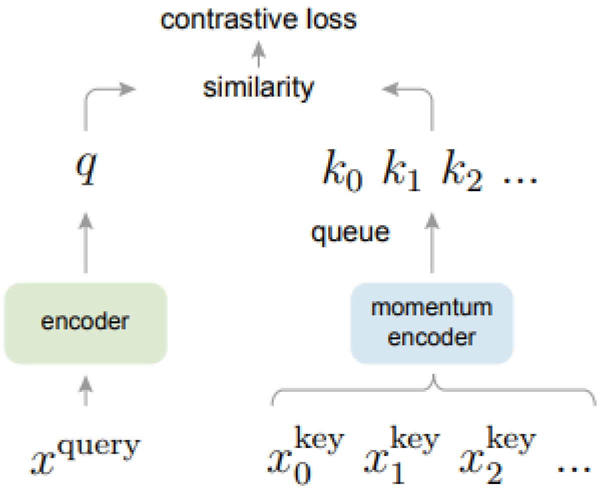
Table 2: An example of semantic similarity after different methods change a sentence's length.

$$dup\_len \in [0, max(2, int(dup\_rate * N))] \quad (3)$$

$$dup\_set = uniform([1, N], num = dup\_len) \quad (4)$$

Momentum Contrast

- 대조 학습은 이론적으로 negative pair가 많을 수록 좋음.
- 그러나 큰 배치 크기가 항상 더 좋은 선택은 아님.
- SimCSE BERT base 경우 배치 크기가 커질 수록 성능이 저하 됨.



$$\theta_{\mathrm{m}} \leftarrow \lambda\theta_{\mathrm{m}} + (1-\lambda)\theta_{\mathrm{e}} \tag{5}$$

ESimCSE

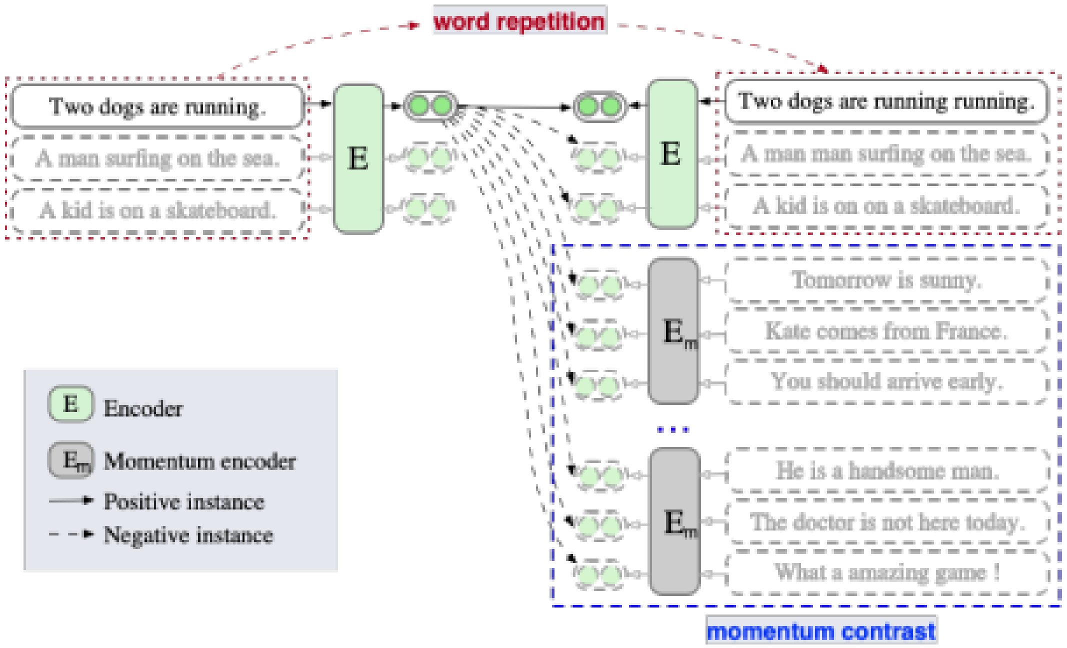


Figure 1: The schematic diagram of the ESimCSE method.

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau} + \sum_{m=1}^{M} e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_m^+)/\tau}} \tag{6}$$

Training

- Model: BERT, RoBERTa
- Dataset: 영어 위키피디아에서 무작위로 추출한 100만개 문장을 사용하여 학습
- Optimizer: Adam
- Batch size = 64
- Temperature: 0.05
- learning rate
  - BERT base: 3e-5
  - 이 외: 1e-5
- dropout
  - base: p=0.1
  - large: p=0.15
- momentum λ: 0.995

Sentence embedding performance

Model	STS12	STS13	STS14	SICK15	STS16	STS-B	SICK-R	Avg.
IS-BERT$_{base}$ △	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-BERT$_{base}$ △	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
ConSERT$_{base}$ ♡	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
BERT$_{base}$-flow◇	63.48	72.14	68.42	73.77	75.37	70.72	63.11	69.57
SG-OPT-BERT$_{base}$ ♠	66.84	80.13	71.23	81.56	77.17	77.23	68.16	74.62
Mirror-BERT$_{base}$ ♯	69.10	81.10	73.00	81.90	75.70	78.00	69.10	75.40
SimCSE-BERT$_{base}$ ♣	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
ESimCSE-BERT$_{base}$	**73.40**	**83.27**	**77.25**	**82.66**	**78.81**	**80.17**	**72.30**	**78.27**
ConSERT$_{large}$ ♡	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
BERT$_{large}$-flow◇	65.20	73.39	69.42	74.92	77.63	72.26	62.50	70.76
SG-OPT-BERT$_{large}$ ♠	67.02	79.42	70.38	81.72	76.35	76.16	70.20	74.46
SimCSE-BERT$_{large}$ ♣	70.88	84.16	76.43	**84.50**	**79.76**	79.26	73.88	78.41
ESimCSE-BERT$_{large}$	**73.21**	**85.37**	**77.73**	84.30	78.92	**80.73**	**74.89**	**79.31**
Mirror-RoBERTa$_{base}$ ♯	66.60	82.70	74.00	82.40	79.70	79.60	69.70	76.40
SimCSE-RoBERTa$_{base}$ ♣	**70.16**	81.77	73.24	81.36	**80.65**	80.22	68.56	76.57
ESimCSE-RoBERTa$_{base}$	69.90	**82.50**	**74.68**	**83.19**	80.30	**80.99**	**70.54**	**77.44**
SimCSE-RoBERTa$_{large}$ ♣	72.86	83.99	75.62	84.77	**81.80**	81.98	71.26	78.90
ESimCSE-RoBERTa$_{large}$	**73.20**	**84.93**	**76.88**	**84.86**	81.21	**82.79**	**72.27**	**79.45**

Table 3: Sentence embedding performance on 7 semantic textual similarity (STS) test sets. ♣ : results from official published model by (Gao et al., 2021).♡ : results from (Yan et al., 2021). ♠ : results from (Kim et al., 2021). ◇ : results from (Li et al., 2020). △ : results are reproduced and reevaluated by (Gao et al., 2021). ♯ : results from (Liu et al., 2021)

Performance on Transfer Tasks

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
SimCSE ♣	81.18	**86.46**	94.45	**88.88**	85.50	89.80	74.43	85.81
ESimCSE	**81.32**	86.22	**94.74**	88.74	**85.50**	**91.00**	**74.90**	**86.06**

Table 11: Results on transfer tasks of different sentence embedding models, in terms of accuracy. ♣ : results from (Gao et al., 2021).

Sentence embedding performance

Model	LD	STS12	STS13	STS14	SICK15	STS16	STS-B	SICK-R	Avg.
SimCSE	> 3	8.93	15.74	11.90	19.68	28.91	21.33	7.86	16.34
	≤ 3	9.29	22.81	19.53	19.92	24.08	22.12	9.53	18.18
ESimCSE	> 3	13.48	23.73	17.14	25.98	34.71	26.22	10.44	21.67
	≤ 3	12.52	28.56	24.13	24.17	29.32	25.63	12.35	22.38

Table 7: The difference between the model predicted cosine similarity and the true label on each dataset's test set. "LD" is short for length difference.

Word Repetition vs Momentum Contrast

Model	STS-B
SimCSE ♣	82.45
+ word repetition	84.09 (+1.64)
+ momentum contrast	83.98 (+1.53)
ESimCSE	**84.85** (+2.40)

Table 4: Improvement on STS-B development sets that word repetition or momentum contrast brings to SimCSE. ♣ : results from official published model by (Gao et al., 2021).

Effect of Sentence-Length-Extension Method

Length-extension Method	STS-B
+Inserting Stop-words	81.72
+Inserting [MASK]	83.08
+Inserting Masked Prediction	84.18
+Word Repetition	84.40
+Sub-word Repetition	**84.85**

Table 5: Effects of sentence-length-extension method.

Batching Sentences of Similar Length in Training

- We divide the training set into two coarse-grained buckets based on whether the sentence length is greater than buc_len, where $buc\_len \in [3, 8]$;
- We divide the training set by sentence length into 6 fine-grained buckets: $\{\leq 3, 4, 5, 6, 7, \geq 8\}$, which we use $buc\_len = 3 \sim 8$ for short.

buc_len	wr	3	4	5
STS-B	84.09	81.92	82.00	82.66
buc_len	6	7	8	$3 \sim 8$
STS-B	82.00	82.13	83.00	82.18

Table 6: Effects of different bucket lengths buc_len. "wr" means using word repetition method instead of bucketing sentences. "$3 \sim 8$" means fine-grained buckets setting: $\{\leq 3, 4, 5, 6, 7, \geq 8\}$.

Will Word Repetition Bring New Bias?

1. We randomly select a sentence as a query, such as q = "I like **this** apple because it **looks very** fresh"
2. We use the query to randomly recall a candidate sentence with 13%-17% overlap tokens, such as s1 = "**This** is a very tall tree and it **looks like** a giant"
3. We apply the word-repetition operation on the overlap tokens in the candidate sentence and produce a word-repeated sentence, such as s2 = "**This this** is a **very very** tall tree and it **looks looks** like a giant."
4. We calculate the similarity of <q, s1 >and <q, s2 >and compare them.

Model	**Sim <q,s1 >**	**Sim <q,s2 >**
SimCSE	26.39	27.07(**+0.68**)
ESimCSE	36.82	36.87(**+0.05**)

Table 8: Effect of repeated words on the average similarity of two sets

Effect of Hyperparameters

- Repetition Rate

dup_rate	0.08	0.12	0.16	0.2
STS-B	83.5	83.62	82.01	83.01
dup_rate	0.24	0.28	0.32	0.36
STS-B	84.24	82.96	**84.85**	83.84

Table 9: Effects of repetition rate dup_rate.

- Momentum Queue Size

Queue Size	**STS-B**
$1 \times batch\_size$	83.83
$1.5 \times batch\_size$	83.81
$2 \times batch\_size$	83.03
$2.5 \times batch\_size$	**84.85**
$3 \times batch\_size$	82.66

Table 10: Effects of queue size of momentum contrast.

결론

- SimCSE 문장 길이에 대한 편향 존재 확인
- 단어 반복 데이터 증강을 통해 편향 개선 및 성능 향상
- 모코를 이용하여 성능 향상

감사합니다.