Self-Guided Contrastive Learning for BERT Sentence Representations

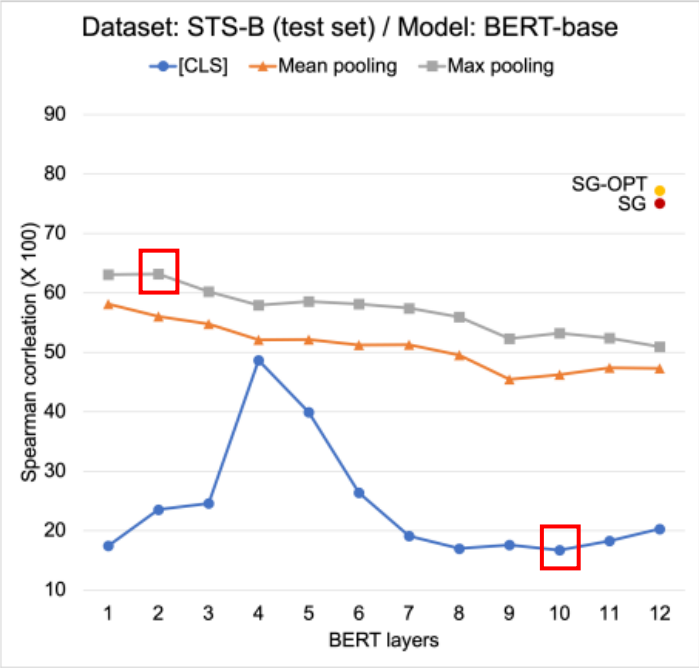
Authors: Taeuk Kim, Kang Min Yoo, Sang-goo Lee

Venue: ACL 2021

1. Introduction

- BERT를 sentence encoder로 사용하는 가장 일반적인 방법은 지도학습으로 fine-tuning 하는 것
  - 이 때 가장 간단하고 효과적인 방법은 마지막층의 CLS embedding은 입력 sequence의 표현으로 사용하는 것

- Labeled data를 사용할 수 없을 때는 어떤 방법으로 sentence embedding을 얻는 것이 최선인지가 명확하지 않음
  - 이 때 가장 일반적인 방법은 마지막 layer에 평균 풀링을 적용하는 것
  - 하지만 이 방법이 언제나 최선은 아닐 수 있음



- 다양한 방법으로 STS 실험
  - Spearman 상관관계 측정
  - 2번째 layer의 max pooling(최고)과 10번째 layer의 CLS embedding(최저) 까지 성능의 범위가 큰 것을 보임
    - 선택한 layer 및 pooling 방법에 따라 성능이 다르게 나오고 그 차이가 큼(불안정함)

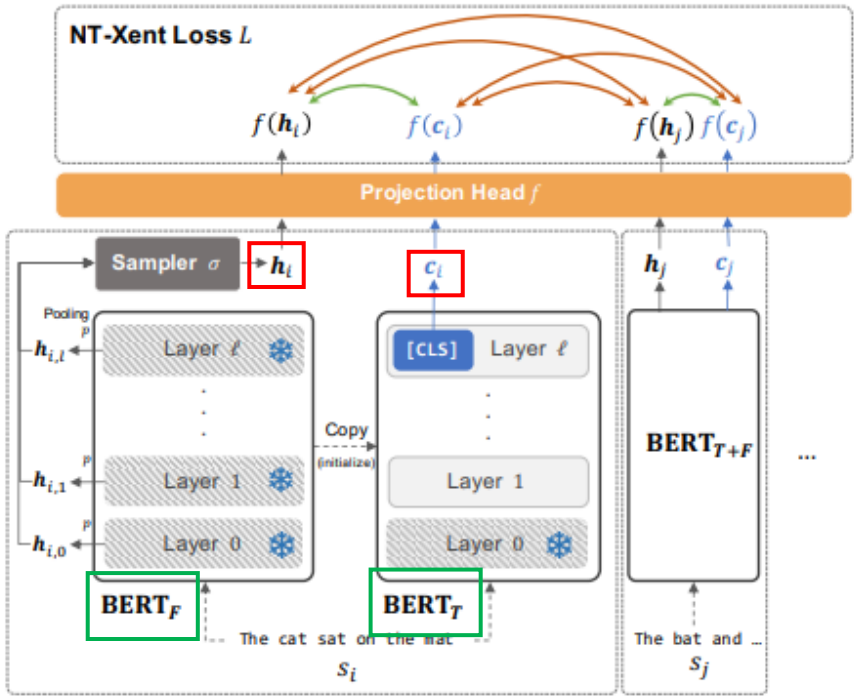
1. Introduction

- 이런 문제를 해결하기 위해 **Self-Guided 방법을 사용하는** Contrastive Learning **방법을 제안**함

- **Contrastive Learning with Self-Guidance**
  - BERT의 hidden representation을 최종 sentence embedding과 가까워야 하는 positive sample로 사용
    - Data augmentation가 필요 없고 training data에 대한 사전 지식도 갖고 있다는 장점이 있음

2. Method

- $\boldsymbol{BERT_F}$ 와 $\boldsymbol{BERT_T}$ 로 구성



- BERT-fixed($\boldsymbol{BERT_F}$)
  - training 하는 동안 training signal을 주기 위해 fix 되어있는 부분
  - Minibatch 내의 sample을 BERT-fixed에 통과시켜 계산한 token level의 hidden representation에 pooling(max pooling)을 씌워 hidden representation$(h_{i,l})$ 계산
  - $(h_{i,l})$에 sampler(uniform sampler)를 거쳐 하나의 hidden representation을 계산(h_i)
    → pivot representation(positive sample로 사용)

- BERT-tuned($\boldsymbol{BERT_T}$)
  - 더 나은 sentence embedding을 위해 fine-tuning 되는 부분
  - 마지막 layer의 [CLS] vector를 sentence embedding (c_i)으로 사용

2. Method

- Loss function 정의

$$L_m^{base} = -\log\left(\phi(\mathbf{x}_m, \mu(\mathbf{x}_m))/Z\right),$$
$$\text{where } \phi(\mathbf{u}, \mathbf{v}) = \exp(g(f(\mathbf{u}), f(\mathbf{v}))/\tau)$$
$$\text{and } Z = \textstyle\sum_{n=1, n\neq m}^{2b} \phi(\mathbf{x}_m, \mathbf{x}_n).$$

$$\mu(\mathbf{x}) = \begin{cases} \mathbf{h}_i & \text{if } \mathbf{x} \text{ is equal to } \mathbf{c}_i. \\ \mathbf{c}_i & \text{if } \mathbf{x} \text{ is equal to } \mathbf{h}_i. \end{cases}$$
$$g(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$$

- g 는 cosine similarity
- f 는 MLP 층의 projection head

→ c_i 와 h_i 가 높은 similarity를 갖도록 학습

- NT-Xent loss function(the normalized temperature-scaled cross entropy loss)을 변형하여 정의

$$-log\frac{\exp(sim(f(X_m), f(\mu(X_m))/\tau)}{\sum_{n=1,n\neq m}^{2b}\exp(sim(f(X_m), f(X_n))/\tau)}$$

새로 정의한 loss function

$$-\log\frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

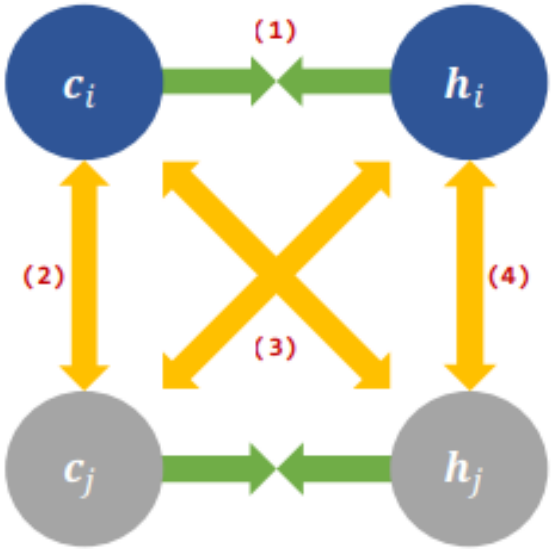
NT-Xent loss function

2. Method

- BERT-fixed와 BERT-tuned가 너무 멀어지는 것을 막기 위해 regularizer loss (L^{reg}) 추가하여 최종 loss function 정의

$$L^{base} = \frac{1}{2b}\sum_{m=1}^{2b} L_m^{base} + \lambda \cdot L^{reg} \qquad L^{reg} = \|\mathrm{BERT}_F - \mathrm{BERT}_T\|_2^2$$

- **Loss function 최적화**
  - NT-Xent loss의 4가지 요인



1. $c_i \rightarrow\leftarrow h_i \,(or\ c_j \rightarrow\leftarrow h_j)$: c_i와 h_i가 일치해야 함
2. $c_i \leftrightarrow c_j$: c_i와 c_j가 멀어져야 함
3. $c_i \leftrightarrow h_j \,(or\ c_j \leftrightarrow h_i)$: c_i와 h_j가 멀어져야 함
4. $h_i \leftrightarrow h_j$: h_i 와 h_j 가 멀어져야 함

2. Method

- **Loss function 최적화**
  - NT-Xent loss 의 4가지 요인들 중 일부는 부정적인 영향을 미칠 수 있음
  - 이를 반영하여 loss function 최적화(개선)

기존: $L_m^{base} = -\log\left(\phi(\mathbf{x}_m, \mu(\mathbf{x}_m))/Z\right),$
$\text{and } Z = \sum_{n=1,n\neq m}^{2b} \phi(\mathbf{x}_m, \mathbf{x}_n).$

  - 첫번째로, h_i를 최적화해야 하는 target 대신 pivot으로서의 역할만 하고 c_i에 더 집중할 수 있도록 수정

$$L_i^{opt1} = -\log\left(\phi(\mathbf{c}_i, \mathbf{h}_i)/\hat{Z}\right),$$
$$\text{where } \hat{Z} = \underbrace{\sum_{j=1,j\neq i}^{b} \phi(\mathbf{c}_i, \mathbf{c}_j)}_{\text{negative}} + \underbrace{\sum_{j=1}^{b} \phi(\mathbf{c}_i, \mathbf{h}_j)}_{\text{positive}}.$$

    - h_i 가 pivot으로만 사용되기 때문에 최적화 대상으로 간주되지 않음 → 4번째 요인이 필요없어짐

  - 두번째로, 2번째 요인이 성능 향상에 중요하지 않다는 것을 발견하여 이 부분을 반영하여 loss function 수정

$$L_i^{opt2} = -\log(\phi(\mathbf{c}_i, \mathbf{h}_i)/\sum_{j=1}^{b} \phi(\mathbf{c}_i, \mathbf{h}_j)).$$

2. Method

$$L_i^{opt2} = -\log(\phi(\mathbf{c}_i, \mathbf{h}_i) / \sum_{j=1}^{b} \phi(\mathbf{c}_i, \mathbf{h}_j)).$$

- 세번째로, multiple view $\{h_{i,k}\}$가 c를 guide하도록 허용함으로써 1번째 요인과 3번째 요인을 다양화하기 위해 loss function 수정

$$L_{i,k}^{opt3} = -\log \frac{\phi(\mathbf{c}_i, \mathbf{h}_{i,k})}{\phi(\mathbf{c}_i, \mathbf{h}_{i,k}) + \sum_{m=1, m\neq i}^{b} \sum_{n=0}^{l} \phi(\mathbf{c}_i, \mathbf{h}_{m,n})}$$

- 따라서, **최종 loss function**은 $L^{opt} = \frac{1}{b(l+1)} \sum_{i=1}^{b} \sum_{k=0}^{l} L_{i,k}^{opt3} + \lambda \cdot L^{reg}$ 이 됨

3. Experiments

- 3-1) STS Tasks
  - SG: 기존 NT-Xent loss를 사용/ SG-OPT: 최적화된 NT-Xent loss를 사용

Models	Pooling	STS-B	SICK-R	STS12	STS13	STS14	STS15	STS16	Avg.
Non-BERT Baselines									
GloVe†	Mean	58.02	53.76	55.14	70.66	59.73	68.25	63.66	61.32
USE†	-	74.92	76.69	64.49	67.80	64.61	76.83	73.18	71.22
BERT-base									
+ No tuning	CLS	20.30	42.42	21.54	32.11	21.28	37.89	44.24	31.40
+ No tuning	Mean	47.29	58.22	30.87	59.89	47.73	60.29	63.73	52.57
+ No tuning	WK	16.07	41.54	16.01	21.80	15.96	33.59	34.07	25.58
+ Flow	Mean-2	71.35$_{\pm 0.27}$	64.95$_{\pm 0.16}$	64.32$_{\pm 0.17}$	69.72$_{\pm 0.25}$	63.67$_{\pm 0.06}$	77.77$_{\pm 0.15}$	69.59$_{\pm 0.28}$	68.77$_{\pm 0.07}$
+ Contrastive (BT)	CLS	63.27$_{\pm 1.48}$	66.91$_{\pm 1.29}$	54.26$_{\pm 1.84}$	64.03$_{\pm 2.35}$	54.28$_{\pm 1.87}$	68.19$_{\pm 0.95}$	67.50$_{\pm 0.96}$	62.63$_{\pm 1.28}$
+ Contrastive (SG)	CLS	75.08$_{\pm 0.73}$	**68.19**$_{\pm 0.36}$	63.60$_{\pm 0.98}$	76.48$_{\pm 0.69}$	67.57$_{\pm 0.57}$	79.42$_{\pm 0.49}$	74.85$_{\pm 0.54}$	72.17$_{\pm 0.44}$
+ Contrastive (SG-OPT)	CLS	**77.23**$_{\pm 0.43}$	68.16$_{\pm 0.50}$	**66.84**$_{\pm 0.73}$	**80.13**$_{\pm 0.51}$	**71.23**$_{\pm 0.40}$	**81.56**$_{\pm 0.28}$	**77.17**$_{\pm 0.22}$	**74.62**$_{\pm 0.25}$
BERT-large									
+ No tuning	CLS	26.75	43.44	27.44	30.76	22.59	29.98	42.74	31.96
+ No tuning	Mean	47.00	53.85	27.67	55.79	44.49	51.67	61.88	48.91
+ No tuning	WK	35.75	38.39	12.65	26.41	23.74	29.34	34.42	28.67
+ Flow	Mean-2	72.72$_{\pm 0.36}$	63.77$_{\pm 0.18}$	62.82$_{\pm 0.17}$	71.24$_{\pm 0.22}$	65.39$_{\pm 0.15}$	78.98$_{\pm 0.21}$	73.23$_{\pm 0.24}$	70.07$_{\pm 0.81}$
+ Contrastive (BT)	CLS	63.84$_{\pm 1.05}$	66.53$_{\pm 2.62}$	52.04$_{\pm 1.75}$	62.59$_{\pm 1.84}$	54.25$_{\pm 1.45}$	71.07$_{\pm 1.11}$	66.71$_{\pm 1.08}$	62.43$_{\pm 1.07}$
+ Contrastive (SG)	CLS	75.22$_{\pm 0.57}$	69.63$_{\pm 0.95}$	64.37$_{\pm 0.72}$	77.59$_{\pm 1.01}$	68.27$_{\pm 0.40}$	80.08$_{\pm 0.28}$	74.53$_{\pm 0.43}$	72.81$_{\pm 0.31}$
+ Contrastive (SG-OPT)	CLS	**76.16**$_{\pm 0.42}$	**70.20**$_{\pm 0.65}$	**67.02**$_{\pm 0.72}$	**79.42**$_{\pm 0.80}$	**70.38**$_{\pm 0.65}$	**81.72**$_{\pm 0.32}$	**76.35**$_{\pm 0.22}$	**74.46**$_{\pm 0.35}$
SBERT-base									
+ No tuning	CLS	73.66	69.71	70.15	71.17	68.89	75.53	70.16	71.32
+ No tuning	Mean	76.98	72.91	70.97	76.53	73.19	79.09	74.30	74.85
+ No tuning	WK	78.38	74.31	69.75	76.92	72.32	81.17	76.25	75.59
+ Flow‡	Mean-2	81.03	74.97	68.95	78.48	**77.62**	81.95	78.94	77.42
+ Contrastive (BT)	CLS	74.67$_{\pm 0.30}$	70.31$_{\pm 0.45}$	71.19$_{\pm 0.37}$	72.41$_{\pm 0.60}$	69.90$_{\pm 0.43}$	77.16$_{\pm 0.48}$	71.63$_{\pm 0.55}$	72.47$_{\pm 0.37}$
+ Contrastive (SG)	CLS	81.05$_{\pm 0.34}$	75.78$_{\pm 0.55}$	73.76$_{\pm 0.76}$	80.08$_{\pm 0.45}$	75.58$_{\pm 0.57}$	83.52$_{\pm 0.43}$	79.10$_{\pm 0.51}$	78.41$_{\pm 0.33}$
+ Contrastive (SG-OPT)	CLS	**81.46**$_{\pm 0.27}$	**76.64** $_{\pm 0.42}$	**75.16**$_{\pm 0.56}$	**81.27**$_{\pm 0.37}$	76.31$_{\pm 0.38}$	**84.71**$_{\pm 0.26}$	**80.33**$_{\pm 0.19}$	**79.41**$_{\pm 0.17}$
SBERT-large									
+ No tuning	CLS	76.01	70.99	69.05	71.34	69.50	76.66	70.08	71.95
+ No tuning	Mean	79.19	73.75	72.27	78.46	74.90	80.99	76.25	76.54
+ No tuning	WK	61.87	67.06	49.95	53.02	46.55	62.47	60.32	57.32
+ Flow‡	Mean-2	81.18	74.52	70.19	80.27	**78.85**	82.97	**80.57**	78.36
+ Contrastive (BT)	CLS	76.71$_{\pm 1.22}$	71.56$_{\pm 1.34}$	69.95$_{\pm 3.57}$	72.66$_{\pm 1.16}$	70.38$_{\pm 2.10}$	77.80$_{\pm 3.24}$	71.41$_{\pm 1.73}$	72.92$_{\pm 1.53}$
+ Contrastive (SG)	CLS	**82.35**$_{\pm 0.15}$	**76.44**$_{\pm 0.41}$	**74.84**$_{\pm 0.57}$	**82.89**$_{\pm 0.41}$	77.27$_{\pm 0.35}$	84.44$_{\pm 0.23}$	79.54$_{\pm 0.49}$	79.68$_{\pm 0.37}$
+ Contrastive (SG-OPT)	CLS	82.05$_{\pm 0.39}$	**76.44**$_{\pm 0.29}$	74.58$_{\pm 0.59}$	**83.79**$_{\pm 0.14}$	76.98$_{\pm 0.19}$	**84.57**$_{\pm 0.27}$	79.87$_{\pm 0.42}$	**79.76**$_{\pm 0.33}$

Models	Pooling	STS-B	SICK-R	STS12	STS13	STS14	STS15	STS16	Avg.
RoBERTa-base									
+ No tuning	CLS	45.41	61.89	16.67	45.57	30.36	55.08	56.98	44.57
+ No tuning	Mean	54.53	62.03	32.11	56.33	45.22	61.34	61.98	53.36
+ No tuning	WK	35.75	54.69	20.31	36.51	32.41	48.12	46.32	39.16
+ Contrastive (BT)	CLS	**79.93**$_{\pm 1.08}$	**71.97**$_{\pm 1.00}$	62.34$_{\pm 2.41}$	78.60$_{\pm 1.74}$	68.65$_{\pm 1.48}$	79.31$_{\pm 0.65}$	77.49$_{\pm 1.29}$	**74.04**$_{\pm 1.16}$
+ Contrastive (SG)	CLS	78.38$_{\pm 0.43}$	69.74$_{\pm 1.00}$	**62.85**$_{\pm 0.88}$	78.37$_{\pm 1.55}$	68.28$_{\pm 0.89}$	**80.42**$_{\pm 0.65}$	**77.69**$_{\pm 0.76}$	73.67$_{\pm 0.62}$
+ Contrastive (SG-OPT)	CLS	77.60$_{\pm 0.30}$	68.42$_{\pm 0.71}$	62.57$_{\pm 1.12}$	**78.96**$_{\pm 0.67}$	**69.24**$_{\pm 0.44}$	79.99$_{\pm 0.44}$	77.17$_{\pm 0.24}$	73.42$_{\pm 0.31}$
RoBERTa-large									
+ No tuning	CLS	12.52	40.63	19.25	22.97	14.93	33.41	38.01	25.96
+ No tuning	Mean	47.07	58.38	33.63	57.22	45.67	63.00	61.18	52.31
+ No tuning	WK	30.29	28.25	23.17	30.92	23.36	40.07	43.32	31.34
+ Contrastive (BT)	CLS	77.05$_{\pm 1.22}$	67.83$_{\pm 1.34}$	57.60$_{\pm 3.57}$	72.14$_{\pm 1.16}$	62.25$_{\pm 2.10}$	71.49$_{\pm 3.24}$	71.75$_{\pm 1.73}$	68.59$_{\pm 1.53}$
+ Contrastive (SG)	CLS	76.15$_{\pm 0.54}$	66.07$_{\pm 0.82}$	**64.77**$_{\pm 2.52}$	71.96$_{\pm 1.53}$	64.54$_{\pm 1.04}$	78.06$_{\pm 0.52}$	75.14$_{\pm 0.94}$	70.95$_{\pm 1.13}$
+ Contrastive (SG-OPT)	CLS	**78.14**$_{\pm 0.72}$	**67.97**$_{\pm 1.09}$	64.29$_{\pm 1.54}$	**76.36**$_{\pm 1.47}$	**68.48**$_{\pm 1.58}$	**80.10**$_{\pm 1.05}$	**76.60**$_{\pm 0.98}$	**73.13**$_{\pm 1.20}$

3. Experiments

- 3-2) Multilingual STS Tasks

Models	Spanish
Baseline (Agirre et al., 2014)	
UMCC-DLSI-run2 (Rank #1)	80.69
MBERT	
+ CLS	12.60
+ Mean pooling	81.14
+ WK pooling	79.78
+ Contrastive (BT)	78.04
+ **Contrastive (SG)**	82.09
+ **Contrastive (SG-OPT)**	82.74

Table 2: SemEval-2014 Task 10 Spanish task.

Models	Arabic (Track 1)	Spanish (Track 3)	English (Track 5)
Baselines			
Cosine baseline (Cer et al., 2017)	60.45	71.17	72.78
ENCU (Rank #1, Tian et al. (2017))	74.40	85.59	85.18
MBERT			
+ CLS	30.57	29.38	24.97
+ Mean pooling	51.09	54.56	54.86
+ WK pooling	50.38	55.87	54.87
+ Contrastive (BT)	54.24	68.16	73.89
+ **Contrastive (SG)**	57.09	78.93	78.24
+ **Contrastive (SG-OPT)**	58.52	80.19	78.03

Table 3: Results on SemEval-2017 Task 1: Track 1 (Arabic), Track 3 (Spanish), and Track 5 (English).

3. Experiments

- 3-3) SentEval and Supervised Fine-tuning
  - MR(Movie Review), CR(product review), SUBJ(subjectivity status), MPQA(opinion-polarity), SST2(movie sentiment analysis), TREC(question-type classification), MRPC(paraphrase detection)

Models	MR	CR	SUBJ	MPQA	SST2	TREC	MRPC	Avg.
BERT-base								
+ Mean	81.46	86.71	95.37	87.90	85.83	90.30	73.36	85.85
+ WK	80.64	85.53	95.27	88.63	85.03	**94.03**	71.71	85.83
+ **SG-OPT**	**82.47**	**87.42**	**95.40**	**88.92**	**86.20**	91.60	**74.21**	**86.60**
BERT-large								
+ Mean	84.38	89.01	95.60	86.69	89.20	90.90	72.79	86.94
+ WK	82.68	87.92	95.32	**87.25**	87.81	91.18	70.13	86.04
+ **SG-OPT**	**86.03**	**90.18**	**95.82**	87.08	**90.73**	**94.65**	**73.31**	**88.26**
SBERT-base								
+ Mean	82.80	89.03	94.07	89.79	88.08	86.93	75.11	86.54
+ WK	82.96	89.33	**95.13**	**90.56**	88.10	**91.98**	**76.66**	**87.82**
+ **SG-OPT**	**83.34**	**89.45**	94.68	89.78	**88.57**	87.30	75.26	86.91

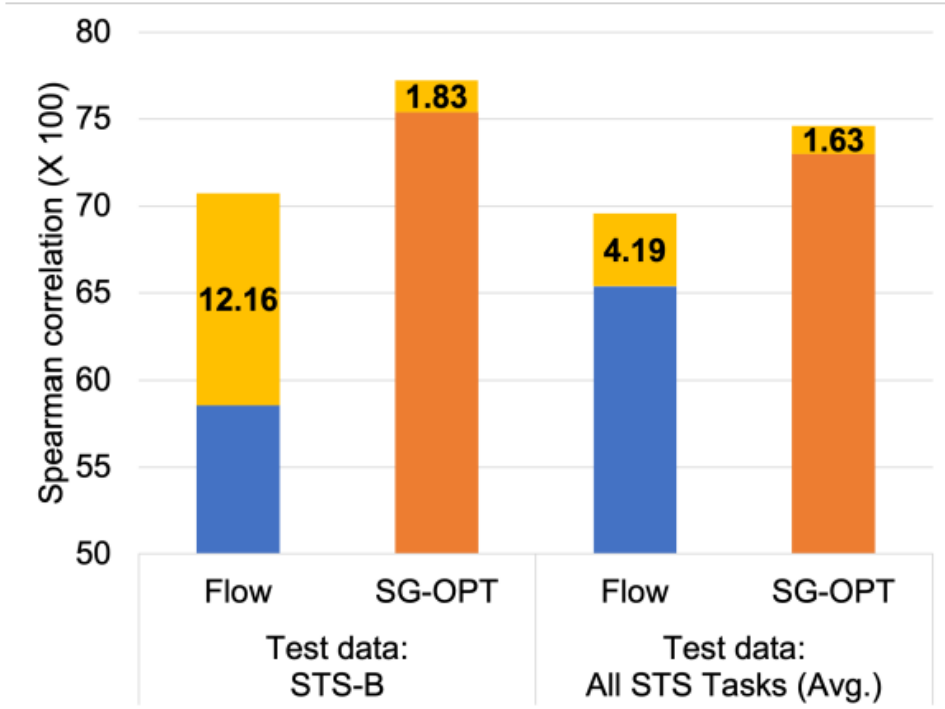
4. Analysis

- 4-1) Ablation Study
  - 기존 NT-Xent loss에 대한 수정(opt1, opt2, opt3) 이 모두 성능에 기여함을 확인

Models	STS Tasks (Avg.)
BERT-base	
+ **SG-OPT** (L^{opt3})	74.62
+ L^{opt2}	73.14 (-1.48)
+ L^{opt1}	72.61 (-2.01)
+ **SG** (L^{base})	72.17 (-2.45)
BERT-base + SG-OPT ($\tau = 0.01, \lambda = 0.1$)	74.62
+ $\tau = 0.1$	70.39 (-4.23)
+ $\tau = 0.001$	74.16 (-0.46)
+ $\lambda = 0.0$	73.76 (-0.86)
+ $\lambda = 1.0$	73.18 (-1.44)
- Projection head (f)	72.78 (-1.84)

4. Analysis

- 4-2) Robustness to Domain Shifts
  - 노란색 부분은 STS와 NLI에서의 성능 차이를 나타냄
  - SG-OPT는 NLI로 훈련할 때 STS에 비해 각각 1.83, 1.63만큼만 낮아지는 반면, Flow는 각각 12.16, 4.19 만큼 낮아짐
  - 즉, SG-OPT가 Flow에 비해 Domain shift에 있어 더 robust함을 보여줌



4. Analysis

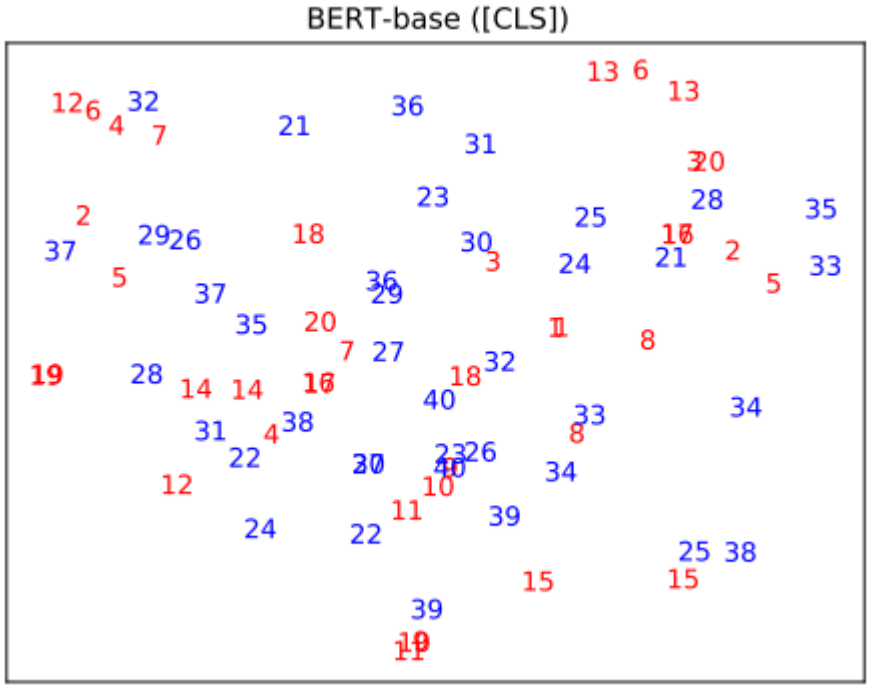
- 4-3) Computational Efficiency
  - STS-B task에서 training 및 inference 시간을 측정
  - Batch size: 16
  - SG-OPT가 Training 시간이 가장 길지만 8분 이내로 적당한 편이고, 훈련이 끝나면 pooling과 같은 후처리가 없기 때문에 inference 시에 효율적임

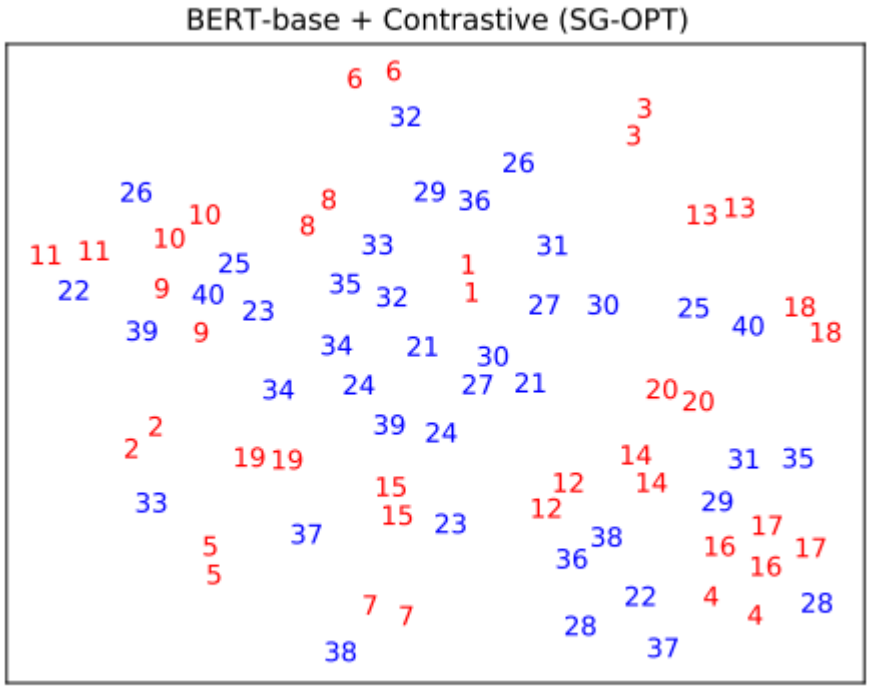
Layer	Elapsed Time	
	Training (sec.)	Inference (sec.)
BERT-base		
+ Mean pooling	-	13.94
+ WK pooling	-	197.03 (≈ 3.3 min.)
+ Flow	155.37 (≈ 2.6 min.)	28.49
+ **Contrastive (SG-OPT)**	455.02 (≈ 7.5 min.)	10.51

Table 6: Computational efficiency tested on STS-B.

4. Analysis

- 4-4) Representation Visualization
  - 빨간 숫자: positive pair/ 파란 숫자: negative pair
  - SG-OPT를 사용한 버전에서 숫자들이 vector space 내에 더 잘 align함을 확인





5. Conclusion

- BERT sentence embedding 개선을 위해 Self-Guidance를 이용한 Contrastive Learning 기법을 제안
- 제안한 방법이 데이터 증강과 같은 절차 없이도 Contrastive Learning 의 이점을 누릴 수 있고, 다른 baseline들보다 high quality의 sentence representation을 생성할 수 있음을 보였음
- 제안한 방법은 training이 끝나면 후처리가 필요하지 않아 inference 면에서 효율적이고, domain shift에 있어 robust함을 보였음

Thank You

감사합니다.