SimCSE: Simple Contrastive Learning of Sentence Embeddings

ADVISOR: Jia-Ling Koh
PRESENTER: Xiao-Yuan Hung
SOURCE: ACL' 22
DATE: 2023/05/09

Outline

1 Introduction
- Problem
- Solution

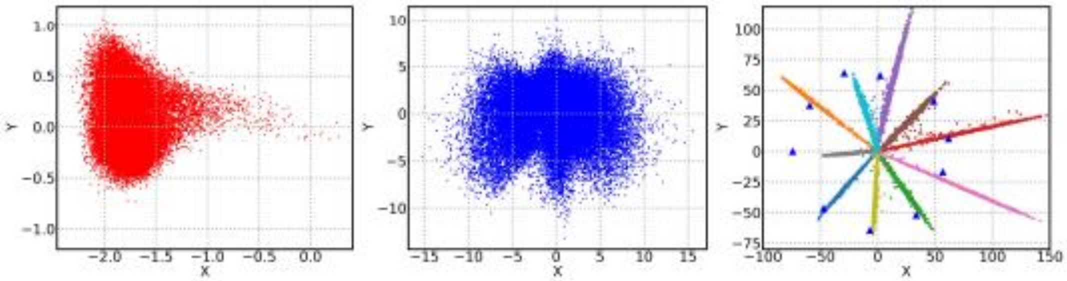
2 Method

3 Experiment

4 Conclusion

Problem

- **Pre-trained embeddings are anisotropy(各向異性)**
  - word embeddings occupy a narrow cone in the vector space

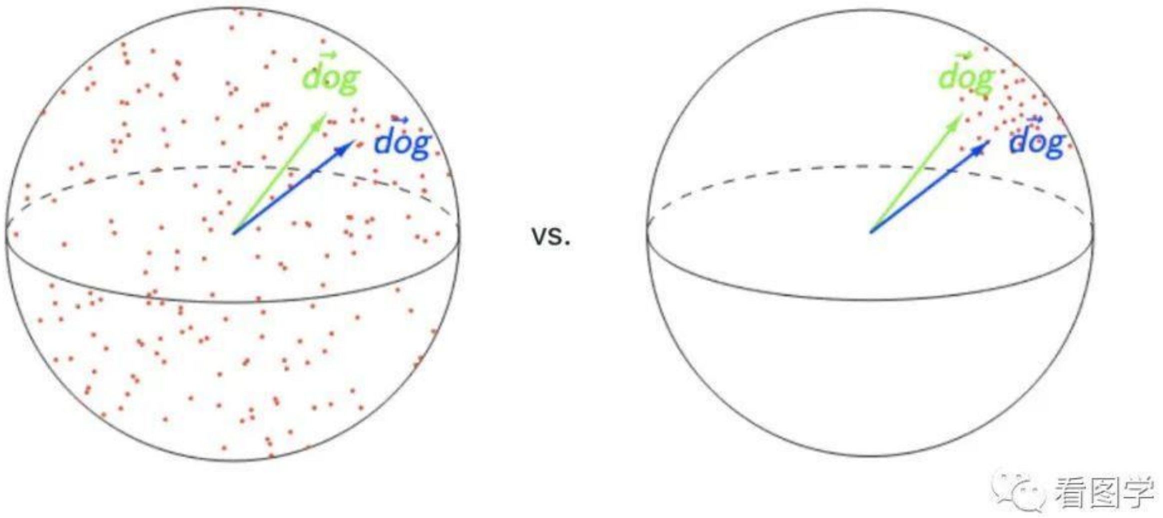


(a) Vanilla Transformer (b) Word2Vec (c) Classification

Gao et al. 2019

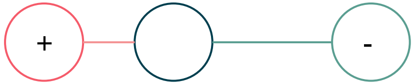
Problem

- **Methods**
  - BERT-Flow
  - BERT-Whitening



isotropic Anisotropic

Solution



- **Pre-trained embedding + Contrastive Learning**

- **Unsupervised**
  - Uses standard dropout as data augmention
- **Supervised**
  - uses entailment + contradiction pairs from NLI datasets

https://www.youtube.com/watch?v=u-OQVQSvx38&t=251s

Outline

1 Introduction

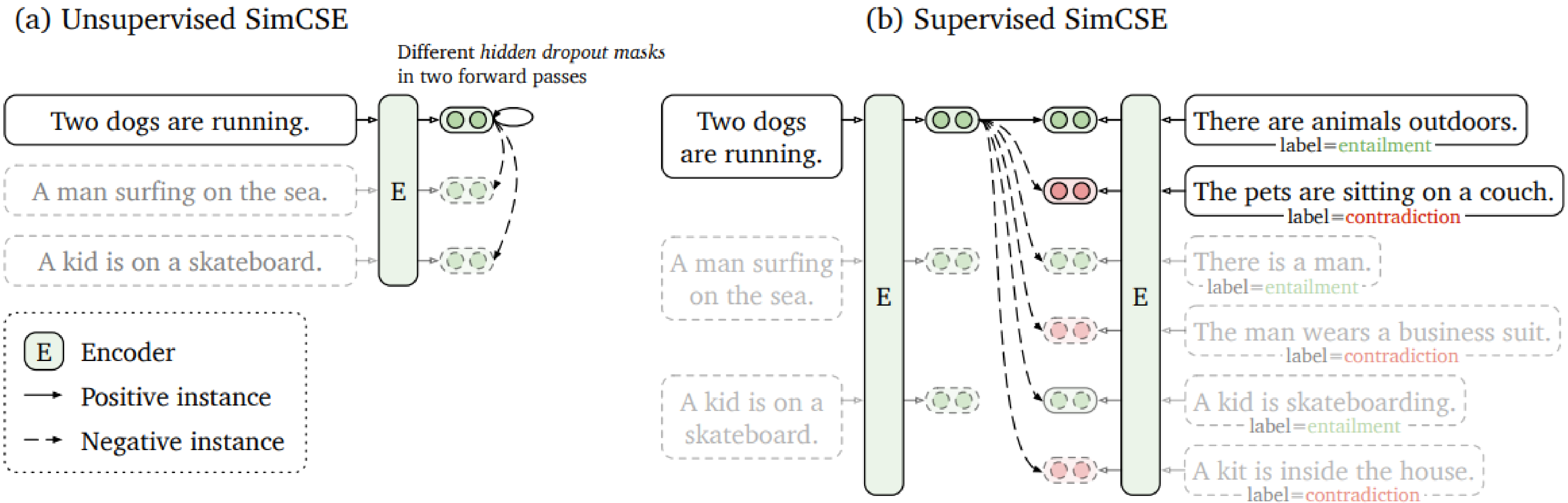
2 Method

3 Experiment

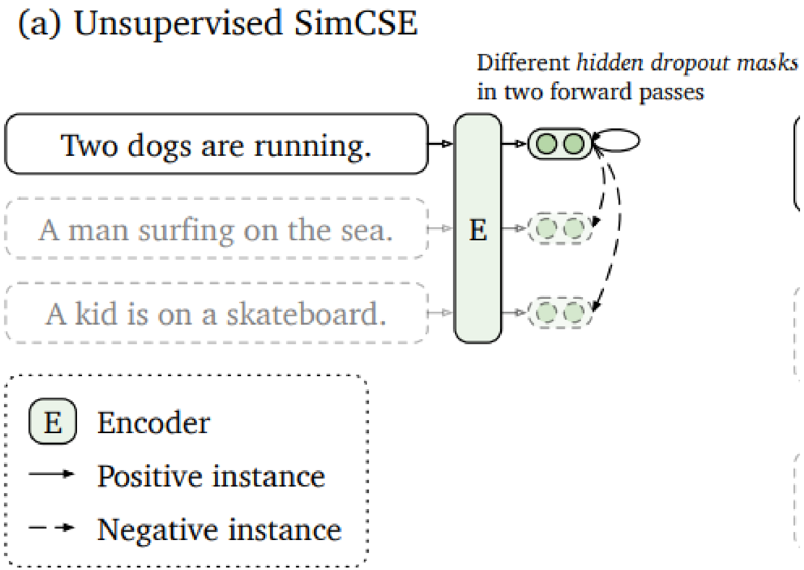
4 Conclusion

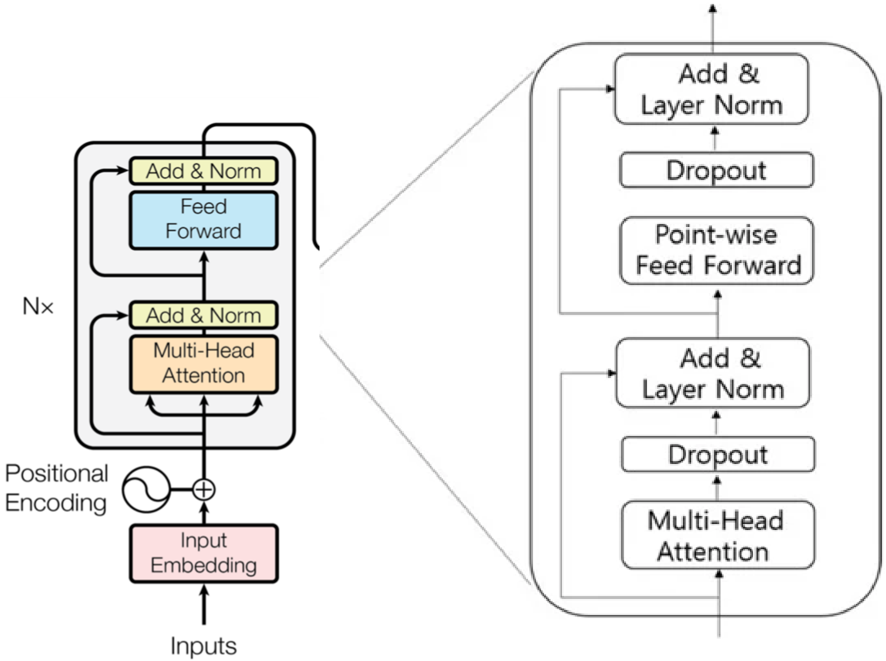
- Unspervised SimCSE
  - method
  - Experiment
- Supervised SimCSE
  - method
  - Experiment

SimCSE

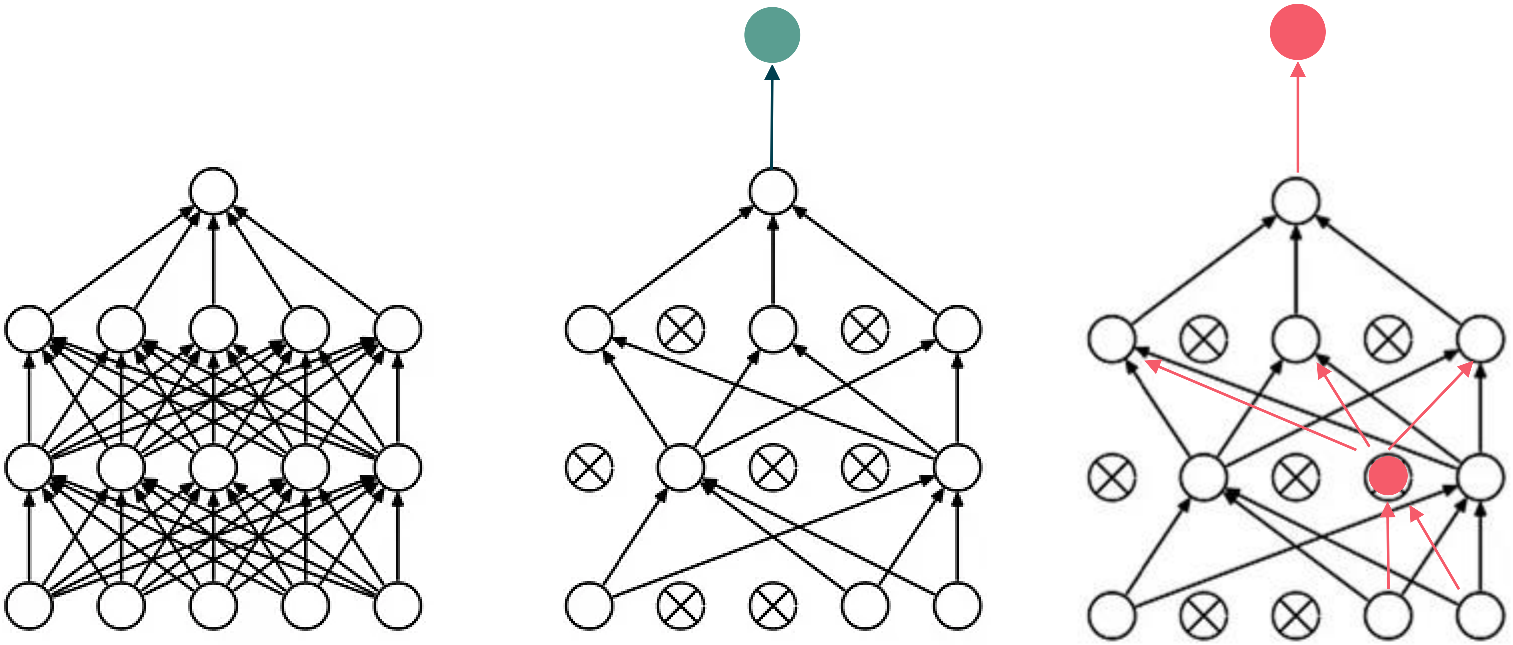

(a) Unsupervised SimCSE
Different *hidden dropout masks* in two forward passes
Two dogs are running.
A man surfing on the sea.
A kid is on a skateboard.
E
E Encoder
Positive instance
Negative instance
(b) Supervised SimCSE
Two dogs are running.
A man surfing on the sea.
A kid is on a skateboard.
E
E
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

Unsupervised : BERT Dropout





Dropout



(a) Standard Neural Net (b) After applying dropout.

Two dogs are running.

https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

Loss function

temperature

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

<-- postive sample

<-- negative sample

encoder sentence

$$\mathbf{h}_i^z = f_\theta(x_i, z)$$

random mask for dropout

sim = $\frac{\mathbf{h}_1^\top \mathbf{h}_2}{\|\mathbf{h}_1\| \cdot \|\mathbf{h}_2\|}$

Comparison of data augmentations

- **STS-B development set**
- **Spearman's correlation**

Different dropout

- SimCSE (dropout) NLP is interesting. —— NLP is interesting.

Compare it to

- Next sentence I do NLP. —— NLP is interesting.
- Synonym replacement The movie is great. —— The movie is fantastic.
- Crop ~~Two dogs~~ are running. —— ~~Two~~ dogs are ~~running.~~
- Delete one word Two dogs are ~~running~~. —— Two dogs ~~are~~ running.

examples of data augmentation

Data augmentation			STS-B
None (unsup. SimCSE)			**82.5**
Crop	*10%*	*20%*	*30%*
	77.8	71.4	63.6
Word deletion	*10%*	*20%*	*30%*
	75.9	72.2	68.2
Delete one word			75.9
w/o dropout			74.2
Synonym replacement			77.4
MLM 15%			62.2

STS-B development set example

sentence1 (string)	sentence2 (string)	similarity_score (float32)
"A man with a hard hat is dancing."	"A man wearing a hard hat is dancing."	5
"A young child is riding a horse."	"A child is riding a horse."	4.75
"A man is feeding a mouse to a snake."	"The man is feeding a mouse to the snake."	5
"A woman is playing the guitar."	"A man is playing guitar."	2.4
"A woman is playing the flute."	"A man is playing a flute."	2.75
"A woman is cutting an onion."	"A man is cutting onions."	2.615
"A man is erasing a chalk board."	"The man is erasing the chalk board."	5
"A woman is carrying a boy."	"A woman is carrying her baby."	2.333
"Three men are playing guitars."	"Three men are on stage playing guitars."	3.75
"A woman peels a potato."	"A woman is peeling a potato."	5

https://huggingface.co/datasets/stsb_multi_mt/viewer/en/dev

Spearman's rank correlation coefficient

$X_{i\ (STS\text{-}B)}$	$Y_{i(data\ augmentations)}$
5	4.75
4.75	3.5
1.25	1.5
3.15	3.75
2.45	1

Spearman's rank correlation coefficient

$X_{i\ (STS\text{-}B)}$	Y_i (data augmentations)	$x_{i\ (rank)}$	$y_{i\ (rank)}$	d_i	d_i^2
1.25	1.5	1	2	-1	1
2.45	1	2	1	1	1
3.15	3.75	3	4	-1	1
4.75	3.5	4	3	1	1
5	4.75	5	5	0	0

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6*4}{5*(5^2-1)} = 0.8$$

Comparison of different unsupervised objectives

- **SimCSE objectives**
  - self-prediction

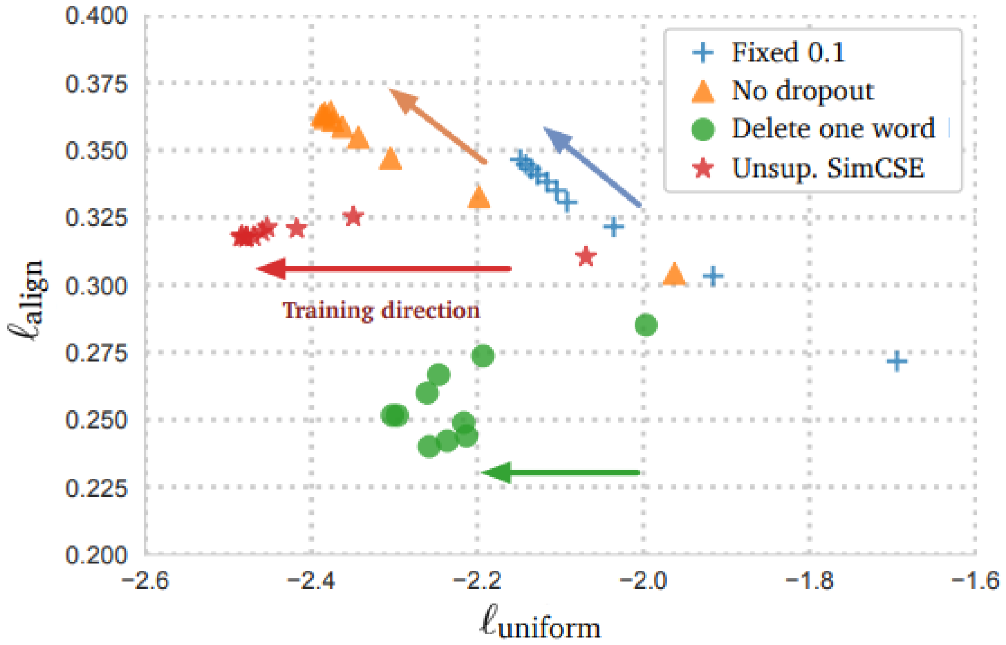
Training objective	f_θ	$(f_{\theta_1}, f_{\theta_2})$
Next sentence	67.1	68.9
Next 3 sentences	67.4	68.8
Delete one word	75.9	73.1
Unsupervised SimCSE	**82.5**	80.7

Effects of different dropout probabilities

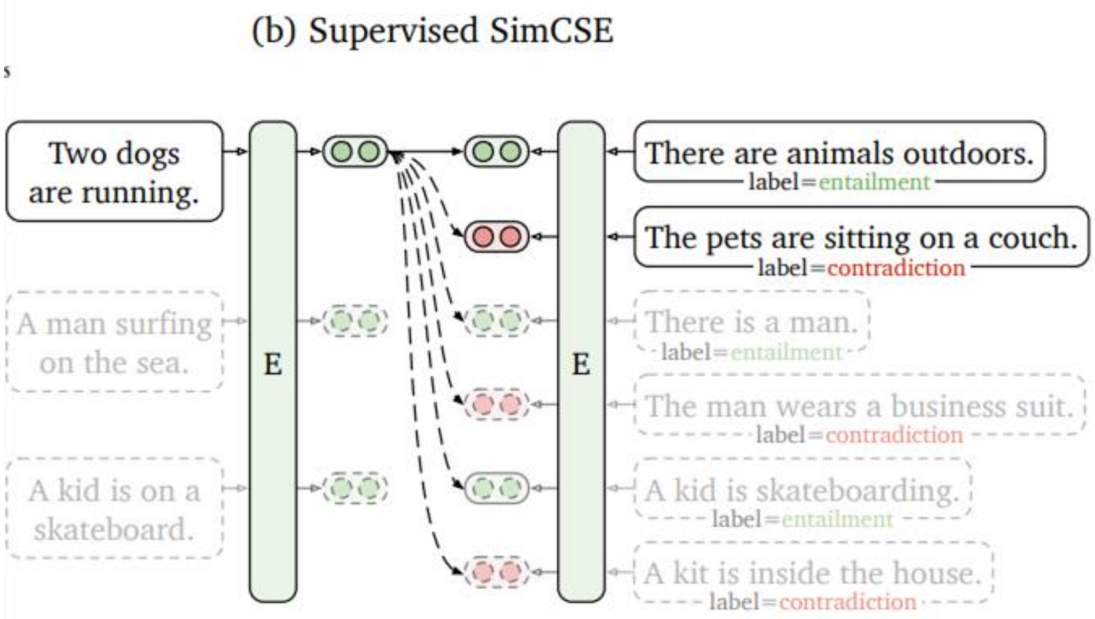
p	*0.0*	*0.01*	*0.05*	*0.1*
STS-B	71.1	72.6	81.1	**82.5**
p	*0.15*	*0.2*	*0.5*	*Fixed 0.1*
STS-B	81.4	80.5	71.0	43.6

ℓ_{align}-$\ell_{uniform}$ plot(unsupervised SimCSE)

- All models improve uniformity
- Unsupervised SimCSE keeps a steady alignment



Supervised : NLI dataset



Choices of labeled data

1. **QQP(Quora question pairs)**
2. **Flickr30k**
   a. each image is annotated with 5 human-written captions
   b. consider any two captions of the same image as a positive pair
3. **ParaNMT**
   a. a large-scale back-translation paraphrase dataset
4. **NLI : SNLI + MNLI**

Dataset	sample	full
Unsup. SimCSE (1m)	-	82.5
QQP (134k)	81.8	81.8
Flickr30k (318k)	81.5	81.4
ParaNMT (5m)	79.7	78.7
SNLI+MNLI		
entailment (314k)	**84.1**	**84.9**
neutral (314k)[8]	82.6	82.9
contradiction (314k)	77.5	77.6
all (942k)	81.7	81.9

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

QQP ↑

ParaNMT-50M↓

Reference Translation	Machine Translation
so, what's half an hour?	half an hour won't kill you.
well, don't worry. i've taken out tons and tons of guys. lots of guys.	don't worry, i've done it to dozens of men.
it's gonna be ...... classic.	yeah, sure. it's gonna be great.
greetings, all!	hello everyone!
but she doesn't have much of a case.	but as far as the case goes, she doesn't have much.
it was good in spite of the taste.	despite the flavor, it felt good.

Table 2: Example paraphrase pairs from PARANMT-50M, where each consists of an English reference translation and the machine translation of the Czech source sentence (not shown).

Relevant Descriptions:
1: A person parasails on the crest of a wave.
2: A windsurfer in the waves of the ocean.
3: A man rides large waves on a wind sail.
4: A man windsurfs in the ocean.
5: A man parasails in the waves.

Given one premise,

- Premise: *There are two dogs running.*

Annotators are required to write hypotheses of

- Entailment: *There are animals outdoors.*
- Contradiction: *The pets are sitting on a couch.*
- Neutral: *The dogs are catching a ball.*

Flickr30k

SNLI+MNLI

Supervised : NLI dataset

Premise

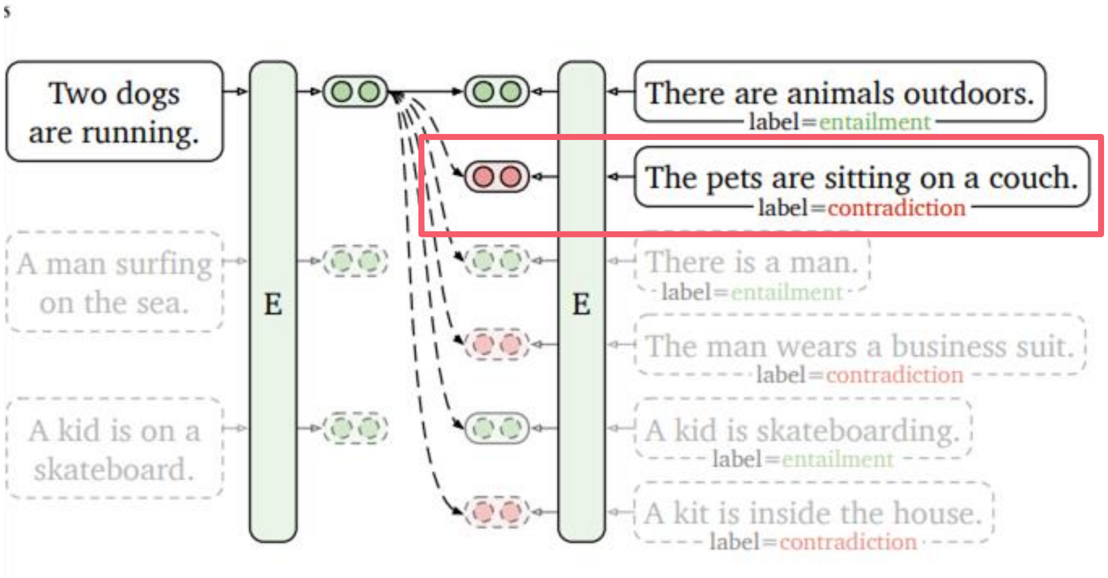
Entailment hypothesis

$$-\log \frac{e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N}\left(e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^+)/\tau}+e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^-)/\tau}\right)}$$

Contradiction hypothesis
+ in-batch negatives

Dataset	sample	full
SNLI+MNLI		
entailment + hard neg.	-	**86.2**
+ ANLI (52k)	-	85.0

(b) Supervised SimCSE



Outline

Dataset

- **STS(2012-2016)**
  - a set of semantic textual similarity datasets
- **STS Benchmark**
  - include text from image captions, news headlines and user forums.
  - include STS(2012-2016)
- **SICK Relatedness**
  - a dataset for compositional distributional semantics

STS12 - Semeval-2012 task 6: A pilot on semantic textual similarity

STS13 - SEM 2013 shared task: Semantic Textual Similarity

STS14 - SemEval-2014 task 10: Multilingual semantic textual similarity

STS15 - SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability

STS16 - SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation

Unsupervised Models

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
$\text{BERT}_{\texttt{base}}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\texttt{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\texttt{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\texttt{base}}$♡	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-$\text{BERT}_{\texttt{base}}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
∗ SimCSE-$\text{BERT}_{\texttt{base}}$	**68.40**	**82.41**	**74.38**	**80.91**	**78.56**	**76.85**	**72.23**	**76.25**
$\text{RoBERTa}_{\texttt{base}}$ (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\texttt{base}}$-whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-$\text{RoBERTa}_{\texttt{base}}$	52.41	75.19	65.52	77.12	78.63	72.41	**68.62**	69.99
∗ SimCSE-$\text{RoBERTa}_{\texttt{base}}$	**70.16**	**81.77**	**73.24**	**81.36**	**80.65**	**80.22**	68.56	**76.57**
∗ SimCSE-$\text{RoBERTa}_{\texttt{large}}$	**72.86**	**83.99**	**75.62**	**84.77**	**81.80**	**81.98**	**71.26**	**78.90**

Supervised Models

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
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
$\text{SBERT}_{\texttt{base}}$♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{SBERT}_{\texttt{base}}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{SBERT}_{\texttt{base}}$-whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
$\text{CT-SBERT}_{\texttt{base}}$	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
∗ $\text{SimCSE-BERT}_{\texttt{base}}$	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
$\text{SRoBERTa}_{\texttt{base}}$♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
$\text{SRoBERTa}_{\texttt{base}}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
∗ $\text{SimCSE-RoBERTa}_{\texttt{base}}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
∗ $\text{SimCSE-RoBERTa}_{\texttt{large}}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

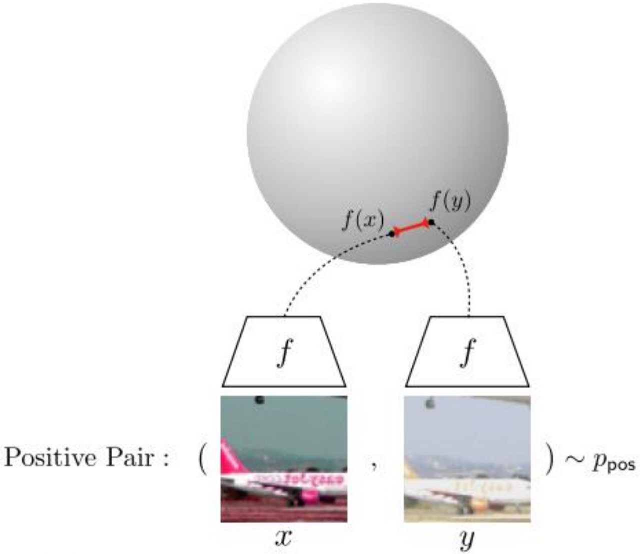
Ablation

Pooler	**Unsup.**	**Sup.**
[CLS]		
w/ MLP	81.7	**86.2**
w/ MLP (train)	**82.5**	85.8
w/o MLP	80.9	**86.2**
First-last avg.	81.2	86.1

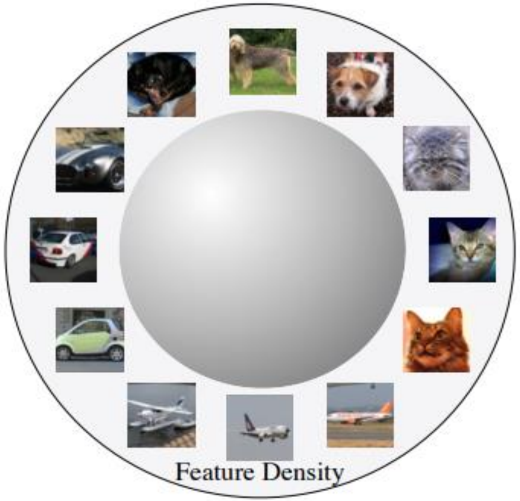
Table 6: Ablation studies of different pooling methods in unsupervised and supervised SimCSE. *[CLS] w/ MLP (train)*: using MLP on [CLS] during training but removing it during testing. The results are based on the development set of STS-B using $\text{BERT}_{\texttt{base}}$.

Two key properties related to the contrastive learning

Wang and Isola (2020)



Alignment: Similar samples have similar features.



Uniformity: Preserve maximal information.

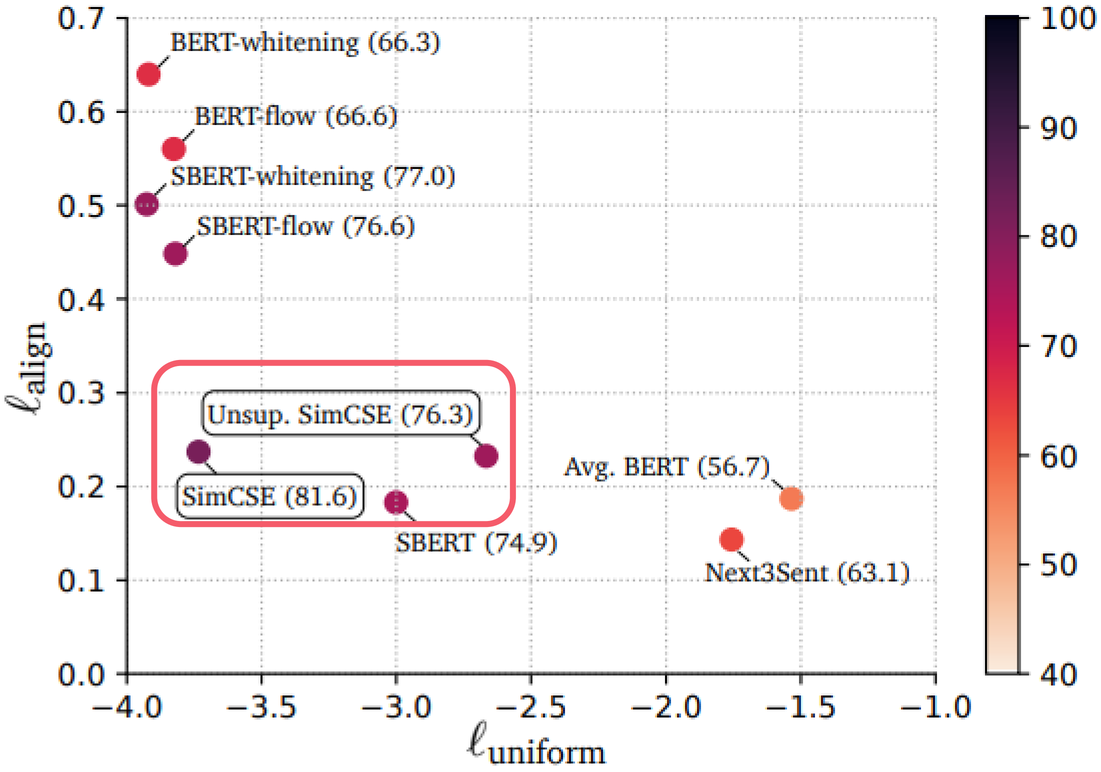
$$\ell_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}} \|f(x) - f(x^+)\|^2.$$

encoder f is perfectly aligned if f(x) = f(y)

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

encoder f is perfectly uniform if the distribution of f(x) for x ~ Pdata

ℓalign-ℓuniform plot of models



Color of points and numbers in brackets represent average STS performance

Case Study

	$\text{SBERT}_{\text{base}}$	Supervised SimCSE-$\text{BERT}_{\text{base}}$
Query: A man riding a small boat in a harbor.		有一個男子在船上
#1	A group of men traveling over the ocean in a small boat.	A man on a moored blue and white boat.
#2	Two men sit on the bow of a colorful boat.	A man is riding in a boat on the water.
#3	A man wearing a life jacket is in a small boat on a lake.	A man in a blue boat on the water.
Query: A dog runs on the green grass near a wooden fence.		有一隻狗在草地上
#1	A dog runs on the green grass near a grove of trees.	The dog by the fence is running on the grass.
#2	A brown and white dog runs through the green grass.	Dog running through grass in fenced area.
#3	The dogs run in the green field.	A dog runs on the green grass near a grove of trees.

Outline

Conclusion

- This paper propose a simple contrastive learning framework that outperforms most existing models.
- Using supervised learning also makes the overall effect better.