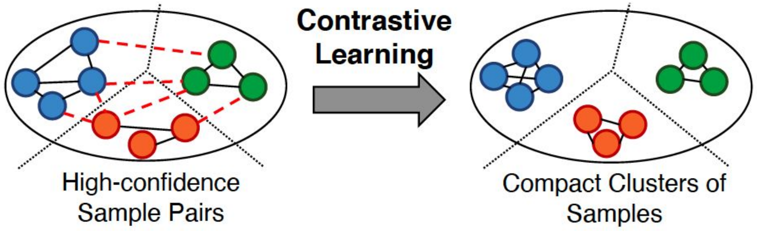
# SimCSE: Simple Contrastive Learning of Sentence Embeddings

Tianyu Gao, Xingcheng Yao, and Danqi Chen

UBC-DLNLP Reading Group, September 23, 2021
Presenter: Chiyu Zhang

# Background




**Contrastive learning** aims to learn effective representation by pulling semantically close neighbors together and pushing apart non-neighbors.

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

where $\tau$ is a temperature hyperparameter and $\mathrm{sim}(\mathbf{h}_1, \mathbf{h}_2)$ is the cosine similarity $\frac{\mathbf{h}_1^\top \mathbf{h}_2}{\|\mathbf{h}_1\| \cdot \|\mathbf{h}_2\|}$. In this work, we encode input sentences using a pre-trained language model such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019): $\mathbf{h} = f_\theta(x)$, and then fine-tune all the parameters using the contrastive learning objective (Eq. 1).

https://arxiv.org/abs/2010.07835

# Contrastive Learning

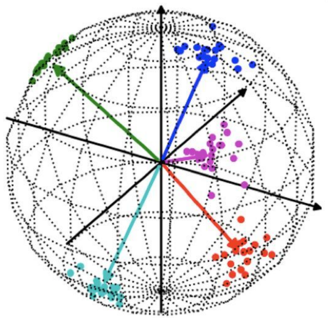
One critical question in contrastive learning is how to construct $(x_i, x_i^+)$ pairs.

**Vision:** Two random transformations of the same image (e.g., cropping, flipping, distortion and rotation)

**Language**: Word deletion, reordering, and substitution.

**This paper:** Dropout noise as data augmentation.

# Contrastive Learning



Two key properties to measure the quality of representation:

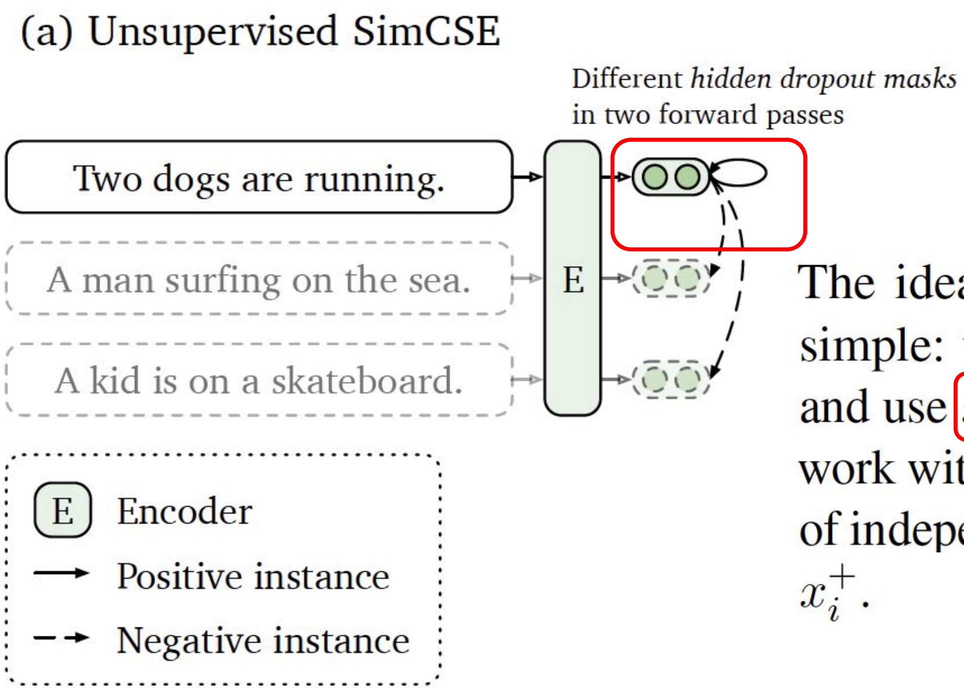
**Alignment:** Given a distribution of positive pairs $p_{pos}$, alignment calculates expected distance between embeddings of the paired instances. **Positive** instances should stay **close.**

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

**Uniformity:** Uniformity measures how well the embeddings are uniformly distributed. **Random** instances should **scatter** on the hypersphere.

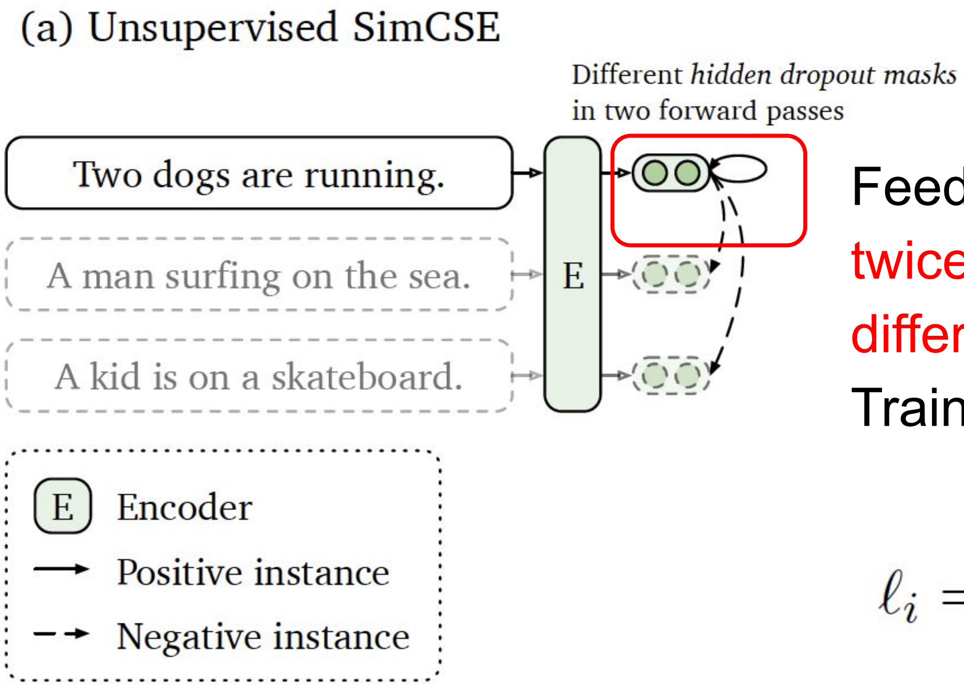
$$\ell_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2},$$

# Unsupervised SimCSE




The idea of unsupervised SimCSE is extremely simple: we take a collection of sentences $\{x_i\}_{i=1}^m$ and use $\boxed{x_i^+ = x_i.}$ The key ingredient to get this to work with identical positive pairs is through the use of independently sampled *dropout masks* for $x_i$ and $x_i^+$.

# Unsupervised SimCSE




Feed the same input to the encoder twice and get two embeddings with different dropout masks *z*, *z'*.

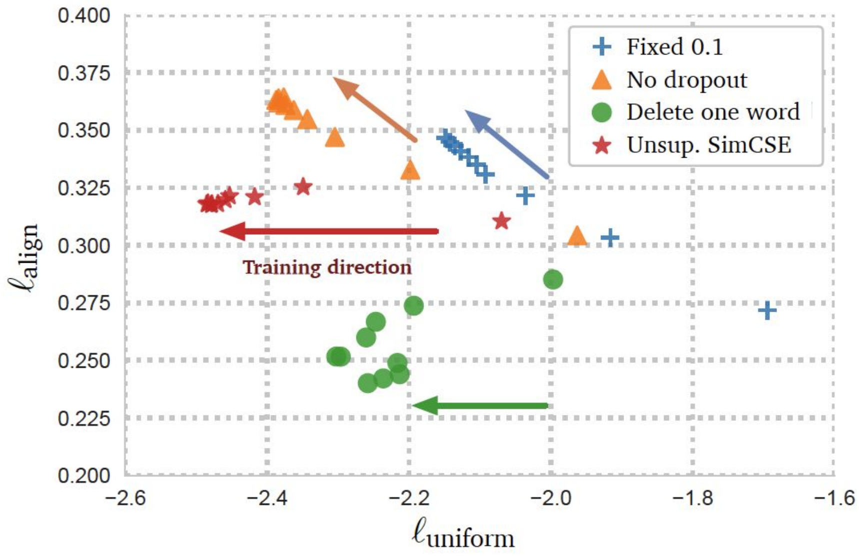
Training objective:

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

# Unsupervised SimCSE

| **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 |

| | | | | |
|---|---|---|---|---|
| $p$ | *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 |

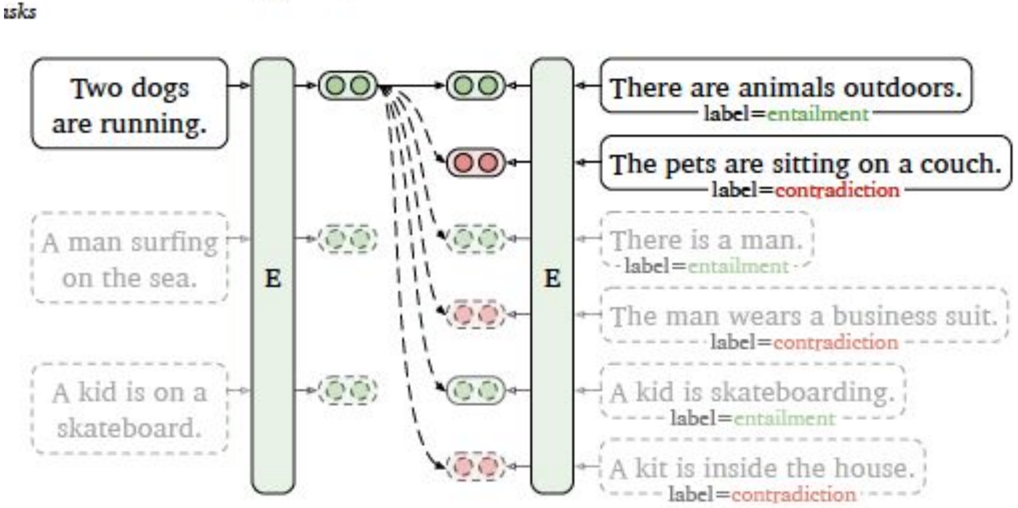


# Supervised SimCSE

Use supervised natural language inference (NLI) datasets:

Given one premise, annotators are required to manually write one sentence that is absolutely true (entailment), one that might be true (neutral), and one that is definitely false (contradiction).

(b) Supervised SimCSE



# Supervised SimCSE

Use supervised natural language inference (NLI) datasets:

Given one premise, annotators are required to manually write one sentence that is absolutely true (entailment), one that might be true (neutral), and one that is definitely false (contradiction).

Formally, we extend $(x_i, x_i^+)$ to $(x_i, x_i^+, x_i^-)$, where $x_i$ is the premise, $x_i^+$ and $x_i^-$ are entailment and contradiction hypotheses. The training objective $\ell_i$ is then defined by ($N$ is mini-batch size):

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} \left( e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau} + e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^-)/\tau} \right)}. \tag{5}$$

# Supervised SimCSE

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

Comparisons of different supervised datasets as positive pairs.

# Experiment

Evaluate on 7 semantic textual similarity tasks

| 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 |
| $\text{IS-BERT}_{\texttt{base}}$ ♡ | 56.77 | 69.24 | 61.21 | 75.23 | 70.16 | 69.21 | 64.25 | 66.58 |
| $\text{CT-BERT}_{\texttt{base}}$ | 61.63 | 76.80 | 68.47 | 77.50 | 76.48 | 74.31 | 69.19 | 72.05 |
| ∗ $\text{SimCSE-BERT}_{\texttt{base}}$ | **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 |
| $\text{DeCLUTR-RoBERTa}_{\texttt{base}}$ | 52.41 | 75.19 | 65.52 | 77.12 | 78.63 | 72.41 | **68.62** | 69.99 |
| ∗ $\text{SimCSE-RoBERTa}_{\texttt{base}}$ | **70.16** | **81.77** | **73.24** | **81.36** | **80.65** | **80.22** | 68.56 | **76.57** |
| ∗ $\text{SimCSE-RoBERTa}_{\texttt{large}}$ | **72.86** | **83.99** | **75.62** | **84.77** | **81.80** | **81.98** | **71.26** | **78.90** |
| *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** |

# Experiment

Add MLM objective

| Model | STS-B | Avg. transfer |
|---|---|---|
| w/o MLM | **86.2** | 85.8 |
| w/ MLM | | |
| $\lambda = 0.01$ | 85.7 | 86.1 |
| $\lambda = 0.1$ | 85.7 | **86.2** |
| $\lambda = 1$ | 85.1 | 85.8 |

Table D.2: Ablation studies of the MLM objective based on the development sets using $\text{BERT}_{\texttt{base}}$.

# Experiment

## Transfer Tasks

- sentence-level objective may not directly benefit transfer tasks.

| Model | MR | CR | SUBJ | MPQA | SST | TREC | MRPC | Avg. |
|---|---|---|---|---|---|---|---|---|
| | | | | *Unsupervised models* | | | | |
| GloVe embeddings (avg.)$^\clubsuit$ | 77.25 | 78.30 | 91.17 | 87.85 | 80.18 | 83.00 | 72.87 | 81.52 |
| Skip-thought$^\heartsuit$ | 76.50 | 80.10 | 93.60 | 87.10 | 82.00 | 92.20 | 73.00 | 83.50 |
| Avg. BERT embeddings$^\clubsuit$ | 78.66 | 86.25 | 94.37 | 88.66 | 84.40 | **92.80** | 69.54 | 84.94 |
| BERT-[CLS]embedding$^\clubsuit$ | 78.68 | 84.85 | 94.21 | 88.23 | 84.13 | 91.40 | 71.13 | 84.66 |
| IS-BERT$_{\text{base}}$$^\heartsuit$ | 81.09 | 87.18 | 94.96 | 88.75 | 85.96 | 88.64 | 74.24 | 85.83 |
| ∗ SimCSE-BERT$_{\text{base}}$ | 81.18 | 86.46 | 94.45 | **88.88** | 85.50 | 89.80 | 74.43 | 85.81 |
| w/ MLM | **82.92** | **87.23** | **95.71** | 88.73 | **86.81** | 87.01 | **78.07** | **86.64** |
| ∗ SimCSE-RoBERTa$_{\text{base}}$ | 81.04 | 87.74 | 93.28 | 86.94 | 86.60 | 84.60 | 73.68 | 84.84 |
| w/ MLM | **83.37** | **87.76** | **95.05** | **87.16** | **89.02** | **90.80** | **75.13** | **86.90** |
| ∗ SimCSE-RoBERTa$_{\text{large}}$ | 82.74 | 87.87 | 93.66 | **88.22** | 88.58 | 92.00 | 69.68 | 86.11 |
| w/ MLM | **84.66** | **88.56** | **95.43** | 87.50 | **89.46** | **95.00** | **72.41** | **87.57** |
| | | | | *Supervised models* | | | | |
| InferSent-GloVe$^\clubsuit$ | 81.57 | 86.54 | 92.50 | 90.38 | 84.18 | 88.20 | 75.77 | 85.59 |
| Universal Sentence Encoder$^\clubsuit$ | 80.09 | 85.19 | 93.98 | 86.70 | 86.38 | 93.20 | 70.14 | 85.10 |
| SBERT$_{\text{base}}$$^\clubsuit$ | **83.64** | **89.43** | 94.39 | **89.86** | **88.96** | **89.60** | 76.00 | **87.41** |
| ∗ SimCSE-BERT$_{\text{base}}$ | 82.69 | 89.25 | **94.81** | 89.59 | 87.31 | 88.40 | 73.51 | 86.51 |
| w/ MLM | 82.68 | 88.88 | 94.52 | 89.82 | 88.41 | 87.60 | **76.12** | 86.86 |
| SRoBERTa$_{\text{base}}$ | 84.91 | 90.83 | 92.56 | 88.75 | 90.50 | 88.60 | **78.14** | 87.76 |
| ∗ SimCSE-RoBERTa$_{\text{base}}$ | 84.92 | **92.00** | **94.11** | **89.82** | 91.27 | 88.80 | 75.65 | 88.08 |
| w/ MLM | **85.08** | 91.76 | 94.02 | 89.72 | **92.31** | **91.20** | 76.52 | **88.66** |
| ∗ SimCSE-RoBERTa$_{\text{large}}$ | 88.12 | 92.37 | 95.11 | 90.49 | 92.75 | 91.80 | 76.64 | 89.61 |
| w/ MLM | **88.45** | **92.53** | **95.19** | **90.58** | **93.30** | **93.80** | **77.74** | **90.23** |

# Ablation Studies

Hard negatives

Pooling methods

| **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 |

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

| **Hard neg** | N/A | Contradiction | | | Contra.+ Neutral |
|---|---|---|---|---|---|
| $\alpha$ | - | 0.5 | 1.0 | 2.0 | 1.0 |
| **STS-B** | 84.9 | 86.1 | **86.2** | **86.2** | 85.3 |

# Analysis: Anisotropy

The anisotropy problem is naturally connected to **uniformity**, both highlighting that embeddings should be **evenly distributed** in the space.

Take a singular spectrum perspective—which is a common practice in analyzing word embeddings.

**Singular value** drops the fastest for vanilla BERT or SBERT embeddings, while SimCSE helps flatten the spectrum distribution.

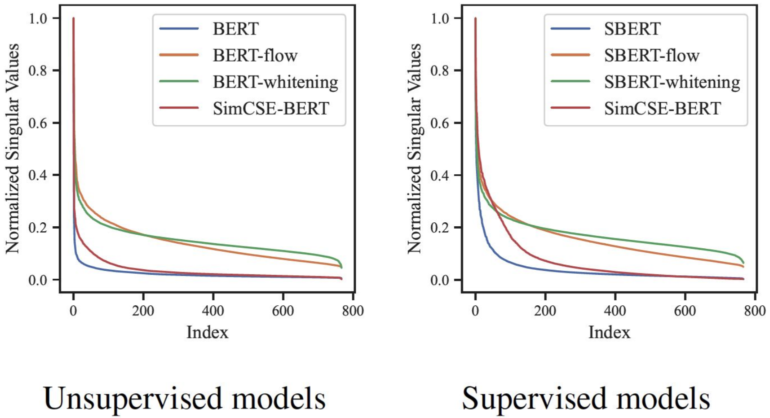



Figure F.1: Singular value distributions of sentence embedding matrix from sentences in STS-B. We normalize the singular values so that the largest one is 1.

# Analysis:

(1) Pre-trained embeddings: good alignment, poor uniformity (i.e., highly anisotropic);

(2) Post-processing methods (BERT-flow and BERT-whitening): improve uniformity, a degeneration in alignment;

(3) Unsupervised SimCSE : improves uniformity and keeping a good alignment

(4) Supervised SimCSE further amends alignment?.

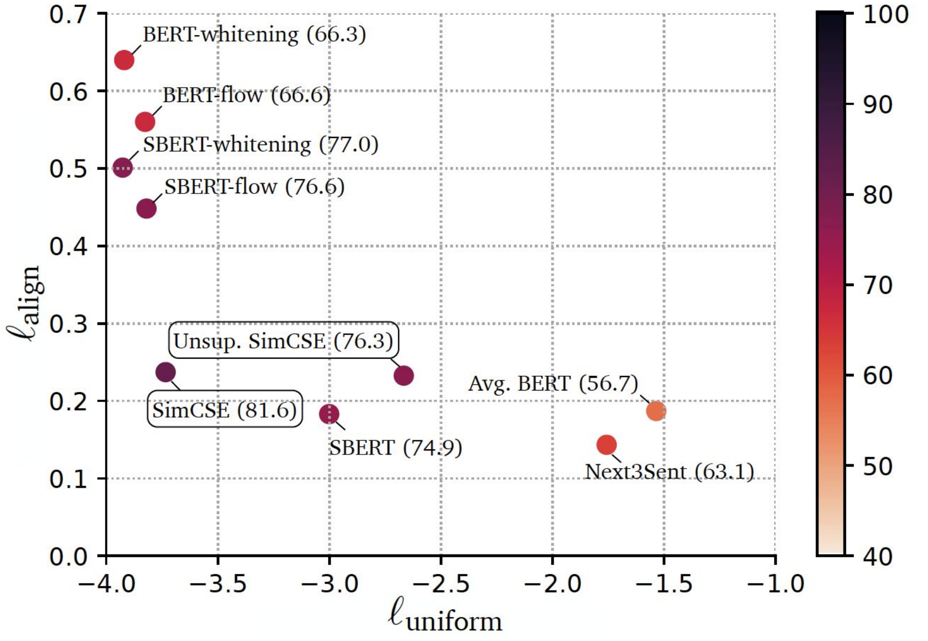



Figure 3: $\ell_{\text{align}}$-$\ell_{\text{uniform}}$ plot of models based on $\text{BERT}_{\text{base}}$. Color of points and numbers in brackets represent average STS performance (Spearman's correlation). *Next3Sent*: "next 3 sentences" from Table 2.

# Reference


- https://arxiv.org/pdf/2104.08821.pdf
- https://openaccess.thecvf.com/content_cvpr_2018/papers/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.pdf
- https://github.com/princeton-nlp/SimCSE
- https://lilianweng.github.io/lil-log/2021/05/31/contrastive-representation-learning.html#unsupervised-sentence-embedding-learning
- https://arxiv.org/abs/2010.07835