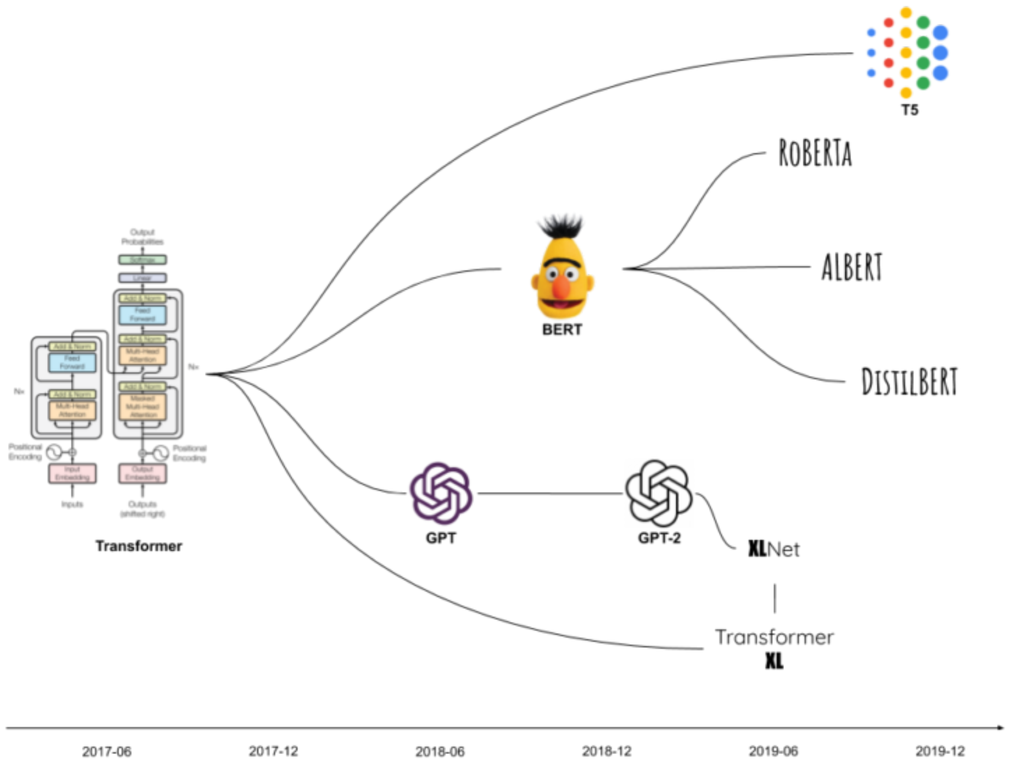
Towards Efficient Pre-training of Language Models

Danqi Chen

@danqi_chen @princeton_nlp

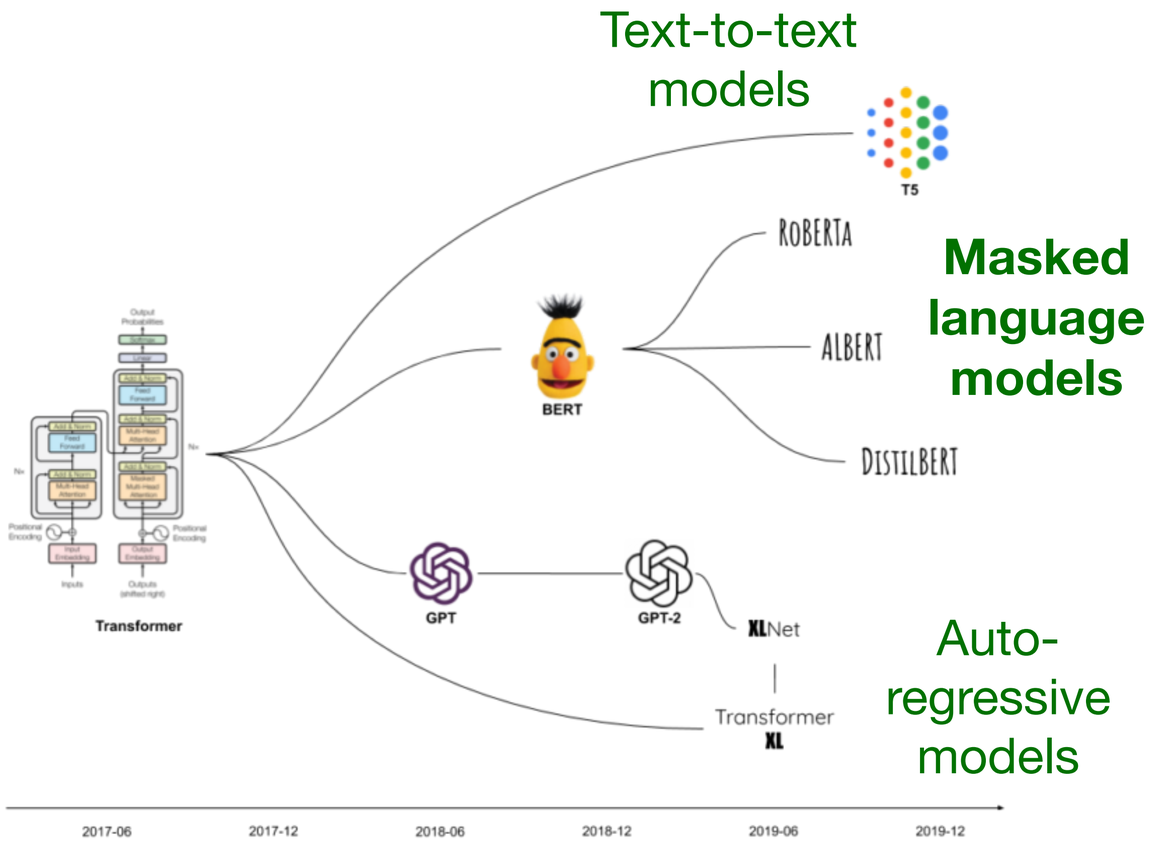
AAAI 2022 Workshop on Self-supervised Learning for Audio and Speech Processing

Pre-trained language models at scale



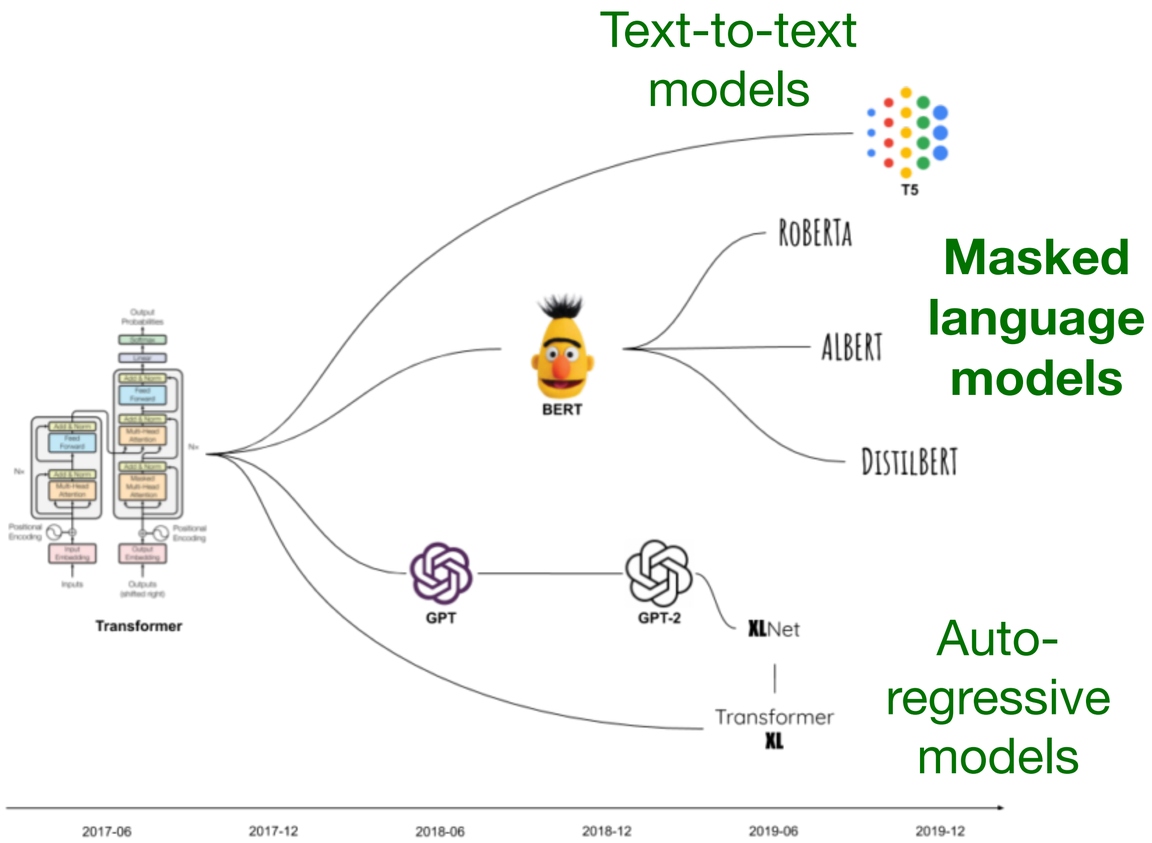
https://www.factored.ai/2021/09/21/an-intuitive-explanation-of-transformer-based-models/

Pre-trained language models at scale

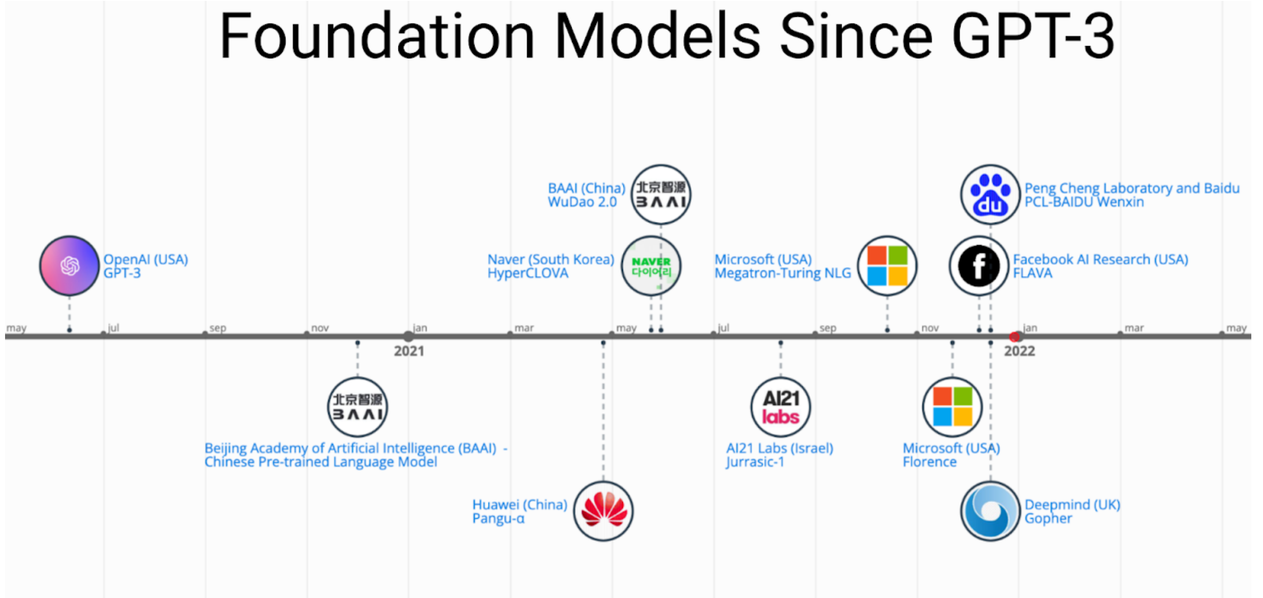


https://www.factored.ai/2021/09/21/an-intuitive-explanation-of-transformer-based-models/

Pre-trained language models at scale



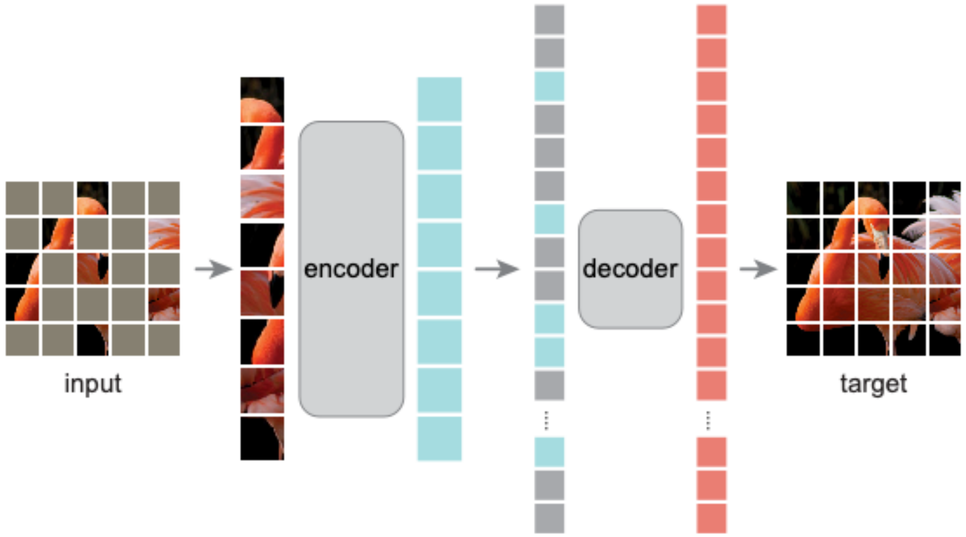
https://www.factored.ai/2021/09/21/an-intuitive-explanation-of-transformer-based-models/



https://lastweekin.ai/p/gpt-3-foundation-models-and-ai-nationalism

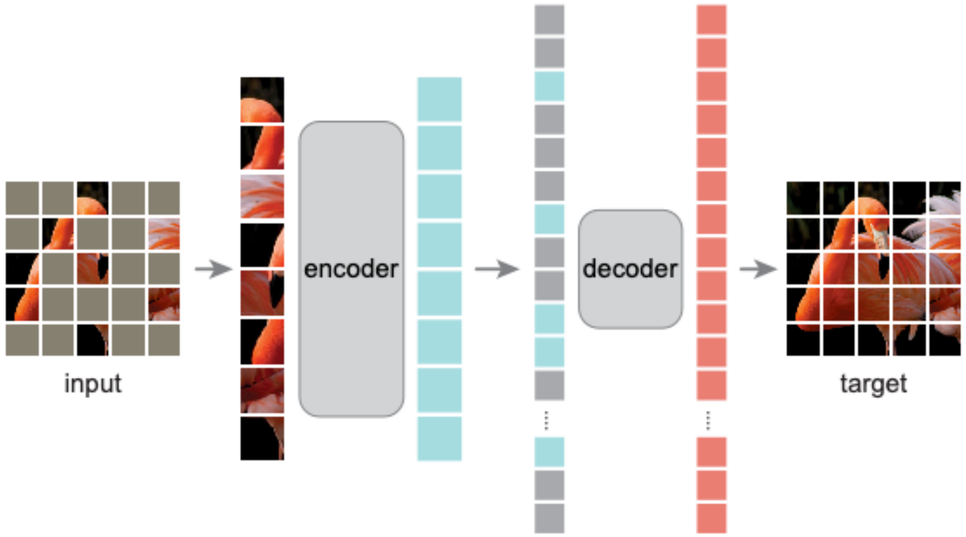
Ideas shared across different modalities

Ideas shared across different modalities

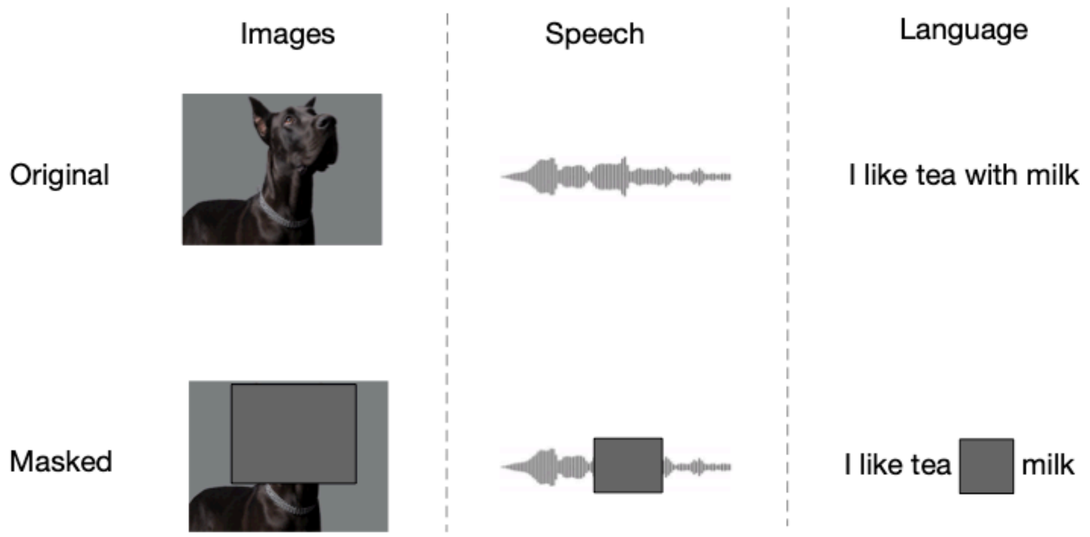


(He et al., 2021)

Ideas shared across different modalities



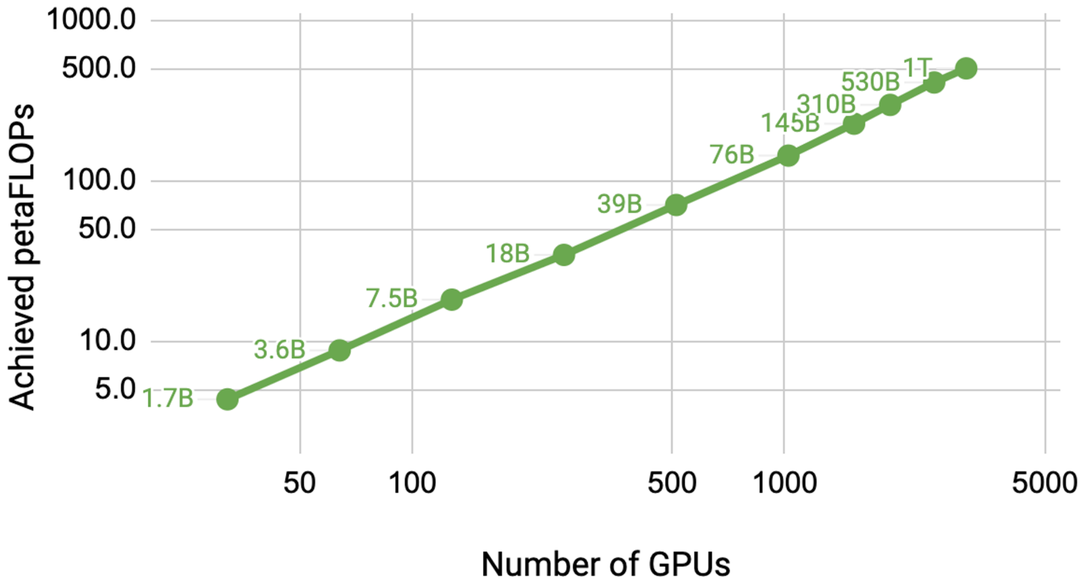
(He et al., 2021)



(Baevski et al., 2022)

This talk

As an academic researcher, what can we do in this pre-training era? What if we have at most **8 GPUs** to spare?

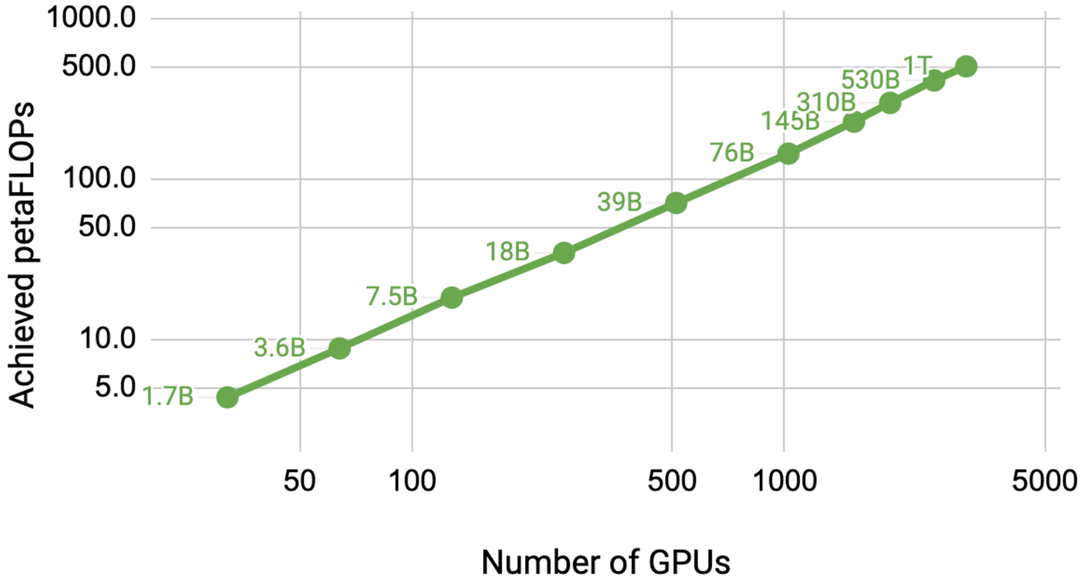


Source: Nvidia Blog

This talk

As an academic researcher, what can we do in this pre-training era? What if we have at most **8 GPUs** to spare?

How can we make pre-training more **accessible** and **affordable**?



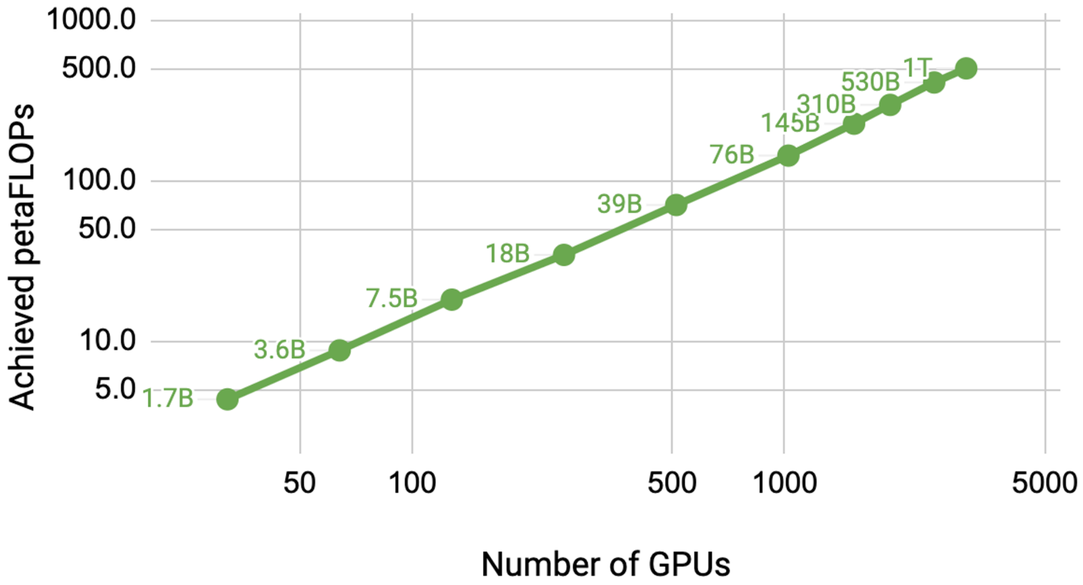
Source: Nvidia Blog

This talk

As an academic researcher, what can we do in this pre-training era? What if we have at most **8 GPUs** to spare?

How can we make pre-training more **accessible** and **affordable**?

Still, we need a lot of more pre-training (low-resource languages, specialized domains)



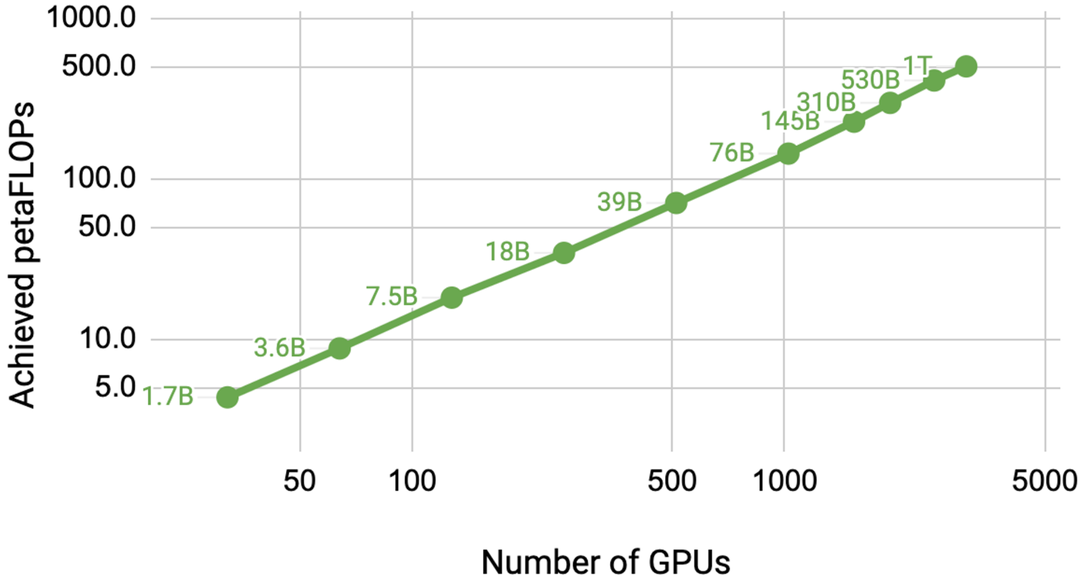
Source: Nvidia Blog

This talk

As an academic researcher, what can we do in this pre-training era? What if we have at most **8 GPUs** to spare?

How can we make pre-training more **accessible** and **affordable**?

Still, we need a lot of more pre-training (low-resource languages, specialized domains)



Source: Nvidia Blog

This talk: two case studies pointing to some promising directions for efficient pre-training

Exploring Optimal Masking Rates and Strategies in Masked Language Models

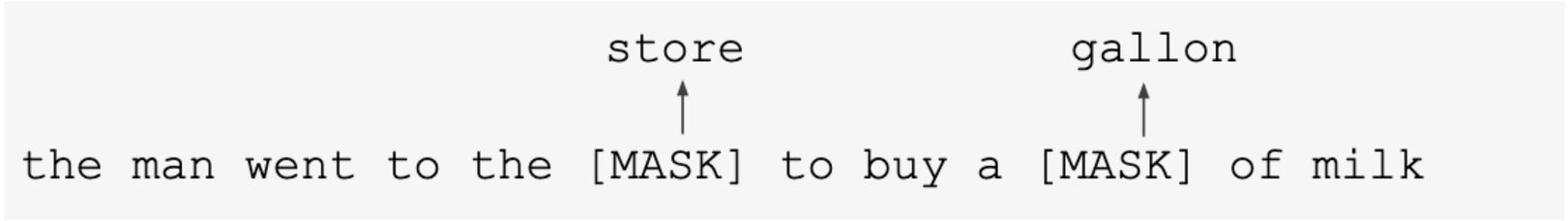


(Wettig et al., arXiv 2022) Should You Mask 15% in Masked Language Modeling?

BERT pre-training

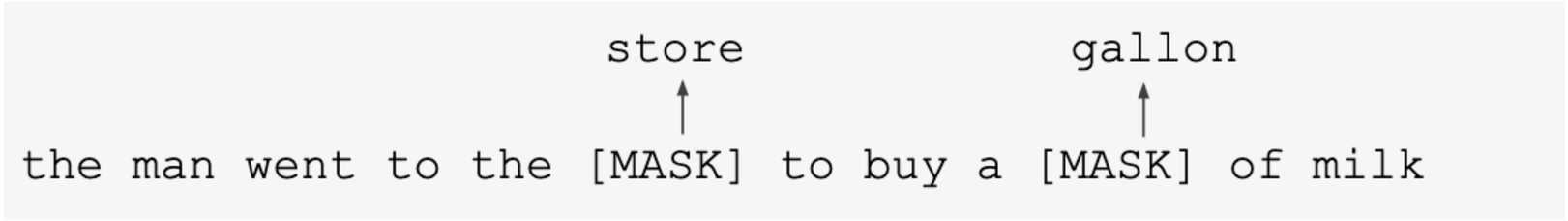
BERT pre-training

- BERT always masks $k = 15\%$ of input words, and predicts the masked words



BERT pre-training

- BERT always masks $k = 15\%$ of input words, and predicts the masked words



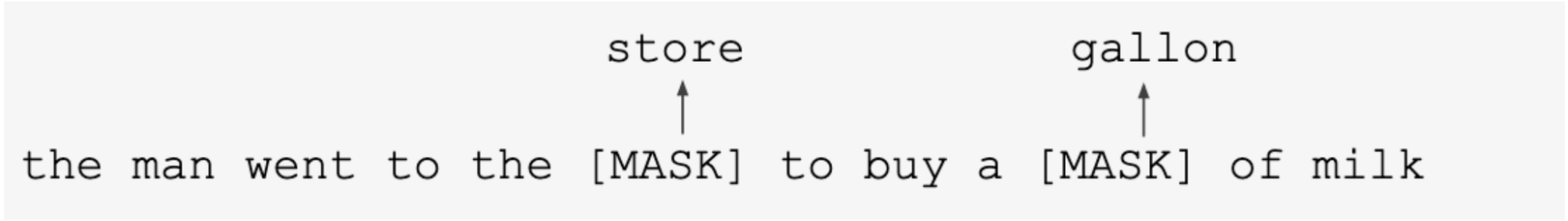
- Common belief:

Too little masking: Too expensive to train
Too much masking: Not enough context

(Credit: Jacob Devlin)

BERT pre-training

- BERT always masks $k = 15\%$ of input words, and predicts the masked words



- Common belief:

Too little masking: Too expensive to train

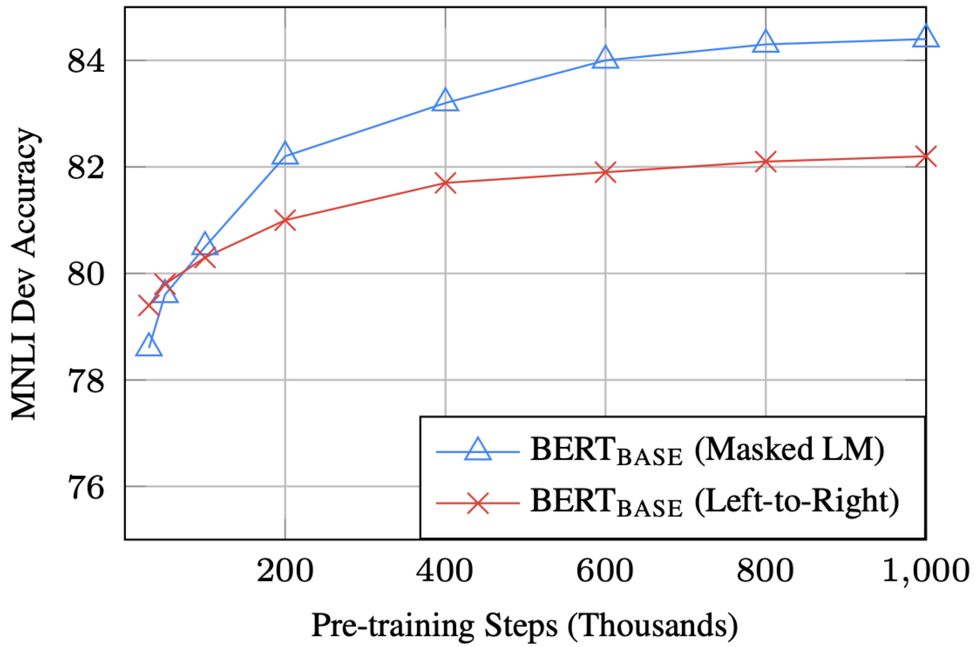
Too much masking: Not enough context

(Credit: Jacob Devlin)

- This 15% masking rate has been used in all the recent MLMs, regardless of **model sizes** and **masking strategies** (span masking, PMI masking)

BERT pre-training

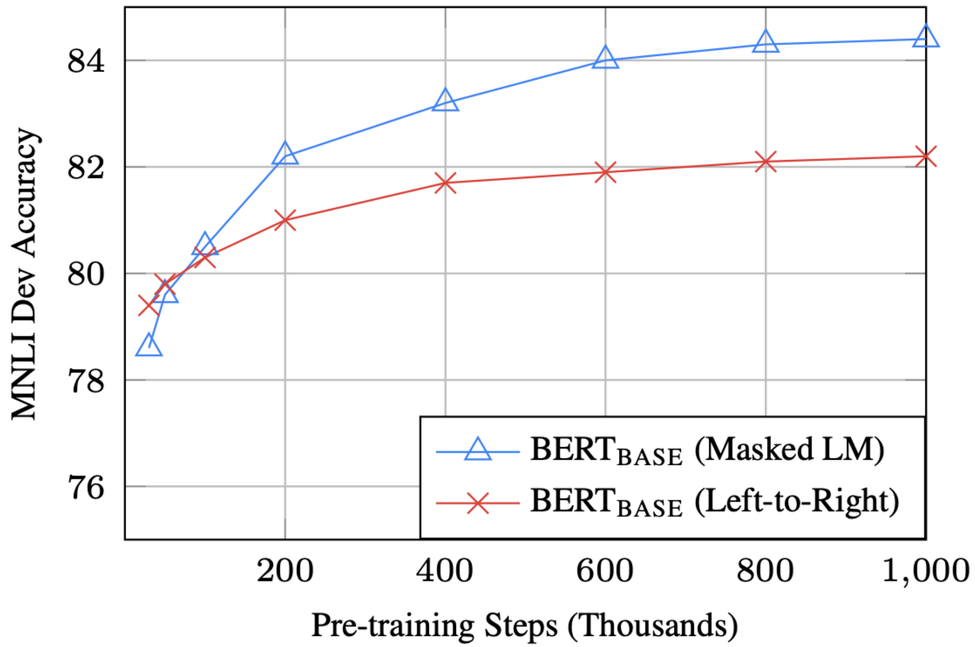
- MLMs are more effective than auto-regressive LMs due to their **bidirectionality**



(Devlin et al., 2019)

BERT pre-training

- MLMs are more effective than auto-regressive LMs due to their **bidirectionality**

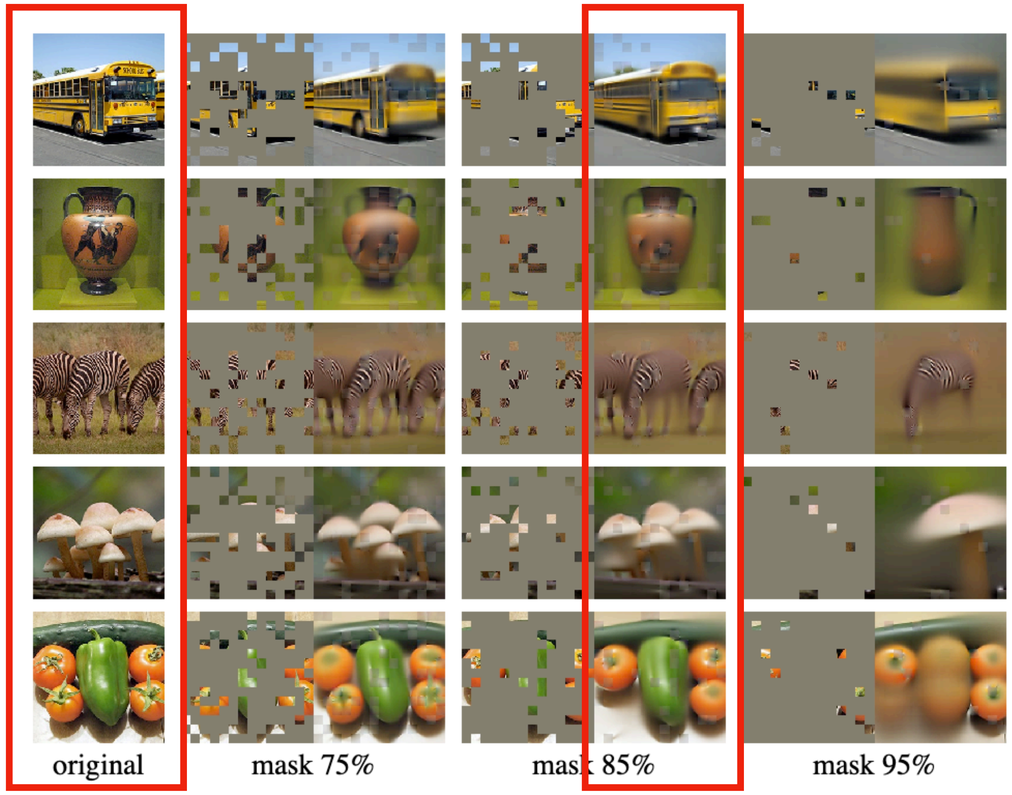


(Devlin et al., 2019)

- However, this k = 15% masking rate has been viewed as a bottleneck for **sample efficiency** because it only learns 15% of tokens per sample

Can we mask more?

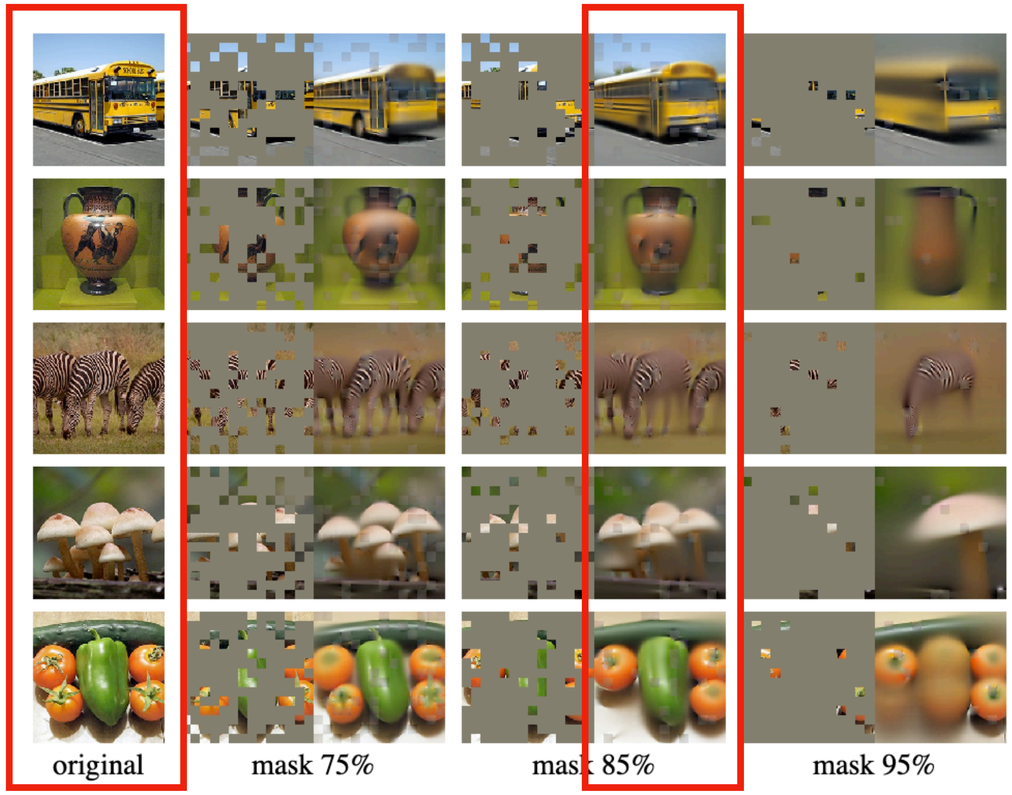
Can we mask more?



(He et al., 2021)

- In images, we can mask 75% to 95% portion of images and still reconstruct the original input plausibly

Can we mask more?



(He et al., 2021)

- In images, we can mask 75% to 95% portion of images and still reconstruct the original input plausibly
- Can we do this for text?

"Information density is different between language and vision."

Can we mask more?

Sentence: We study high mask ing rates in pre-training language models .

	Pre-training	
m	Example	PPL

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Can we mask more?

Sentence: We study high mask ing rates in pre-training language models .

	Pre-training	
m	Example	PPL
15%	We study high [redacted] ing rates [redacted] pre-training language models .	17.7

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Can we mask more?

Sentence: We study high mask ing rates in pre-training language models .

	Pre-training	
m	Example	PPL
40%	We study high [MASK] [MASK] rates [MASK] pre- [MASK] [MASK] models .	69.4
15%	We study high [MASK] ing rates [MASK] pre-training language models .	17.7

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Can we mask more?

Sentence: We study high mask ing rates in pre-training language models .

	Pre-training	
m	Example	PPL
80%	We [▮▮▮▮] high [▮▮▮▮] [▮▮] [▮▮▮] [▮] [▮▮▮▮▮▮▮] [▮▮▮▮▮▮] models [▮]	1141.4
40%	We study high [▮▮▮▮] [▮▮] rates [▮] pre- [▮▮▮▮▮] [▮▮▮▮▮▮] models .	69.4
15%	We study high [▮▮▮▮] ing rates [▮] pre-training language models .	17.7

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Can we mask more?

Sentence: We study high mask ing rates in pre-training language models .

*: Fine-tuning on downstream tasks

Pre-training			Downstream		
m	Example	PPL	MNLI	QNLI	SQuAD[2]
80%	We █ high █ █ █ █ █ █ models █	1141.4	80.8	87.9	86.2
40%	We study high █ █ rates █ pre-█ █ models .	69.4	**84.5**	**91.6**	**89.8**
15%	We study high █ ing rates █ pre-training language models .	17.7	84.2	90.9	88.0

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Can we mask more?

Sentence: We study high mask ing rates in pre-training language models .

*: Fine-tuning on downstream tasks

	Pre-training		Downstream		
m	Example	PPL	MNLI	QNLI	SQuAD[2]
80%	We high models	1141.4	80.8	87.9	86.2
40%	We study high rates pre- models .	69.4	**84.5**	**91.6**	**89.8**
15%	We study high ing rates pre-training language models .	17.7	84.2	90.9	88.0

- These results build on <u>an efficient pre-training recipe</u> (Izsak et al., 2021): training a BERT-large model for 24 hours on 8 RTX 2080 GPUs, performance at least good as BERT-base

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Can we mask more?

Sentence: We study high mask ing rates in pre-training language models .

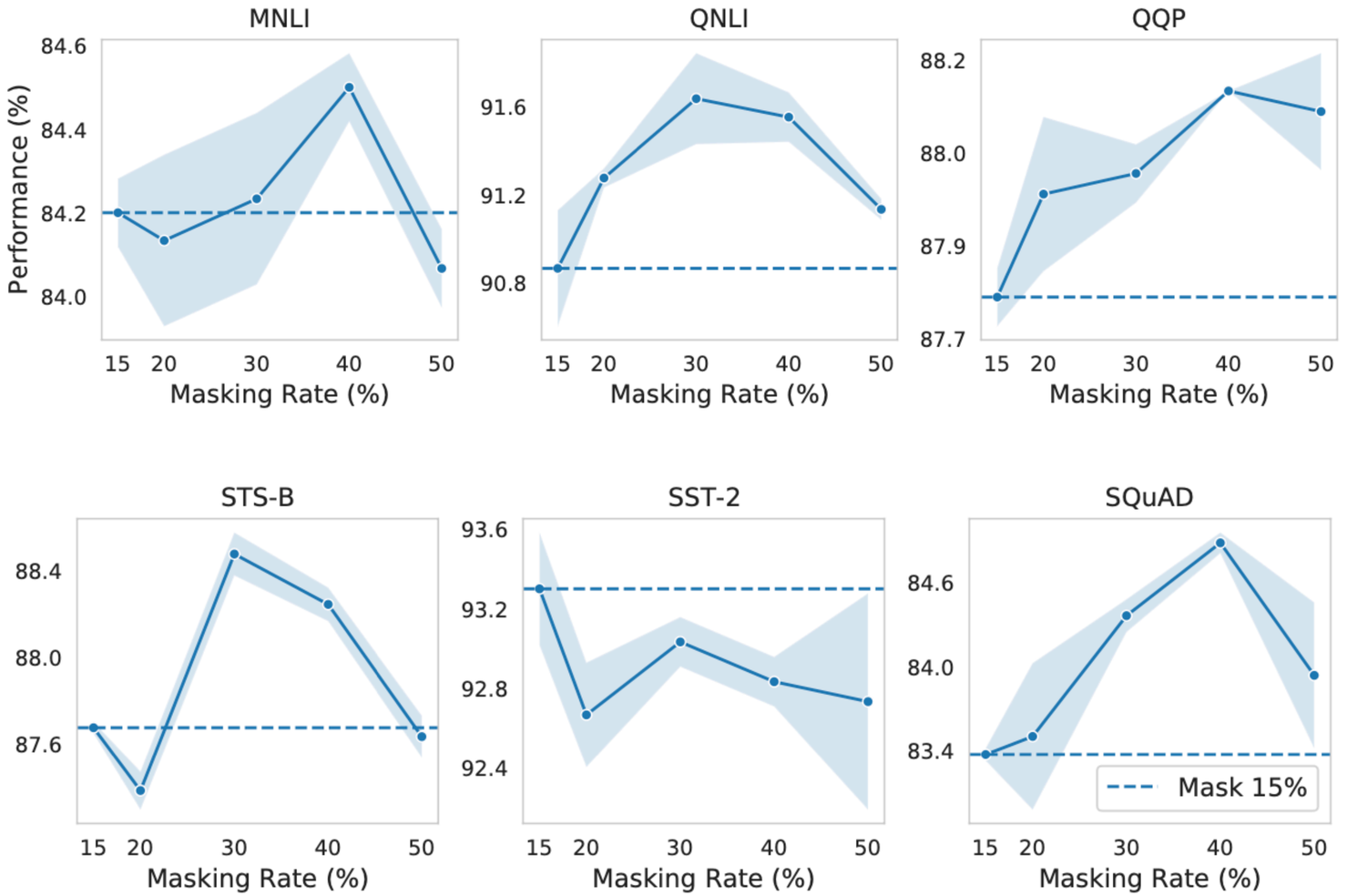
*: Fine-tuning on downstream tasks

	Pre-training		Downstream		
m	Example	PPL	MNLI	QNLI	SQuAD[2]
80%	We high models	1141.4	80.8	87.9	86.2
40%	We study high rates pre- models .	69.4	**84.5**	**91.6**	**89.8**
15%	We study high ing rates pre-training language models .	17.7	84.2	90.9	88.0

- These results build on an efficient pre-training recipe (Izsak et al., 2021): training a BERT-large model for 24 hours on 8 RTX 2080 GPUs, performance at least good as BERT-base
- The ability to reconstruct the original inputs may **NOT correlate** with representations for downstream task fine-tuning.

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Can we mask more?

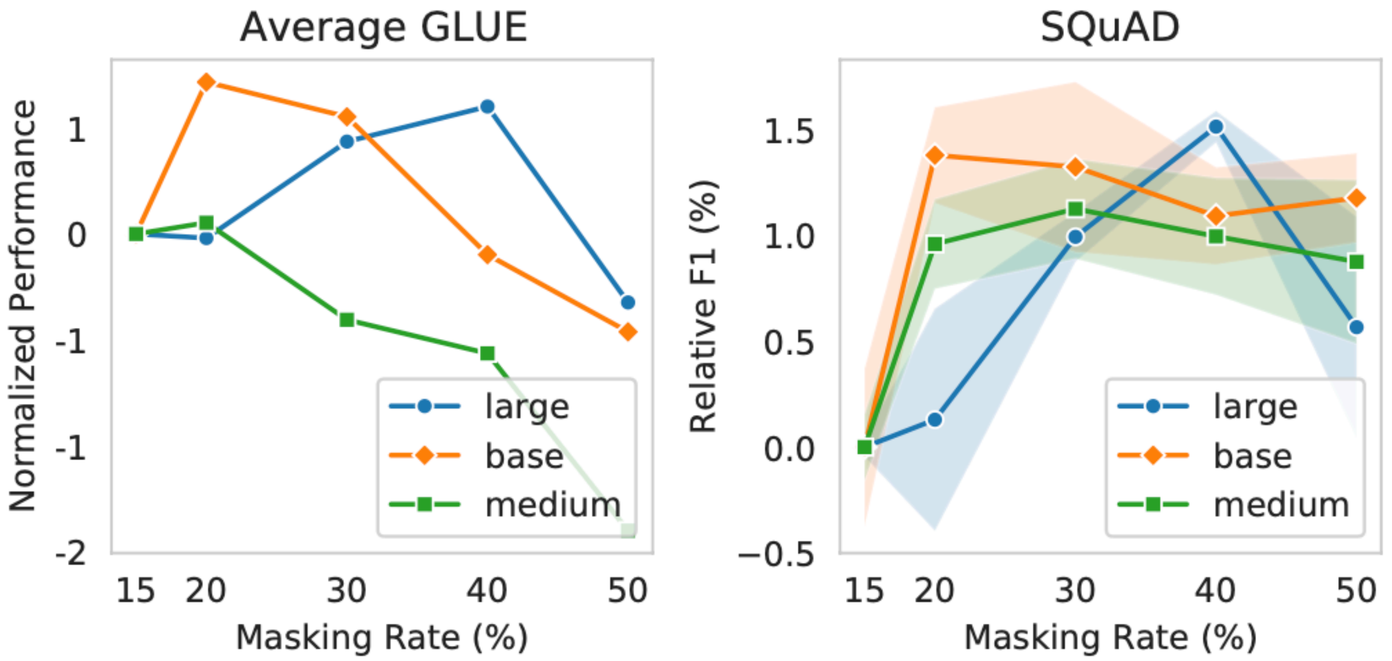


40% masking rate is consistently better than **15%** masking rate

The improvements are even **larger** when combined with **efficient training recipes**

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

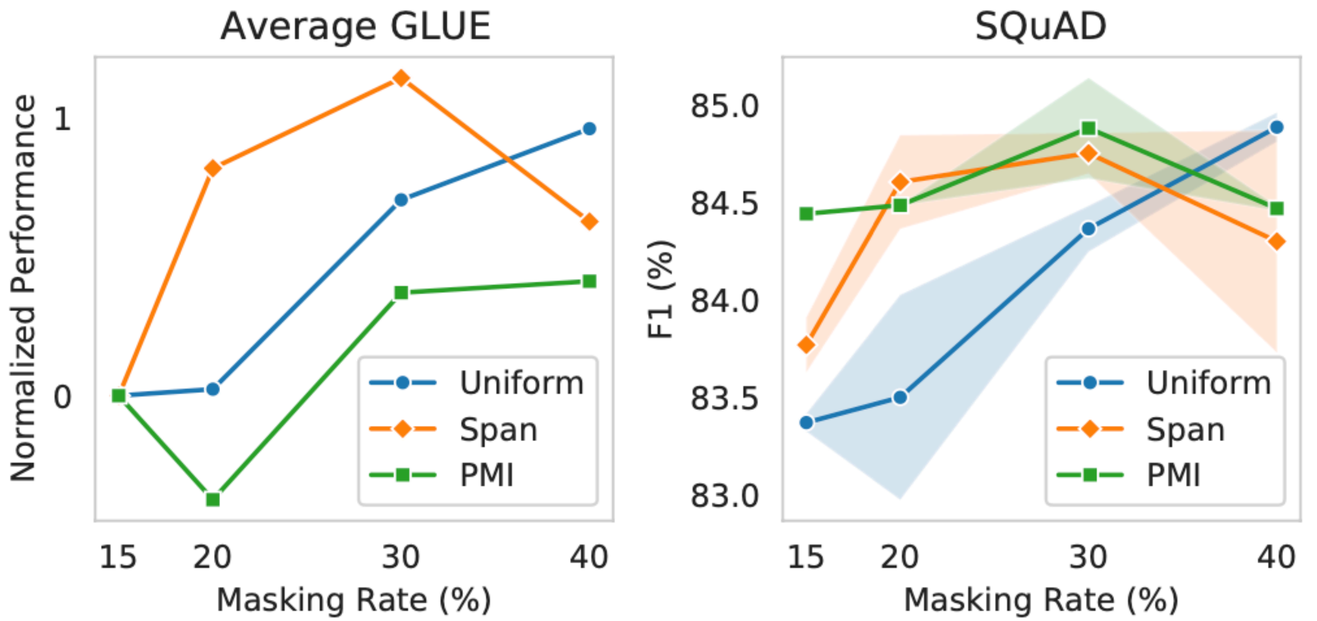
Other key findings



Larger models should adopt **larger masking rates**

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

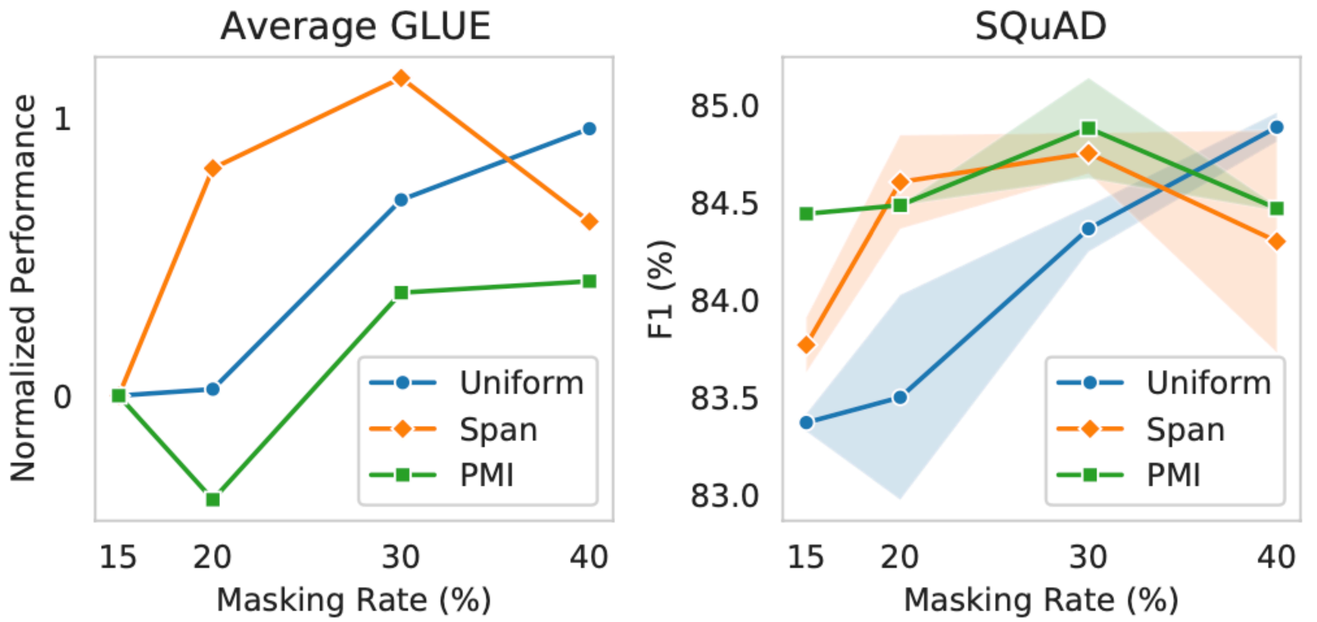
Other key findings



Uniform masking is competitive as **span masking** (Joshi et al., 2020) or **PMI masking** (Levine et al., 2021) with **high masking rates**

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

Other key findings



Uniform masking is competitive as **span masking** (Joshi et al., 2020) or **PMI masking** (Levine et al., 2021) with **high masking rates**

Bonus: The 80-10-10 corruption strategy may be not important too (see the paper)!

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

What can we learn from this?

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

What can we learn from this?

- The less we **corrupt**, the better.
- The more we **predict**, the better.

corruption rate prediction rate

m_{corr}	m_{pred}	MNLI	QNLI	QQP	STS-B	SST-2
40%	40%	84.5	91.6	88.1	88.2	92.8
40%	20%	83.7↓	90.6↓	87.8↓	87.5↓	92.9
20%	20%	84.1↓	91.3↓	87.9↓	87.4↓	92.7↓
20%	40%	85.7↑	92.0↑	87.9↓	88.6↑	93.4↑
10%	40%	86.3↑	92.3↑	88.3↑	88.9↑	93.2↑
5%	40%	86.9↑	92.2↑	88.5↑	88.6↑	93.9↑

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

What can we learn from this?

- The less we **corrupt**, the better.
- The more we **predict**, the better.

Q: Can we **disentangle** the **corruption rate** and **prediction rate** to enable more efficient pre-training?

corruption rate prediction rate

m_{corr}	m_{pred}	MNLI	QNLI	QQP	STS-B	SST-2
40%	40%	84.5	91.6	88.1	88.2	92.8
40%	20%	83.7↓	90.6↓	87.8↓	87.5↓	92.9
20%	20%	84.1↓	91.3↓	87.9↓	87.4↓	92.7↓
20%	40%	85.7↑	92.0↑	87.9↓	88.6↑	93.4↑
10%	40%	86.3↑	92.3↑	88.3↑	88.9↑	93.2↑
5%	40%	86.9↑	92.2↑	88.5↑	88.6↑	93.9↑

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

What can we learn from this?

- The less we **corrupt**, the better.
- The more we **predict**, the better.

Q: Can we **disentangle** the **corruption rate** and **prediction rate** to enable more efficient pre-training?

corruption rate prediction rate

m_{corr}	m_{pred}	MNLI	QNLI	QQP	STS-B	SST-2
40%	40%	84.5	91.6	88.1	88.2	92.8
40%	20%	83.7↓	90.6↓	87.8↓	87.5↓	92.9
20%	20%	84.1↓	91.3↓	87.9↓	87.4↓	92.7↓
20%	40%	85.7↑	92.0↑	87.9↓	88.6↑	93.4↑
10%	40%	86.3↑	92.3↑	88.3↑	88.9↑	93.2↑
5%	40%	86.9↑	92.2↑	88.5↑	88.6↑	93.9↑

- Why do we want high masking rates?

Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

What can we learn from this?

- The less we **corrupt**, the better.
- The more we **predict**, the better.

Q: Can we **disentangle** the **corruption rate** and **prediction rate** to enable more efficient pre-training?

corruption rate prediction rate

m_{corr}	m_{pred}	MNLI	QNLI	QQP	STS-B	SST-2
40%	40%	84.5	91.6	88.1	88.2	92.8
40%	20%	83.7↓	90.6↓	87.8↓	87.5↓	92.9
20%	20%	84.1↓	91.3↓	87.9↓	87.4↓	92.7↓
20%	40%	85.7↑	92.0↑	87.9↓	88.6↑	93.4↑
10%	40%	86.3↑	92.3↑	88.3↑	88.9↑	93.2↑
5%	40%	86.9↑	92.2↑	88.5↑	88.6↑	93.9↑

- Why do we want high masking rates?

Q: Can we **only encode** the unmasked (50-60%) tokens to reduce the number of computations?

He et al., 2021 show that they can achieve 4.1x speed-up with 75% input masked

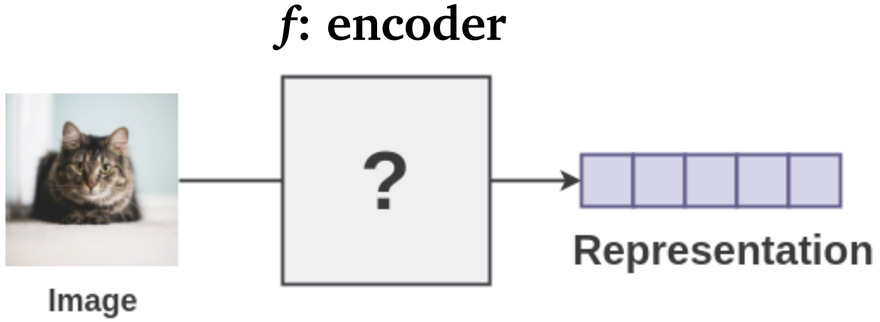
Wettig et al., 2022. Should You Mask 15% in Masked Language Modeling?

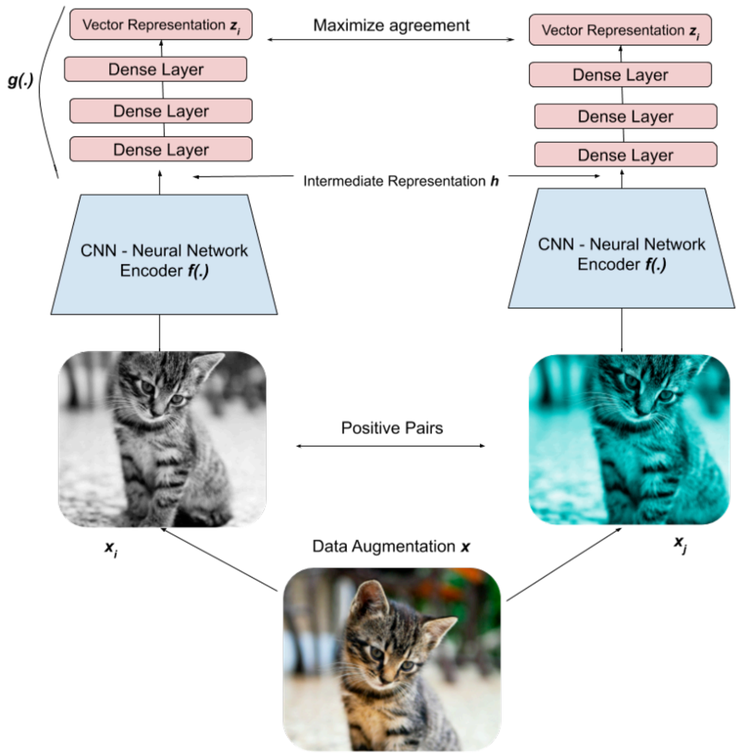
Contrastive Learning Eases the Anisotropy Problem in MLMs

(Gao et al., EMNLP 2021) SimCSE: Simple Contrastive Learning of Sentence Embeddings

Contrastive learning of visual representations

SimCLR (Chen et al., 2020), MoCo (He et al., 2020), SwAV (Caron et al., 2020) and many others

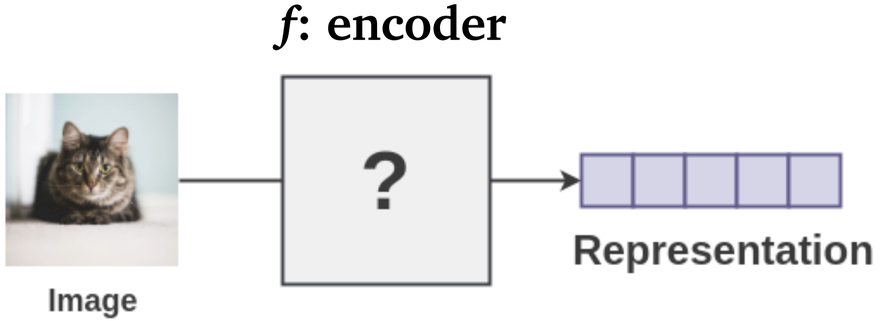




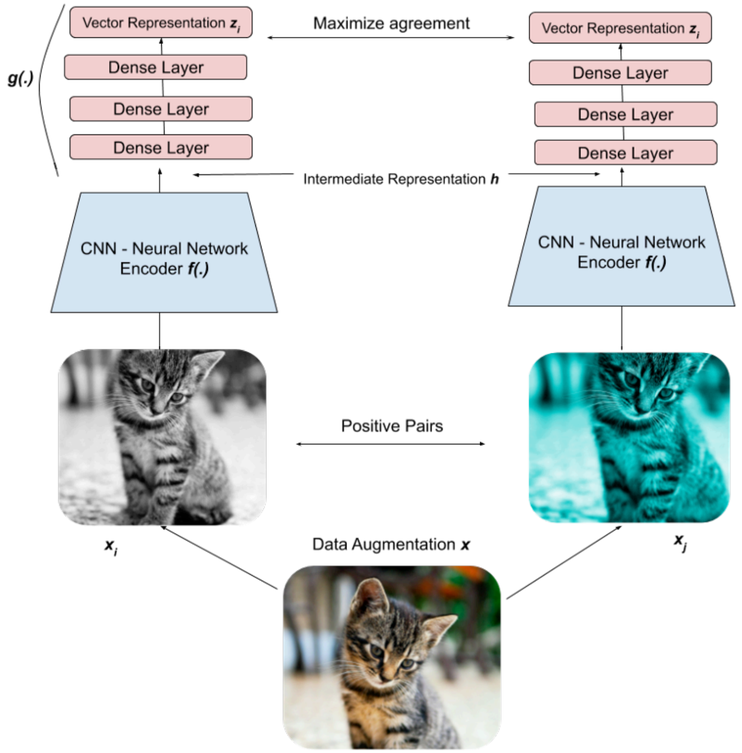
(Chen et al., 2020)

Contrastive learning of visual representations

SimCLR (Chen et al., 2020), MoCo (He et al., 2020), SwAV (Caron et al., 2020) and many others



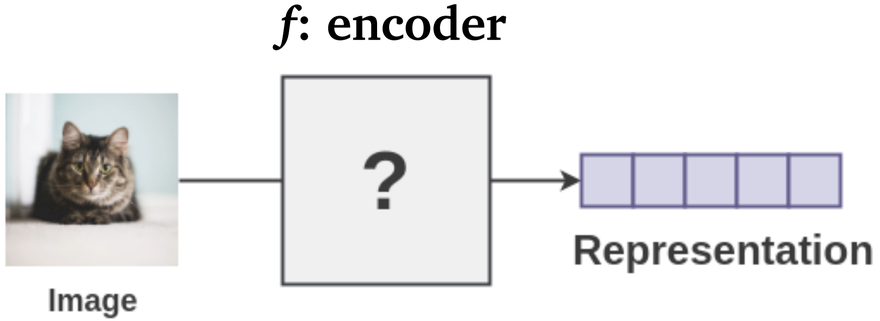
Our goal: learning good sentence representations

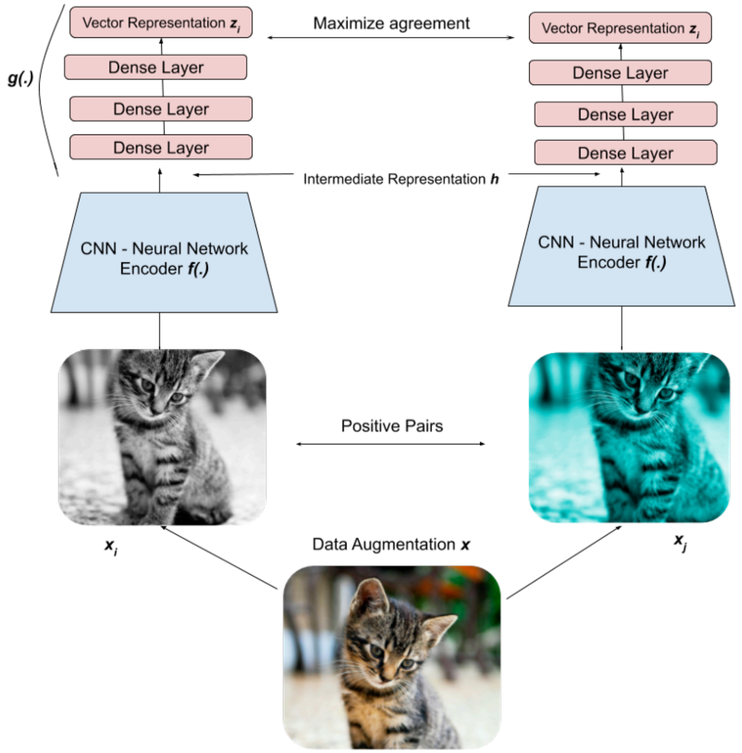


(Chen et al., 2020)

Contrastive learning of visual representations

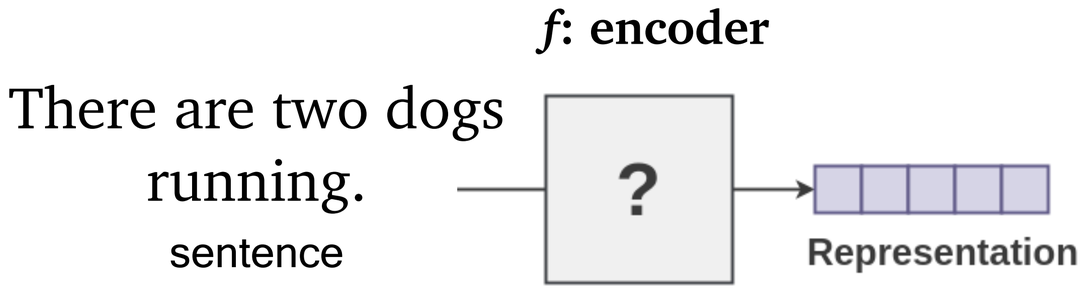
SimCLR (Chen et al., 2020), MoCo (He et al., 2020), SwAV (Caron et al., 2020) and many others





(Chen et al., 2020)

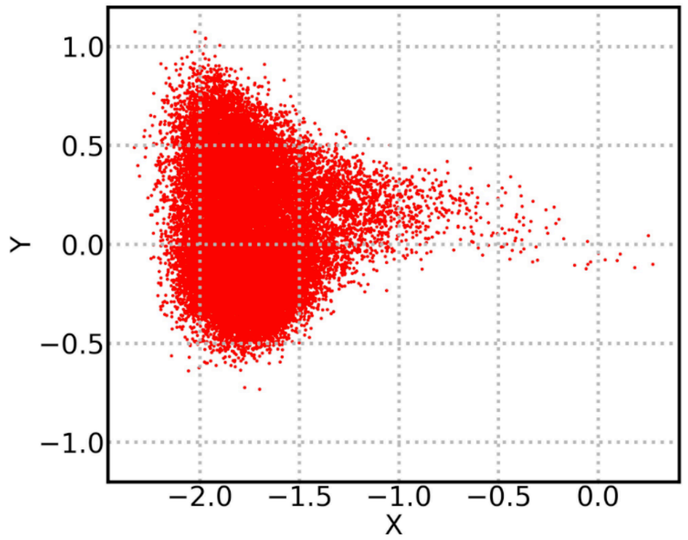
Our goal: learning good sentence representations



MLM representations are anisotropic

MLM representations are anisotropic

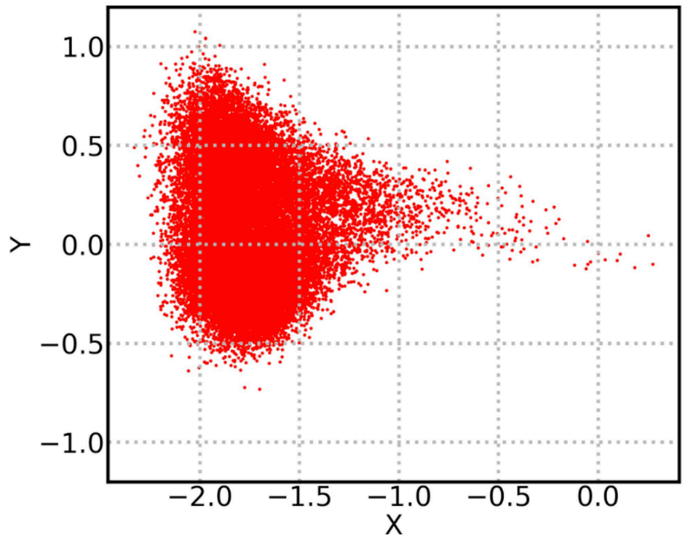
Issue: pre-trained representations are **highly anisotropic** (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020)



(Gao et al., 2019)

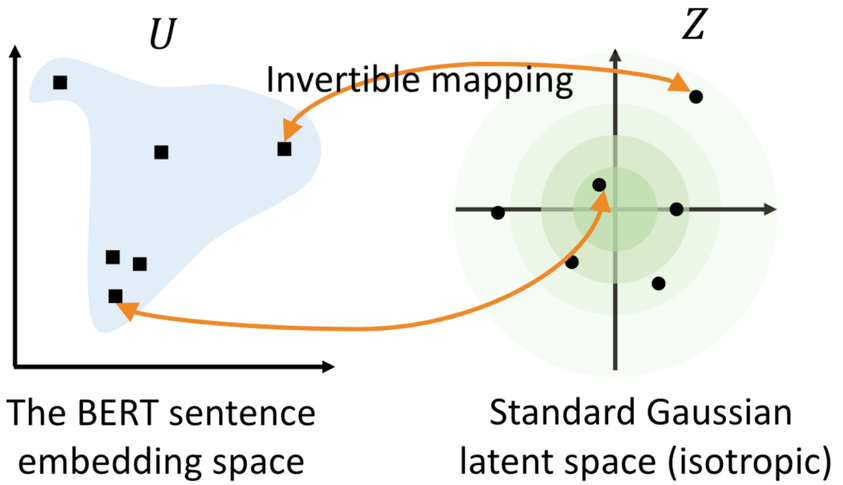
MLM representations are anisotropic

Issue: pre-trained representations are **highly anisotropic** (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020)



(Gao et al., 2019)

Solution: post-processing and mapping embeddings to an isotropic space



BERT-flow (Li et al., 2020)

BERT-whitening (Su et al., 2021)

Our approach: SimCSE

Our approach: SimCSE

A simple contrastive learning framework for sentence representations:

- **Unsupervised SimCSE: only uses *dropout* as data augmentation**
- See supervised SimCSE in the paper!

Our approach: SimCSE

A simple contrastive learning framework for sentence representations:

- **Unsupervised SimCSE: only uses *dropout* as data augmentation**
- See supervised SimCSE in the paper!

InfoNCE loss (van den Oord et al. 2018)

N: batch size
(in-batch negatives)

$$\mathcal{L}_N = -\mathbb{E}_X \left(\log \frac{\exp(\text{sim}(f(x), f(x^+)))}{\exp(\text{sim}(f(x), f(x^+))) + \sum_{j=1}^{N-1} \exp(\text{sim}(f(x), f(x_j)))} \right)$$

x: a sentence, $f(\cdot)$: BERT encoder "[CLS]" + fine-tuning

Unsupervised SimCSE

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

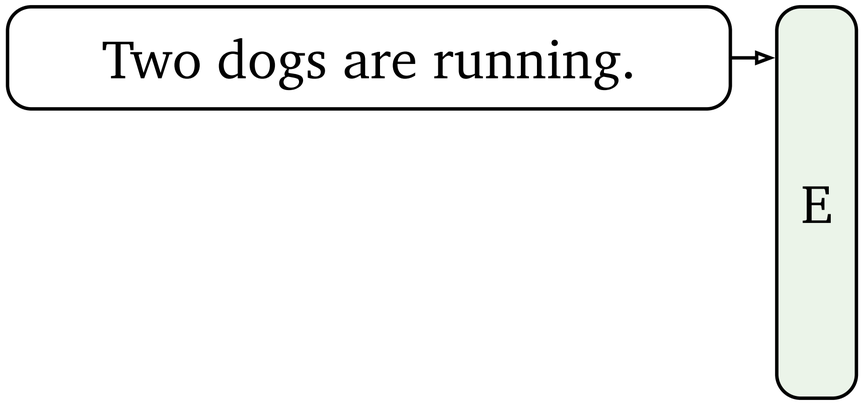
Negative pairs: embeddings of other sentences (in-batch negatives)

Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Unsupervised SimCSE

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

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

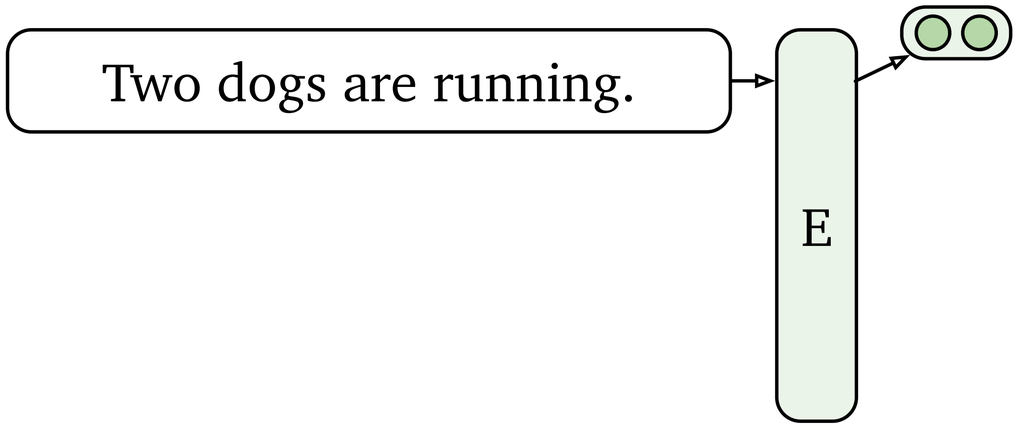


Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Unsupervised SimCSE

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

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

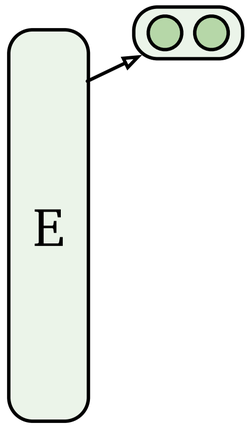


Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Unsupervised SimCSE

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

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

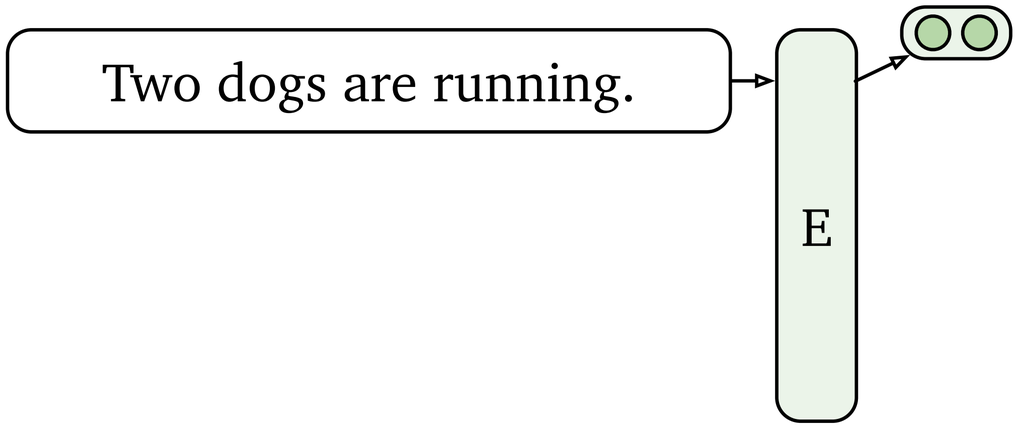


Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Unsupervised SimCSE

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

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

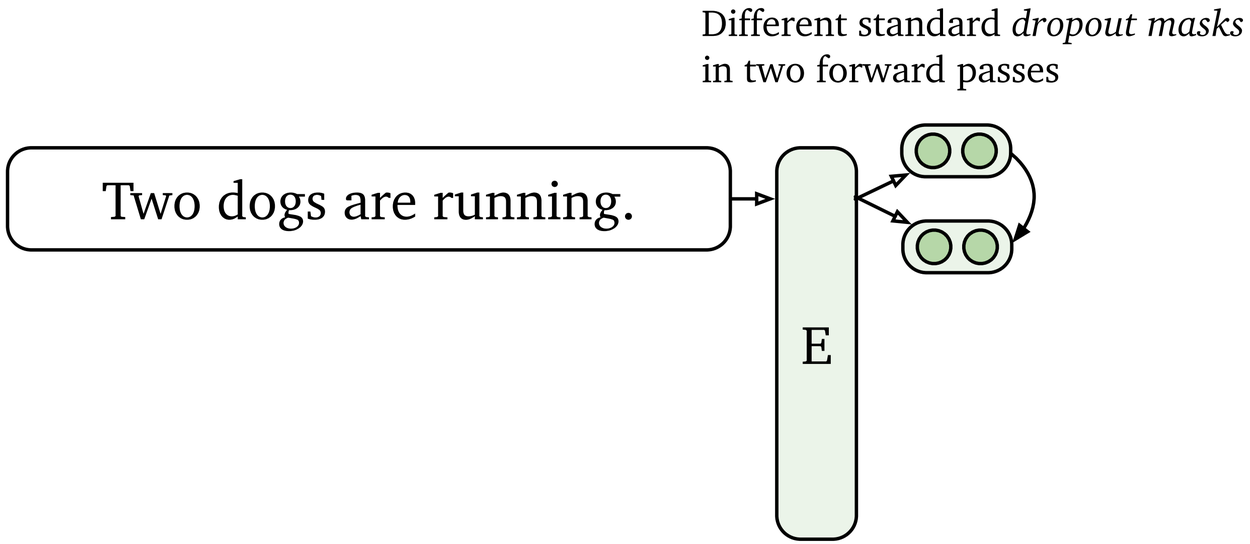


Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Unsupervised SimCSE

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

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

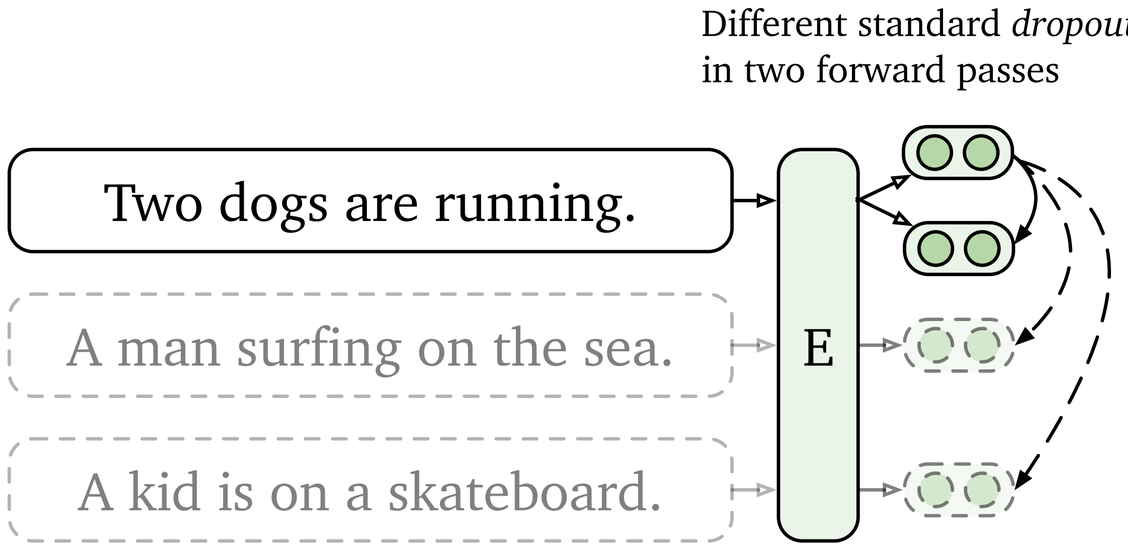


Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Unsupervised SimCSE

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

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

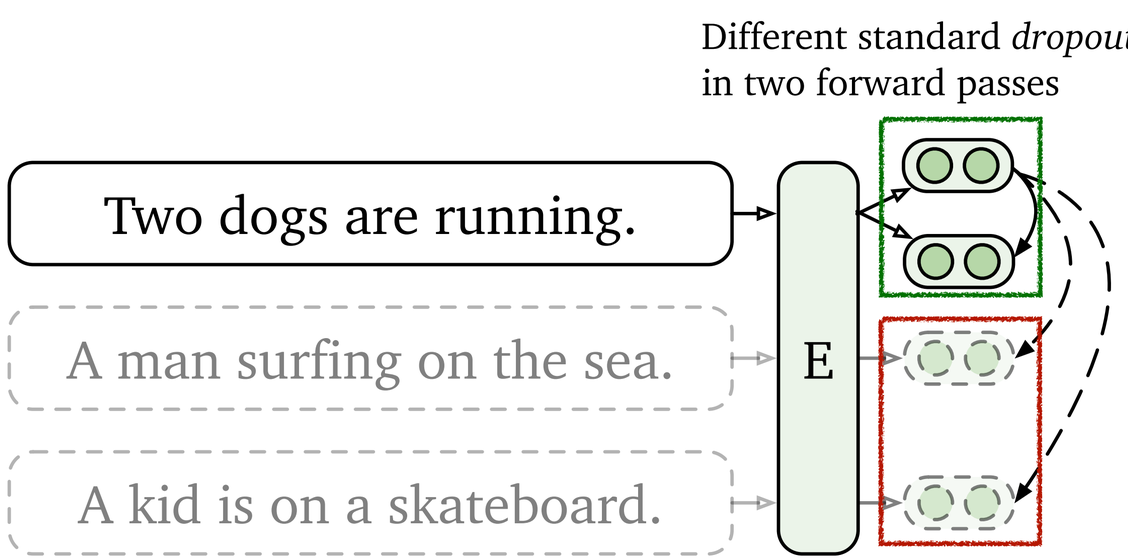


Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Unsupervised SimCSE

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

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



Same sentence with different dropout

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

sim: cosine similarity

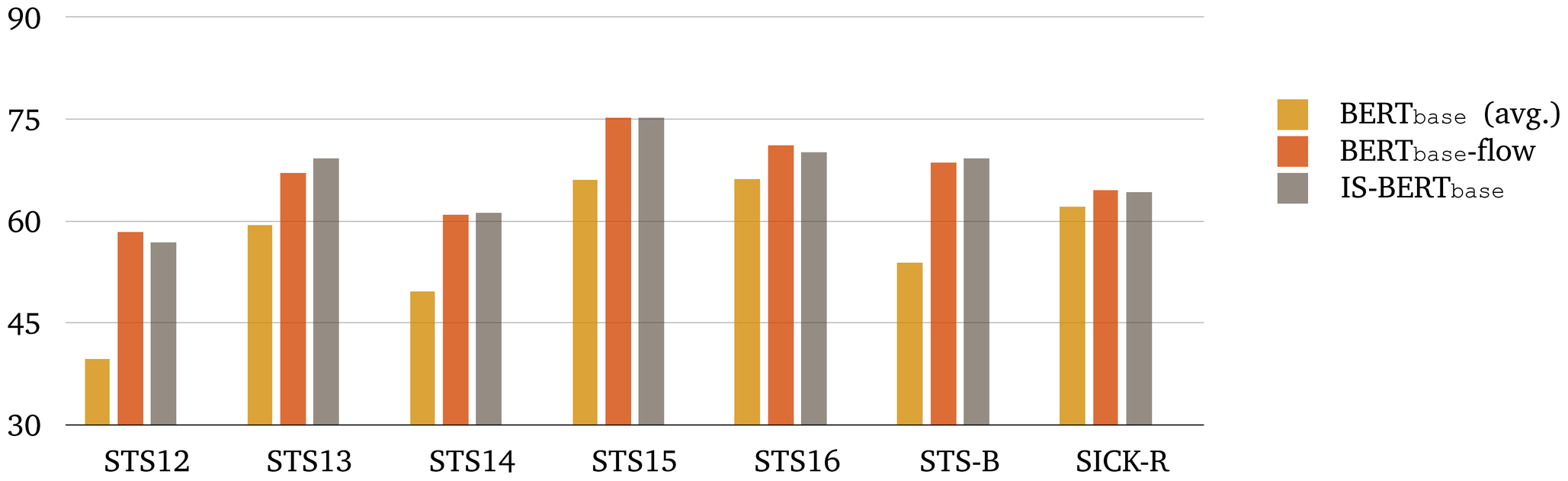
τ: temperature term

In-batch negatives

Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Evaluation on STS tasks

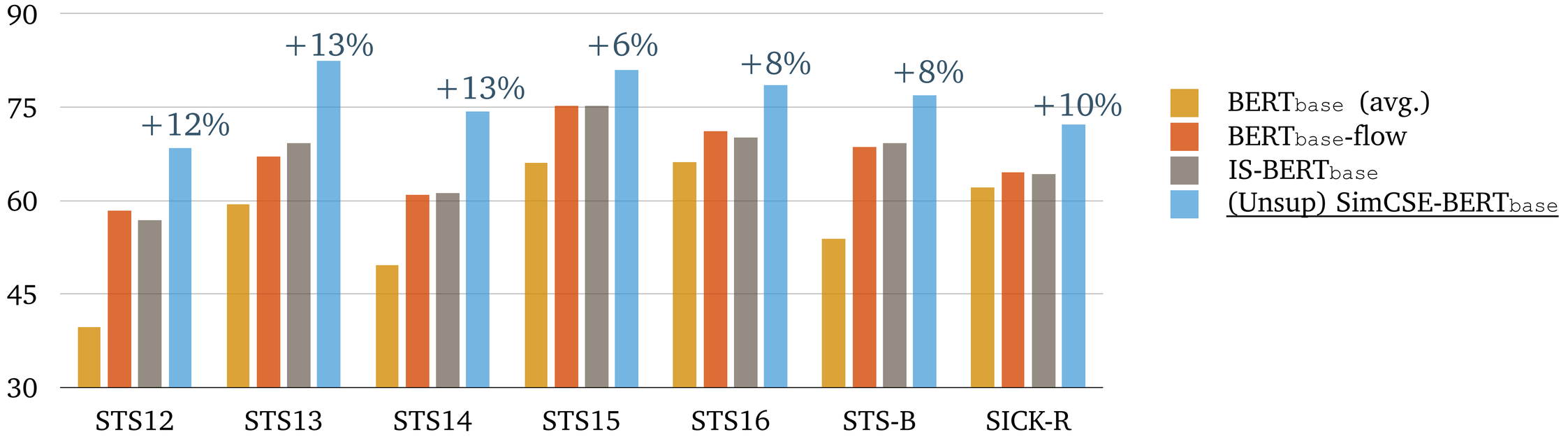
Semantic textual similarity (STS) tasks: Spearman's correlation



Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Evaluation on STS tasks

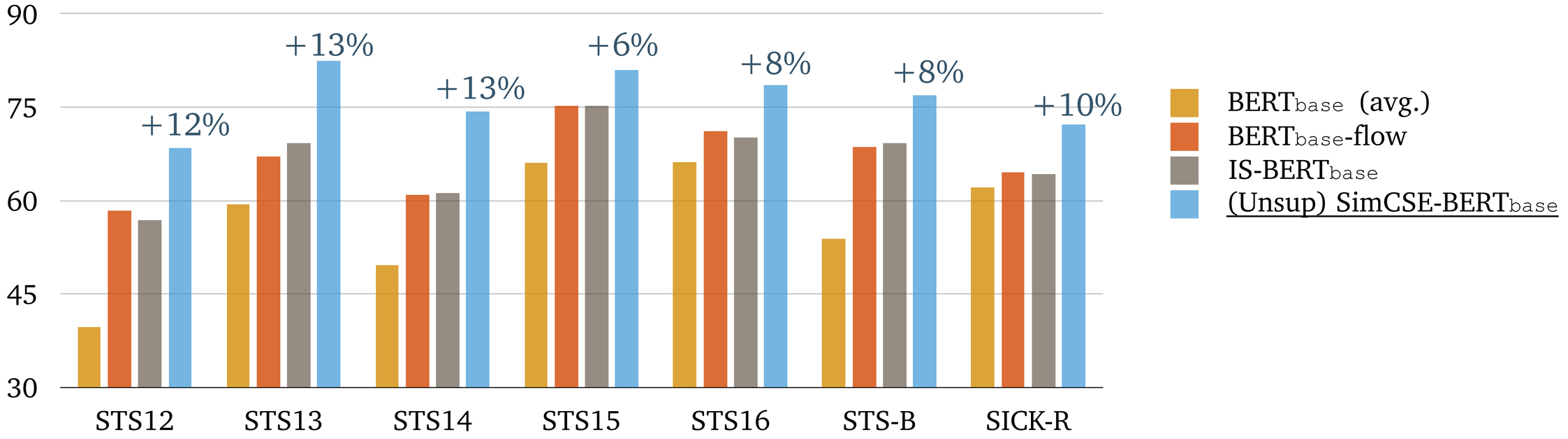
Semantic textual similarity (STS) tasks: Spearman's correlation



Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Evaluation on STS tasks

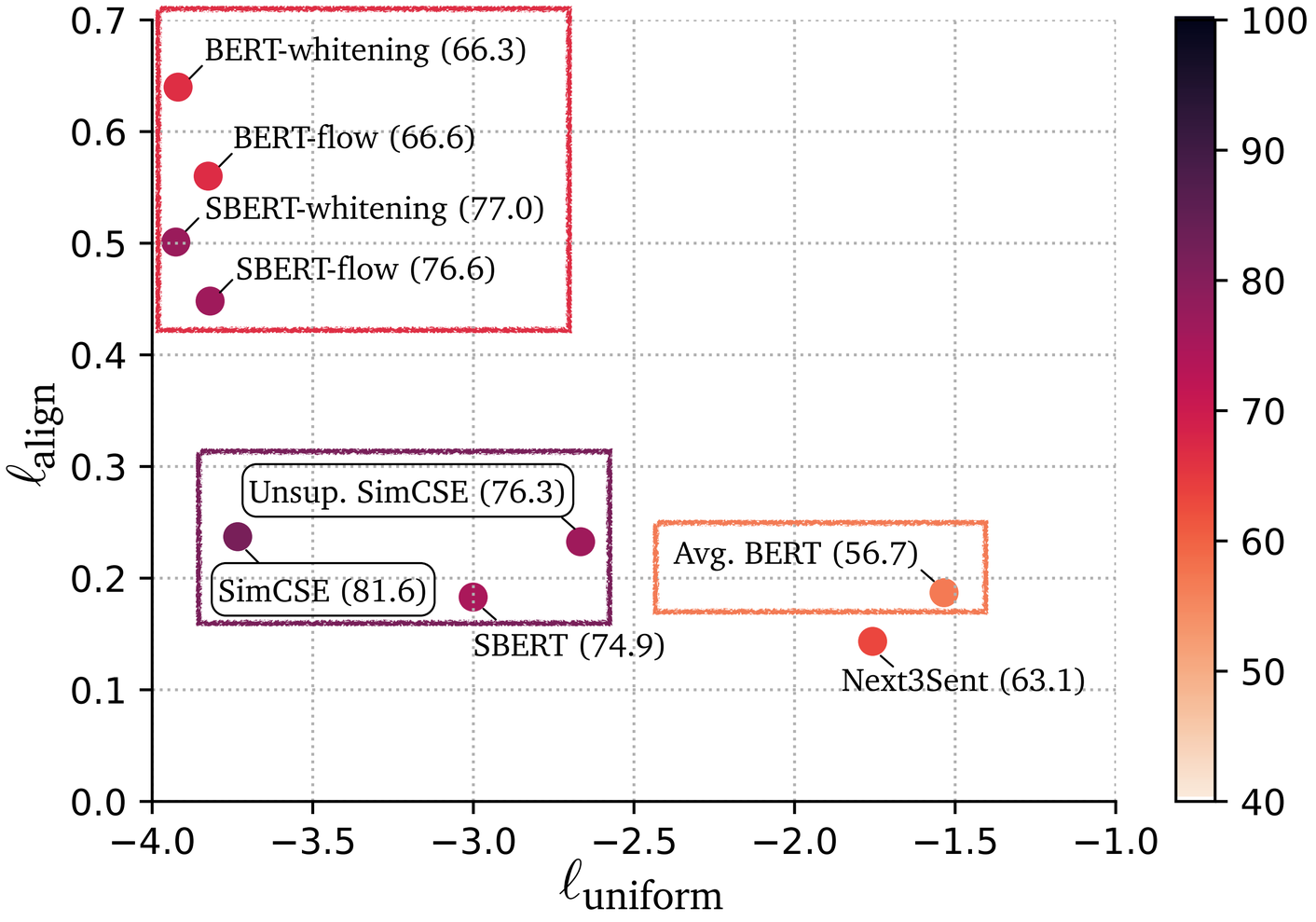
Semantic textual similarity (STS) tasks: Spearman's correlation



Note: SimCSE is very cheap (batch size = 64, 1 hour on 1 GPU)

Gao et al., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings

Alignment vs uniformity



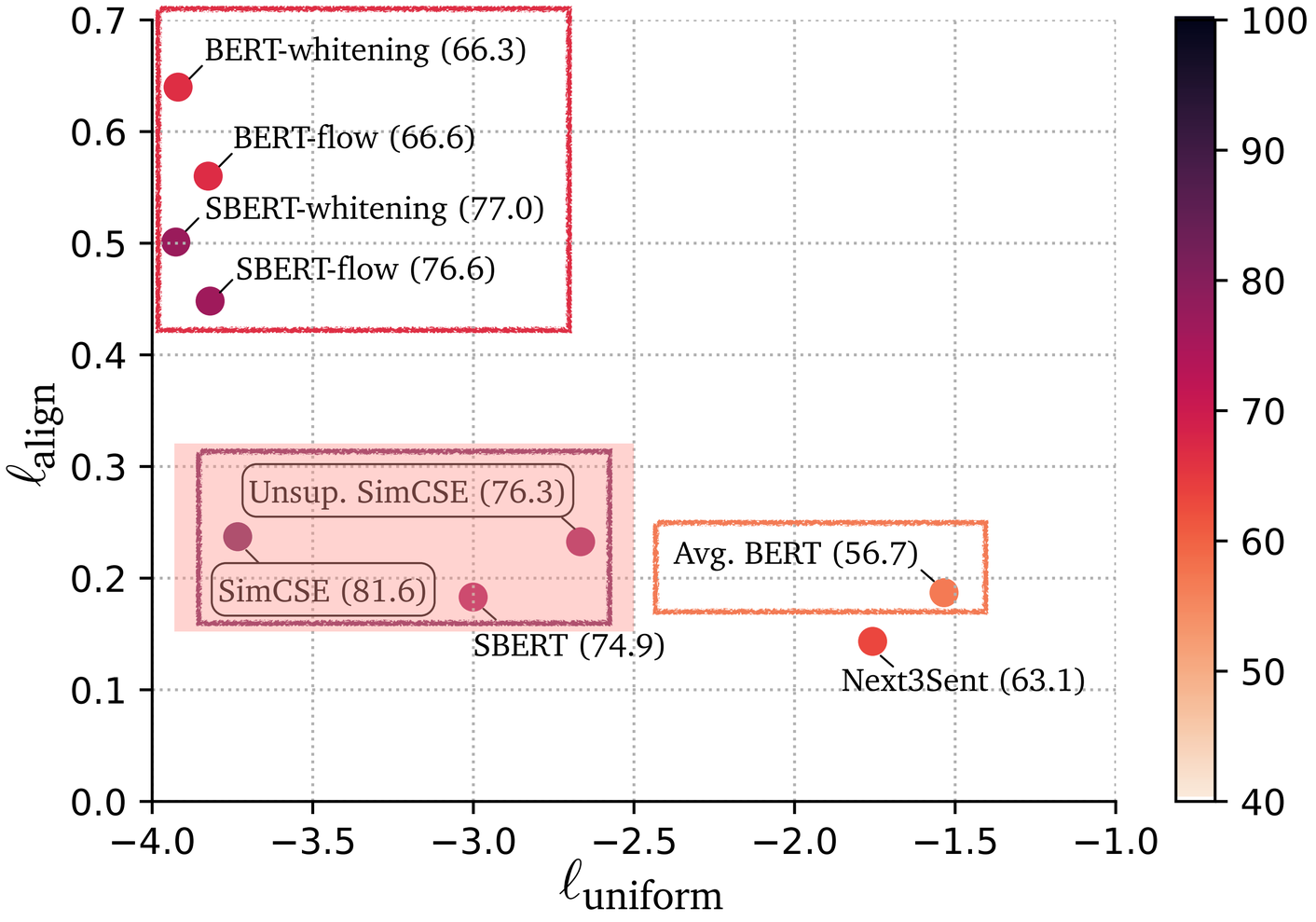
Alignment = how well positive pairs are aligned

Uniformity = how well the embeddings are uniformly distributed

(Wang and Isola, 2020)

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

Alignment vs uniformity



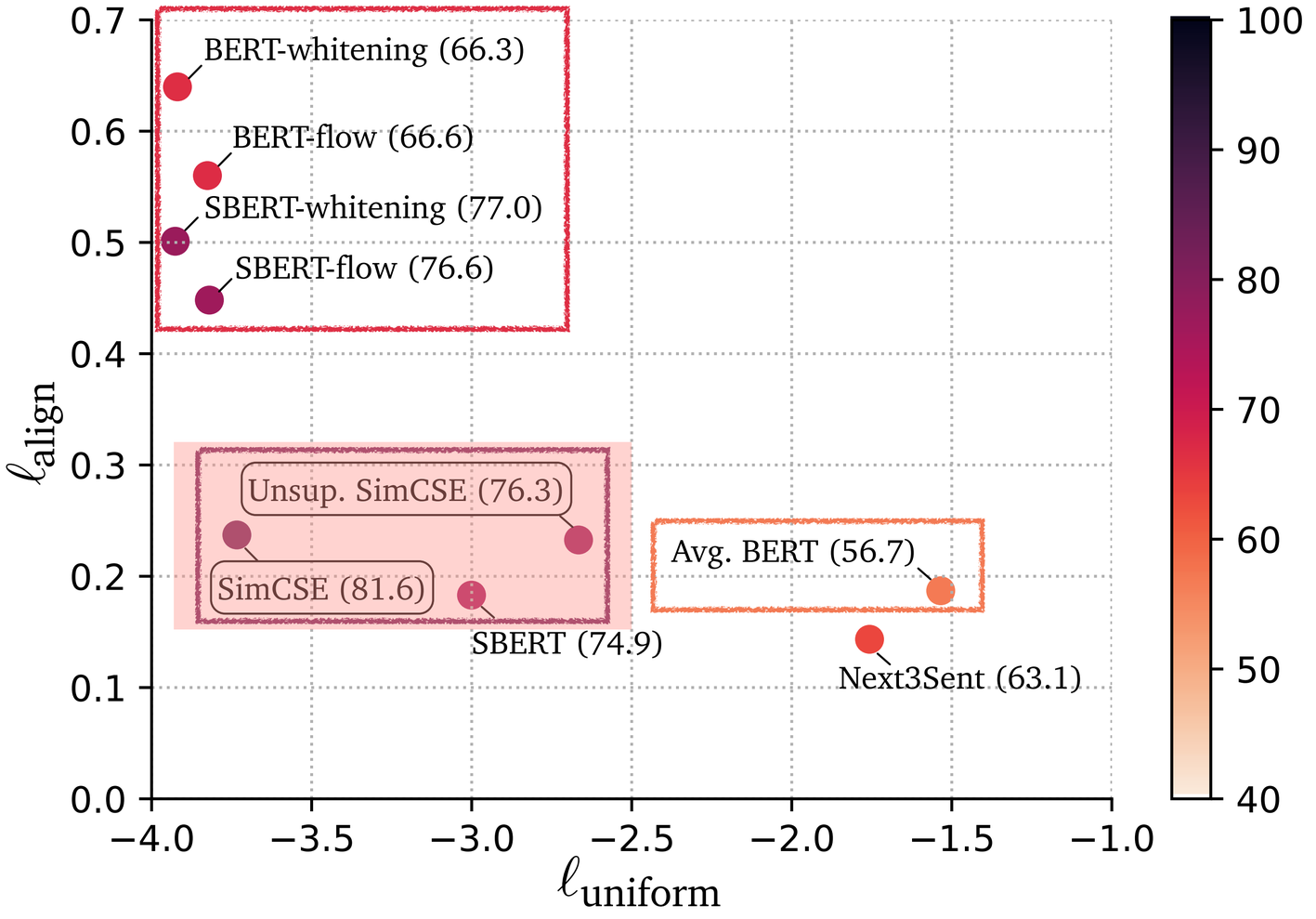
Alignment = how well positive pairs are aligned

Uniformity = how well the embeddings are uniformly distributed

(Wang and Isola, 2020)

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

Alignment vs uniformity



Alignment = how well positive pairs are aligned

Uniformity = how well the embeddings are uniformly distributed

(Wang and Isola, 2020)

We also theoretically show that contrastive objective can improve the isotropy by inherently flattening the singular value distribution of the embedding space (see the paper).

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

What can we learn from this?

What can we learn from this?

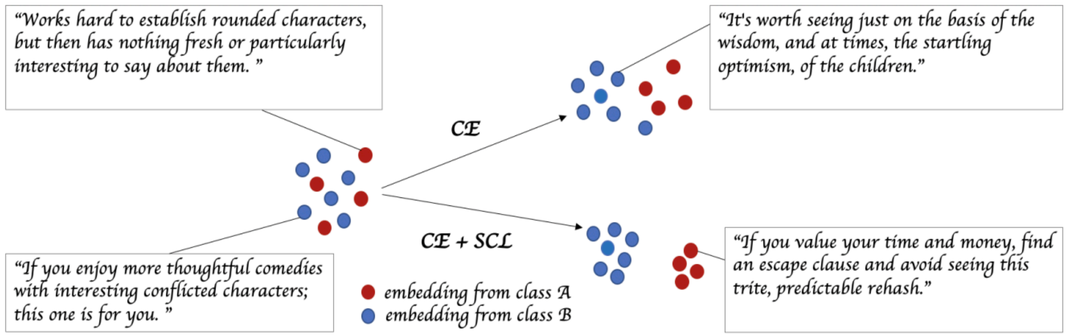
- Pre-trained BERT representations are anistropic and contrastive learning can ease this problem.

What can we learn from this?

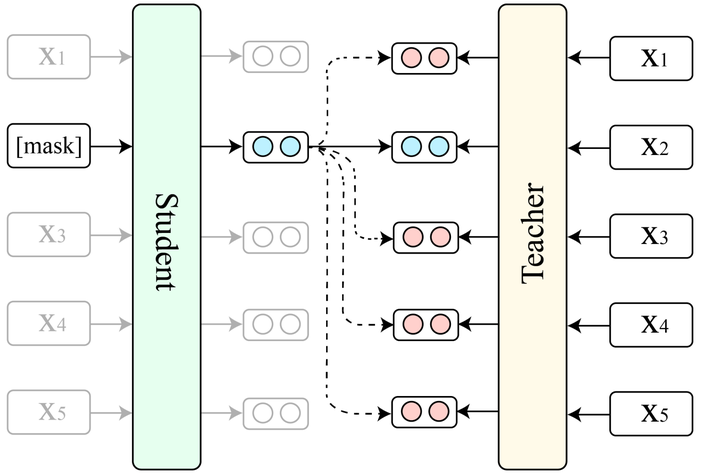
- Pre-trained BERT representations are anistropic and contrastive learning can ease this problem.
- Next question: Shall we fix the problem at pre-training or fine-tuning stage?

What can we learn from this?

- Pre-trained BERT representations are anistropic and contrastive learning can ease this problem.
- Next question: Shall we fix the problem at pre-training or fine-tuning stage?



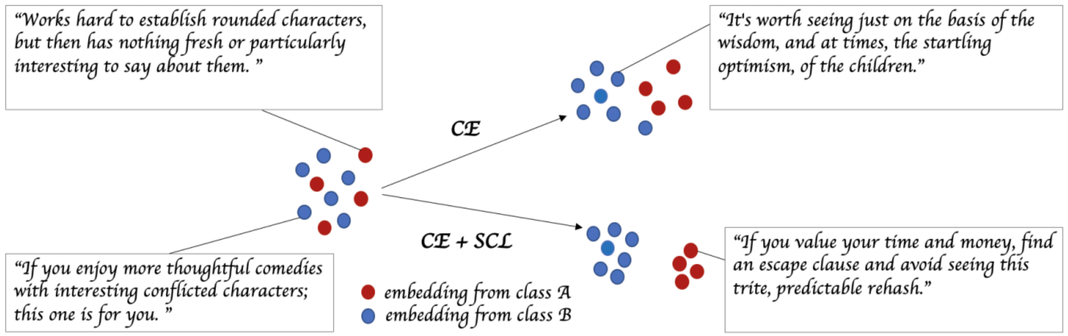
(Gunel et al., 2021): contrastive fine-tuning



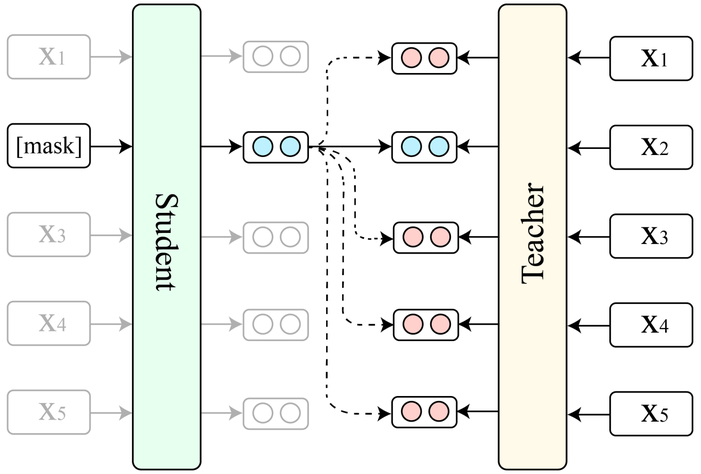
(Su et al., 2021):
MLM + contrastive pre-training

What can we learn from this?

- Pre-trained BERT representations are anistropic and contrastive learning can ease this problem.
- Next question: Shall we fix the problem at pre-training or fine-tuning stage?



(Gunel et al., 2021): contrastive fine-tuning



(Su et al., 2021):
MLM + contrastive pre-training

Q: Can we do contrastive pre-training from the scratch?
Would that improve the efficiency of pre-training?

Summary

How can we make pre-training
more **accessible** and **affordable**?

Summary

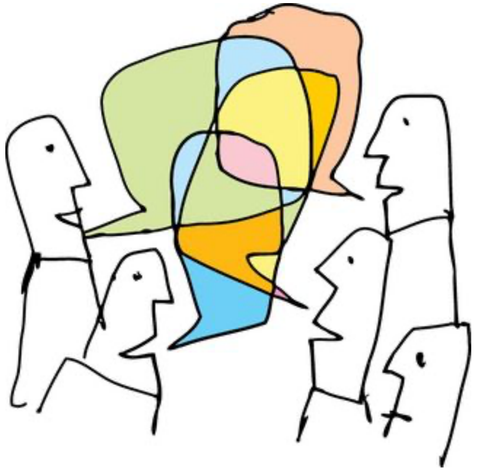
How can we make pre-training
more **accessible** and **affordable**?

- Use **high masking rates** + efficient implementation
- **Contrastive** pre-training

Summary

How can we make pre-training
more **accessible** and **affordable**?

- Use **high masking rates** + efficient implementation
- **Contrastive** pre-training



Thanks!

danqic@cs.princeton.edu