InfoCSE: Information-aggregated Contrastive Learning of Sentence Embeddings

EMNLP 2022

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, Songlin Hu

발제자: 윤예준

InfoCSE: https://arxiv.org/pdf/2210.06432.pdf

01. 연구배경

대조학습: 의미가 유사한 것들은 더 가깝게 만들고 유사하지 않은 것들은 더 멀게 하는 학습 방법

SimCSE

- Positive pair
  - Dropout을 noise로 사용
    - 동일한 문장을 인코더에 두 번 전달
    - 서로 다른 dropout mask가 적용되어 두 개의 다른 임베딩을 얻음

- Negative pair
  - 같은 mini-batch 내의 다른 문장들의 임베딩을 negative pair로 사용

$$H = Enc\left(x, z\right), H^{+} = Enc\left(x, z^{+}\right) \quad (1)$$

$$h = Pooler\left(H\right), h^{+} = Pooler\left(H^{+}\right) \quad (2)$$

$$\mathcal{L}^{\mathrm{cl}} = -\log \frac{\exp(\mathrm{sim}\left(h, h^{+}\right)/\tau)}{\sum_{h' \in B} \exp(\mathrm{sim}\left(h, h'^{+}\right)/\tau)} \quad (3)$$

SimCSE: https://arxiv.org/pdf/2104.08821.pdf

01. 연구배경

- SimCSE는 좋은 결과를 얻었고 다른 sentence와 구별할 수 있는 sentence embedding을 학습할 수 있지만, sentence embedding이 sentence의 semantic을 잘 포함하고 있다고 하기엔 충분하지 않다고 함.

- Sentence embedding이 sentence의 semantic과 sufficiently equivalent하다면, 원래 문장을 재구성 할 수 있어야 함.

- 그러나 실험에 따르면 SimCSE에서 MLM objective는 STS task의 성능을 일관되게 떨어뜨림.

- 이는 MLM objective optimization의 gradients가 인코더 매개변수를 과도하게 업데이트하여 대조학습 task를 방해하기 때문.

- 따라서 contrastive sentence embedding learning에 sentence reconstruction task를 조합하는 것은 쉬운 일이 아님.

Model	STS-B
SimCSE-BERT$_{base}$	86.2
w/ MLM	
$\lambda = 0.01$	85.7
$\lambda = 0.1$	85.7
$\lambda = 1$	85.1

Table 1: Table from SimCSE (Gao et al., 2021). The masked language model (MLM) objective brings a consistent drop to the SimCSE model in semantic textual similarity tasks. "w/" means "with", λ is the balance hyperparameter for MLM loss.

$$\mathcal{L}^{\text{cl}} = -\log \frac{\exp(\text{sim}(h, h^{+})/\tau)}{\sum_{h' \in B} \exp(\text{sim}(h, h'^{+})/\tau)} \quad (3)$$

$$\mathcal{L}^{\text{mlm}} = \sum_{j \in masked} CE\left(H^{j}W, \hat{x}^{j}\right) \quad (4)$$

sentence reconstruction task로 contrastive sentence embedding learning을 개선한 InfoCSE 제안

02. 제안 방법

Symbol definition

- Sentences $x \in X$
- 12-layer Transformer blocks BERT model: Enc
- Sentence x of length l, we append a special [CLS] token to it, and then feed it into BERT for encoding.
- The output of each layer is a vector list of length l+ 1: hidden states
- The Last layer's hidden states: H
- The vector at [CLS] position: h
- The hidden state of the 6th layer : M
- The other hidden states of the 6th layer except the [CLS] position: $M_{>0}$

02. 제안 방법

1. Pre-training of The Auxiliary Network

- Auxiliary network의 6계층은 BERT 하위 6계층과 공유됨
- 두 개의 MLM objective를 동시에 최적화

$$\mathcal{L}^{\mathrm{mlm}} = \sum_{j \in masked} CE\left(H^j W, \hat{x}^j\right) \qquad (4)$$

$$\widetilde{H} = [h, M^{>0}]$$

$$\mathcal{L}^{\mathrm{aux}} = \sum_{j \in masked} CE\left(\widetilde{H}^j W, \hat{x}^j\right) \qquad (5)$$

$$\mathcal{L}^{\mathrm{pretrain}} = \mathcal{L}^{\mathrm{aux}} + \mathcal{L}^{\mathrm{mlm}} \qquad (6)$$

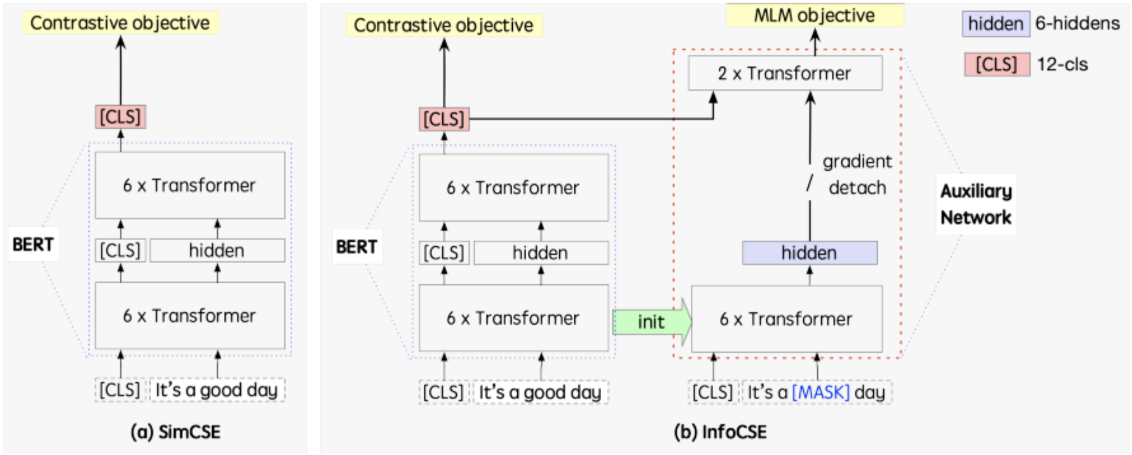

Figure 1: Comparison of InfoCSE and SimCSE structures. SimCSE learns sentence representations through contrastive learning on the [CLS] output embeddings of the BERT model. In addition to contrastive learning, InfoCSE designs an auxiliary network for sentence reconstruction with the [CLS] embeddings, enabling to learn better sentence representations.

02. 제안 방법

2. Joint Training of MLM and Contrastive Learning

- Auxiliary network의 6계층은 BERT 하위 6계층과 공유하지 않음
- 이때 Auxiliary network의 6계층은 frozen
- sentence x는 두 번 복사됨
    - positive pair: x^+
    - masked input: $\hat{x}$

$$\mathcal{L}^{\mathrm{cl}} = -\log \frac{\exp(\mathrm{sim}(h, h^+)/\tau)}{\sum_{h' \in B} \exp(\mathrm{sim}(h, h'^+)/\tau)} \qquad (3)$$

$$C = [h, Detach(M^{>0})]$$

$$\mathcal{L}^{\mathrm{joint}} = \mathcal{L}^{\mathrm{cl}} + \mathcal{L}^{\mathrm{aux}} * \lambda \qquad (7)$$

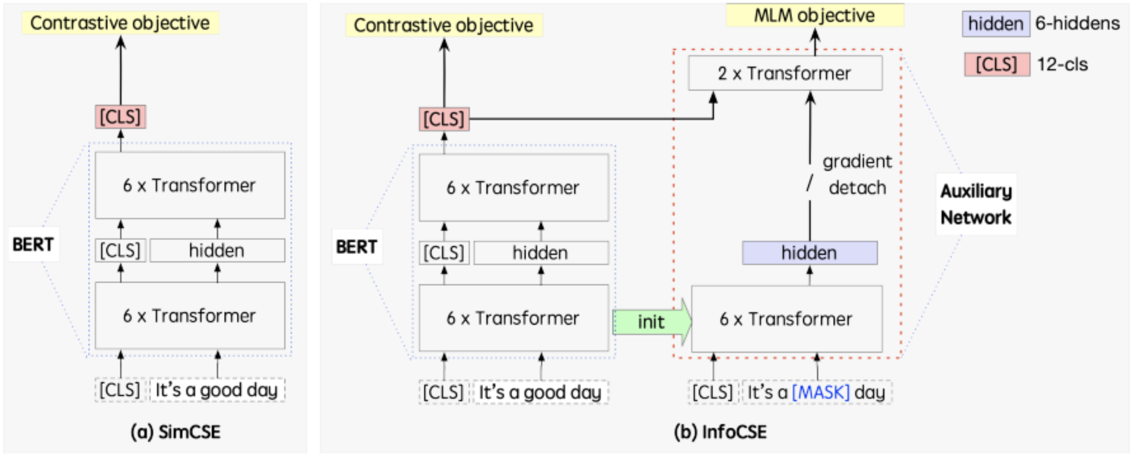

Figure 1: Comparison of InfoCSE and SimCSE structures. SimCSE learns sentence representations through contrastive learning on the [CLS] output embeddings of the BERT model. In addition to contrastive learning, InfoCSE designs an auxiliary network for sentence reconstruction with the [CLS] embeddings, enabling to learn better sentence representations.

03. 실험 결과

Setup
- Pretraining the auxiliary network
  - Dataset: Bookcorpus, Wikipedia

- Jointly optimizing contrastive objective and the auxiliary MLM objective
  - Dataset:1-million sentences randomly drawn from English Wikipedia for training

- 모든 평가는 BERT output만 사용

03. 실험 결과

STS(semantic textual similarity) task

- 7개의 STS task에서 실험 진행

Model	STS12	STS13	STS14	SICK15	STS16	STS-B	SICK-R	Avg.
GloVe embeddings(avg.)	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
BERT$_{base}$(first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT$_{base}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT$_{base}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
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
DiffCSE-BERT$_{base}$ ¶	72.28	84.43	76.47	83.90	80.54	80.59	71.23	78.49
InfoCSE-BERT$_{base}$	70.53	**84.59**	76.40	**85.10**	**81.95**	**82.00**	71.37	**78.85**
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
DiffCSE-BERT$_{large}$ †	72.11	84.99	76.19	85.09	78.65	80.34	73.93	78.76
InfoCSE-BERT$_{large}$	71.89	**86.17**	77.72	**86.20**	**81.29**	**83.16**	74.84	**80.18**

Table 2: 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). ⋆ : results from (Wu et al., 2021a). ¶ : results from (Chuang et al., 2022). †: The original paper does not report the results of BERT-large, so we use the official public code to perform a grid search on important hyperparameters for the best results.

03. 실험 결과

Open-domain Retrieval Tasks

- embedding의 zero-shot performance 측정

- BEIR(benchmarking information retrieval) 이용
    - 18개 dataset 중 14개 사용

Dataset	SimCSE		ESimCSE		DiffCSE		InfoCSE	
	base	large	base	large	base	large	base	large
trec-covid	0.2750	0.2264	0.2291	0.2829	0.2368	0.2291	**0.3937**	<u>0.3166</u>
nfcorpus	0.1048	0.1356	0.1149	0.1483	0.1204	0.1470	0.1358	**0.1576**
nq	0.1628	0.1671	0.0935	0.1705	0.1188	0.1556	**0.2023**	<u>0.1790</u>
fiqa	0.0985	0.0975	0.0731	**0.1117**	0.0924	<u>0.1027</u>	0.0991	0.1000
arguana	0.2796	0.2078	<u>0.3376</u>	0.2604	0.2500	0.2572	0.3244	**0.4133**
webis-touche2020	**0.1342**	0.0878	0.0786	0.1057	0.0912	0.0781	<u>0.0935</u>	0.0920
quora	0.7375	0.7511	0.7411	0.7615	0.7491	0.7471	<u>0.8241</u>	**0.8268**
cqadupstack	0.1349	0.1082	0.1276	0.1196	0.1197	0.1160	**0.2097**	<u>0.1881</u>
dbpedia-entity	0.1662	0.1495	0.1260	0.1650	0.1537	0.1571	**0.2101**	<u>0.1838</u>
scidocs	0.0611	0.0688	0.0657	0.0796	0.0673	0.0699	<u>0.0837</u>	**0.0859**
climate-fever	**0.1420**	0.1065	0.0796	0.1302	0.1019	<u>0.1087</u>	0.0937	0.0840
scifact	0.2492	0.2541	0.3013	0.2875	0.2666	0.2811	<u>0.3269</u>	**0.3801**
hotpotqa	0.2382	0.1896	0.1213	0.1970	0.1730	0.2068	**0.3177**	<u>0.2781</u>
fever	**0.2916**	0.1776	0.0756	0.1689	0.1416	0.1849	**0.1978**	0.1252
average	0.2197	0.1948	0.1832	0.2135	0.1916	0.2030	**0.2509**	<u>0.2436</u>

Table 3: Zero-shot evaluation results on the BEIR benchmark. All scores denote **nDCG@10**. The best score on a given dataset is marked in **bold**, and the second best is <u>underlined</u>.

03. 실험 결과

- Pre-training of The Auxiliary Network

Model	STS-B
InfoCSE	**85.49**
w/o pre-training	83.73

Table 4: Development set results of STS-B for InfoCSE with or without auxiliary network pre-training. “w/o" denotes without.

- Joint Training of MLM and Contrastive Learning

Model	STS-B
InfoCSE	**85.49**
w/o MLM loss	82.45
w/o Contrastive loss	40.00

Table 5: Development set results of STS-B for InfoCSE variants, where we vary the objective. “w/o" denotes without.

- The Impact of Gradient Detach in Joint Training

Model	STS-B
InfoCSE	**85.49**
w/o gradient detach	84.41

Table 6: Development set results of STS-B for InfoCSE with or without gradient detach. “w/o" denotes without.

03. 실험 결과

- The Impact of Mask Rate

Mask Rate	*10%*	*15%*	*20%*	*25%*
STS-B	83.97	84.62	84.74	85.08
Mask Rate	*30%*	*35%*	*40%*	*45%*
STS-B	84.19	84.70	**85.49**	84.13

Table 7: Development set results of STS-B when we vary the mask rate.

- The Impact of Coefficient λ

λ	0.	5e-6	1e-5	5e-5
STS-B	82.45	84.50	**85.49**	84.7
λ	1e-4	5e-3	1e-2	5e-2
STS-B	84.29	80.48	76.15	75.27

Table 8: Development set results of STS-B when we vary the coefficient λ.

- The Impact of Pooler

Model	**cls**	**cls_before_pooler**
SimCSE	81.72	82.45
DiffCSE	83.90	84.56
InfoCSE	**85.08**	**85.49**

Table 9: Development set results of STS-B where we vary the pooler choice. "cls" denotes using the representation of [CLS] token; "cls_before_pooler" denotes using the representation of [CLS] token without the extra linear+activation.

03. 실험 결과

- Transfer Tasks

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
GloVe embeddings (avg.)	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip-thought	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
Avg. BERT embeddings	78.66	86.25	94.37	88.66	84.40	92.80	69.54	84.94
BERT-[CLS] embedding	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS-BERT$_{base}$ △	81.09	87.18	94.96	88.75	85.96	88.64	74.24	85.83
SimCSE ♣	81.18	86.46	94.45	88.88	85.50	89.80	74.43	85.81
w/MLM ♣	82.92	87.23	95.71	88.73	86.81	87.01	78.07	86.64
InfoCSE	81.76	86.57	94.90	88.86	87.15	90.60	76.58	86.63

Table 10: Results on transfer tasks of different sentence embedding models, in terms of accuracy. ♣ : results from official published model by (Gao et al., 2021).♡ : results from (Yan et al., 2021). △ : results are reproduced and reevaluated by (Gao et al., 2021). ¶ : results from (Chuang et al., 2022).

03. 실험 결과

- In-domain Retrieval Task
  - In-domain에서 2758개 문장 사용

Model	R@1	R@5	R@10
SimCSE	74.23	94.85	95.88
DiffCSE	79.38	96.91	98.97
InfoCSE	**80.41**	**100.00**	**100.00**

Table 11: In-domain retrieval results."R@" denotes recall.

- Compatibility of Different Auxiliary Objectives

Model	STSB	Avg.
SimCSE	82.45	76.25
w/ RTD (DiffCSE)	84.56	78.27
w/ MLM (InfoCSE)	85.49	78.49
w/ RTD + MLM	**85.83**	**79.39**

Table 12: The comparison of the improvement brought by different auxiliary objectives to SimCSE. "w/" denotes without. "STS-B" denotes the best result on the STS-B development set. "Avg." denotes the corresponding average result on 7 semantic textual similarity (STS) test sets.

감사합니다.