HYU AI Lab Seminar #15

Supervised Contrastive Learning For Pre-trained Language Model Fine-tuning

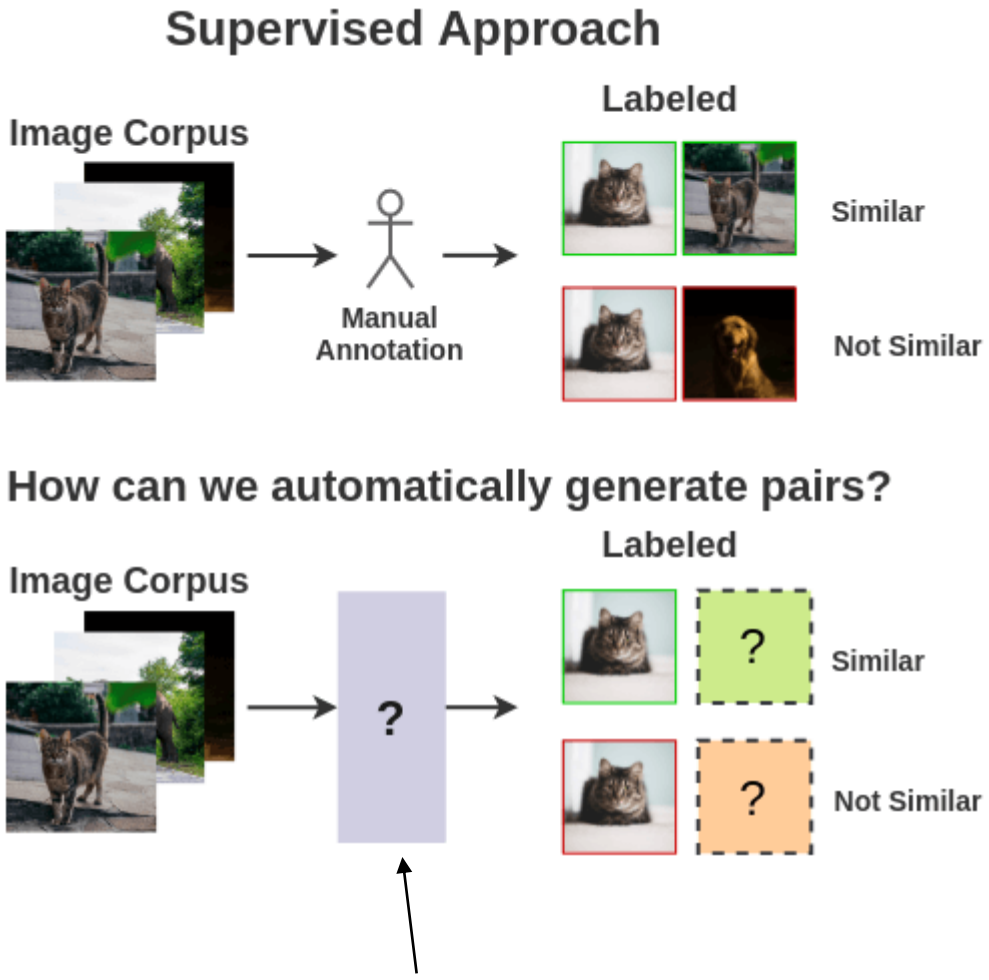
Beliz Gunnel, Jingfei Du, Alexis Conneau, Ves Stoyanov
Stanford University, Facebook AI

최원혁

"대조하는"

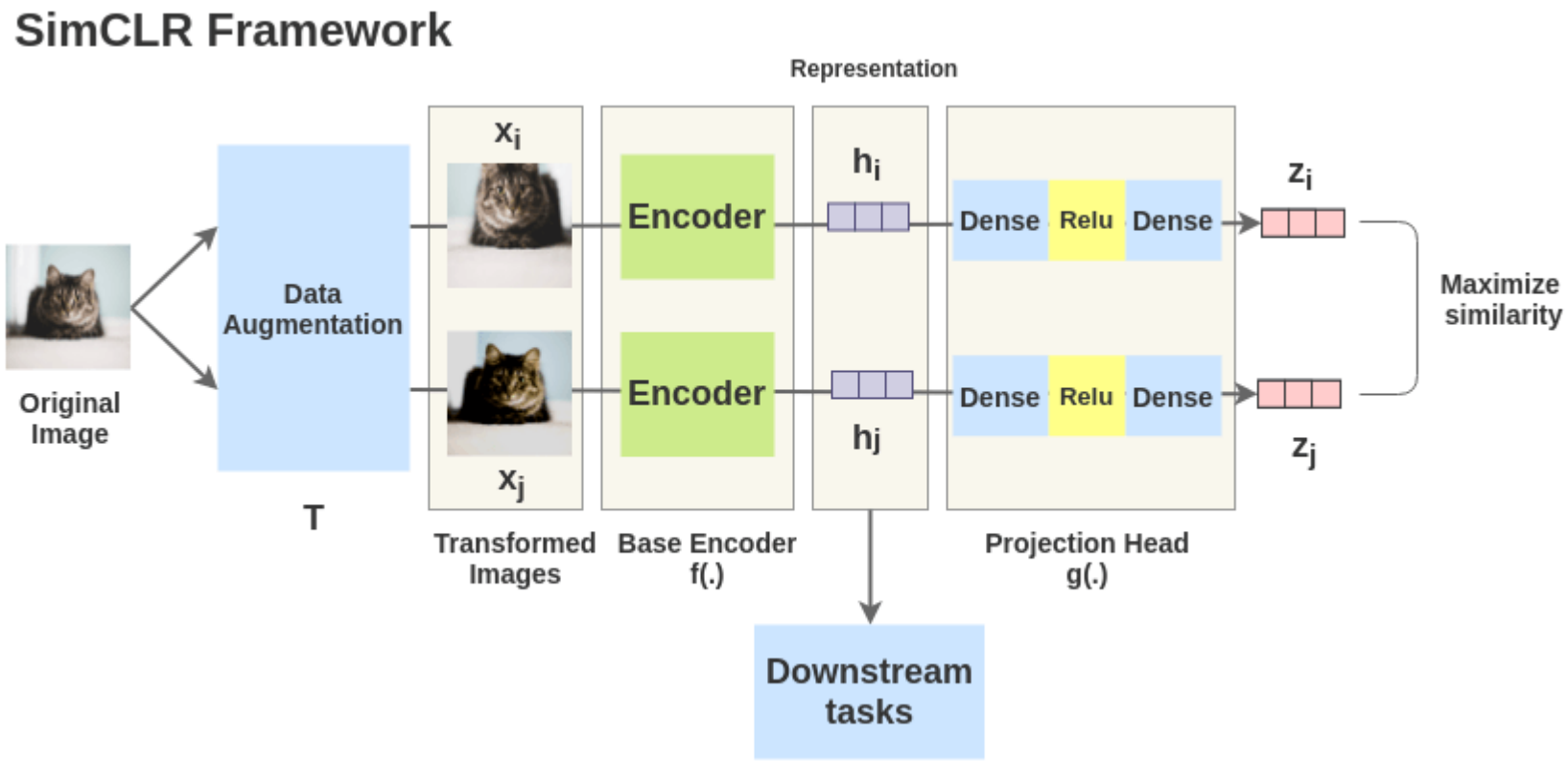
Contrastive Learning 이란?

- Self-supervised Learning의 한 방법
- 서로 다른 Input이 유사한지 아닌지 학습



Augmentation을 통해 self-supervised learning이 가능하다

SCL



Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

SCL

Abstract

- Add Supervised Contrastive Learning(SCL) term to the fine-tuning objective
  - ✓ Improves performance on several natural language understanding tasks from GLUE benchmark
  - ✓ Improve few-shot learning setting (20, 100, 1000 labeled examples)
  - ✓ Robust to the noise
  - ✓ Better generalization ability to related tasks

- Does not require any specialized architectures, memory banks, data augmentation of any kind

SCL

Approach

$$\mathcal{L} = (1 - \lambda) \cdot \mathcal{L}_{CE} + \lambda \mathcal{L}_{SCL} \tag{1}$$

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i) \tag{2}$$

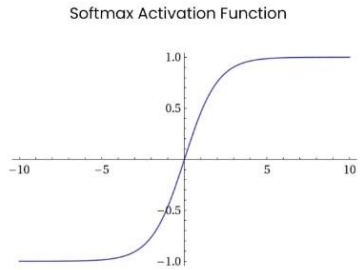
$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \log \frac{\exp\left(\Phi(x_i) \cdot \Phi(x_j)/\tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp\left(\Phi(x_i) \cdot \Phi(x_k)/\tau\right)} \tag{3}$$

N_{y_i} : total number of examples in the batch have the same label as y_i

τ : scalar temperature parameter

λ : scalar weighting hyperparameter

Φ(·) : L2 norm embedding final encoder hidden layer before softmax projection



SCL

Experiment

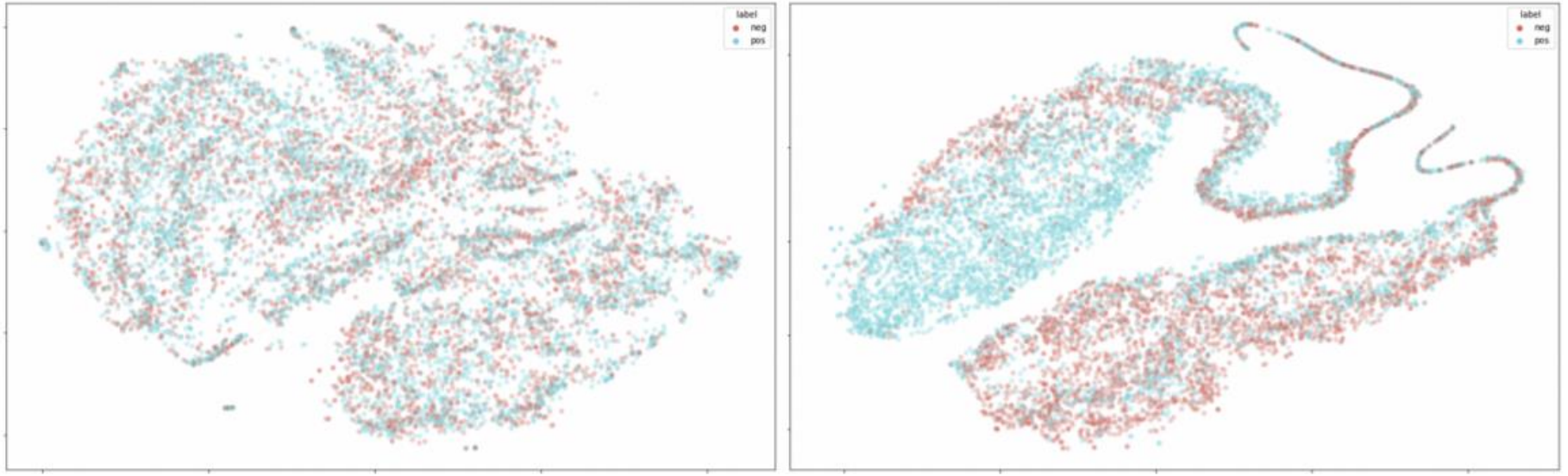


Figure 2: tSNE plots of learned CLS embedding on SST-2 test set where we have 20 labeled examples, comparing CE with and without SCL term. Blue: positive examples; red: negative examples.

Dataset	Task	Domain	#Train	#Classes
SST-2	sentiment analysis	movie reviews	67k	2
CoLA	grammatical correctness	linguistic publications	8.5k	2
MRPC	paraphrase	news	3.7k	2
RTE	textual entailment	news/Wikipedia	2.5k	2
QNLI	question answering/textual entailment	Wikipedia	105k	2
MNLI	textual entailment	multi-domain	393k	3

Table 1: GLUE Benchmark datasets used for evaluation.

- Run each experiment with 10 different seeds, pick the top model out of 10 seeds based on validation accuracy
- For few-shot learning, sample 10 different training set samples based on the total number of examples N
  - ✓ Taking the label distribution of the original training set into account
- Use fairseq library RoBERTa-Large model

GLUE BENCHMARK FULL DATASET RESULTS

Model	Loss	SST-2	CoLA	MRPC	RTE	QNLI	MNLI	Avg
$\text{RoBERTa}_{\text{Large}}$	CE	94.7	86.4	87.3	85.0	94.5	**90.0**	89.7
$\text{RoBERTa}_{\text{Large}}$	CE + SCL	**95.9**	**87.3**	**87.8**	**85.6**	**95.4**	89.9	**90.3**

Table 2: Results on the GLUE benchmark. We compare fine-tuning RoBERTa-Large with CE with and without SCL using the full training set of each task.

Model	Loss	Bsz	SST-2	CoLA	QNLI
$\text{RoBERTa}_{\text{Base}}$	CE + SCL	16	93.9	83.4	92.1
$\text{RoBERTa}_{\text{Base}}$	CE + SCL	64	94.2	84.8	92.7
$\text{RoBERTa}_{\text{Base}}$	CE + SCL	256	**94.3**	**84.9**	**92.9**

Table 3: Ablation study fine-tuning RoBERTa-Base with CE+SCL using the full training set of each task, increasing the batch size (Bsz).

GLUE BENCHMARK FEW-SHOT LEARNING RESULTS

Model	Loss	N	SST-2	QNLI	MNLI
$\text{RoBERTa}_{\text{Large}}$	CE	20	85.9±2.1	65.0±2.0	39.3±2.5
$\text{RoBERTa}_{\text{Large}}$	CE + SCL	20	**88.1±3.3**	**75.7±4.8**	**42.7±4.6**
$\text{RoBERTa}_{\text{Large}}$	CE	100	91.1±1.3	81.4±0.4	59.2±2.1
$\text{RoBERTa}_{\text{Large}}$	CE + SCL	100	**92.8±1.3**	**82.5±0.4**	**61.1±3.0**
$\text{RoBERTa}_{\text{Large}}$	CE	1000	94.0±0.6	89.2±0.6	81.4±0.2
$\text{RoBERTa}_{\text{Large}}$	CE + SCL	1000	**94.1±0.5**	**89.8±0.4**	**81.5±0.2**

여기에 수식을 입력하십시오.

Table 4: Few-shot learning results on the GLUE benchmark where we have N=20, 100, 1000 labeled examples for training. Reported results are the mean and the standard deviation of the test accuracies of the top 3 models based on validation accuracy out of 10 random training set samples.

ROBUSTNESS ACROSS AUGMENTED NOISY TRAINING DATASETS

Dataset	Loss	Original	T=0.3	T=0.5	T=0.7	T=0.9	Average
SST-2	CE	91.1±1.3	92.0±1.3	91.4±1.0	**91.7±1.3**	90.0±0.5	91.3±1.2
SST-2	CE + SCL	**92.8±1.3**	**92.6±0.9**	**91.5±1.0**	91.2±0.6	**91.5±1.0**	**91.7±1.0**
QNLI	CE	81.9±0.4	81.1±2.3	80.0±2.9	78.9±3.7	75.9±4.0	79.0±3.5
QNLI	CE + SCL	**82.5±0.4**	**82.7±1.9**	**81.9±2.5**	**81.3±0.6**	**80.1±2.5**	**81.5±2.0**
MNLI	CE	59.2±2.1	54.0±1.1	55.3±2.4	54.6±2.2	47.0±1.8	52.7±3.9
MNLI	CE + SCL	**61.1±3.0**	**61.2±2.3**	**62.1±0.9**	**62.3±1.1**	**53.0±2.1**	**59.7±4.3**

Table 5: Results on the GLUE benchmark for robustness across noisy augmented training sets. Average shows the average performance across augmented training sets.

Dataset	Type	Sentence
SST-2	Original	As possibly the best actor working in movies today.
SST-2	Augmented (T=0.3)	As perhaps the best actor who now stars in films.
SST-2	Original	The young stars are too cute; the story and ensuing complications are too manipulative.
SST-2	Augmented (T=0.9)	The babies are too cute, the image and complications that follow too manipulative.
QNLI	Original	Brain tissue is naturally soft, but can be stiffened with what liquid?
QNLI	Augmented (T=0.3)	Brain tissue is omitted naturally, but with what fluid it can be stiffened?
QNLI	Original	In March 1968, CBS and Sony formed CBS/Sony Records, a Japanese business joint venture.
QNLI	Augmented (T=0.9)	CBS was founded by CBS and Sony Records in March 1962, a Japanese company.
MNLI	Original	However, the link did not transfer the user to a comment box particular to the rule at issue.
MNLI	Augmented (T=0.3)	However, the link did not send the user to a comment field specifically for the rule.
MNLI	Original	Tenants could not enter the apartment complex due to a dangerous chemical spill.
MNLI	Augmented (T=0.9)	Tenants were banned from entering the medical property because of a blood positive substance.

Table 6: Sample of augmented examples with different noise levels for the robustness experiment shown in Table 5. Higher temperature (T) corresponds to more noise in the augmented training set.

GENERALIZATION ABILITY OF TASK MODELS

Model	Loss	N	Amazon-2	Yelp-2
$\text{RoBERTa}_{\text{Large}}$	CE	40	87.4±6.4	90.8±2.2
$\text{RoBERTa}_{\text{Large}}$	CE + SCL	40	**90.3±0.6**	**91.2±0.4**

Table 7: Generalization of the SST-2 task model (fine-tuned using the full training set) to related tasks (Amazon-2, Yelp-2) where there are 20 labeled examples for each class.

- Fine-tuning on SST-2 full train set
- Transfer this model to two related single sentence sentiment analysis binary classification task
  - ✓ Amazon-2
  - ✓ Yelp-2
- Sample 20 labeled examples for each class, and follow the few-shot learning experimental setup

Reference

- Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning 리뷰, HOYA012'S RESEARCH BLOG, https://hoya012.github.io/blog/byol/
- PR-231: A Simple Framework for Contrastive Learning of Visual Representations, JinWon Lee, https://youtu.be/FWhM3juUM6s
- Self-Supervised Learning(Algorithm&application), Seokho Moon, file:///C:/Users/wonhyuk/Desktop/20201120_Self_Supervised_Representation_Learning_Seokho.pdf
- Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.
- Gunel, Beliz, et al. "Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning." *arXiv preprint arXiv:2011.01403* (2020).
- Temperature Scaling, Calibration: On Calibration of Modern Neural Netwoks, Curaai00's Deep Learning Blog, https://curaai00.tistory.com/10
- The Illustrated SimCLR Framework, Amit Chaudhary, https://amitness.com/2020/03/illustrated-simclr/
- SimCLR을 이용한 향상된 자기주도 및 반주도 학습, 시나브로의 테크산책, https://brunch.co.kr/@synabreu/76

End