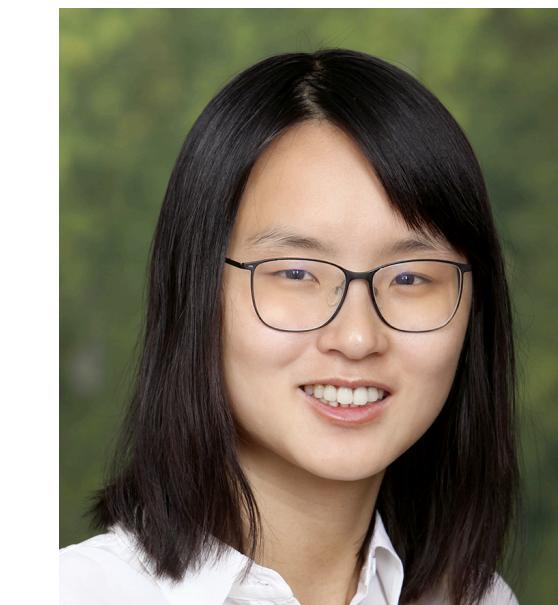




SimCSE: Simple Contrastive Learning of Sentence Embeddings



Tianyu Gao*
Princeton University

Xingcheng Yao*
Tsinghua University

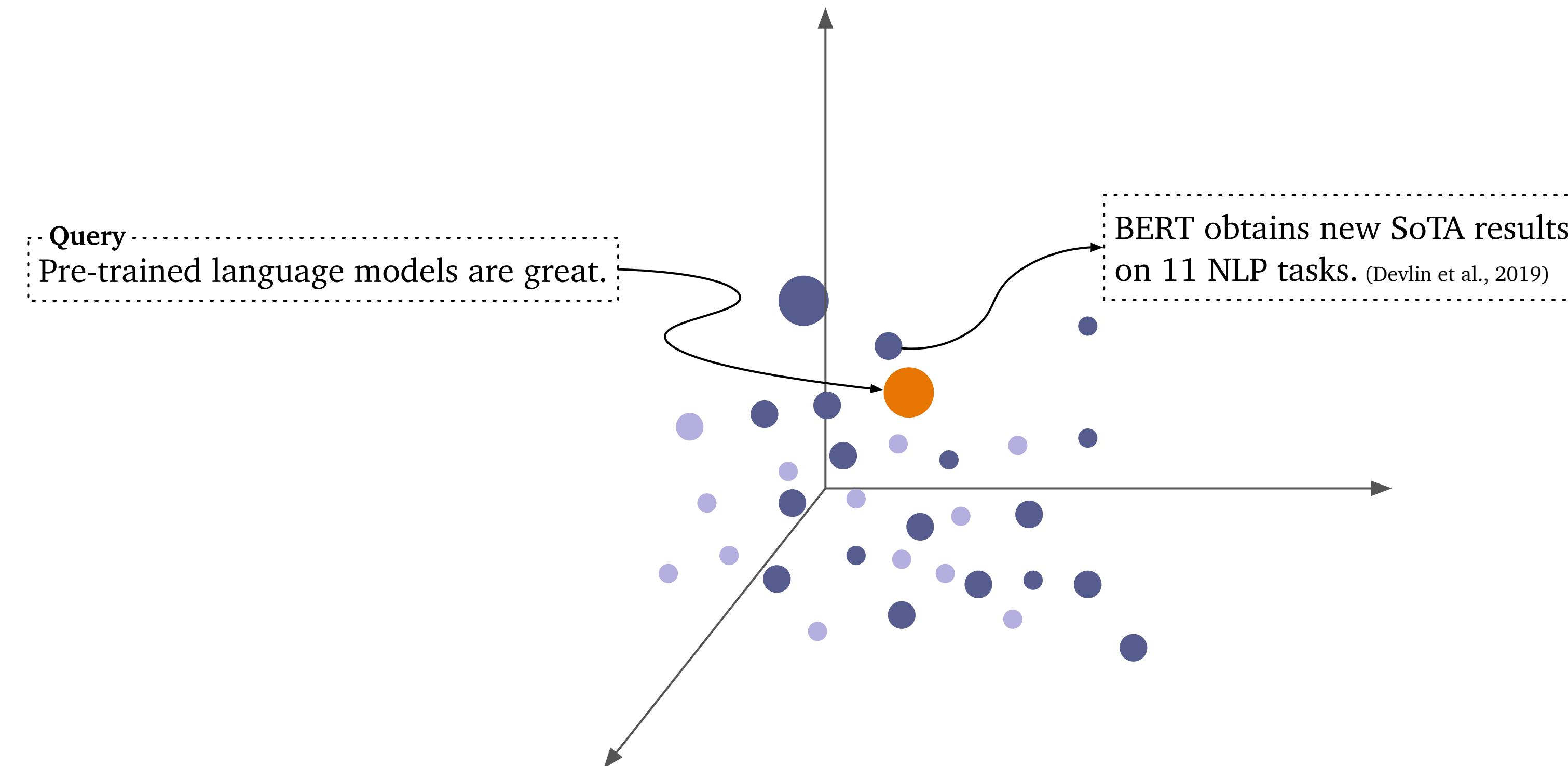
Danqi Chen
Princeton University

* equal contribution

Sentence Embeddings

Learning universal representations of **sentences** has wide applications in NLP

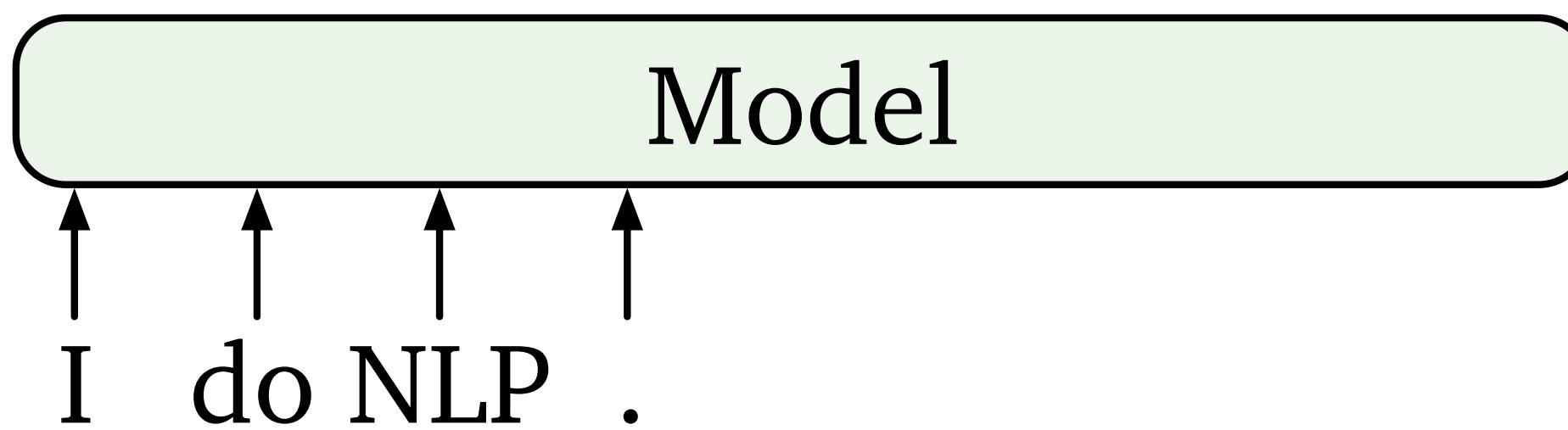
- Zero-shot retrieval
- Sentence clustering
- ...



Previous Approaches: Next Sentence Prediction

Use **current sentence** to predict **next sentence**

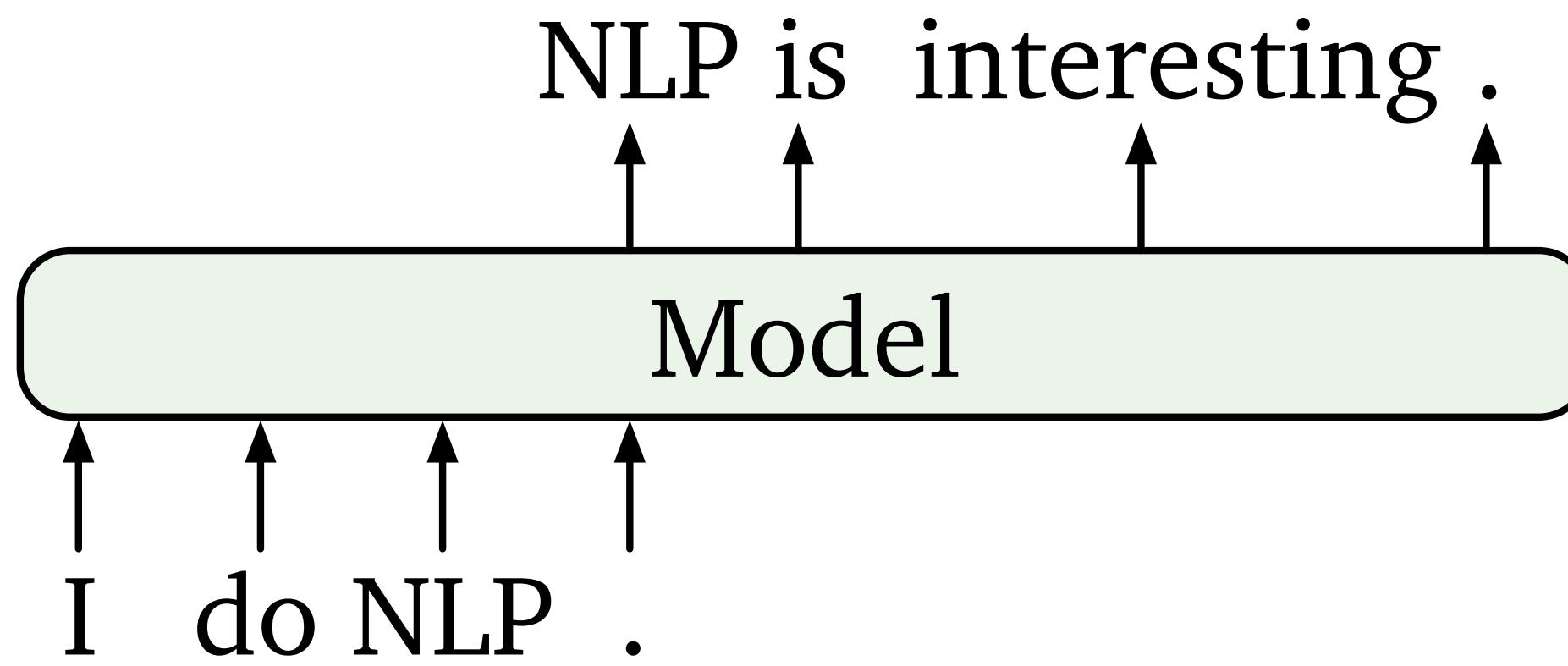
(Kiros et al., 2015; Logeswaran et al., 2018)



Previous Approaches: Next Sentence Prediction

Use **current sentence** to predict **next sentence**

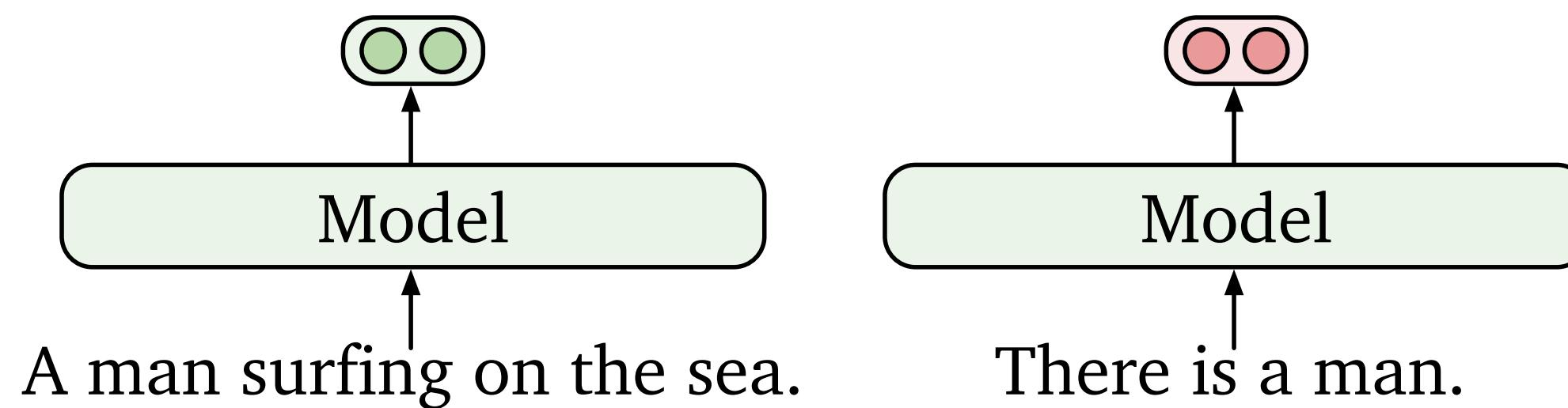
(Kiros et al., 2015; Logeswaran et al., 2018)



Previous Approaches: NLI Supervision

Use **natural language inference** (NLI) datasets as extra supervision

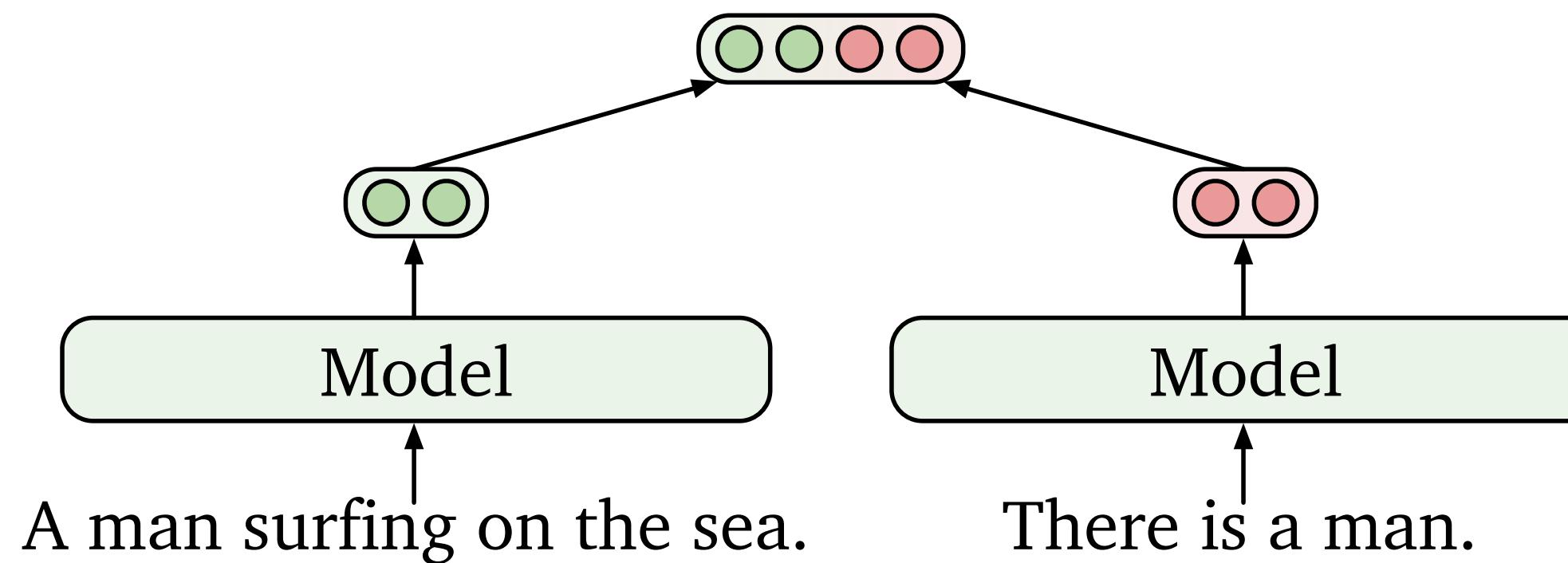
(Conneau et al., 2017; Reimers and Gurevych, 2019)



Previous Approaches: NLI Supervision

Use **natural language inference** (NLI) datasets as extra supervision

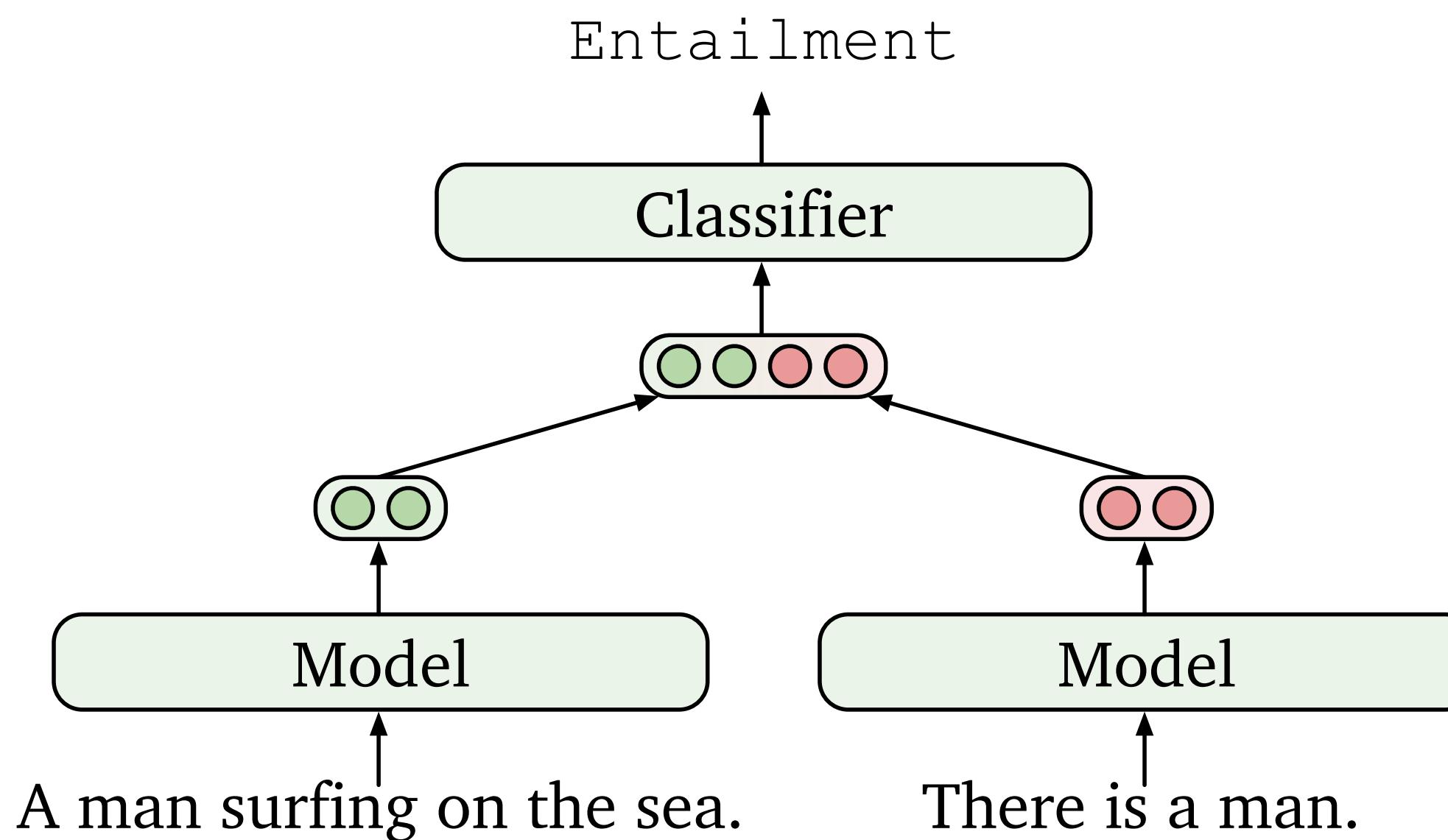
(Conneau et al., 2017; Reimers and Gurevych, 2019)



Previous Approaches: NLI Supervision

Use **natural language inference** (NLI) datasets as extra supervision

(Conneau et al., 2017; Reimers and Gurevych, 2019)



Previous Approaches: Data Augmentation

Maximize agreement between different views of the same sentence (**data augmentation**)
(Wu et al., 2020; Meng et al., 2021; Giorgi et al., 2021)

BERT from Sesame Street is so cute.

Previous Approaches: Data Augmentation

Maximize agreement between different views of the same sentence (**data augmentation**)
(Wu et al., 2020; Meng et al., 2021; Giorgi et al., 2021)

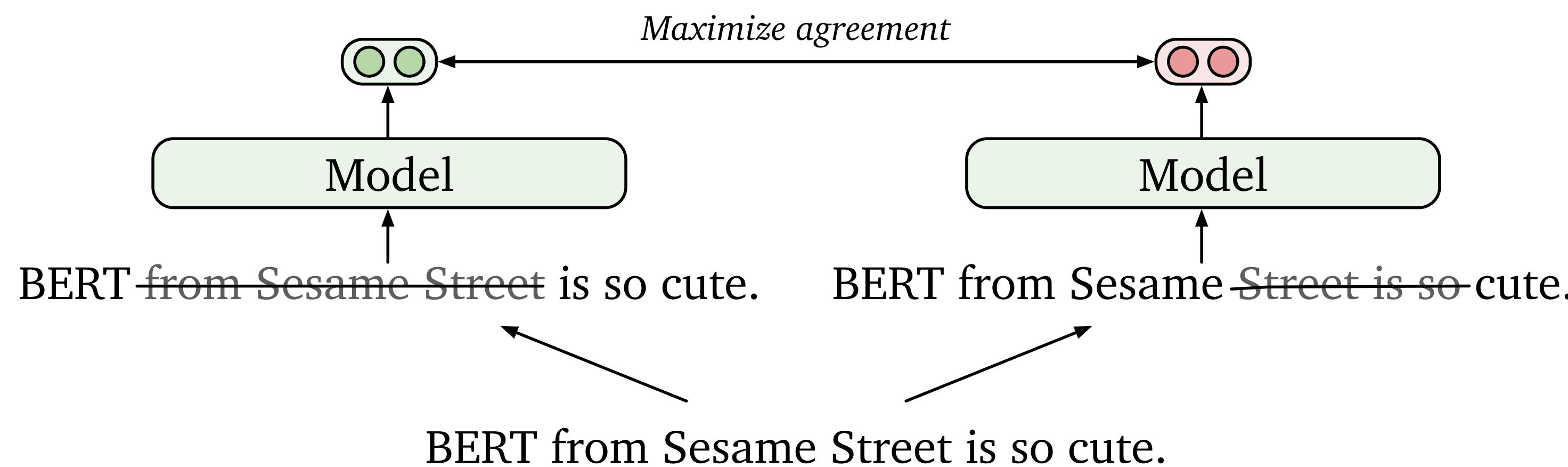
BERT from Sesame Street is so cute. BERT from Sesame Street is so cute.

The diagram illustrates two different views of the same sentence, "BERT from Sesame Street is so cute." The first view, on the left, shows a red arrow pointing to the word "Street". The second view, on the right, shows a blue arrow pointing to the word "cute". This visual representation highlights how data augmentation can create multiple variations of a single sentence to improve model performance.

Previous Approaches: Data Augmentation

Maximize agreement between different views of the same sentence (**data augmentation**)

(Wu et al., 2020; Meng et al., 2021; Giorgi et al., 2021)



SimCSE: Simple Contrastive Learning of Sentence Embeddings

SimCSE: Simple Contrastive Learning of Sentence Embeddings



Pre-trained embeddings + contrastive learning = SoTA sentence embeddings!

SimCSE: Simple Contrastive Learning of Sentence Embeddings



Pre-trained embeddings + contrastive learning = SoTA sentence embeddings!



- Unsupervised SimCSE: only uses standard **dropout** as data augmentation

SimCSE: Simple Contrastive Learning of Sentence Embeddings



Pre-trained embeddings + contrastive learning = SoTA sentence embeddings!



- Unsupervised SimCSE: only uses standard **dropout** as data augmentation
- Supervised SimCSE: uses **entailment** + **contradiction** pairs from NLI datasets

SimCSE: Simple Contrastive Learning of Sentence Embeddings



Pre-trained embeddings + contrastive learning = SoTA sentence embeddings!



- Unsupervised SimCSE: only uses standard **dropout** as data augmentation
- Supervised SimCSE: uses **entailment** + **contradiction** pairs from NLI datasets



Why does this work?

SimCSE: Simple Contrastive Learning of Sentence Embeddings



Pre-trained embeddings + contrastive learning = SoTA sentence embeddings!



- Unsupervised SimCSE: only uses standard **dropout** as data augmentation
- Supervised SimCSE: uses **entailment** + **contradiction** pairs from NLI datasets



Why does this work?

- The contrastive objective regularizes pre-trained embeddings' space to be more **uniform**

SimCSE: Simple Contrastive Learning of Sentence Embeddings



Pre-trained embeddings + contrastive learning = SoTA sentence embeddings!



- Unsupervised SimCSE: only uses standard **dropout** as data augmentation
- Supervised SimCSE: uses **entailment** + **contradiction** pairs from NLI datasets



Why does this work?

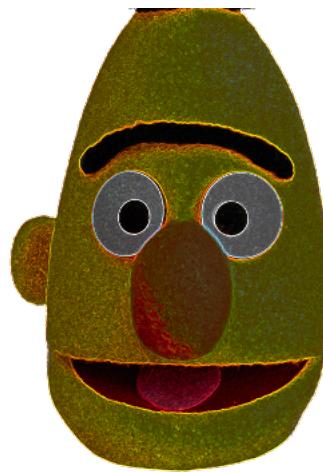
- The contrastive objective regularizes pre-trained embeddings' space to be more **uniform**
- It better **aligns** semantically close pairs with supervised signals

Contrastive Learning

Main idea: Pulling semantically close neighbors together and pushing apart non-neighbors (Hadsell et al., 2006)

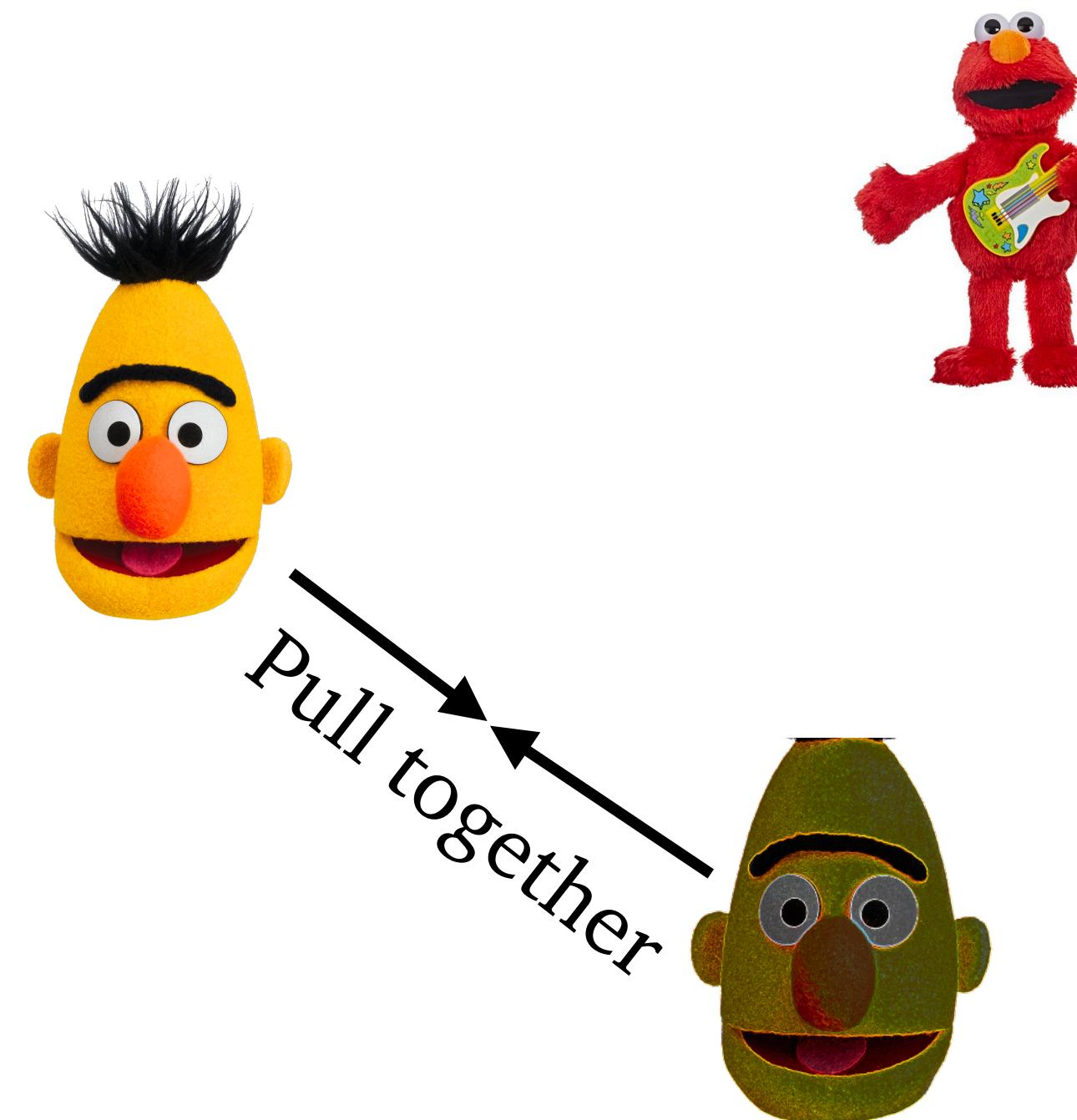
Contrastive Learning

Main idea: Pulling semantically close neighbors together and pushing apart non-neighbors (Hadsell et al., 2006)



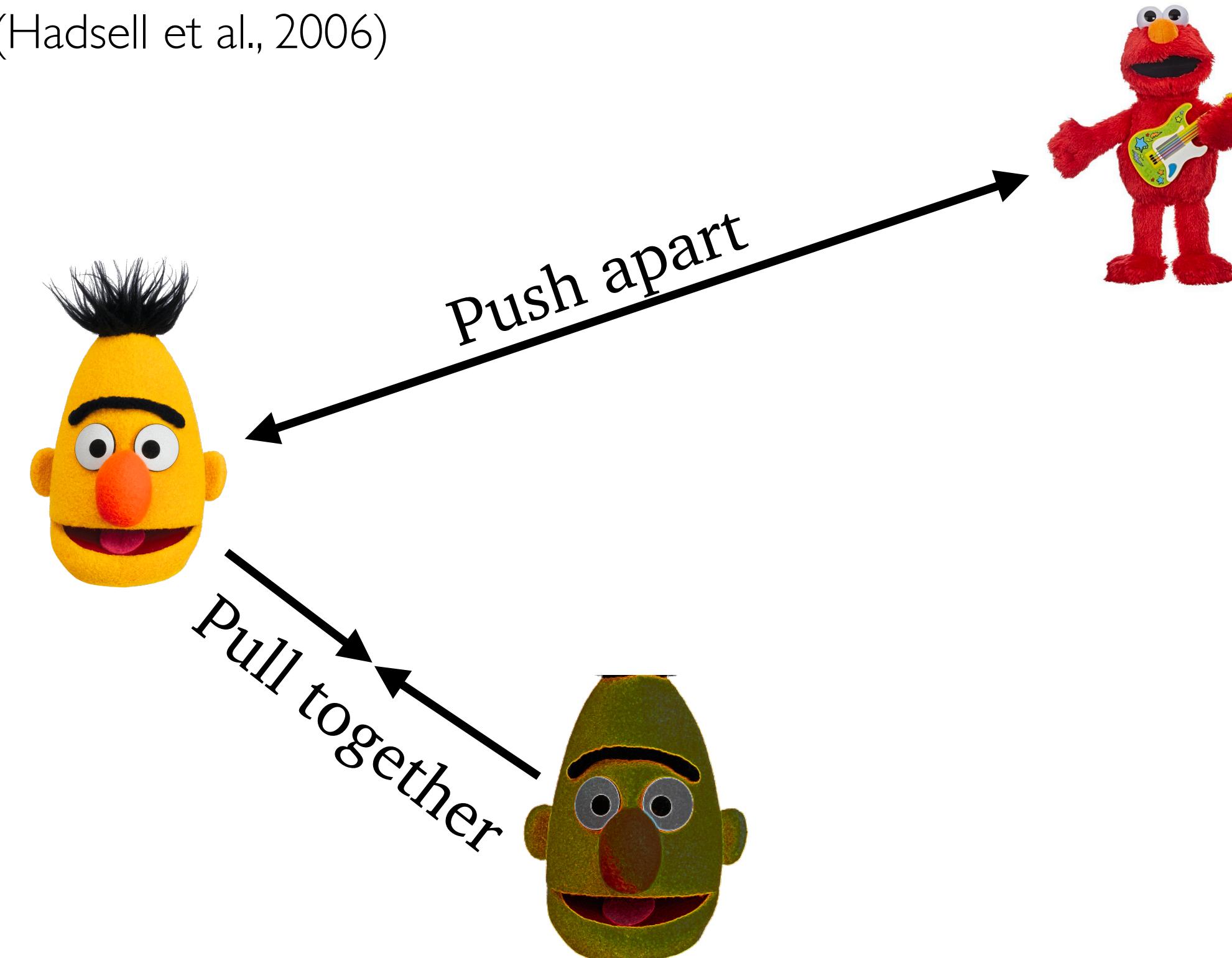
Contrastive Learning

Main idea: Pulling semantically close neighbors together and pushing apart non-neighbors (Hadsell et al., 2006)



Contrastive Learning

Main idea: Pulling semantically close neighbors together and pushing apart non-neighbors (Hadsell et al., 2006)



Contrastive Learning

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

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

Contrastive Learning

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

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

Similarity function Sentence embedding Temperature

Contrastive Learning

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

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

Similarity function Sentence embedding Temperature

$e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}$

\leftarrow Positive pairs

\leftarrow Negative pairs

Contrastive Learning

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

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

Similarity function Sentence embedding Temperature

$e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}$

\leftarrow Positive pairs

\leftarrow Negative pairs

What should be used as positives and negatives?

Contrastive Learning

InfoNCE loss (Oord et al., 2018; Chen et al., 2020)

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

Similarity function Sentence embedding Temperature

$e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}$

\leftarrow Positive pairs

\leftarrow Negative pairs

What should be used as positives and negatives?

Previous work: **data augmentation**

Unsupervised SimCSE

Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

Unsupervised SimCSE

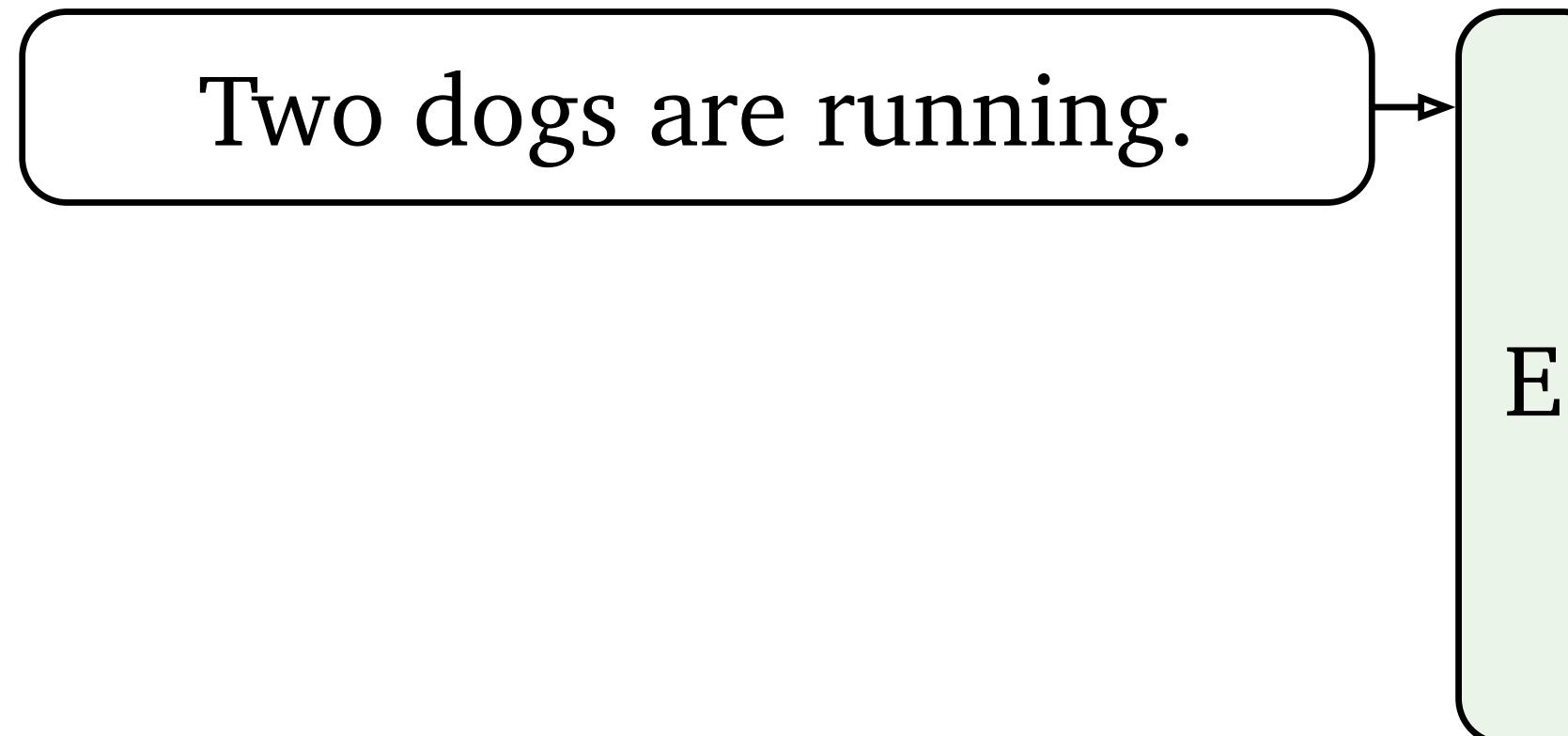
Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)

Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

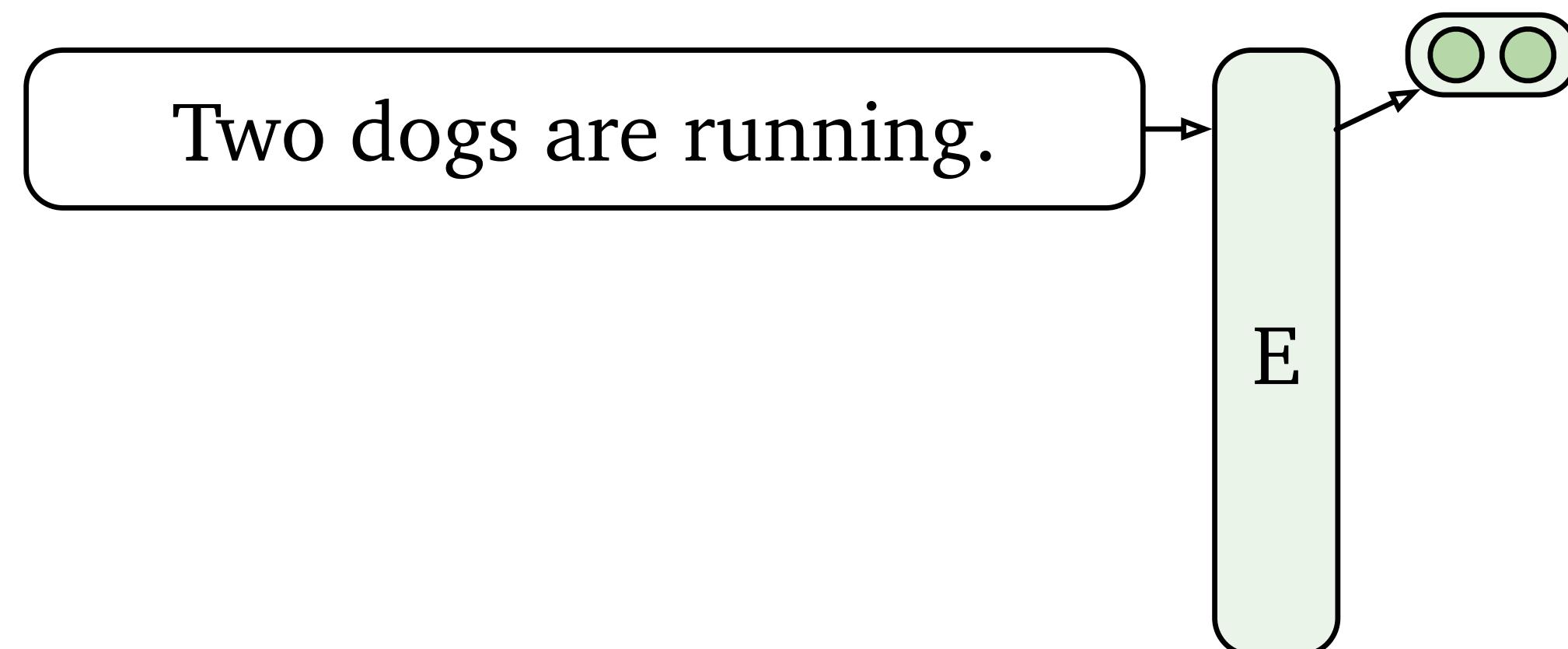
Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

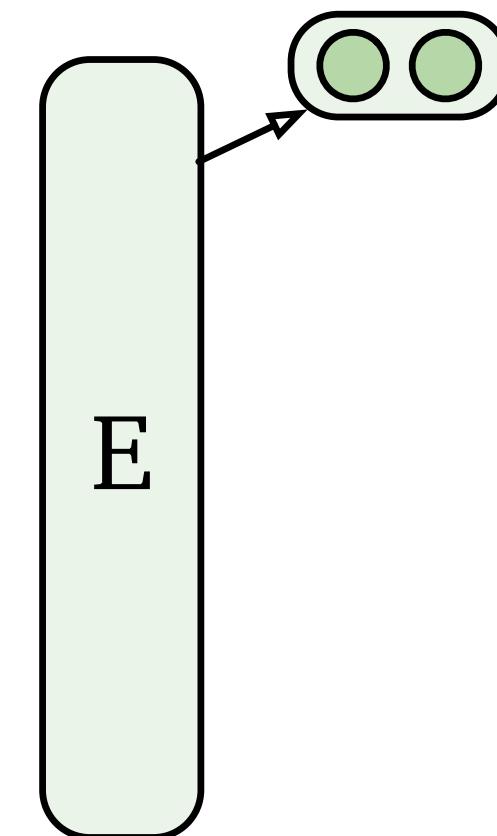
Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

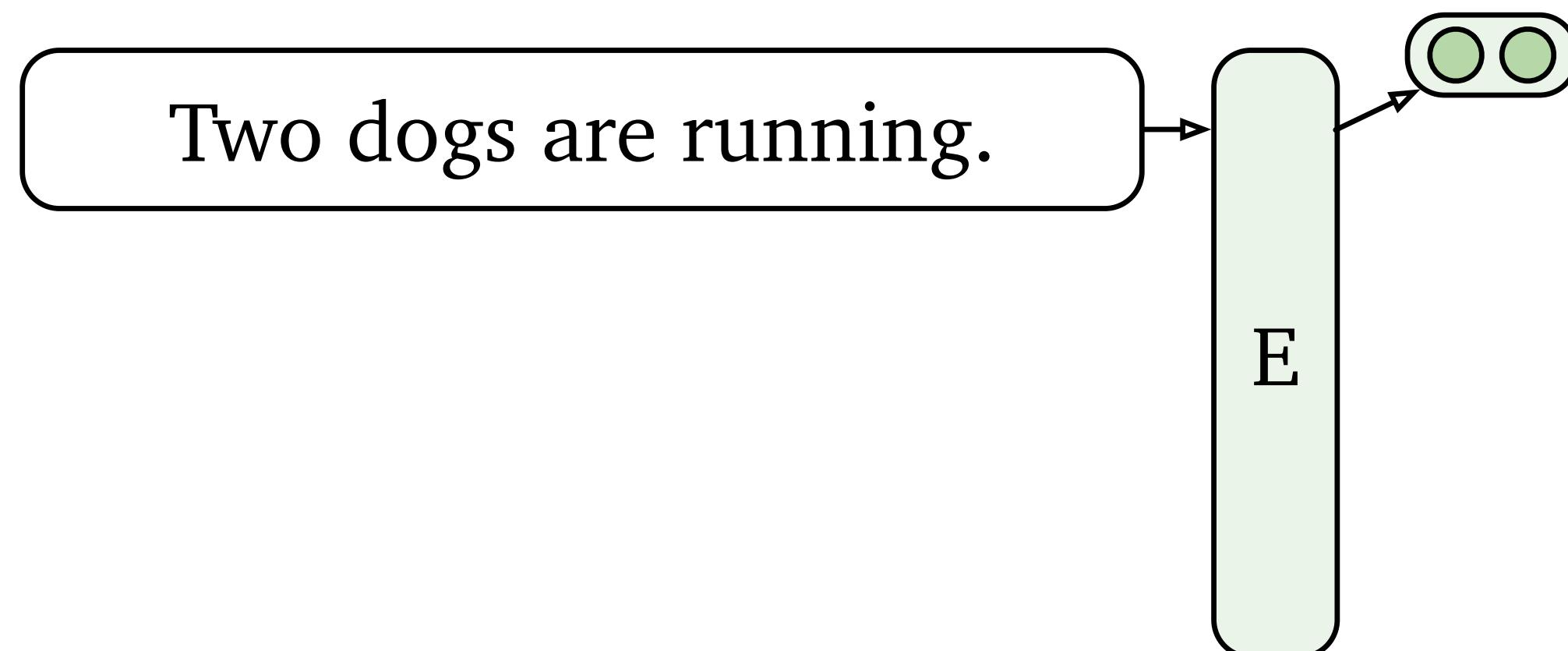
Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

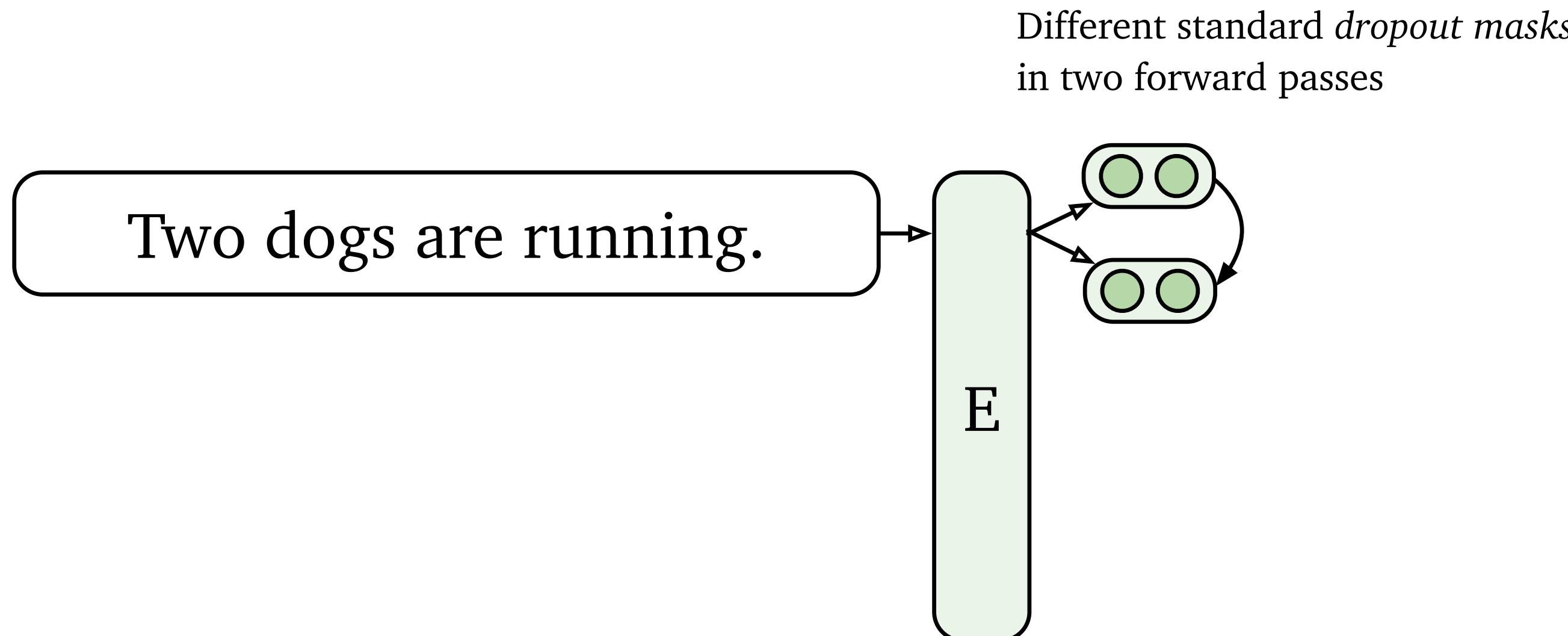
Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

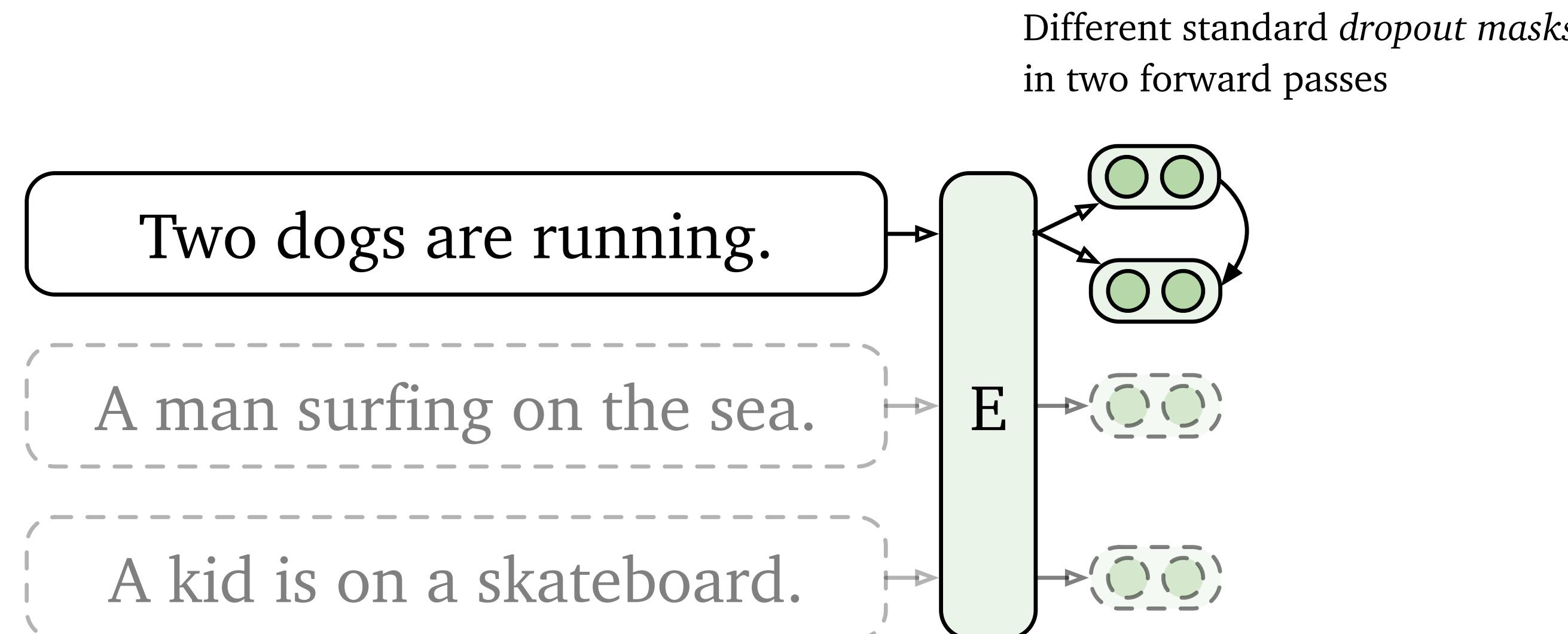
Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

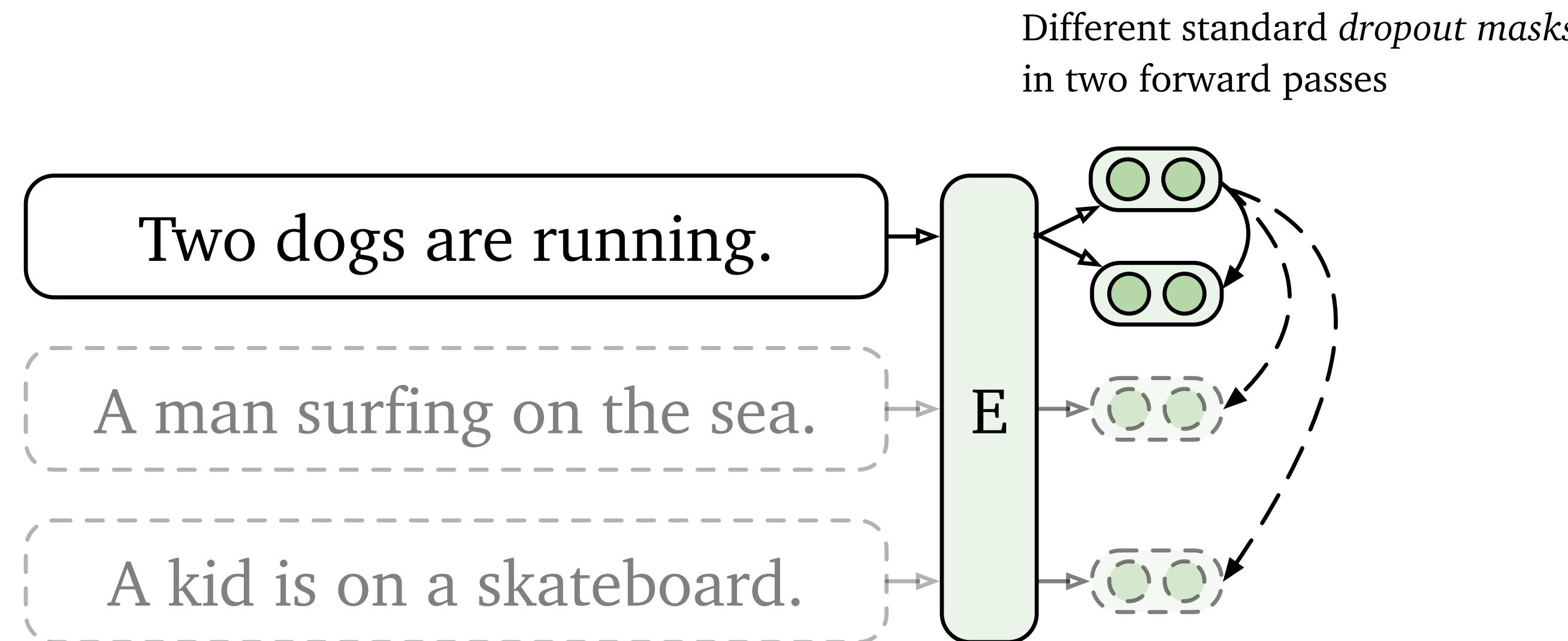
Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

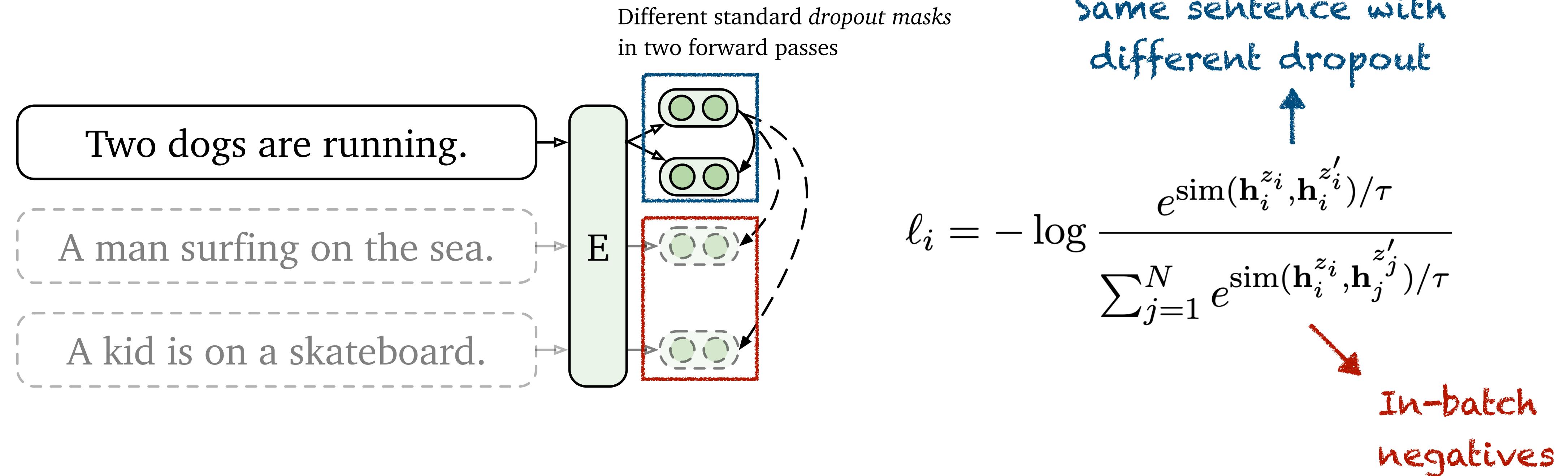
Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Unsupervised SimCSE

Positive pairs: embeddings of the **same sentence** with **different standard dropout masks**

Negative pairs: embeddings of other sentences from the same batch (in-batch negatives)



Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

Given one premise,

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

Given one premise,

- Premise: *There are two dogs running.*

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

Given one premise,

- Premise: *There are two dogs running.*

Annotators are required to write hypotheses of

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

Given one premise,

- Premise: *There are two dogs running.*

Annotators are required to write hypotheses of

- Entailment: *There are animals outdoors.*

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

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.*

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

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.*

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

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.*

Positive pairs

Supervised SimCSE

NLI datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018)

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.*

Positive pairs

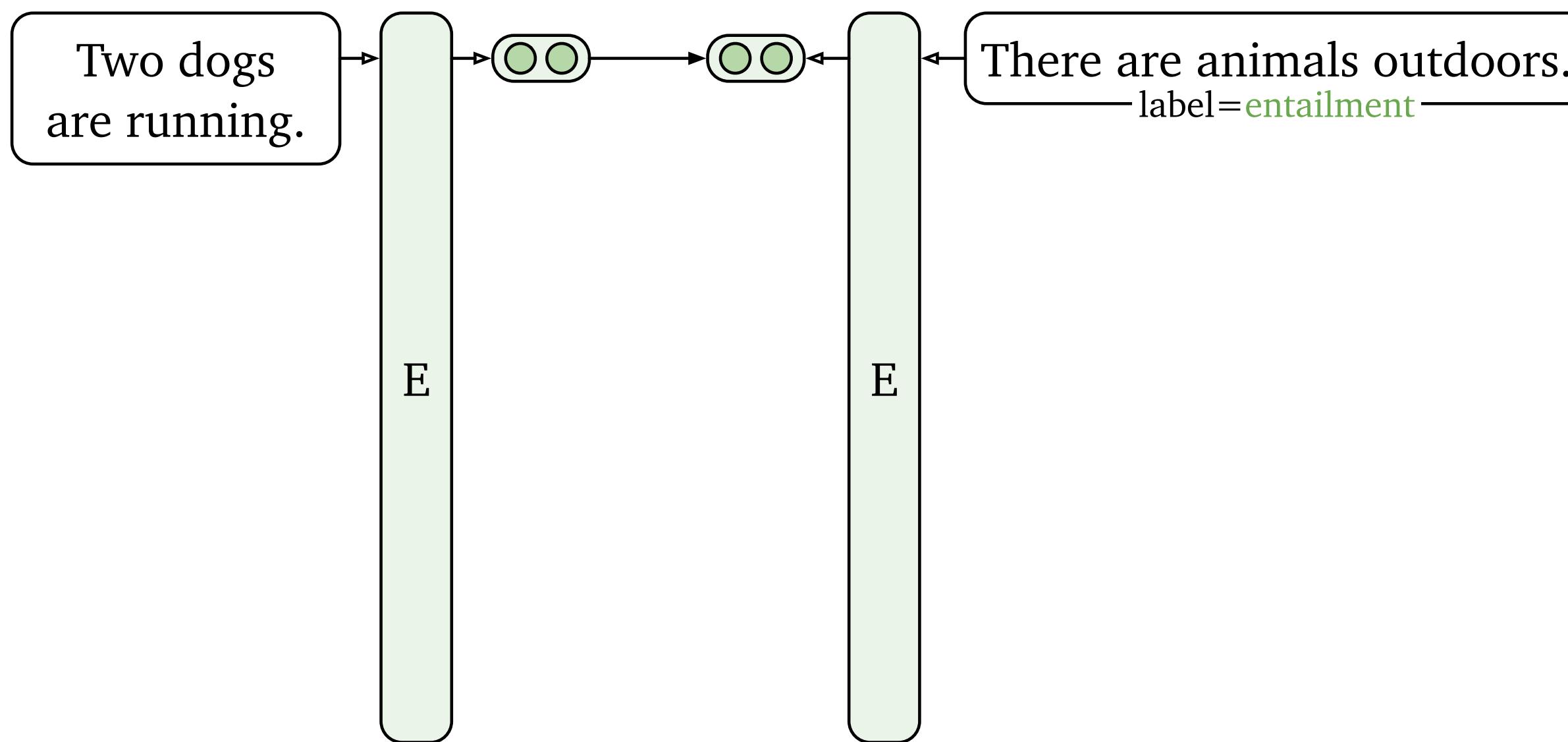
Hard negatives

Supervised SimCSE

- **Positive** pairs = **entailment** (premise, hypothesis) pairs
- **Negative** pairs = **contradiction** (premise, hypothesis) pairs + in-batch negatives

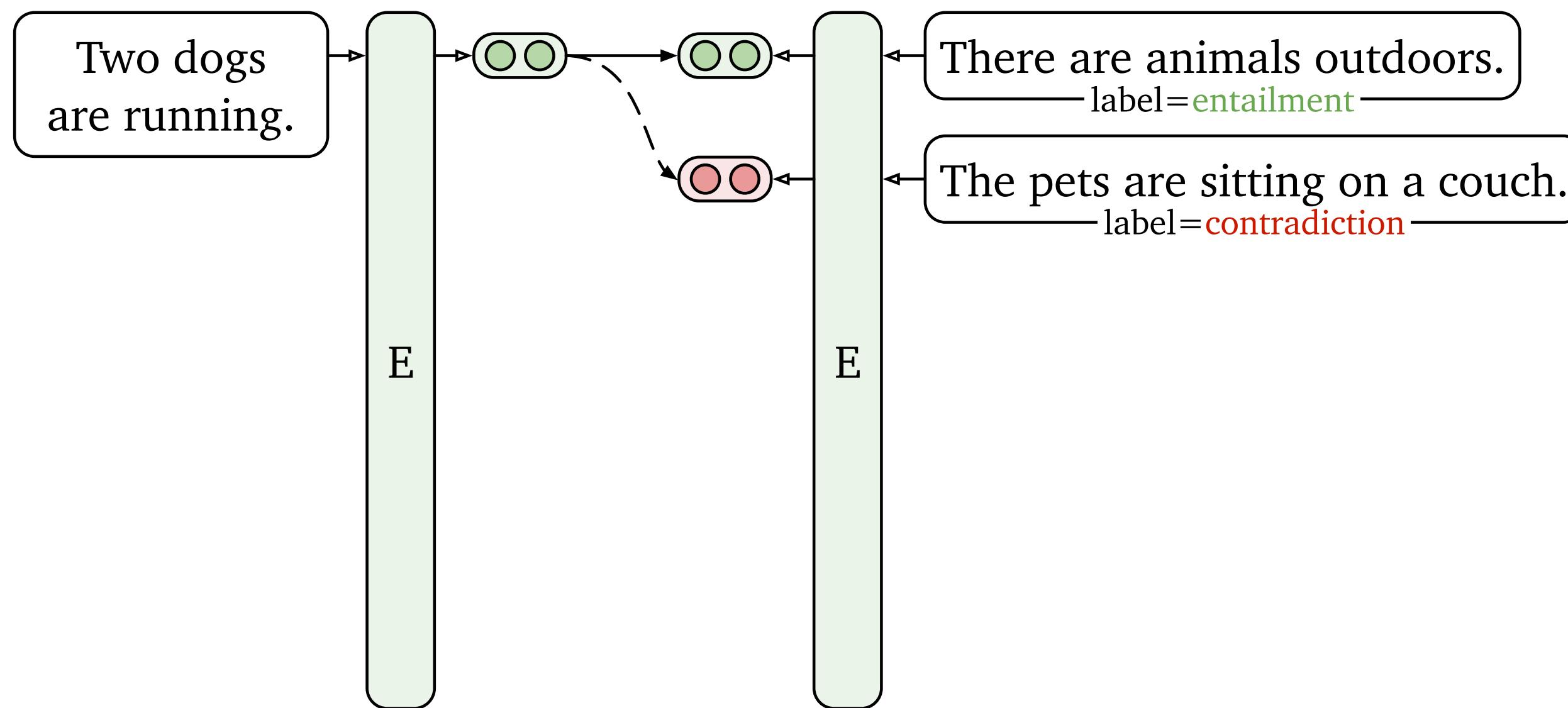
Supervised SimCSE

- **Positive** pairs = **entailment** (premise, hypothesis) pairs
- **Negative** pairs = **contradiction** (premise, hypothesis) pairs + in-batch negatives



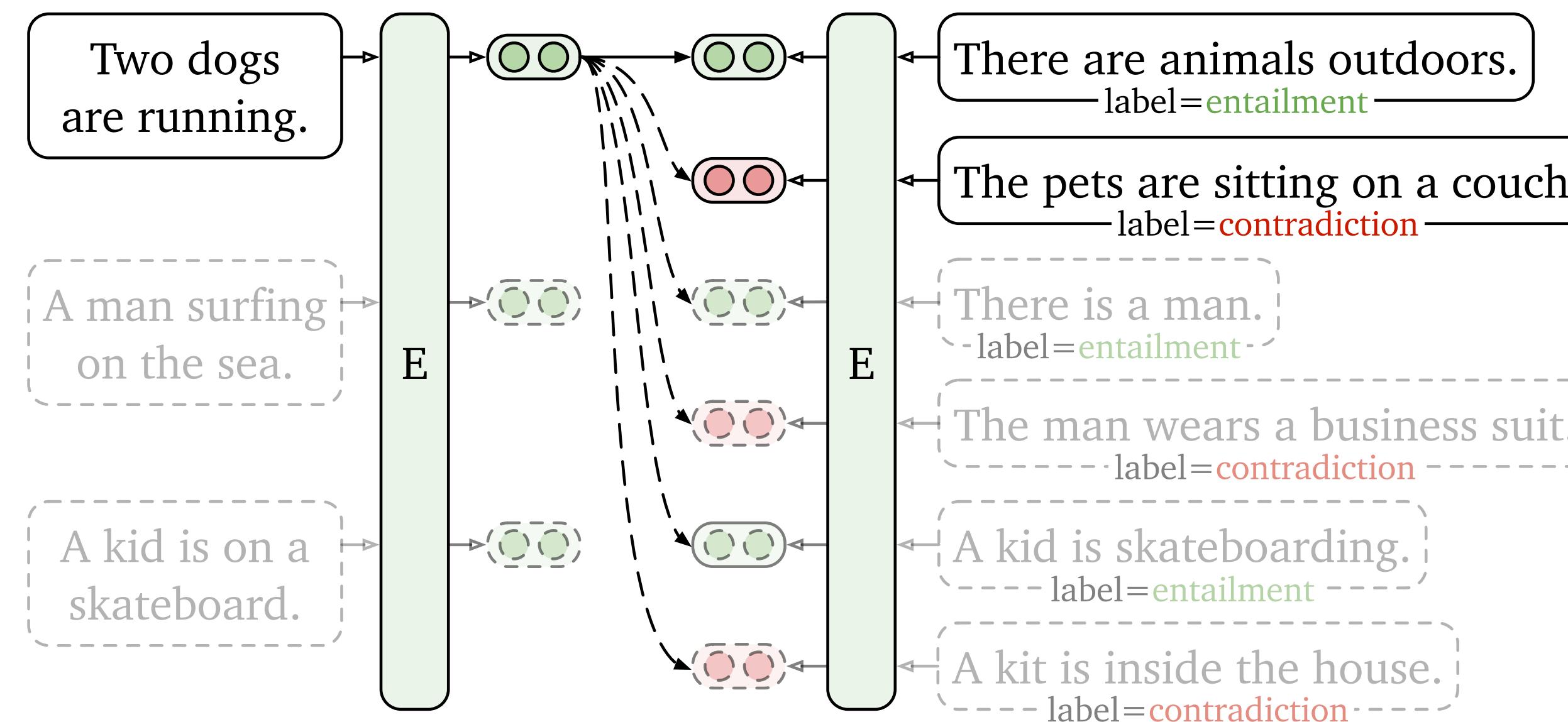
Supervised SimCSE

- **Positive** pairs = **entailment** (premise, hypothesis) pairs
- **Negative** pairs = **contradiction** (premise, hypothesis) pairs + in-batch negatives



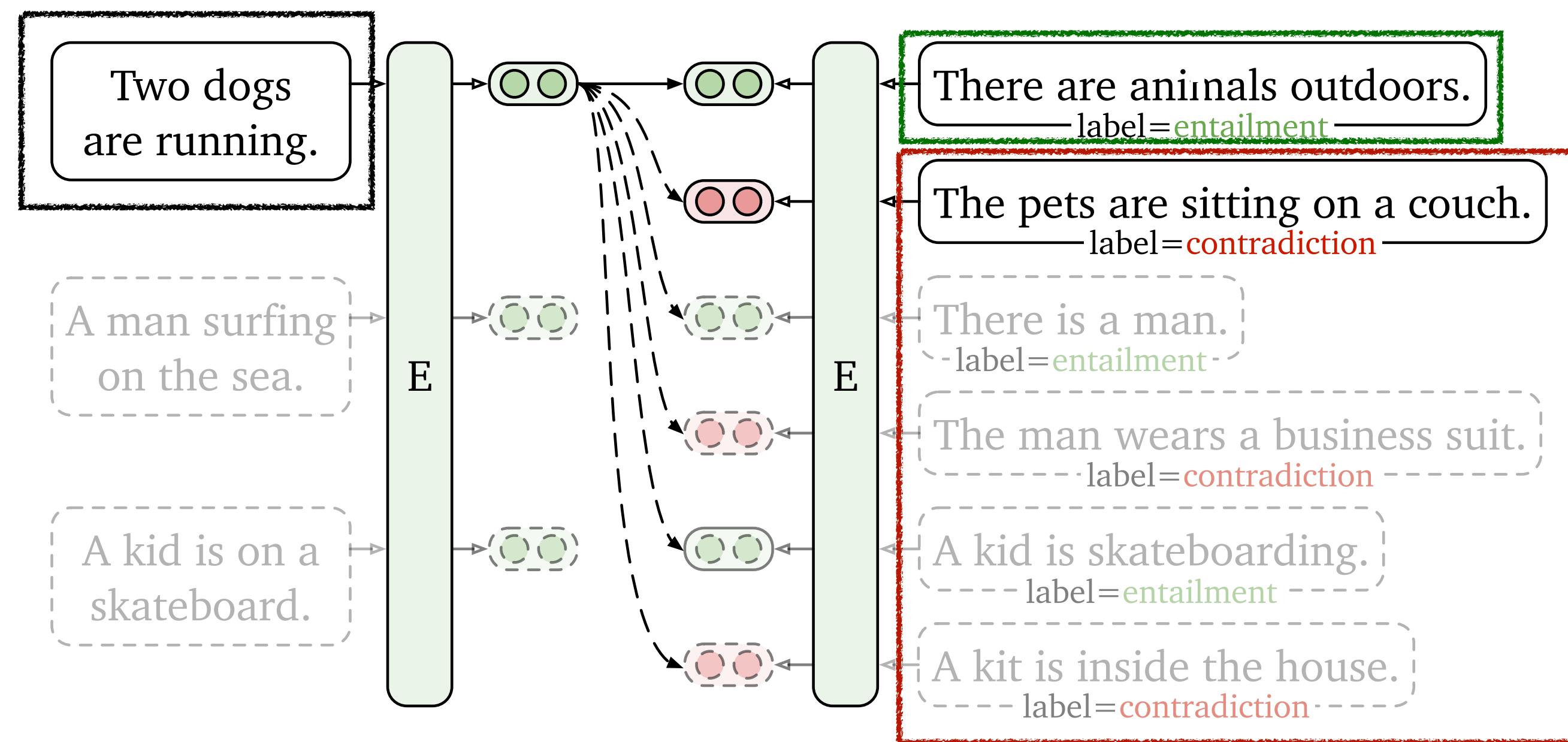
Supervised SimCSE

- Positive pairs = **entailment** (premise, hypothesis) pairs
- Negative pairs = **contradiction** (premise, hypothesis) pairs + in-batch negatives



Supervised SimCSE

- Positive pairs = **entailment** (premise, hypothesis) pairs
- Negative pairs = **contradiction** (premise, hypothesis) pairs + in-batch negatives



Premise **Entailment hypothesis**

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

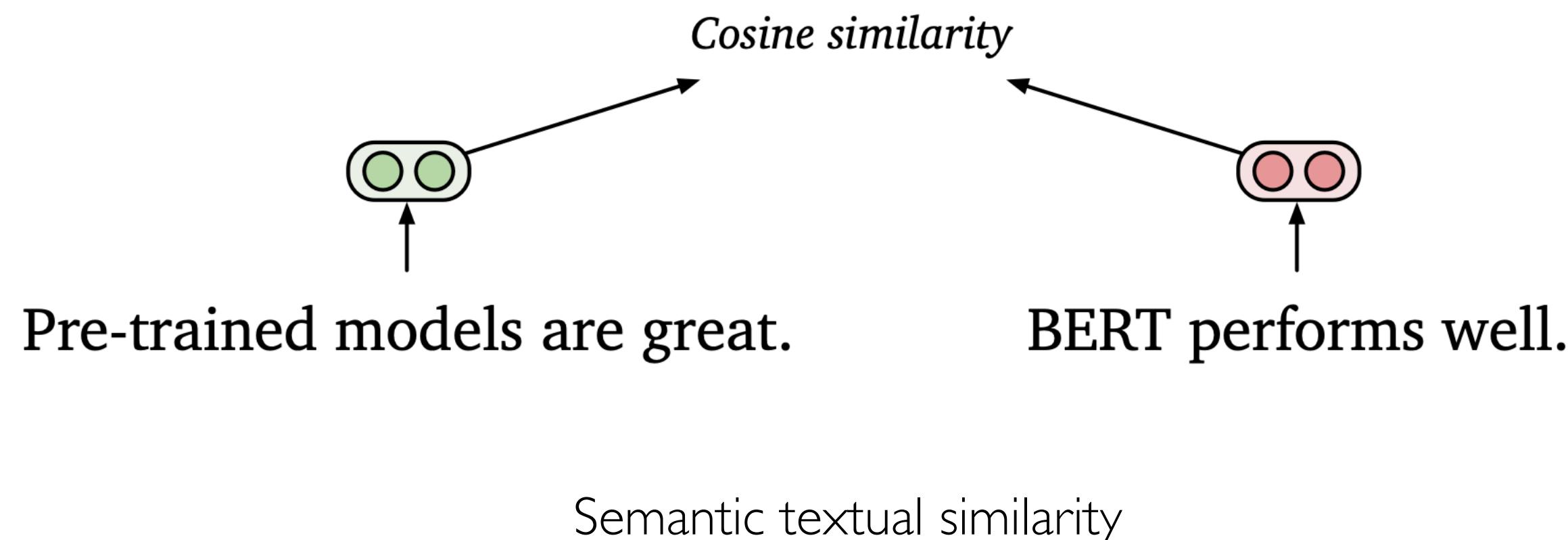
Contradiction hypothesis + in-batch negatives

Evaluation

- All sentence embeddings are **fixed**
- **Semantic textual similarity** (STS) tasks
 - Given sentence pairs, regress the similarity scores (Pearson/Spearman's correlation)
- **Transfer** tasks
 - Train linear classifiers on text classification tasks (accuracy)

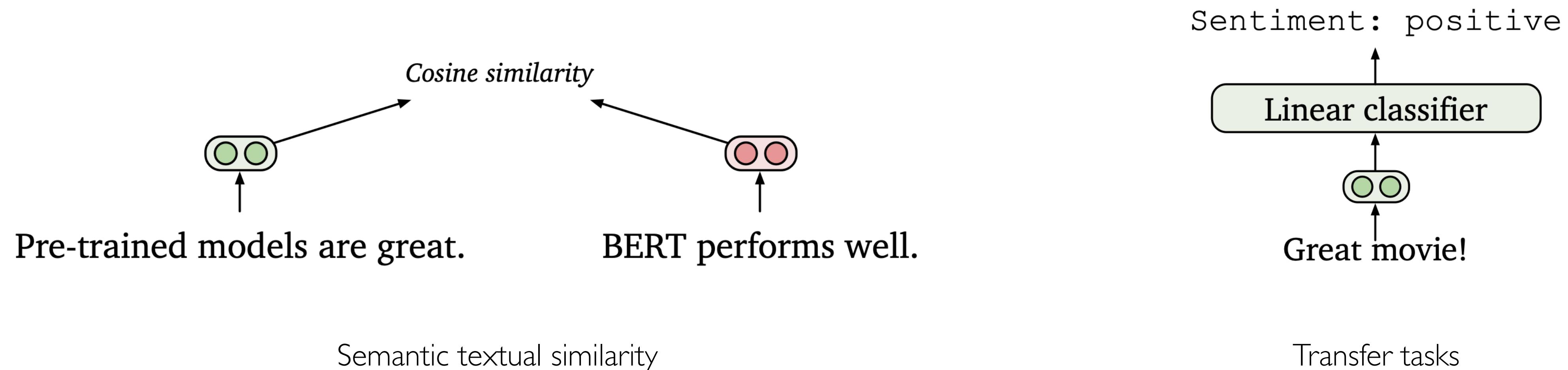
Evaluation

- All sentence embeddings are **fixed**
- **Semantic textual similarity** (STS) tasks
 - Given sentence pairs, regress the similarity scores (Pearson/Spearman's correlation)
- **Transfer** tasks
 - Train linear classifiers on text classification tasks (accuracy)



Evaluation

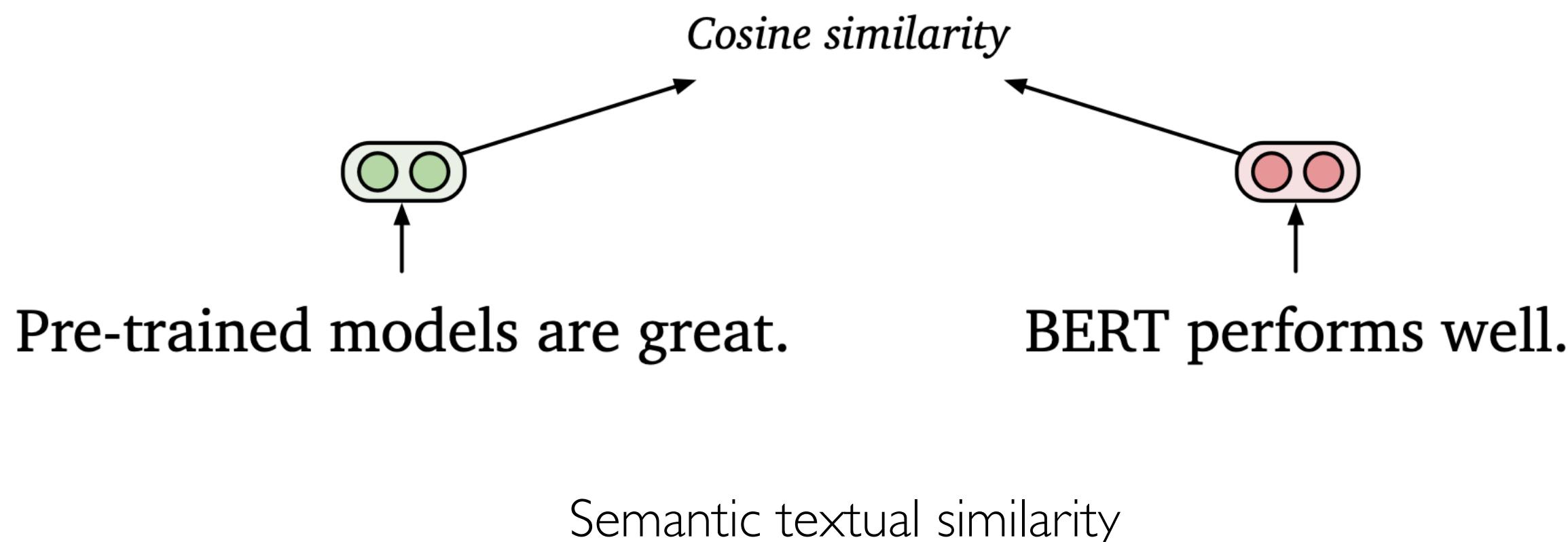
- All sentence embeddings are **fixed**
- **Semantic textual similarity** (STS) tasks
 - Given sentence pairs, regress the similarity scores (Pearson/Spearman's correlation)
- **Transfer** tasks
 - Train linear classifiers on text classification tasks (accuracy)



Evaluation

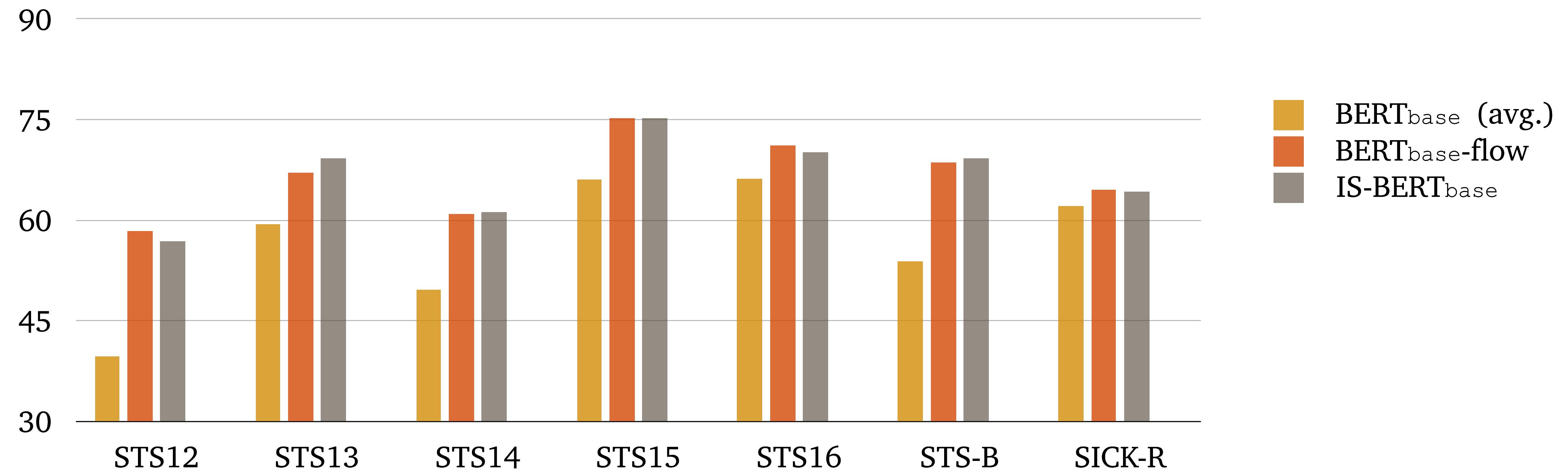
- All sentence embeddings are **fixed**
- **Semantic textual similarity (STS) tasks**
 - Given sentence pairs, regress the similarity scores (Pearson/Spearman's correlation)
- **Transfer tasks**
 - Train linear classifiers on text classification tasks (accuracy)

We mainly use this



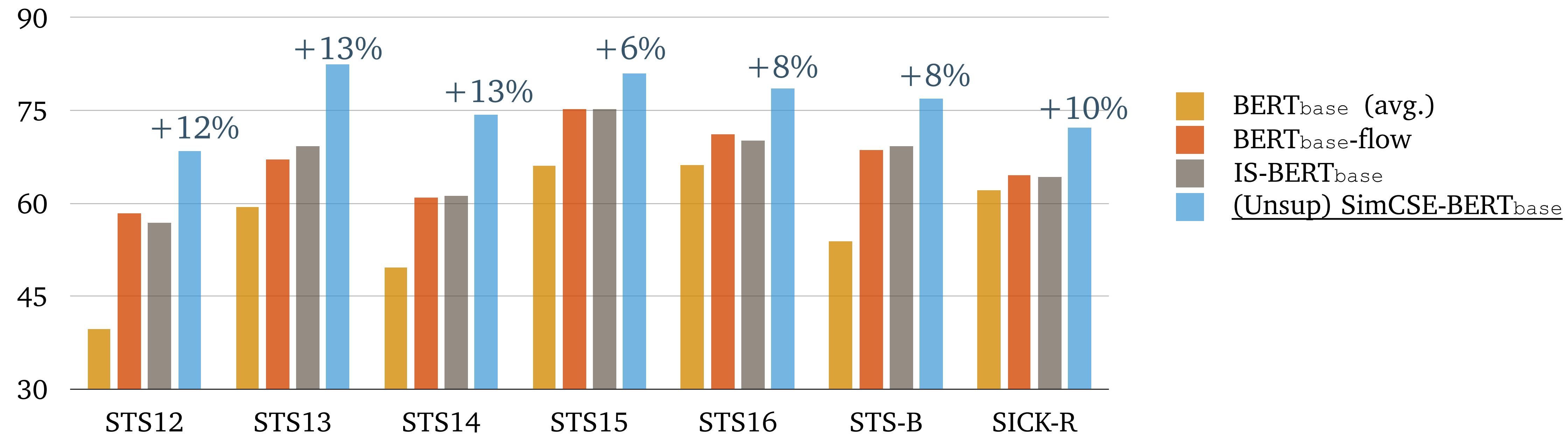
Unsupervised SimCSE: Main Results

Semantic textual similarity (STS) tasks: Spearman's correlation



Unsupervised SimCSE: Main Results

Semantic textual similarity (STS) tasks: Spearman's correlation

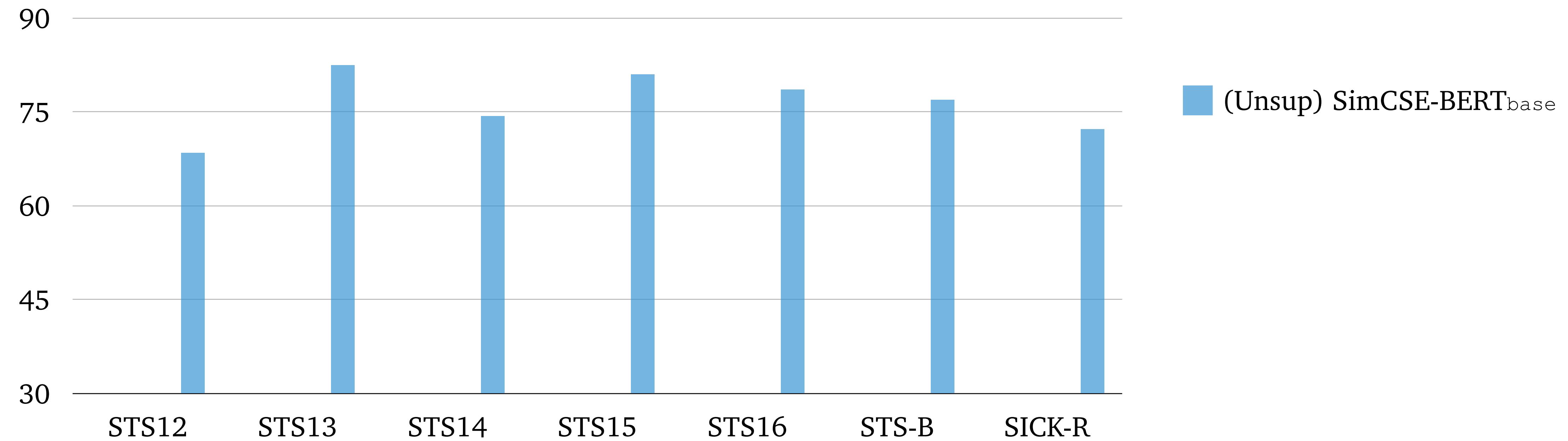


~10% higher than previous SOTA

~20% higher than avg. BERT embeddings

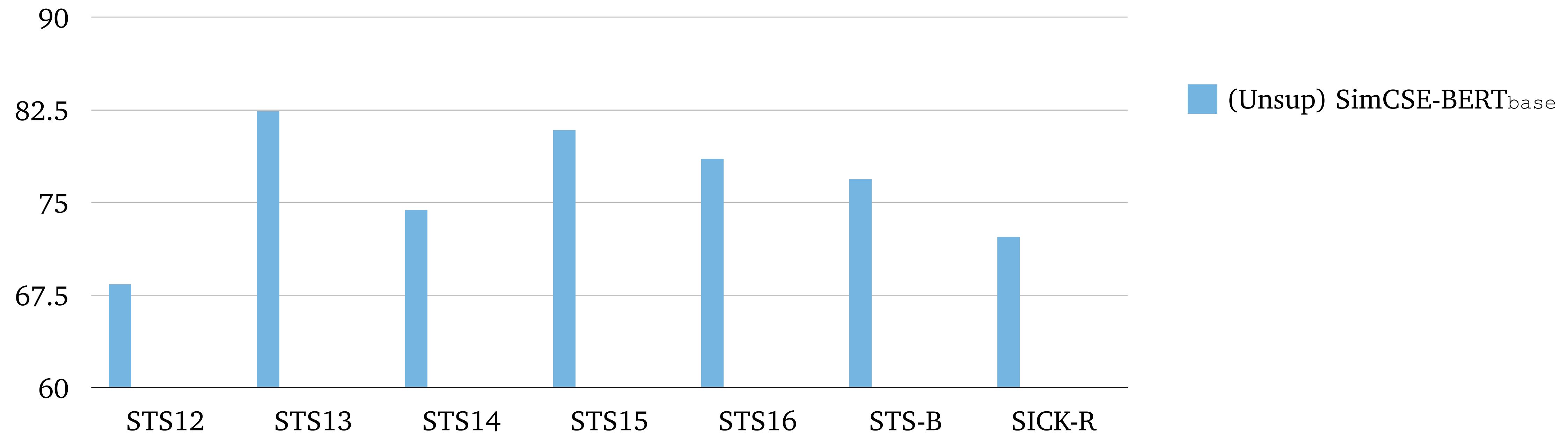
Supervised SimCSE: Main Results

Semantic textual similarity (STS) tasks: Spearman's correlation



Supervised SimCSE: Main Results

Semantic textual similarity (STS) tasks: Spearman's correlation



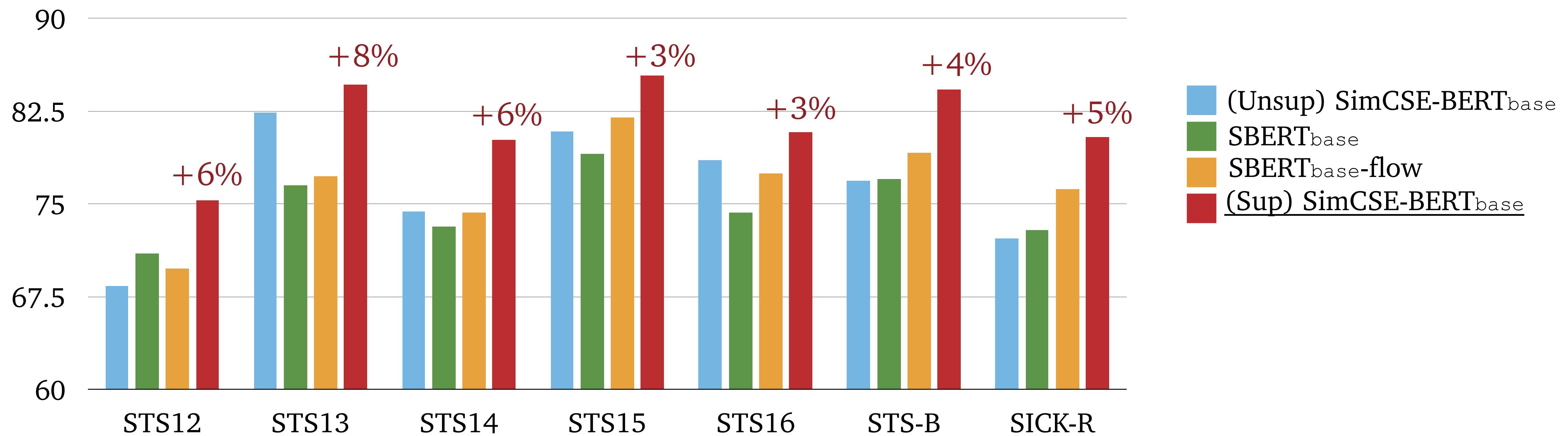
Supervised SimCSE: Main Results

Semantic textual similarity (STS) tasks: Spearman's correlation



Supervised SimCSE: Main Results

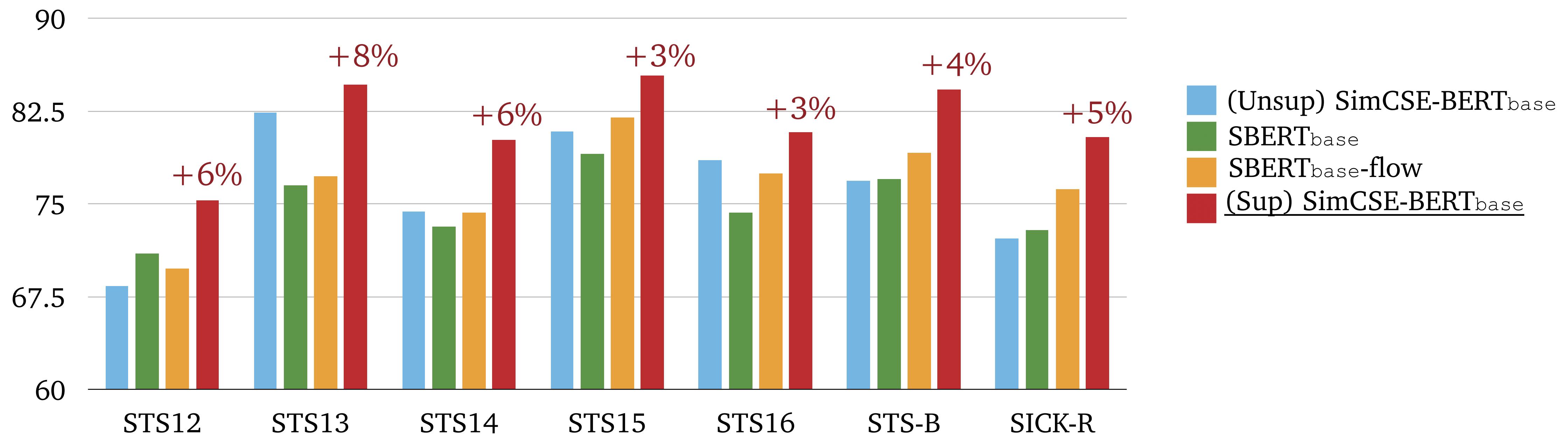
Semantic textual similarity (STS) tasks: Spearman's correlation



Even unsupervised SimCSE matches supervised SentenceBERT
6.7% higher than SentenceBERT using the same NLI datasets

Supervised SimCSE: Main Results

Semantic textual similarity (STS) tasks: Spearman's correlation



Even unsupervised SimCSE matches supervised SentenceBERT
6.7% higher than SentenceBERT using the same NLI datasets

See more results in the paper

Why Does This Work?

Different unsupervised positive pairs

Why Does This Work?

Different unsupervised positive pairs

- SimCSE (dropout)

Different dropout

NLP is interesting. —— NLP is interesting.

Why Does This Work?

Different unsupervised positive pairs

- SimCSE (dropout)

Compare it to

Different dropout

NLP is interesting. —— NLP is interesting.

Why Does This Work?

Different unsupervised positive pairs

- SimCSE (dropout)

Compare it to

- Next sentence

Different dropout

NLP is interesting. —— NLP is interesting.

I do NLP. —— NLP is interesting.

Why Does This Work?

Different unsupervised positive pairs

- SimCSE (dropout)

Compare it to

- Next sentence
- Synonym replacement

Different dropout

NLP is interesting. —— NLP is interesting.

I do NLP. —— NLP is interesting.

The movie is great. —— The movie is fantastic.

Why Does This Work?

Different unsupervised positive pairs

- SimCSE (dropout)

Compare it to

- Next sentence
- Synonym replacement
- Crop

Different dropout

NLP is interesting. —— NLP is interesting.

I do NLP. —— NLP is interesting.

The movie is great. —— The movie is fantastic.

~~Two dogs are running.~~ —— ~~Two dogs are running.~~

Why Does This Work?

Different unsupervised positive pairs

- SimCSE (dropout)

Compare it to

- Next sentence
- Synonym replacement
- Crop
- Delete one word

Different dropout

NLP is interesting. —— NLP is interesting.

I do NLP. —— NLP is interesting.

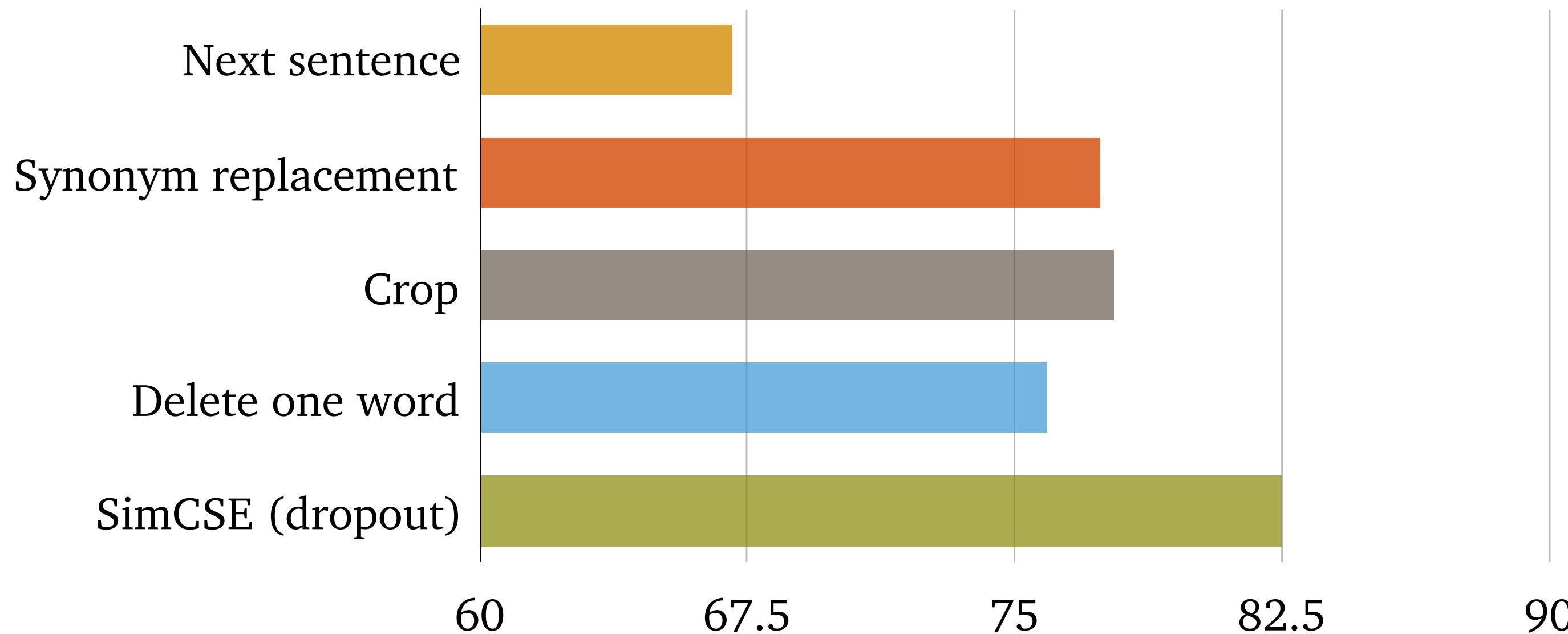
The movie is great. —— The movie is fantastic.

~~Two dogs are running.~~ —— ~~Two dogs are running.~~

Two dogs are running. —— Two dogs are running.

Why Does This Work?

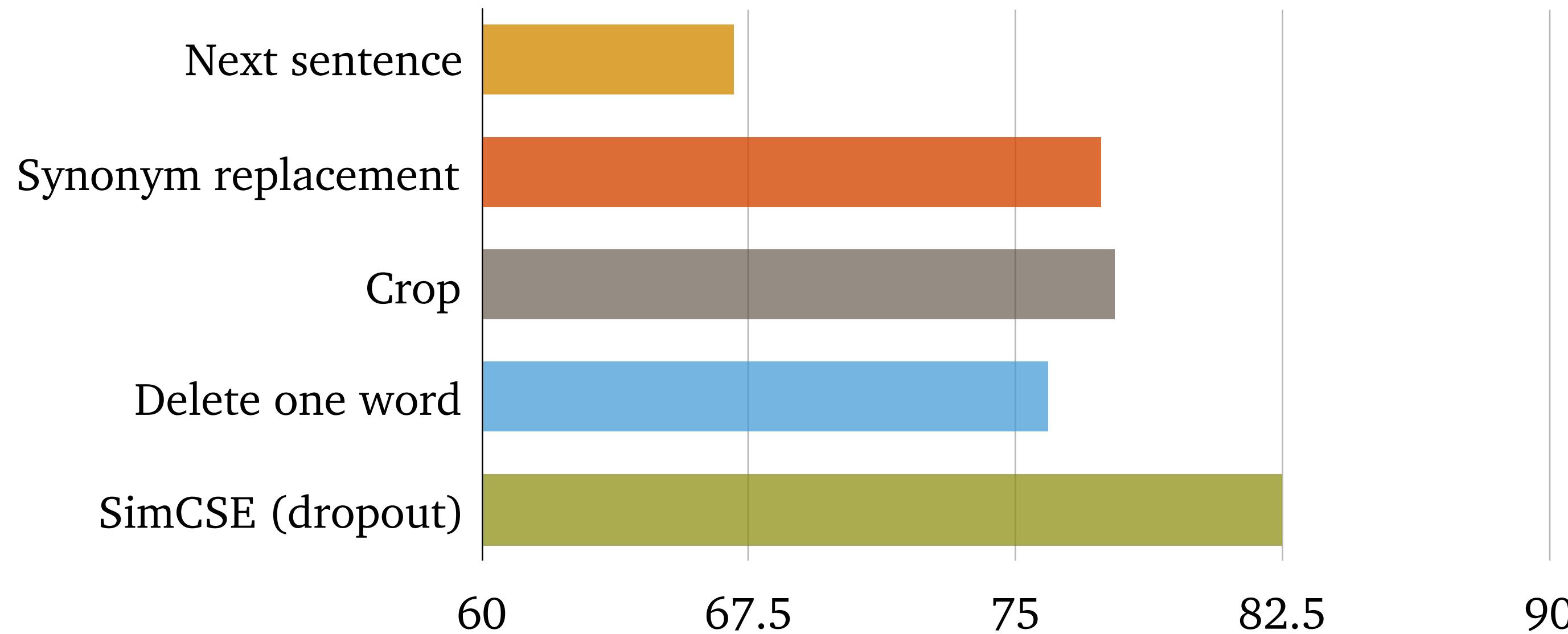
Unsupervised SimCSE



Default setting: 1 million sentences randomly sampled from English Wikipedia, N=64, evaluated on STS-B development set (Spearman's correlation)

Why Does This Work?

Unsupervised SimCSE



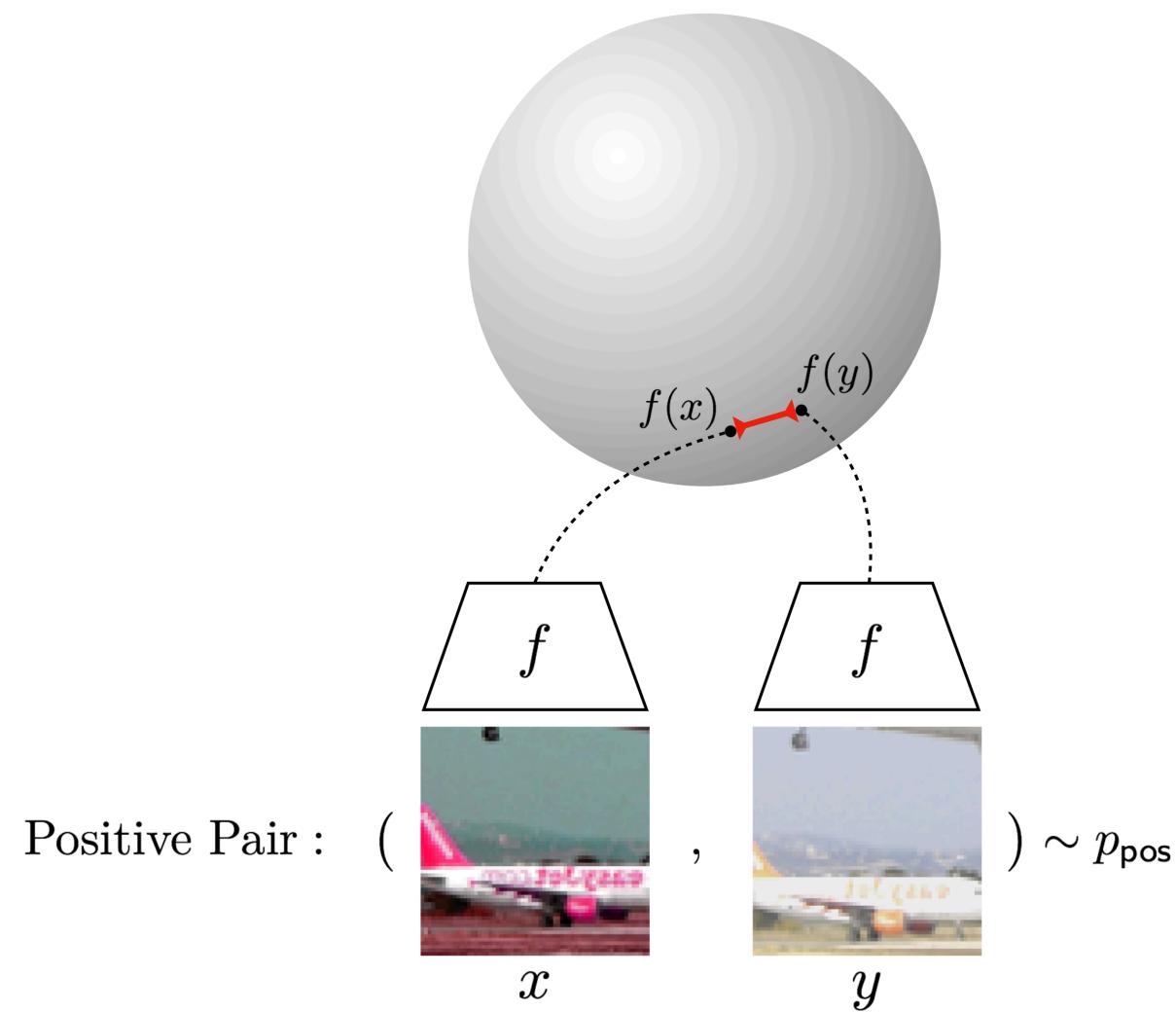
Default setting: 1 million sentences randomly sampled from English Wikipedia, N=64, evaluated on STS-B development set (Spearman's correlation)

- Predicting sentence itself ≫ predicting next sentence
- Using dropout as data augmentation ≫ discrete data augmentation (!!)

Alignment vs. Uniformity

Alignment vs. Uniformity

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

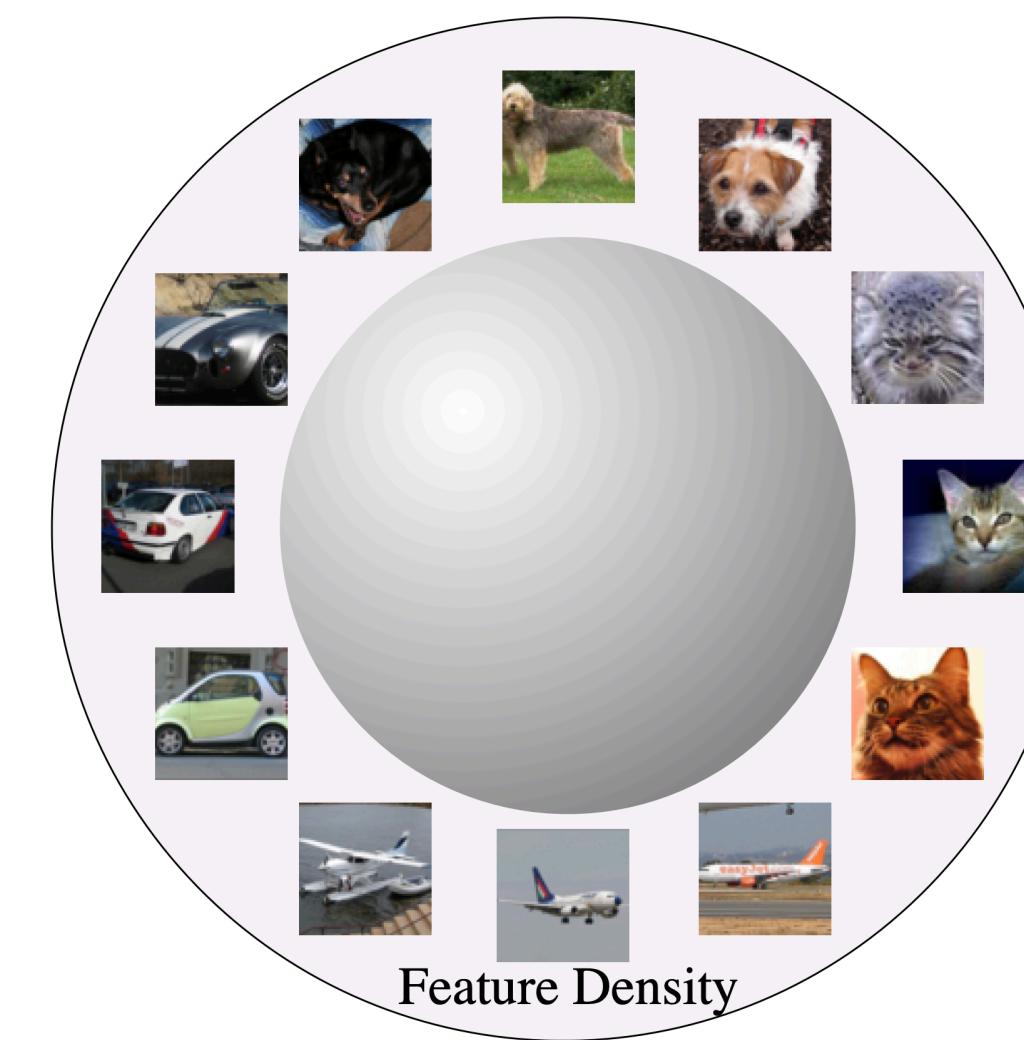
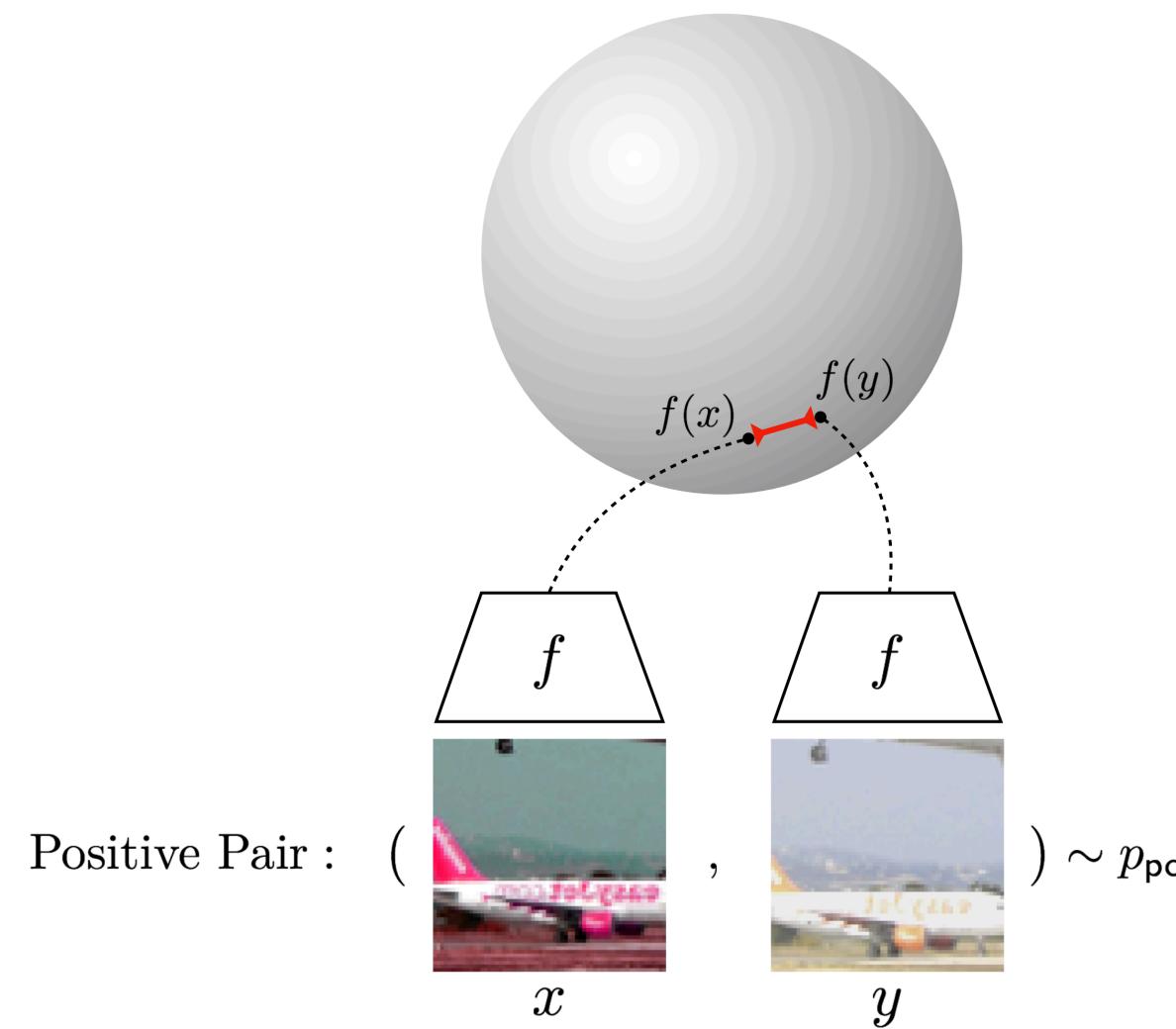


Alignment: how well positive pairs are aligned

Alignment vs. Uniformity

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

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



Alignment: how well positive pairs are aligned

Uniformity: how well the embeddings are uniformly distributed

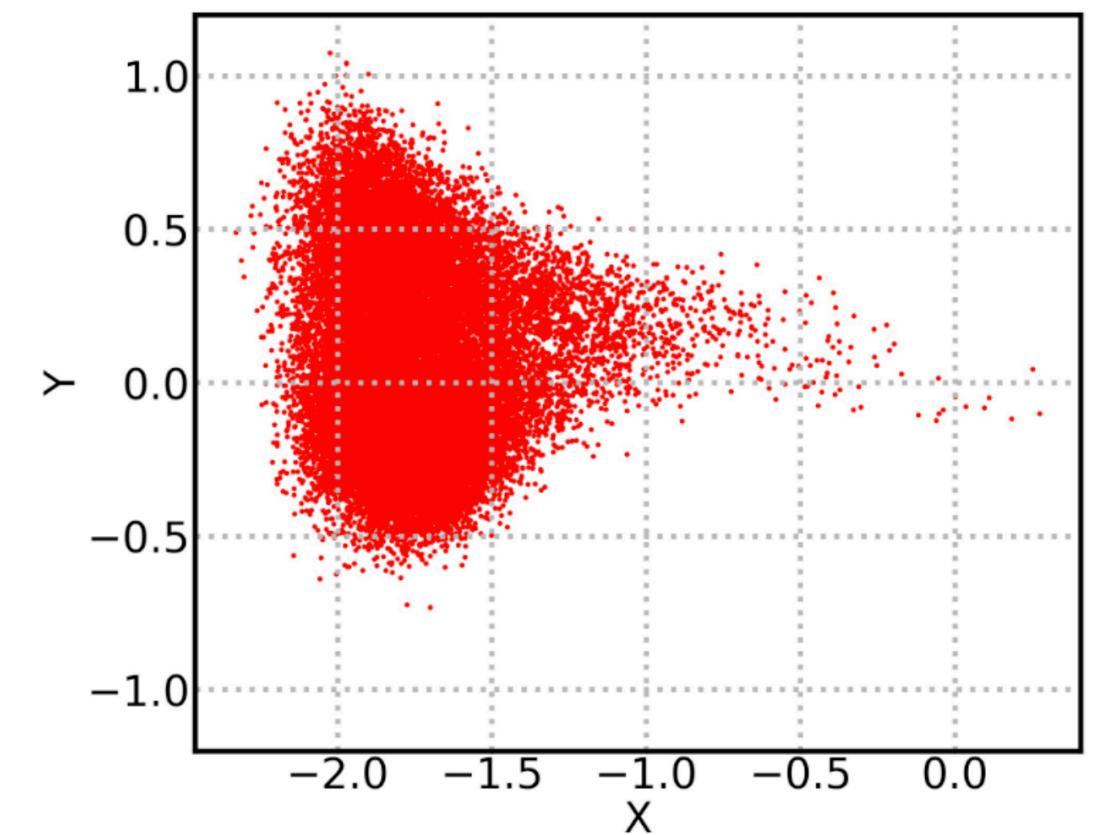
Alignment vs. Uniformity

Sentence embeddings from **pre-trained language models?**

Alignment vs. Uniformity

Sentence embeddings from **pre-trained language models**?

- Pre-trained embeddings **well encode** sentence semantics but they are highly **anisotropic** (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020)

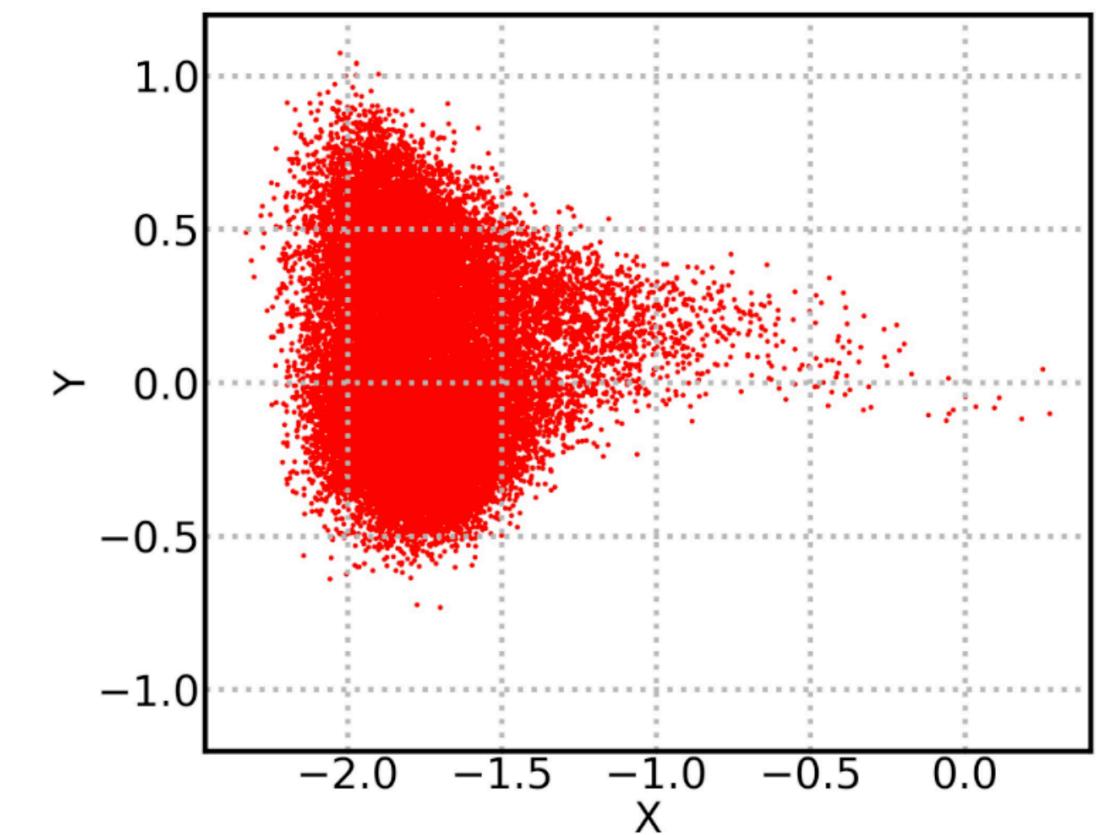


Alignment vs. Uniformity

Sentence embeddings from **pre-trained language models**?

- Pre-trained embeddings **well encode** sentence semantics but they are highly **anisotropic** (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020)

Good alignment



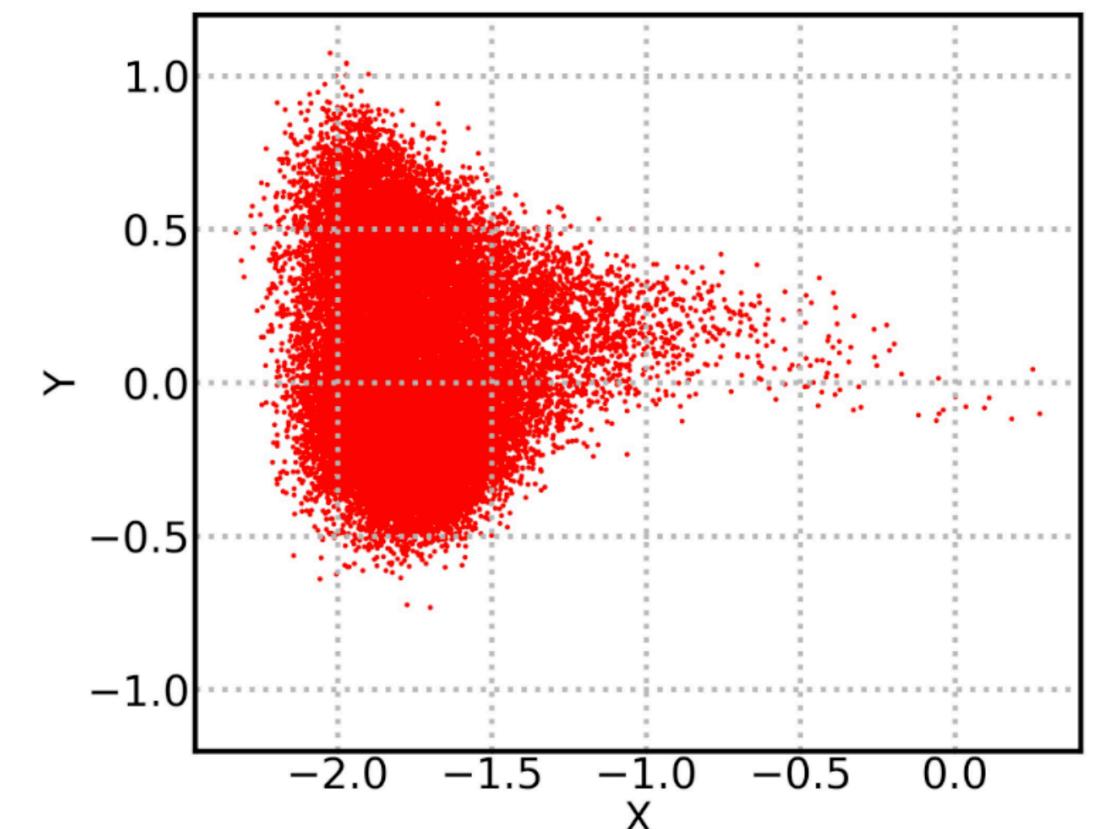
Alignment vs. Uniformity

Sentence embeddings from **pre-trained language models**?

- Pre-trained embeddings **well encode** sentence semantics but they are highly **anisotropic** (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020)

Good alignment

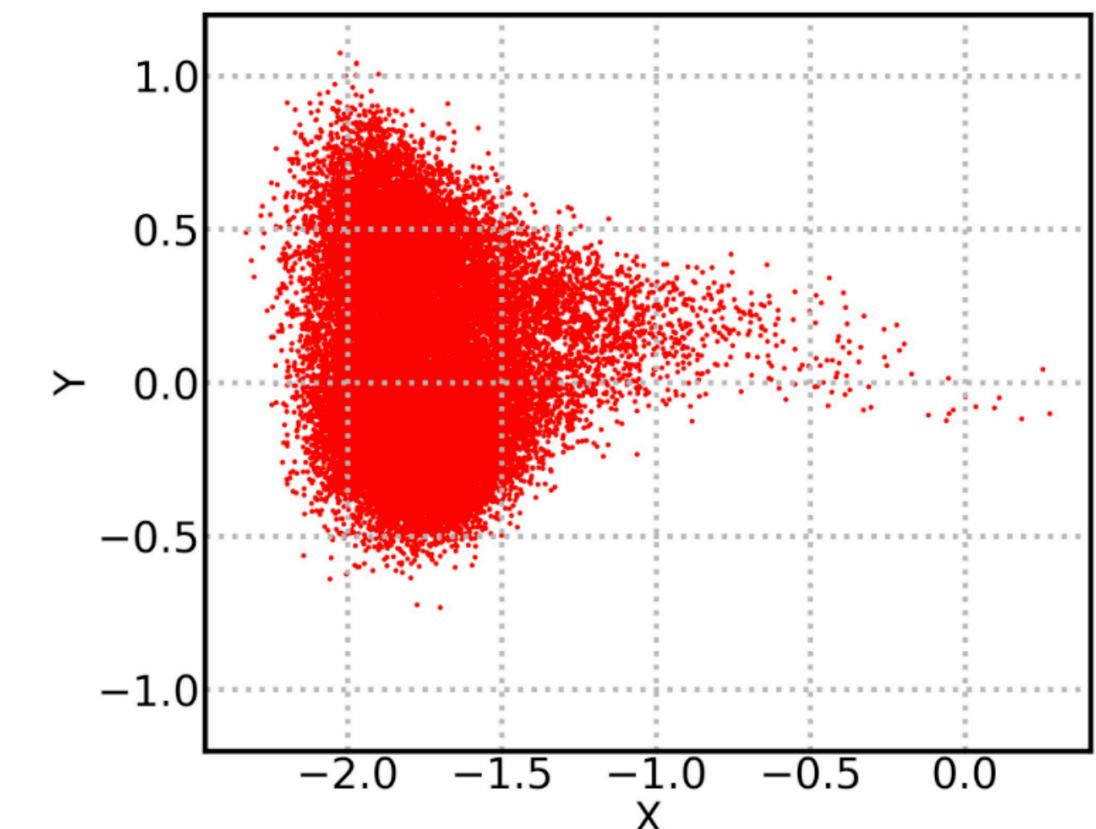
Bad uniformity



Alignment vs. Uniformity

Sentence embeddings from **pre-trained language models**?

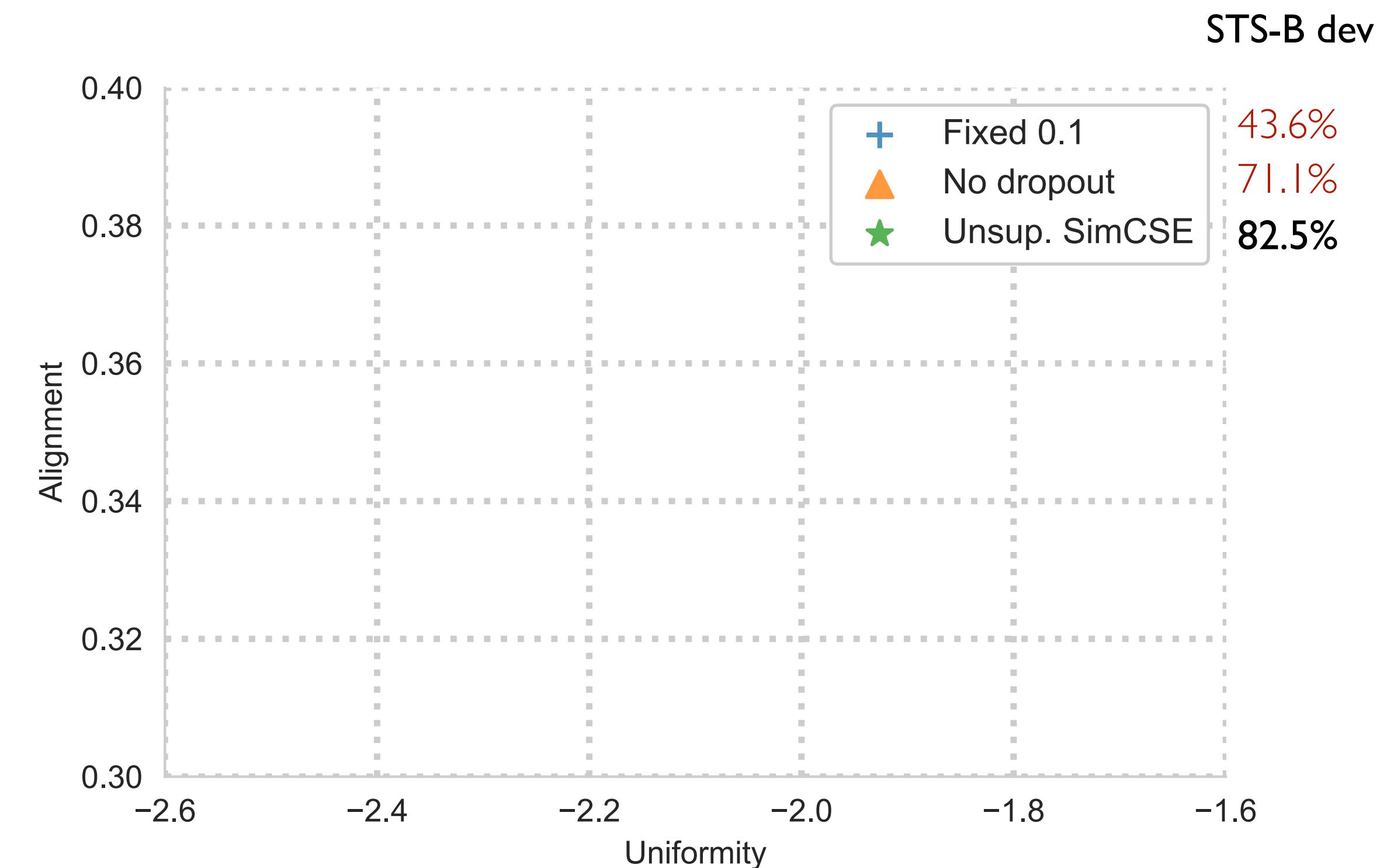
- Pre-trained embeddings **well encode** sentence semantics but they are highly **anisotropic** (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020)
 - Post-processing methods aim to improve uniformity
 - BERT-flow (Li et al., 2020)
 - BERT-whitening (Su et al., 2021)
- Good alignment* *Bad uniformity*



Unsupervised SimCSE: A Deep Dive

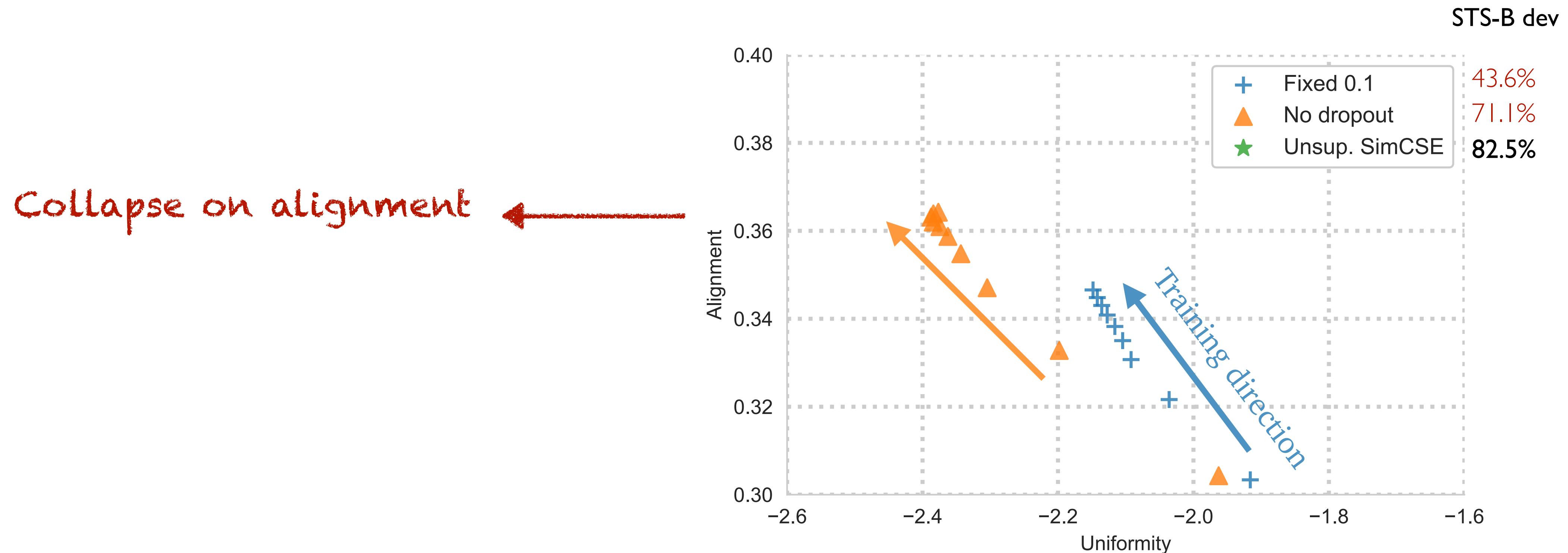
Two variants:

- **Fixed 0.1**
 - Standard dropout (rate=0.1)
 - Same dropout masks for positives
- **No dropout**
 - Dropout rate=0



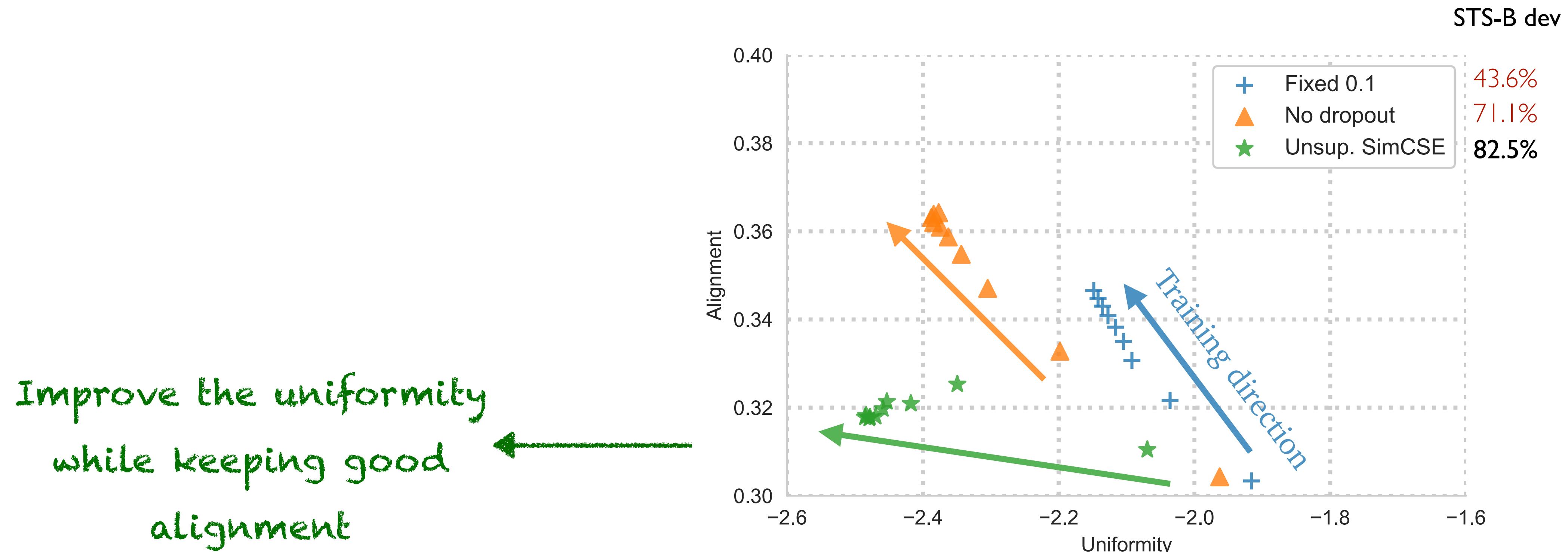
$l_{\text{uniform}}, l_{\text{align}}$: the lower, the better

Unsupervised SimCSE: A Deep Dive



$l_{\text{uniform}}, l_{\text{align}}$: the lower, the better

Unsupervised SimCSE: A Deep Dive

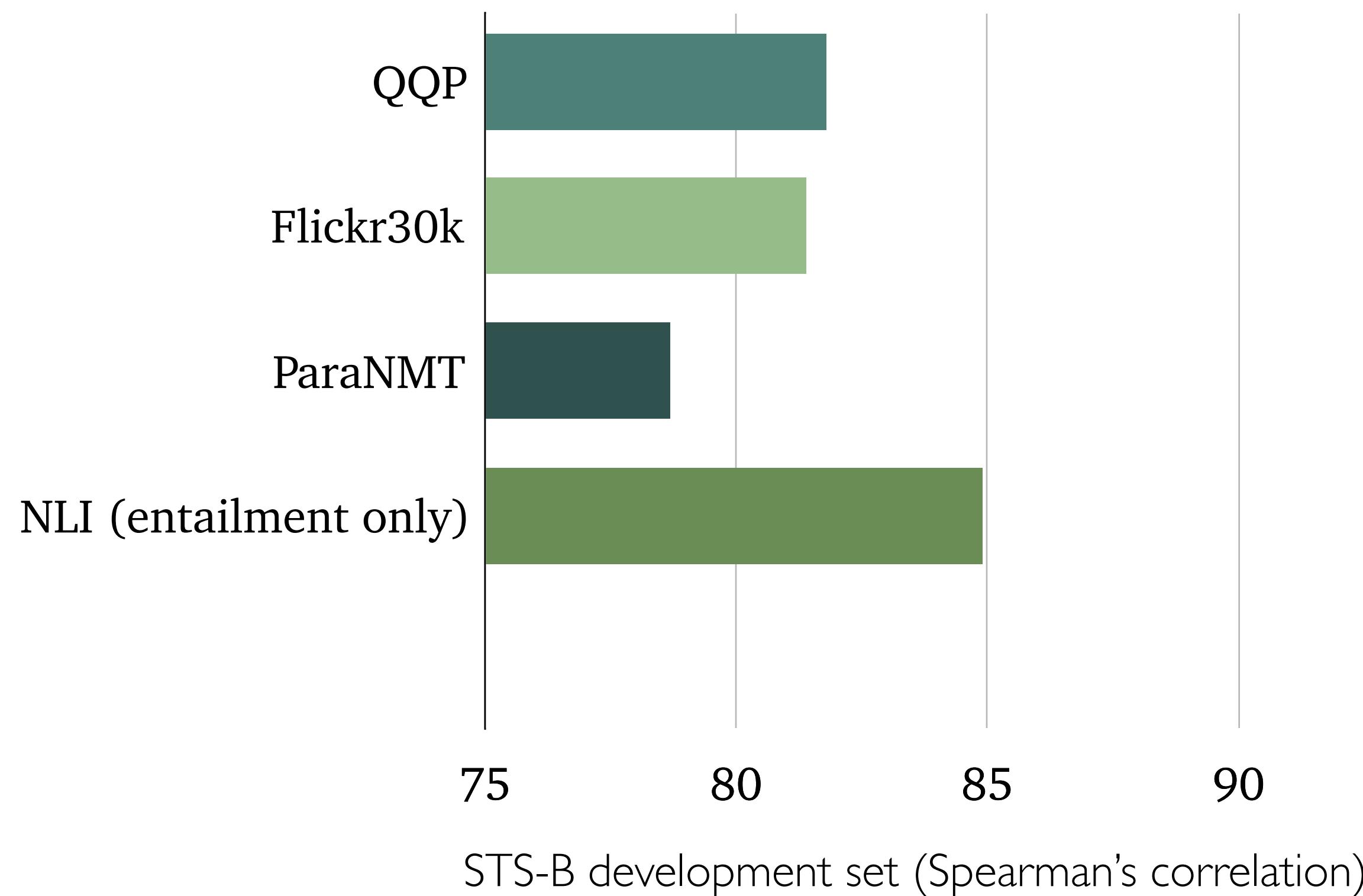


$l_{\text{uniform}}, l_{\text{align}}$: the lower, the better

Why Does This Work?

Supervised SimCSE

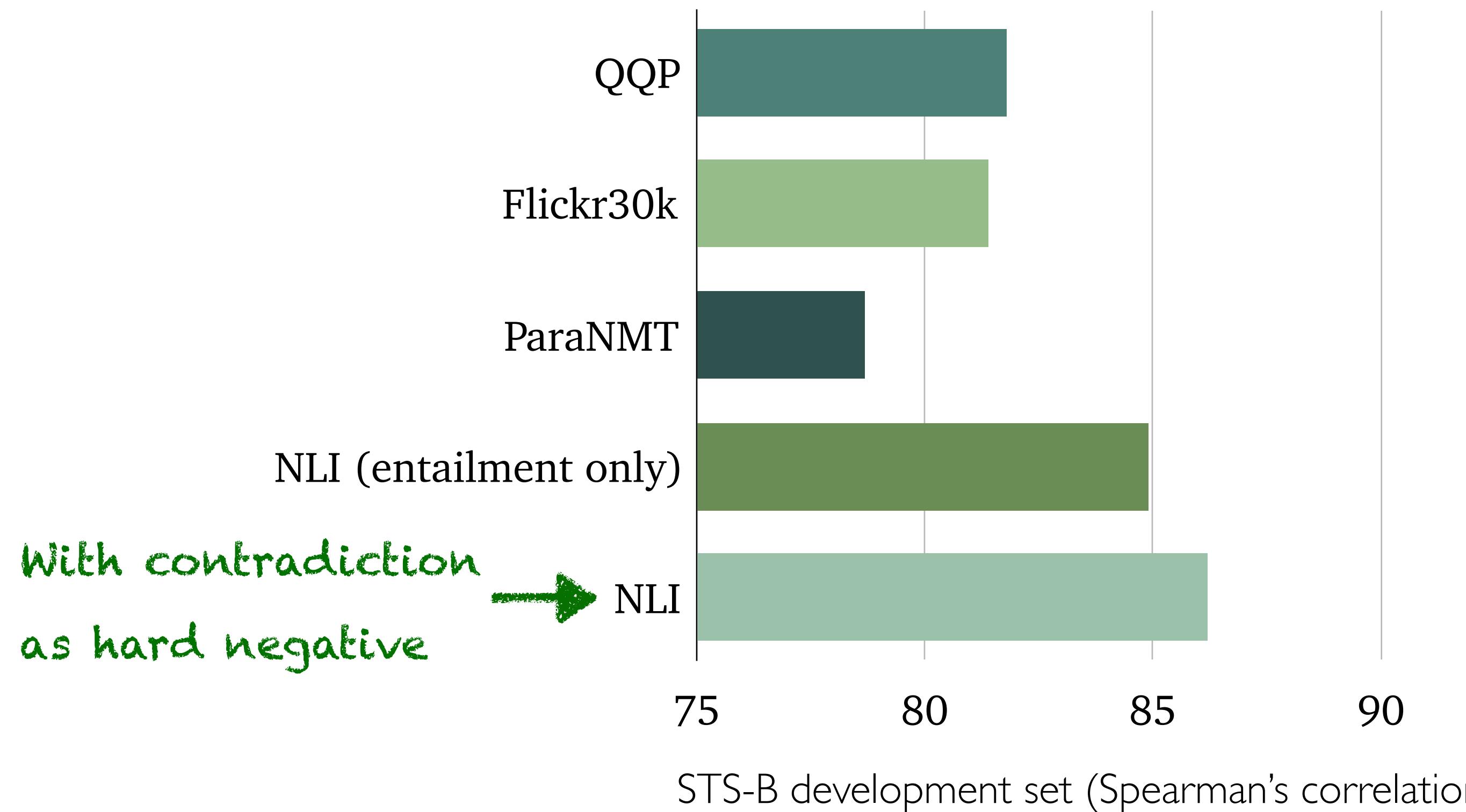
- Why is NLI a good dataset for positive pairs?



Why Does This Work?

Supervised SimCSE

- Why is NLI a good dataset for positive pairs?

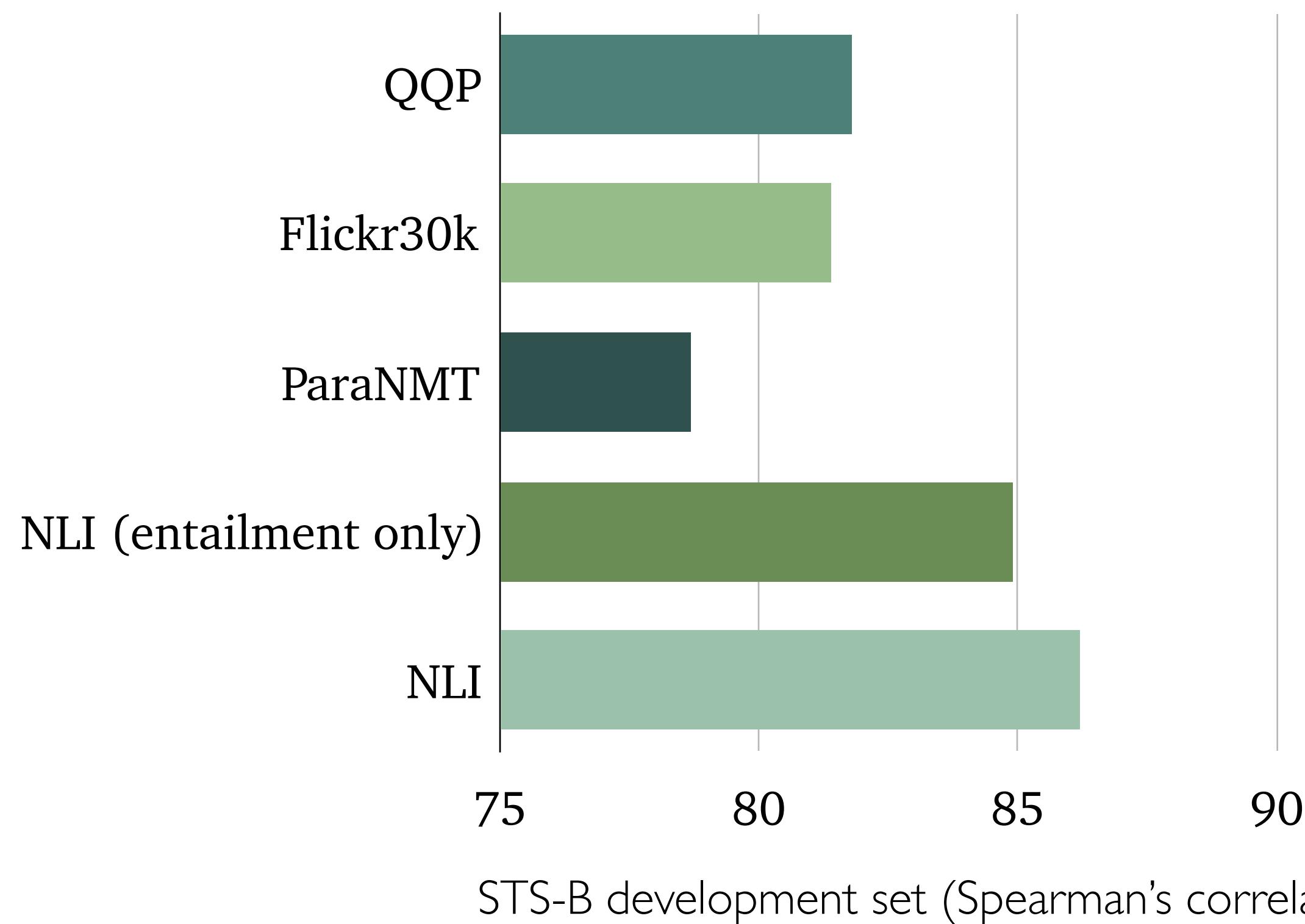


Why Does This Work?

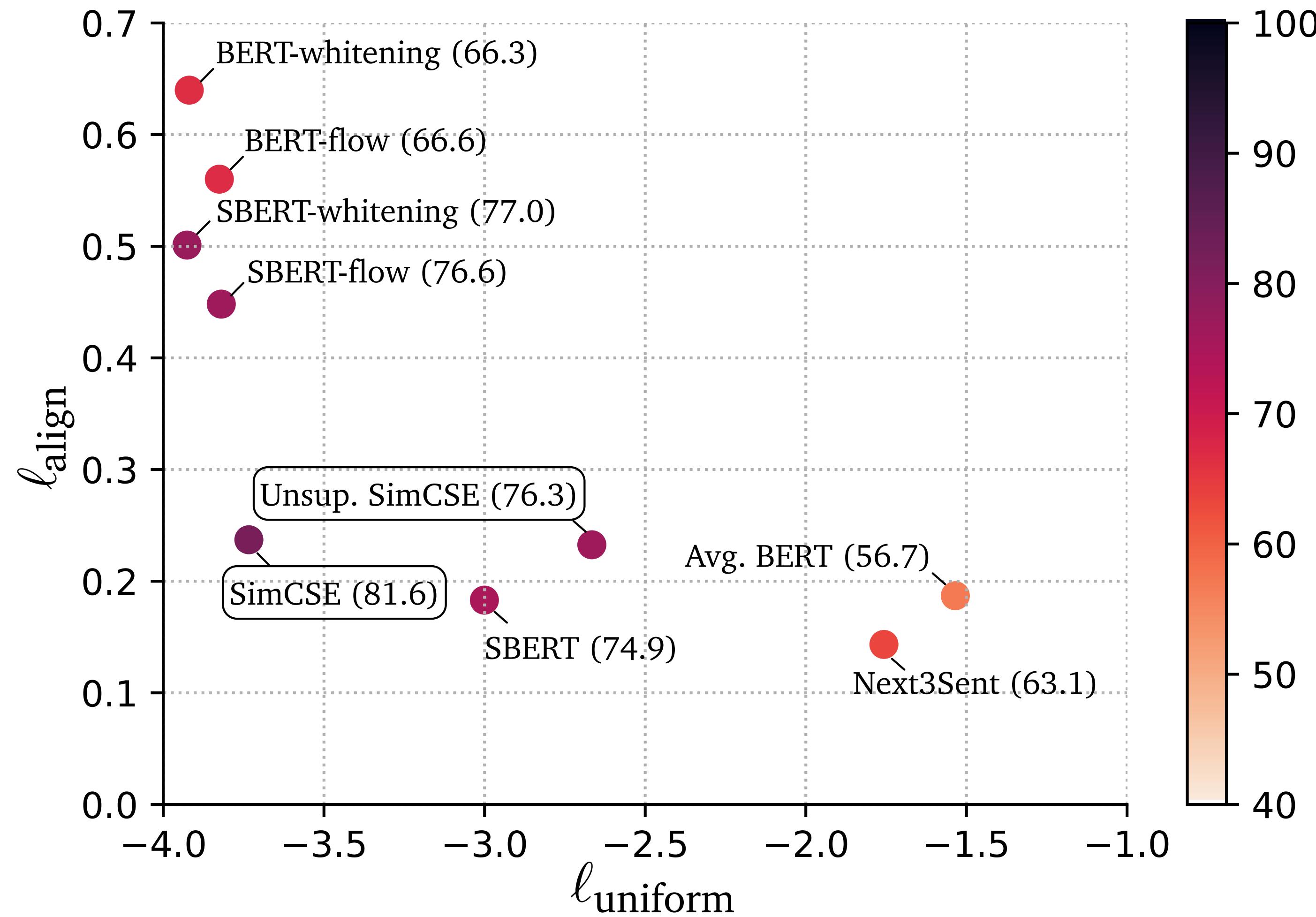
Supervised SimCSE

- Why is NLI a good dataset for positive pairs?

High annotation quality and
smaller lexical overlap

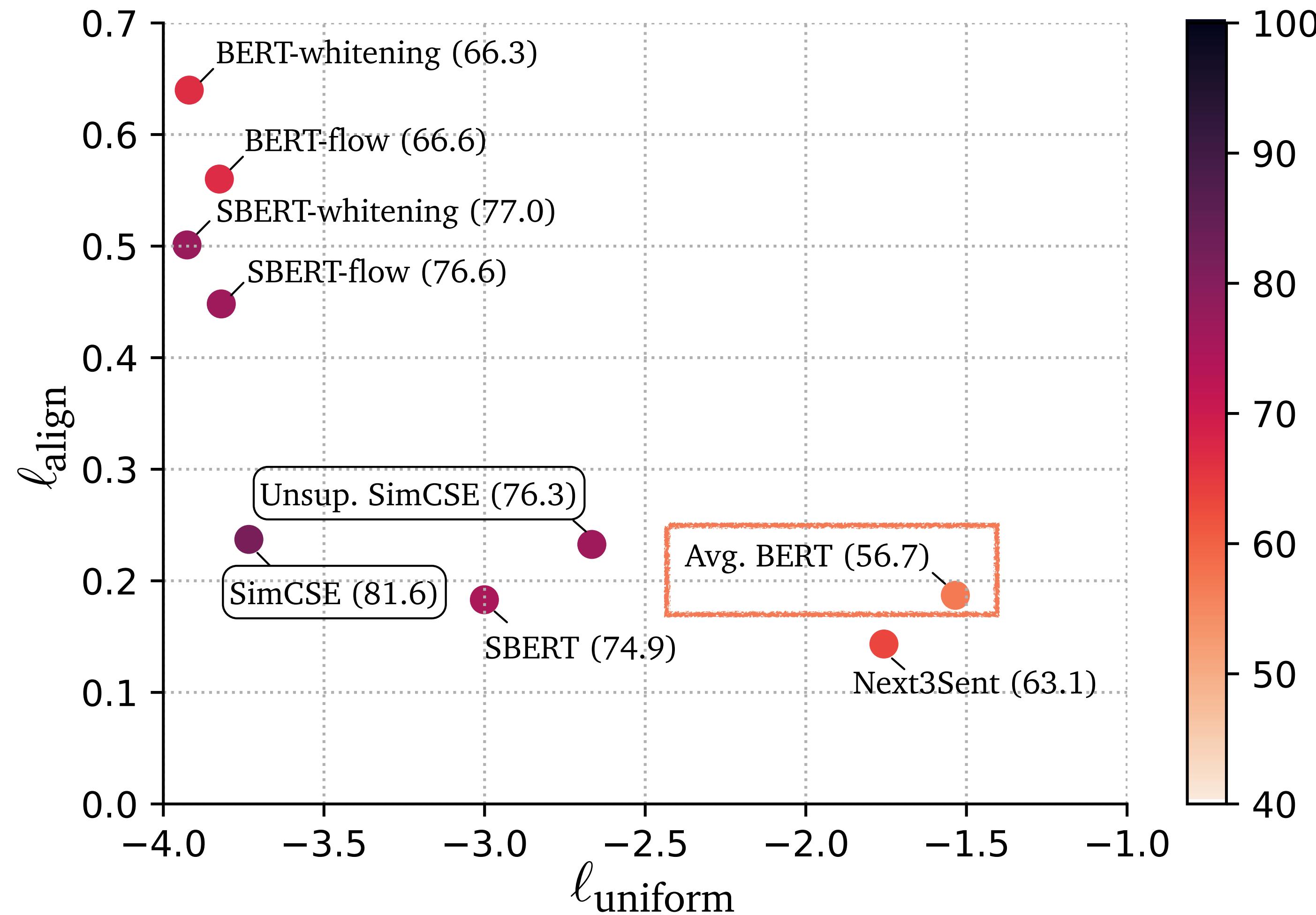


Comparison of All Models



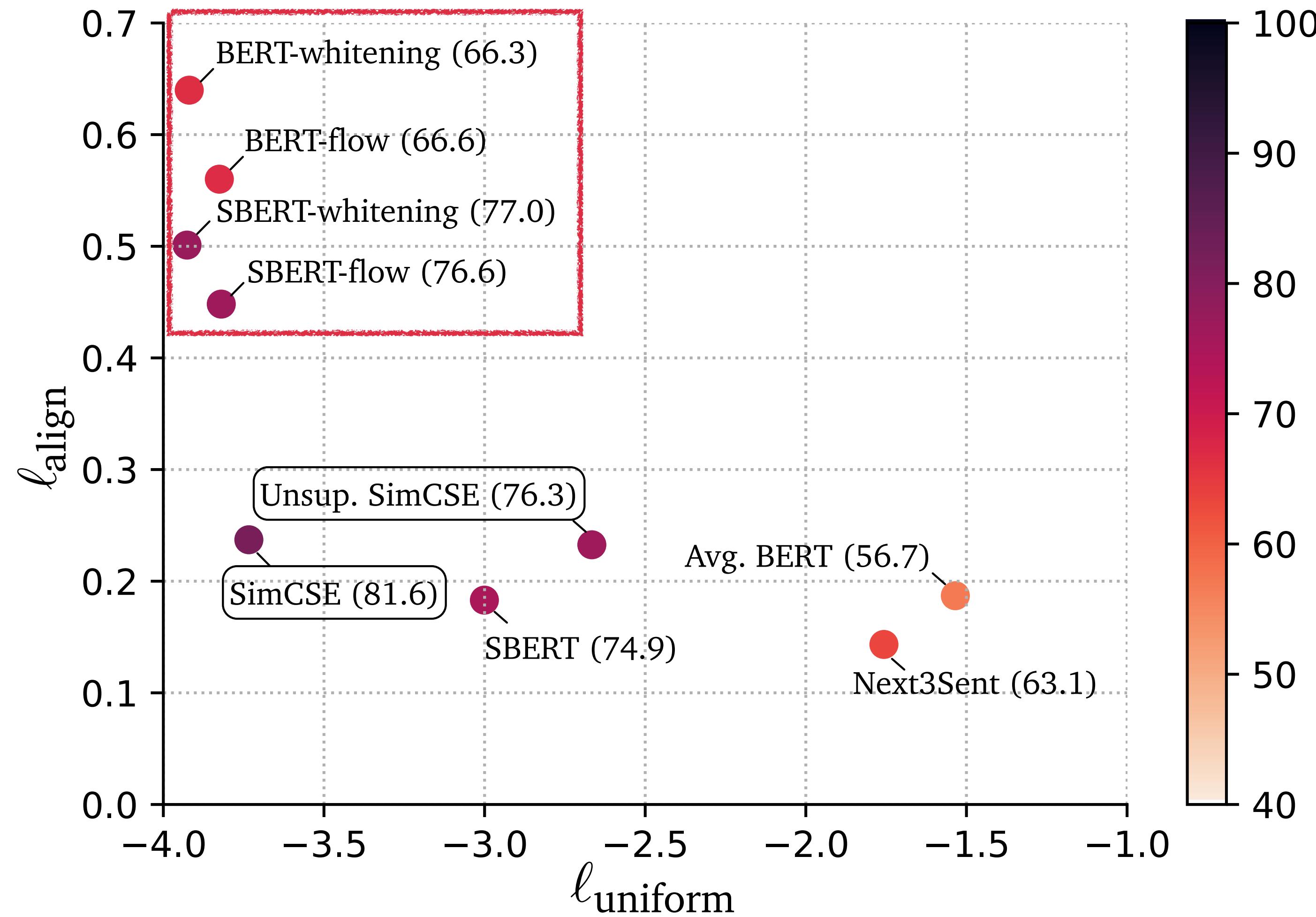
$\ell_{\text{uniform}}, \ell_{\text{align}}$: the lower, the better

Comparison of All Models



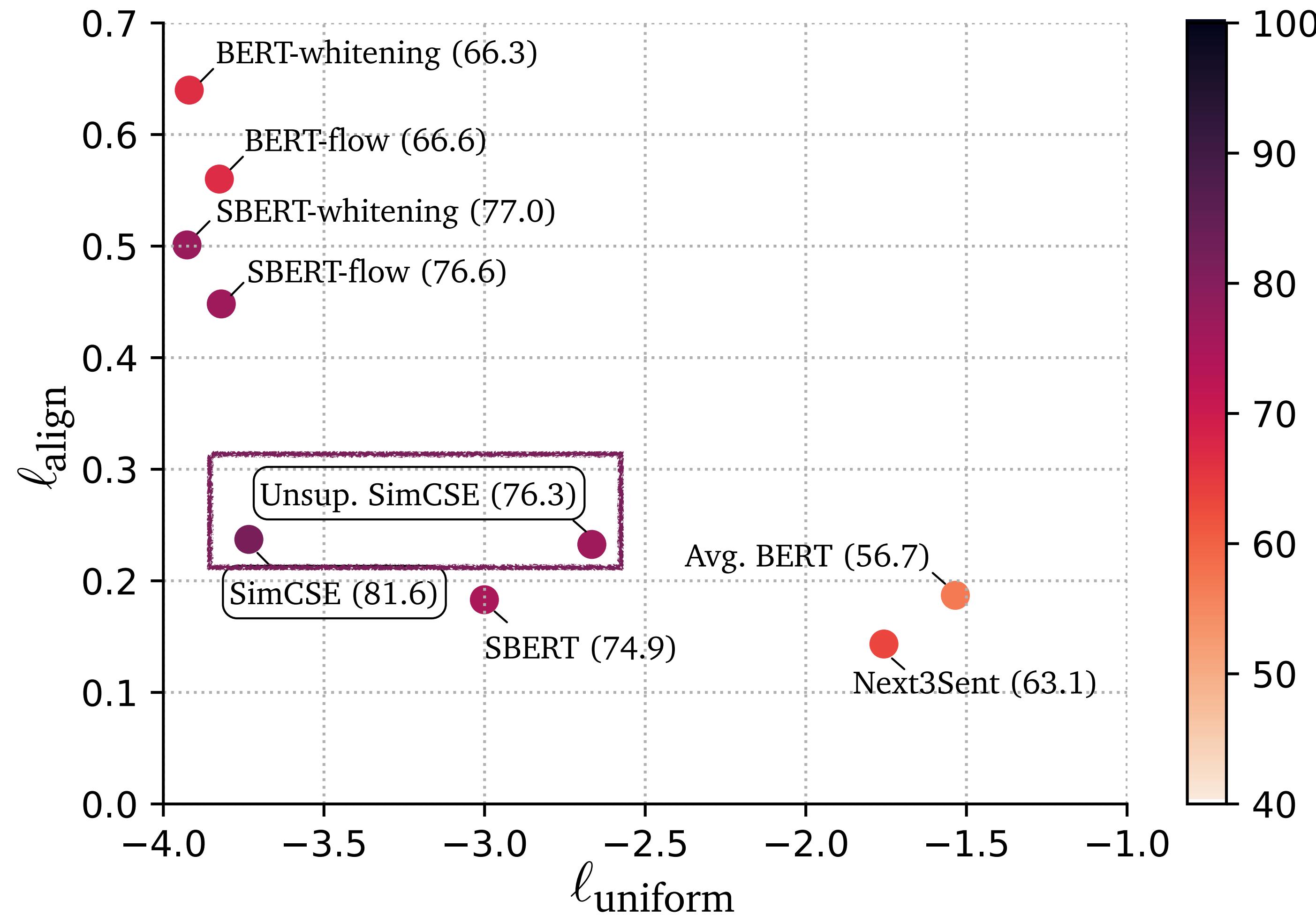
$\ell_{\text{uniform}}, \ell_{\text{align}}$: the lower, the better

Comparison of All Models



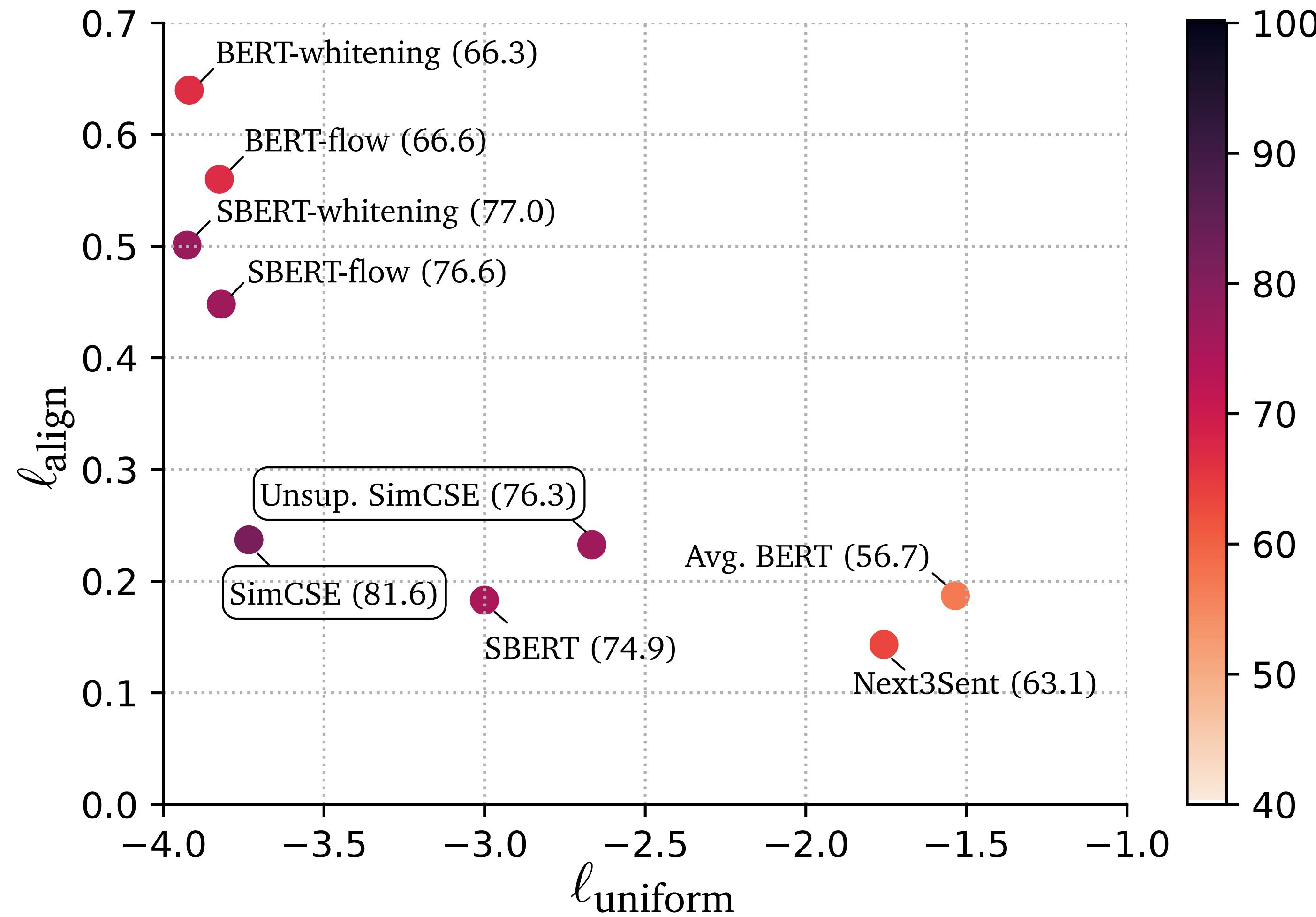
$\ell_{\text{uniform}}, \ell_{\text{align}}$: the lower, the better

Comparison of All Models



$\ell_{\text{uniform}}, \ell_{\text{align}}$: the lower, the better

Comparison of All Models



More theoretical analysis in the paper!

Summary

Summary

SimCSE

State-of-the-art sentence embeddings with contrastive learning

Summary

SimCSE

State-of-the-art sentence embeddings with contrastive learning

- Unsupervised SimCSE: standard **dropout** as positive pairs
- Supervised SimCSE: **entailment** from NLI as positives and **contradiction** as hard negatives

Summary

SimCSE

State-of-the-art sentence embeddings with contrastive learning

- Unsupervised SimCSE: standard **dropout** as positive pairs
- Supervised SimCSE: **entailment** from NLI as positives and **contradiction** as hard negatives

Why

Summary

SimCSE

State-of-the-art sentence embeddings with contrastive learning

- Unsupervised SimCSE: standard **dropout** as positive pairs
- Supervised SimCSE: **entailment** from NLI as positives and **contradiction** as hard negatives

Why

- Use **alignment** and **uniformity** to analyze different models
- **Theoretically** show that contrastive objective improves pre-trained embeddings' uniformity

Q & A

Code: <https://github.com/princeton-nlp/SimCSE>

Contact: tianyug@cs.princeton.edu



@gaotianyu1350



@princeton_nlp