Contrastive learning

Giacomo Camposampiero

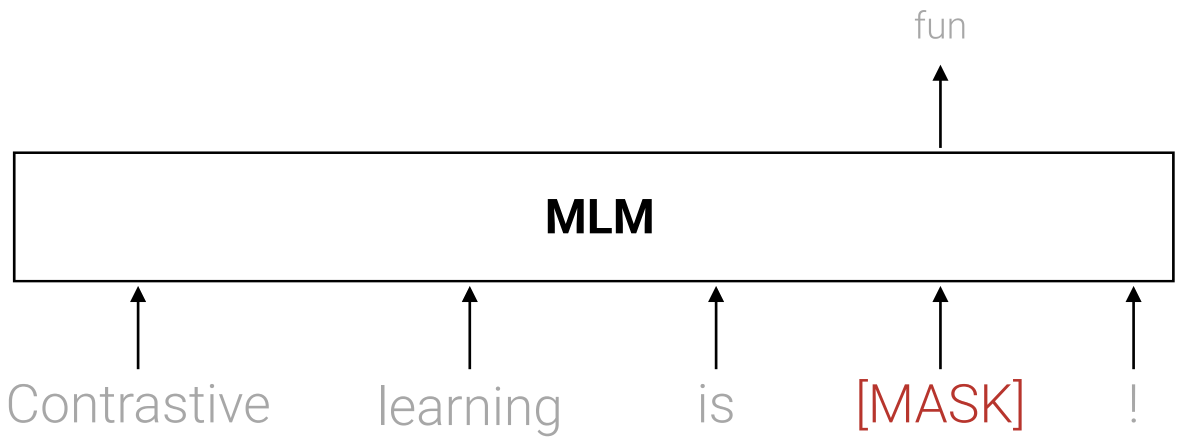
Seminar in Deep Neural Networks, 03.05.2022

Introduction

Self-supervised Learning

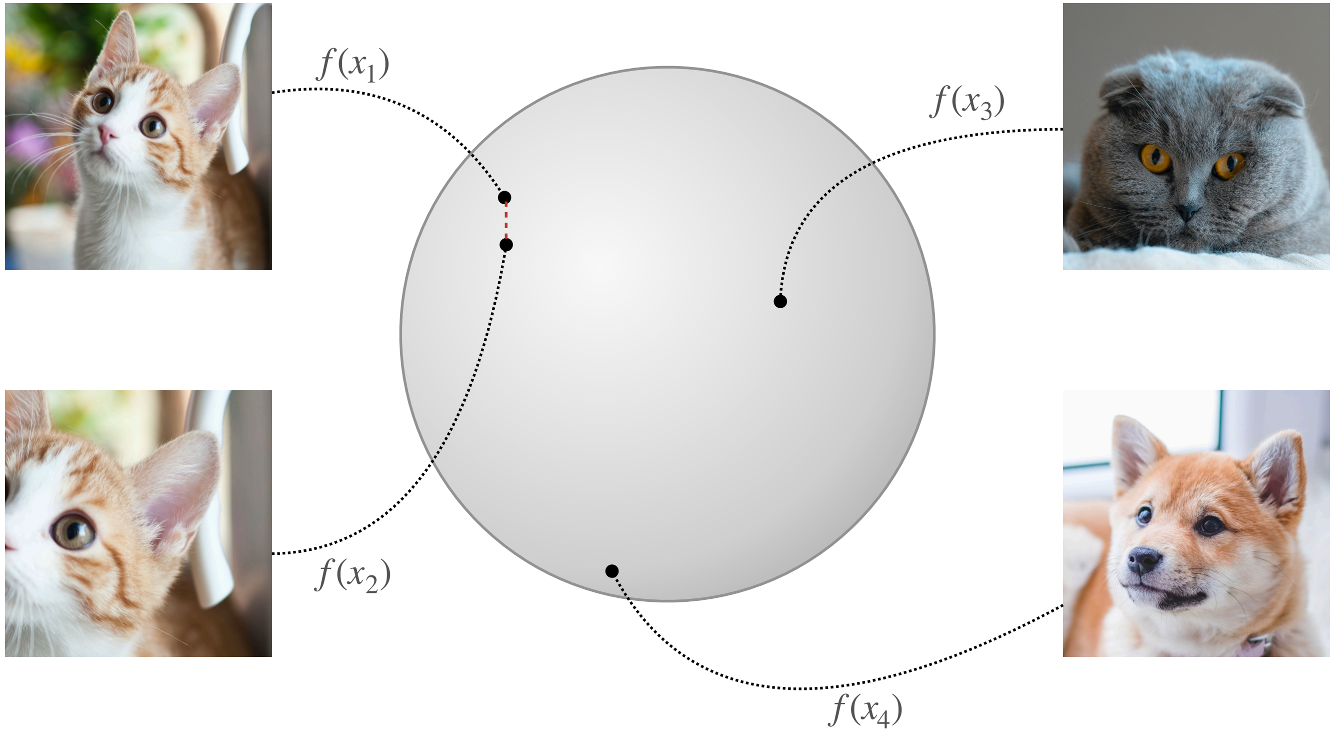
Popular approach for representation learning, middle-ground between supervised and unsupervised

- unlabeled dataset (can leverage huge datasets)
- supervised training task (dummy task)



Contrastive Learning

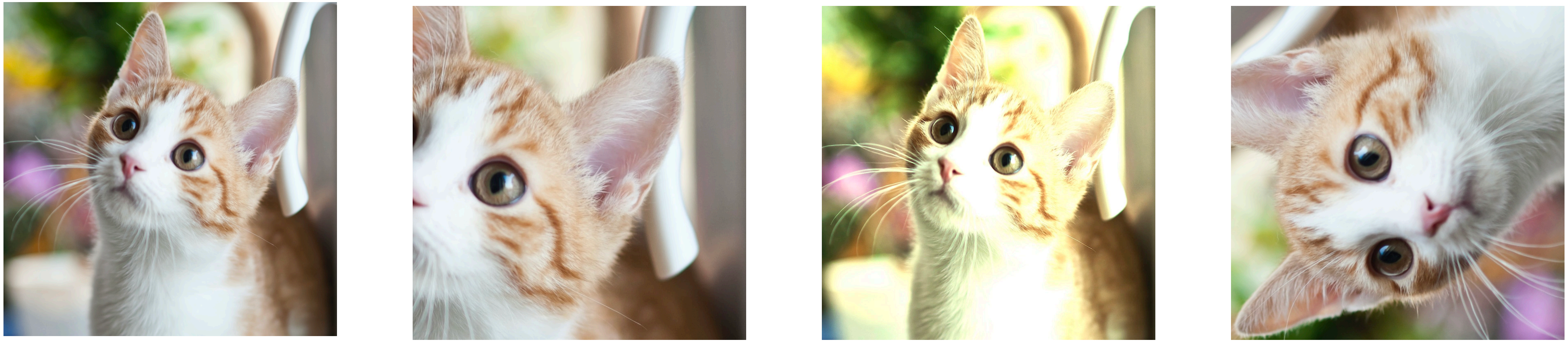
In the latent space, positive pairs stay together while negative are pushed away



⚠️ in the self-supervised setting, positive pairs are augmentations of the same sample

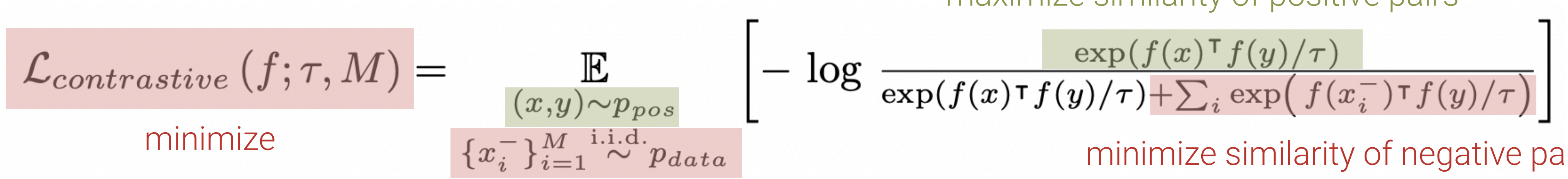
Intuition behind Contrastive Learning

The encoder learns to extract most shared informations between positive pairs (preserving semantic information), while remaining invariant to other noise factors



Formalization of Contrastive Learning

General contrastive loss defined by (Wang and Isola 2020)



maximize similarity of positive pairs

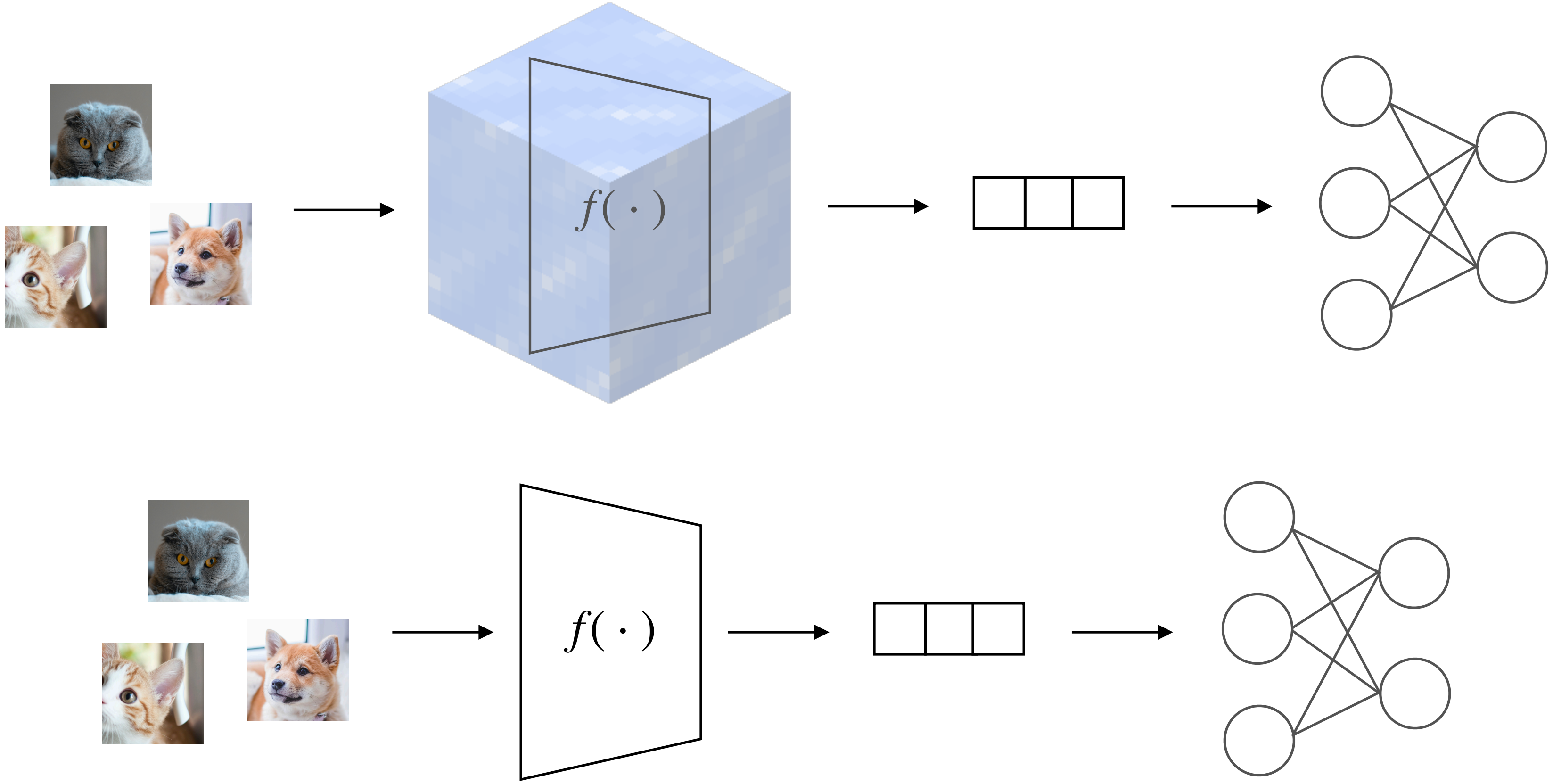
$$\mathcal{L}_{contrastive}(f;\tau,M) = \mathop{\mathbb{E}}_{\substack{(x,y)\sim p_{pos} \\ \{x_i^-\}_{i=1}^{M} \overset{\text{i.i.d.}}{\sim} p_{data}}} \left[-\log \frac{\exp(f(x)^\intercal f(y)/\tau)}{\exp(f(x)^\intercal f(y)/\tau) + \sum_i \exp\left(f(x_i^-)^\intercal f(y)/\tau \right)} \right]$$

minimize

minimize similarity of negative pairs

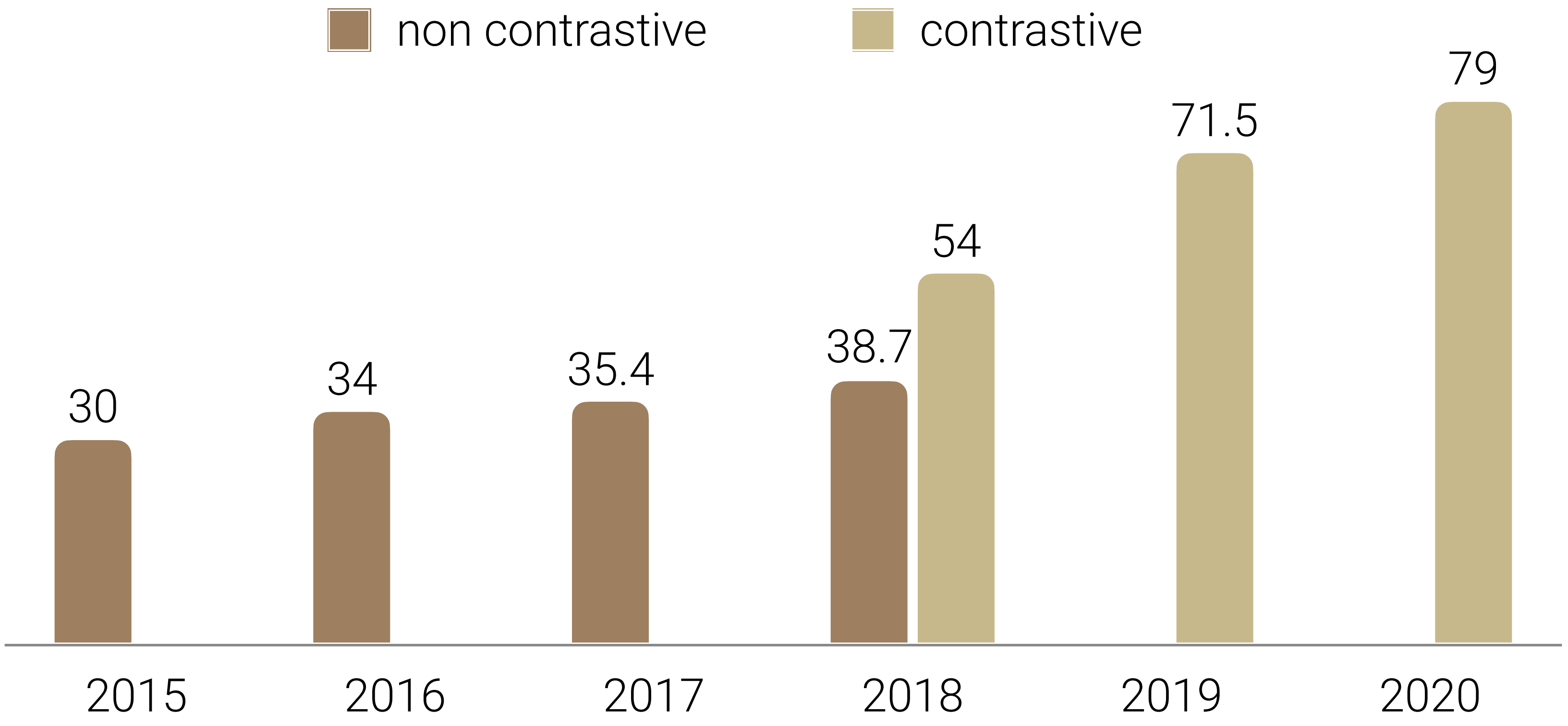
How is it used?

Contrastive learning is used primarily to learn representations



Does it work?

Turns out that yes, it works!



Best-scoring self-supervised approaches on the ImageNet Linear Benchmark.
Courtesy of Yonglong Tian, Contrastive Learning: A General Self-supervised Learning Approach

Contrastive learning in Computer Vision

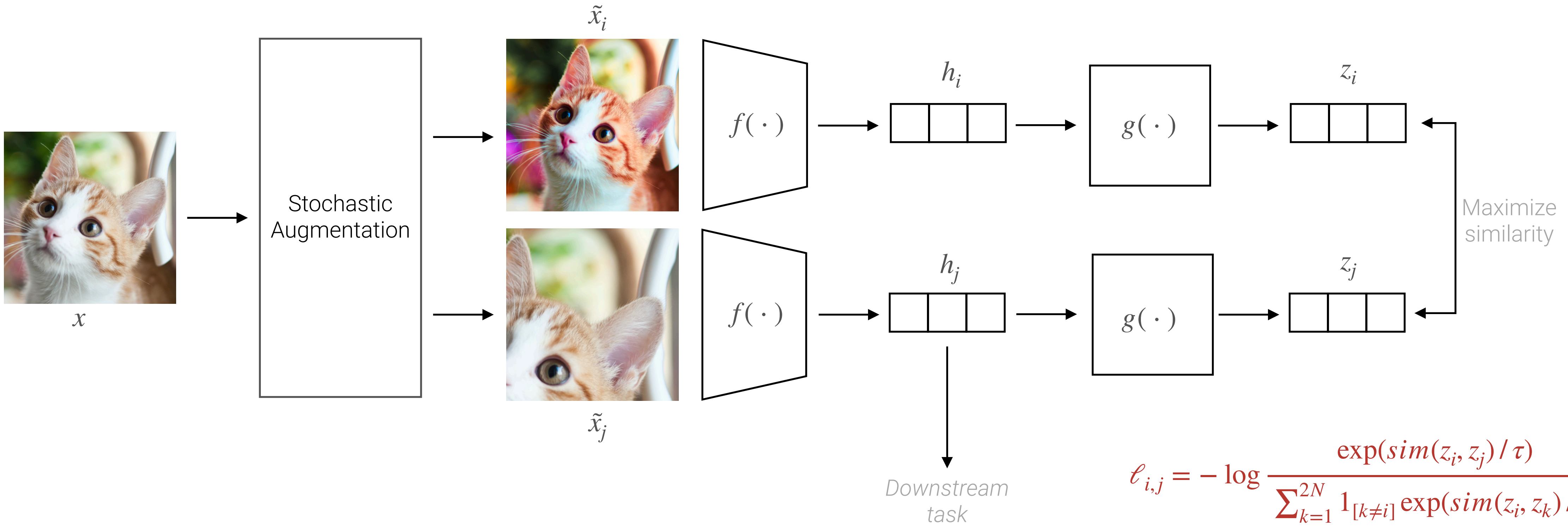
Contrastive Learning originated and flourished in the Computer Vision field

Why? It’s easy to augment samples retaining semantic meaning!

- random cropping
- random color distortions
- random Gaussian blur

SimCLR (Chen et al. 2020)

Simple framework to learn visual representations without human supervision, using parallel augmentation



$$\ell_{i,j} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(sim(z_i, z_k)/\tau)}$$

Introduction

Theoretical Understanding of CL

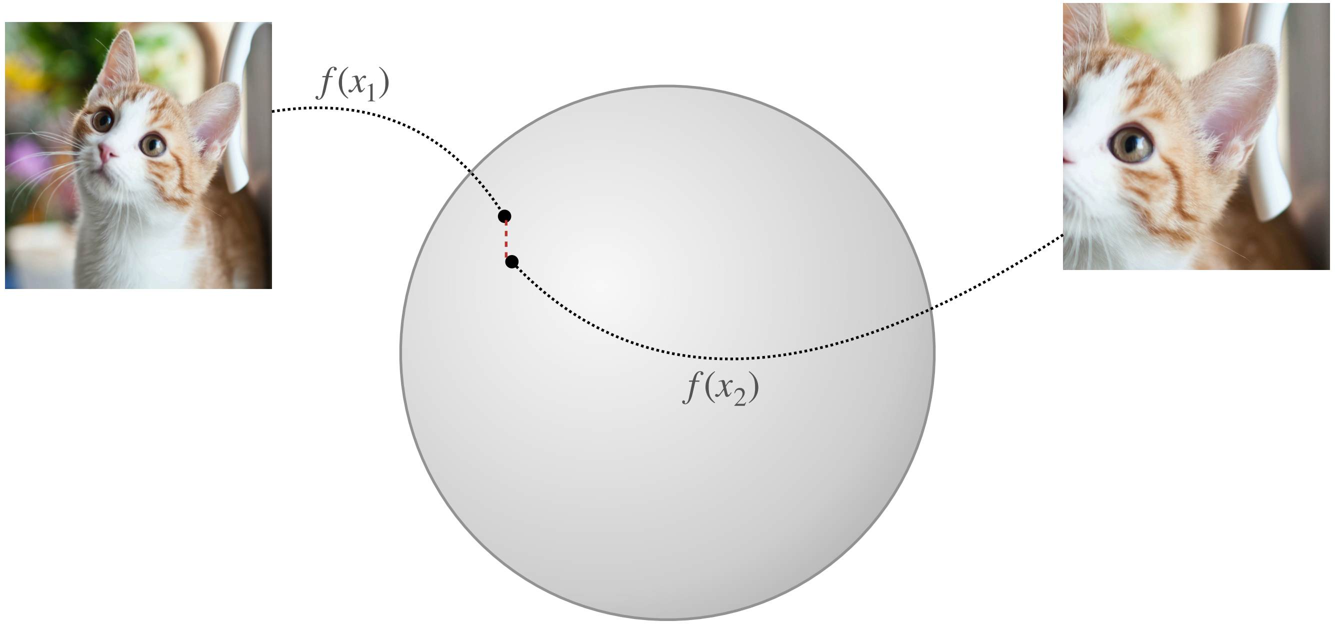
Alignment and Uniformity on the Hypersphere (Wang & Isola 2020)

A recent work from Wang and Isola backed contrastive learning with a theoretical explanation. In their work, they identify two main metrics for evaluating embeddings

- **alignment**
- **uniformity**

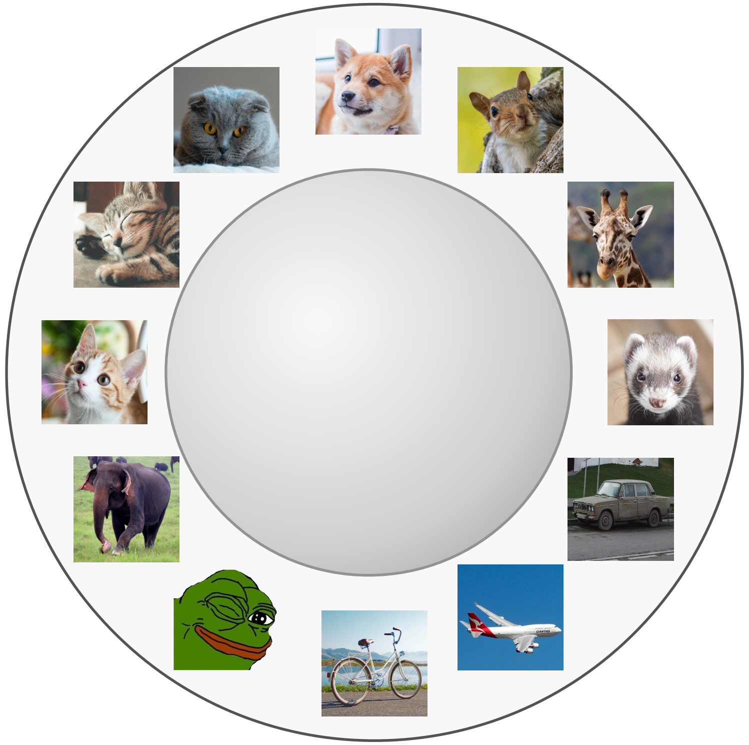
Alignment and Uniformity on the Hypersphere (Wang & Isola 2020)

Alignment, that measures the noise-invariance property



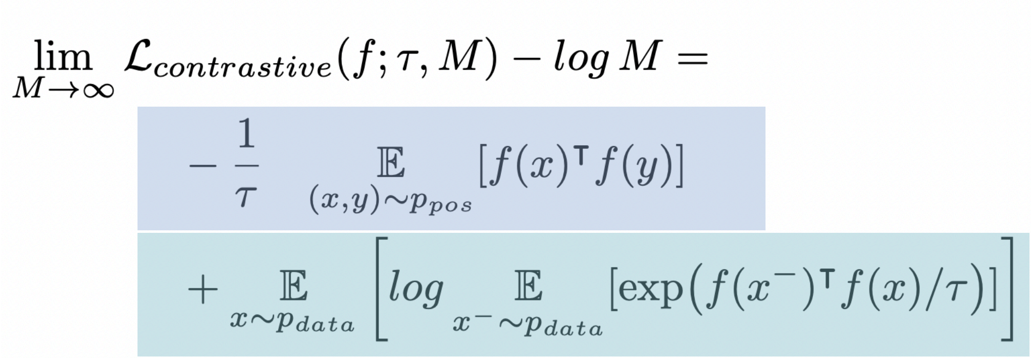
Alignment and Uniformity on the Hypersphere (Wang & Isola 2020)

Uniformity, that measures how uniformly distributed are the feature vectors in the hypersphere



Alignment and Uniformity on the Hypersphere (Wang & Isola 2020)

For fixed $\tau > 0$, as the number of samples $M \to \infty$, the (normalized) contrastive loss converges to

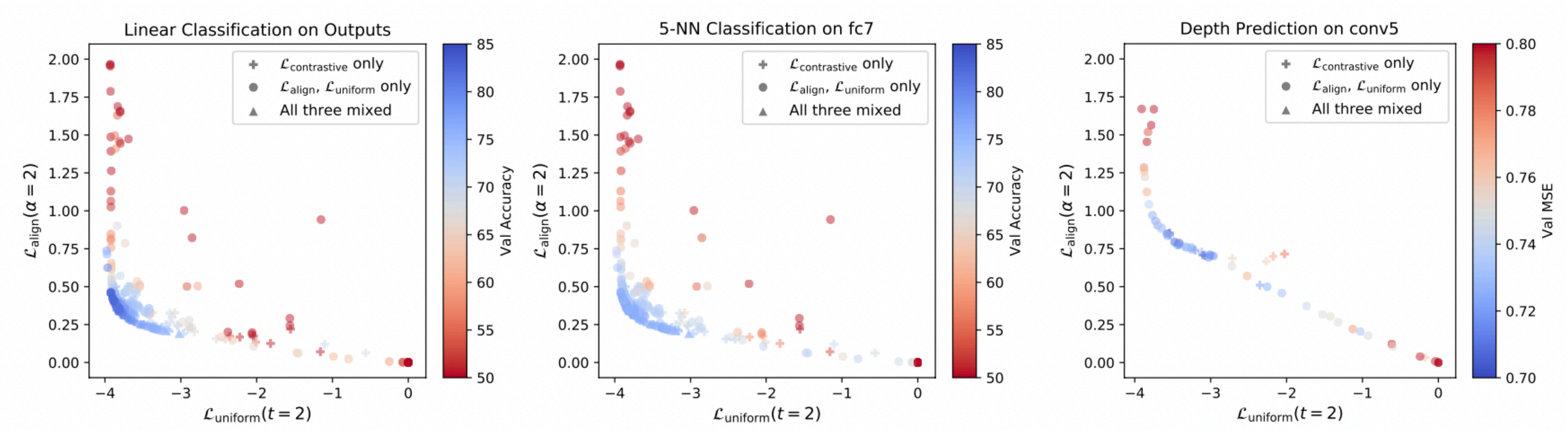


$$\lim_{M\to\infty} \mathcal{L}_{contrastive}(f;\tau,M) - log\, M = -\frac{1}{\tau}\mathop{\mathbb{E}}_{(x,y)\sim p_{pos}}\left[f(x)^\intercal f(y)\right] + \mathop{\mathbb{E}}_{x\sim p_{data}}\left[log \mathop{\mathbb{E}}_{x^-\sim p_{data}}\left[\exp\big(f(x^-)^\intercal f(x)/\tau\big)\right]\right]$$

Alignment and Uniformity on the Hypersphere (Wang & Isola 2020)

The authors also **empirically** verified their claims and found out that

- alignment and uniformity loss strongly agree with downstream task performance



Understanding Contrastive Representation Learning through Alignment and Uniformity on the hypersphere, Wang & Isola 2020

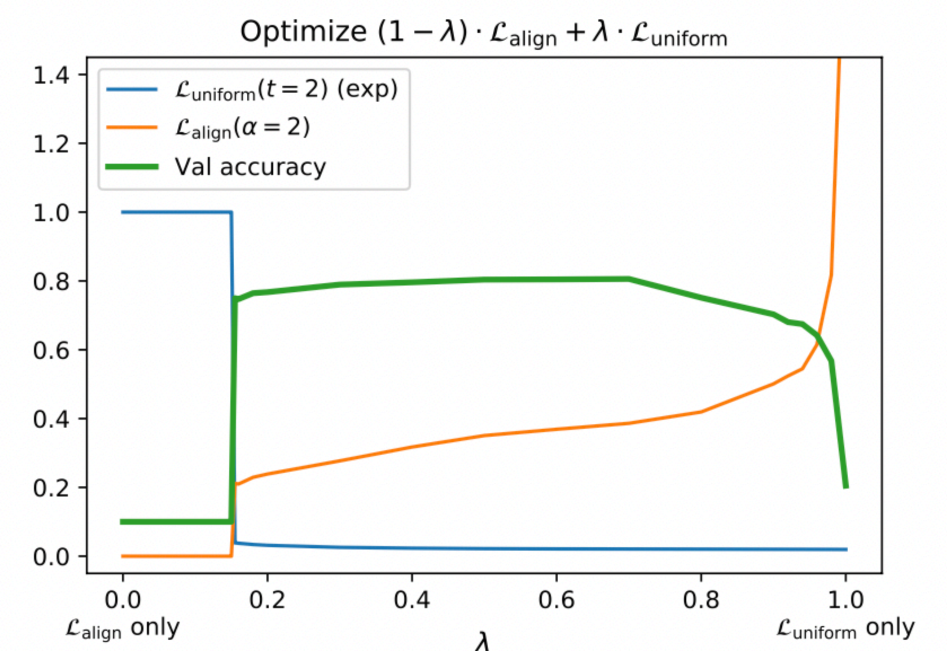
Alignment and Uniformity on the Hypersphere (Wang & Isola 2020)

- alignment and uniformity are valid properties across many representation learning variants

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Alignment and Uniformity on the Hypersphere (Wang & Isola 2020)

- both alignment and uniformity are necessary for good representations



Understanding Contrastive Representation Learning through Alignment and Uniformity on the hypersphere, Wang & Isola 2020

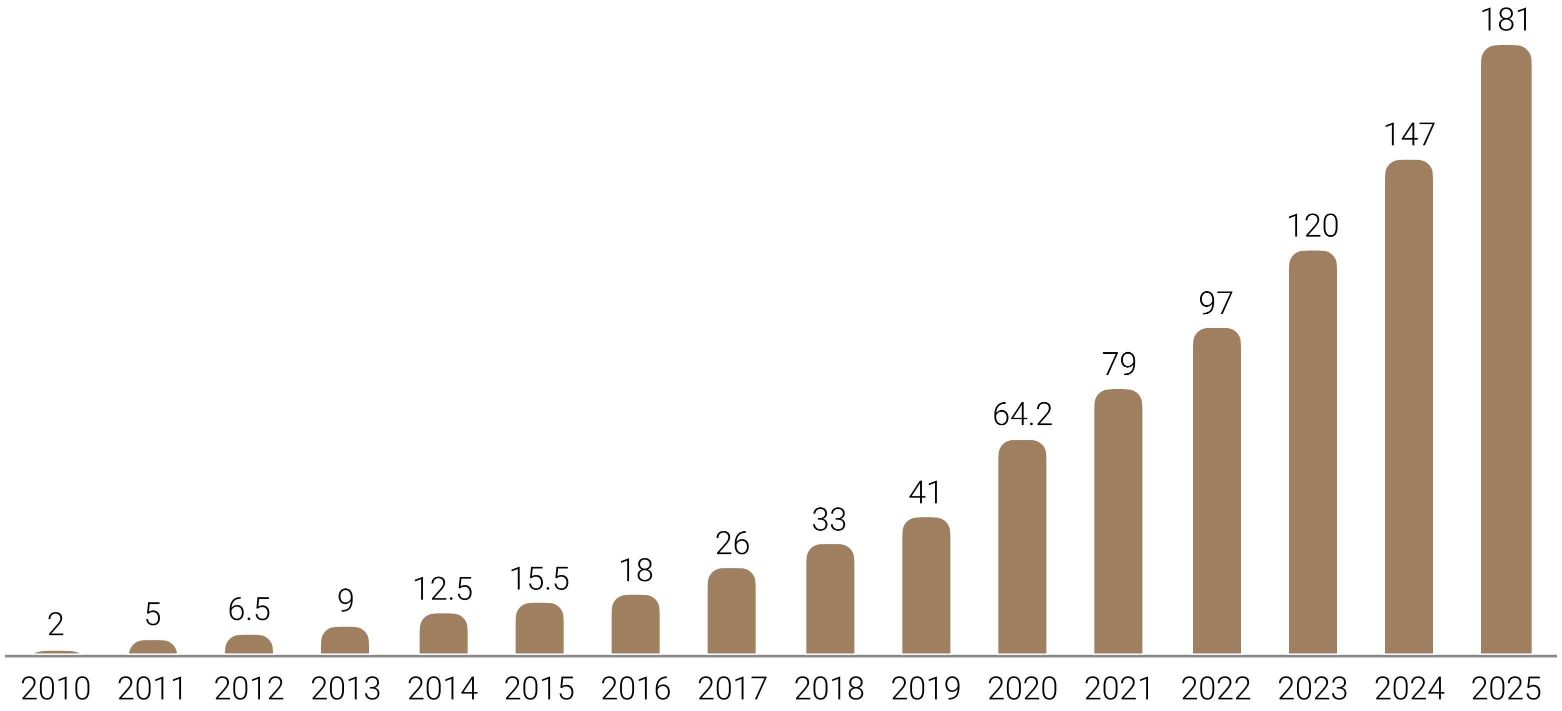
Introduction

Theoretical Understanding of CL

CL in Natural Language Processing

Contrastive Learning in Natural Language Processing

In NLP, self-supervised learning has been around for a while to leverage the huge quantity of unlabeled textual data created by humanity.



Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 [zettabytes, 10^{21}], *Sources*: IDC; Seagate; Statista estimates

Contrastive Learning in Natural Language Processing

As a result, a lot of self-supervised formulations to learn text representations have been developed in the last few decades

- center word prediction
- next sentence prediction
- masked language modeling

What about **Contrastive Learning**?

Contrastive Learning in Natural Language Processing

Still not very established, but recent success of CL in other areas stimulated research about it.

Main problem: data augmentation

Contrastive Learning in Natural Language Processing

Main approaches used for augmentation

- back-translation (Fang et al. 2019)
- lexical edits (Wei and Zhou, 2019)
- cutoff (Shen et al. 2020)
- **dropout**

SimCSE (Gao et al. 2021)

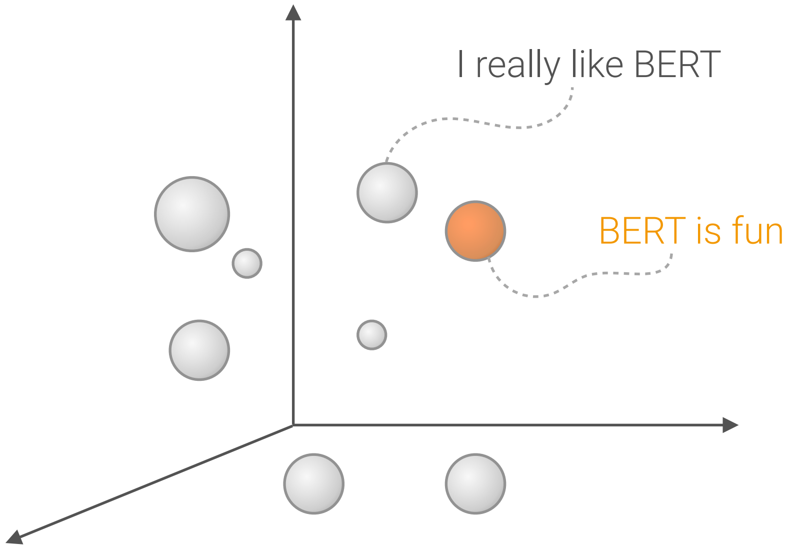
Simple contrastive learning framework to learn sentence embeddings, using **dropout** as augmentation technique during training

$$+ \quad -\log \frac{\exp\left(\mathrm{sim}(h_i^{z_i}, h_i^{z'_i})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}(h_i^{z_i}, h_j^{z'_j})/\tau\right)} \quad = \text{SoTA}$$

SimCSE (Gao et al. 2021)

Why do we need sentence embeddings?

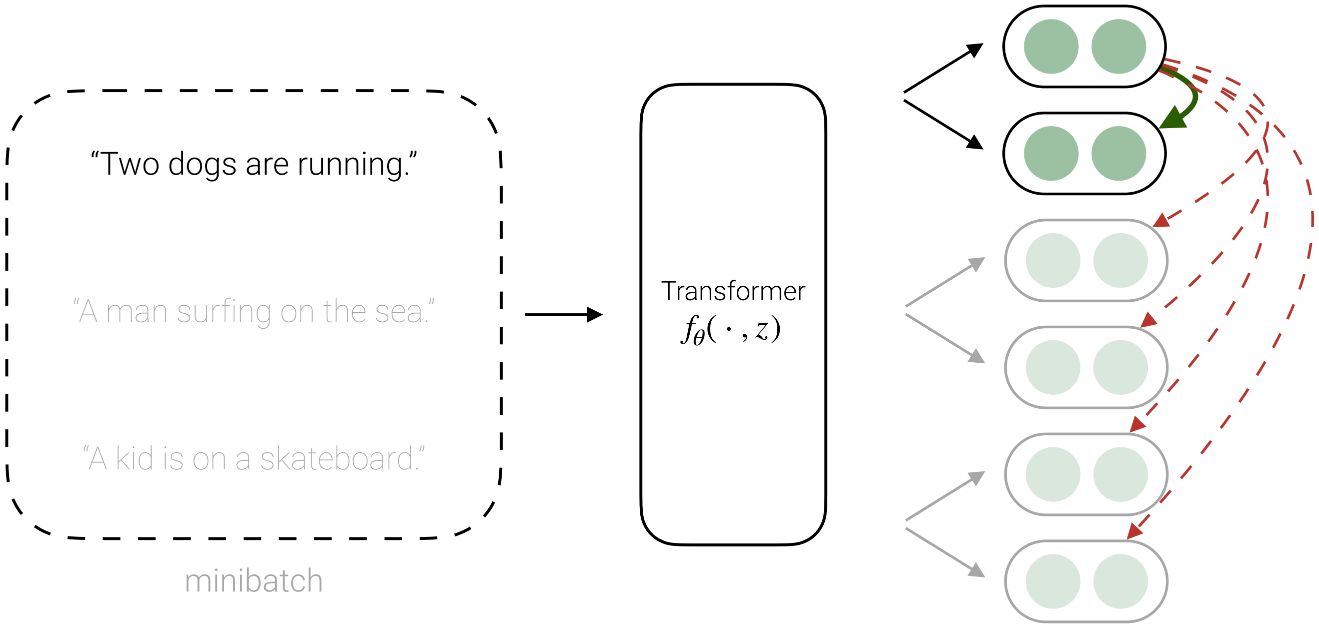
- zero-shot retrieval
- sentence clustering



Gao et al. EMNLP 2021

SimCSE (Gao et al. 2021)

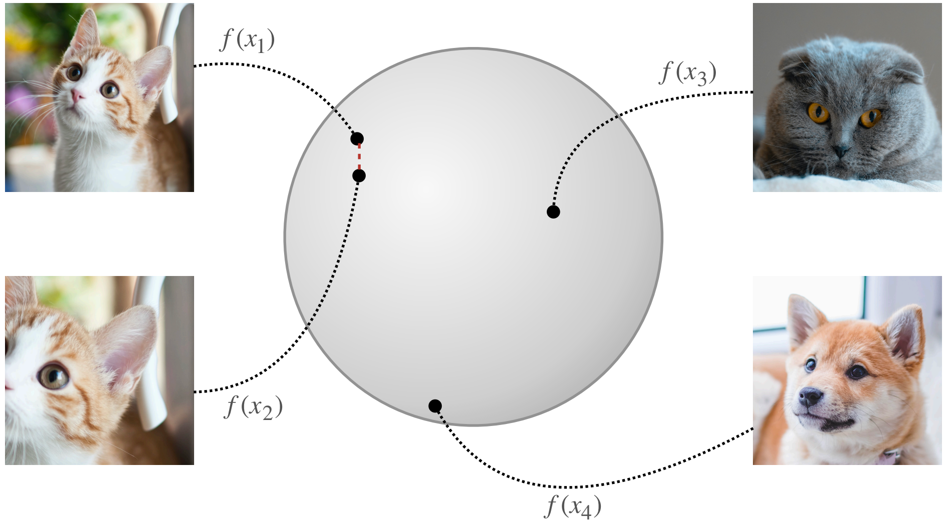
The **self-supervised** framework is very similar to the other examples we have already seen before



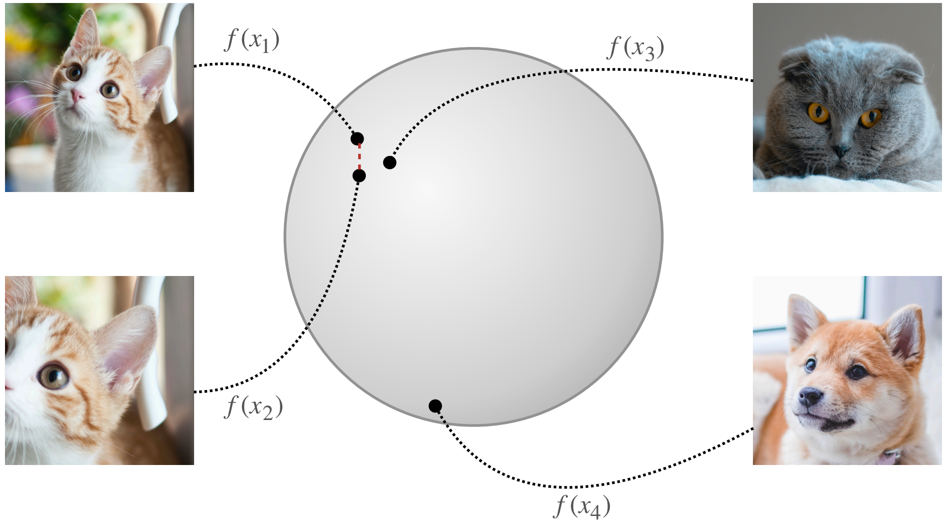
$$\ell_i = -\log \frac{\exp\left(\mathrm{sim}(h_i^{z_i}, h_i^{z'_i})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}(h_i^{z_i}, h_j^{z'_j})/\tau\right)}$$

Brief detour: Supervised Contrastive Learning

It’s just contrastive learning with labeled data!



Unsupervised setting



Supervised setting

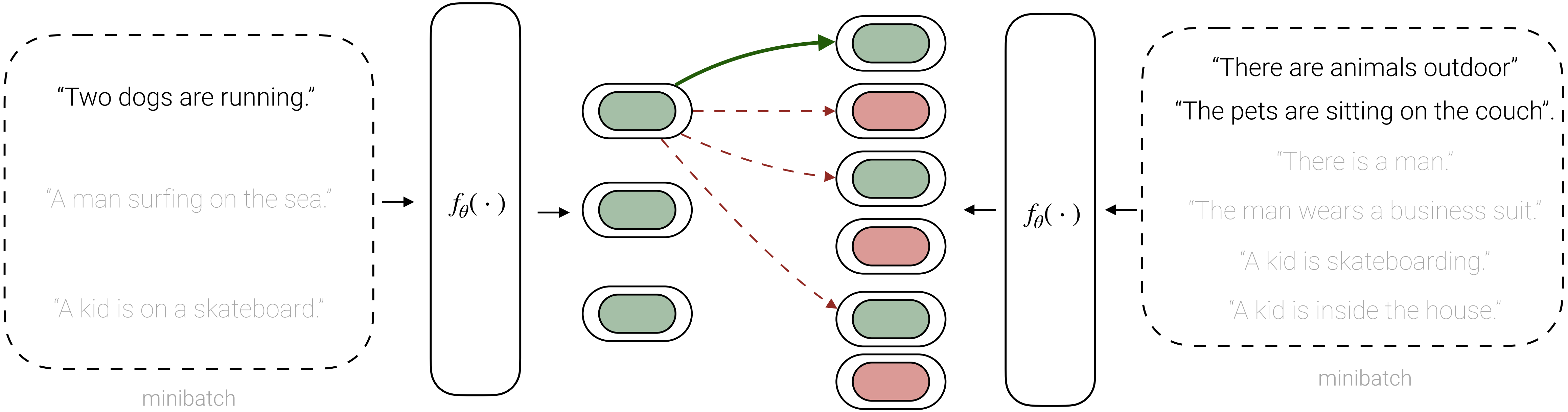
Brief detour: Natural Language Inference task

Natural Language Inference is the task of determining whether an hypothesis can be inferred from a premise

Premise	Hypothesis	Label
Two dogs are running.	There are animals outdoor	entailment
Two dogs are running.	The pets are sitting on the couch”	contradiction
Two dogs are running.	Today it’s sunny.	neutral

SimCSE (Gao et al. 2021)

The authors also experiment using a **supervised contrastive** framework using NLI datasets for training

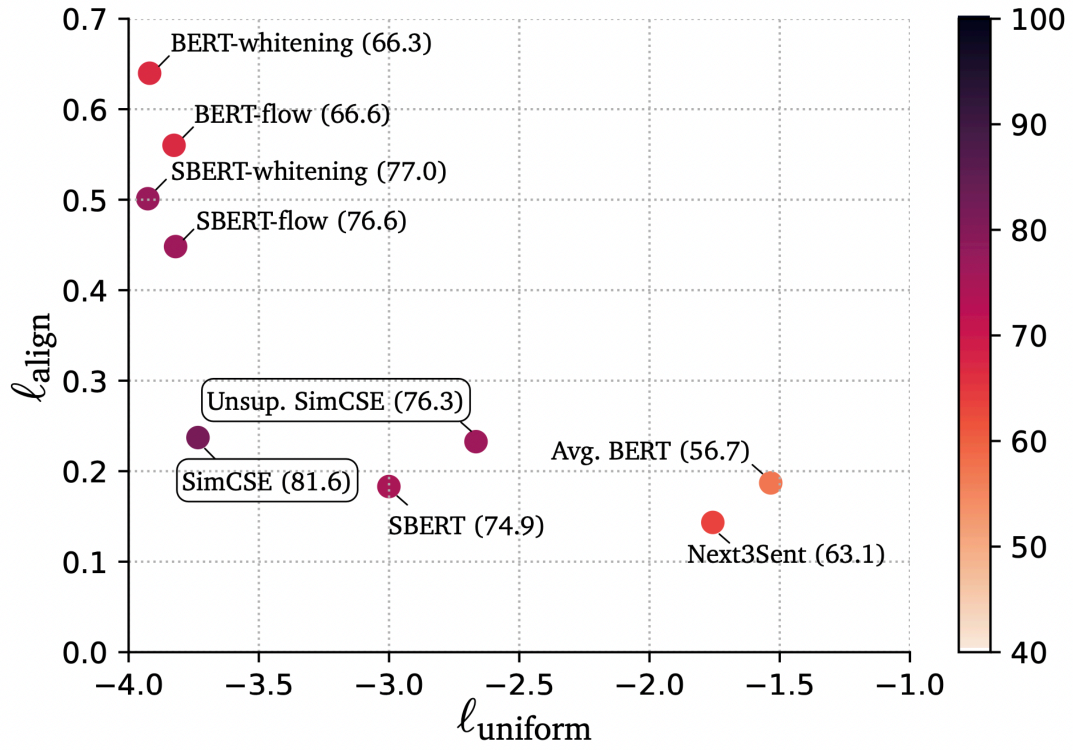


"Two dogs are running."
"A man surfing on the sea."
"A kid is on a skateboard."
minibatch
$f_\theta(\cdot)$
$f_\theta(\cdot)$
"There are animals outdoor"
"The pets are sitting on the couch".
"There is a man."
"The man wears a business suit."
"A kid is skateboarding."
"A kid is inside the house."
minibatch

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

SimCSE (Gao et al. 2021)

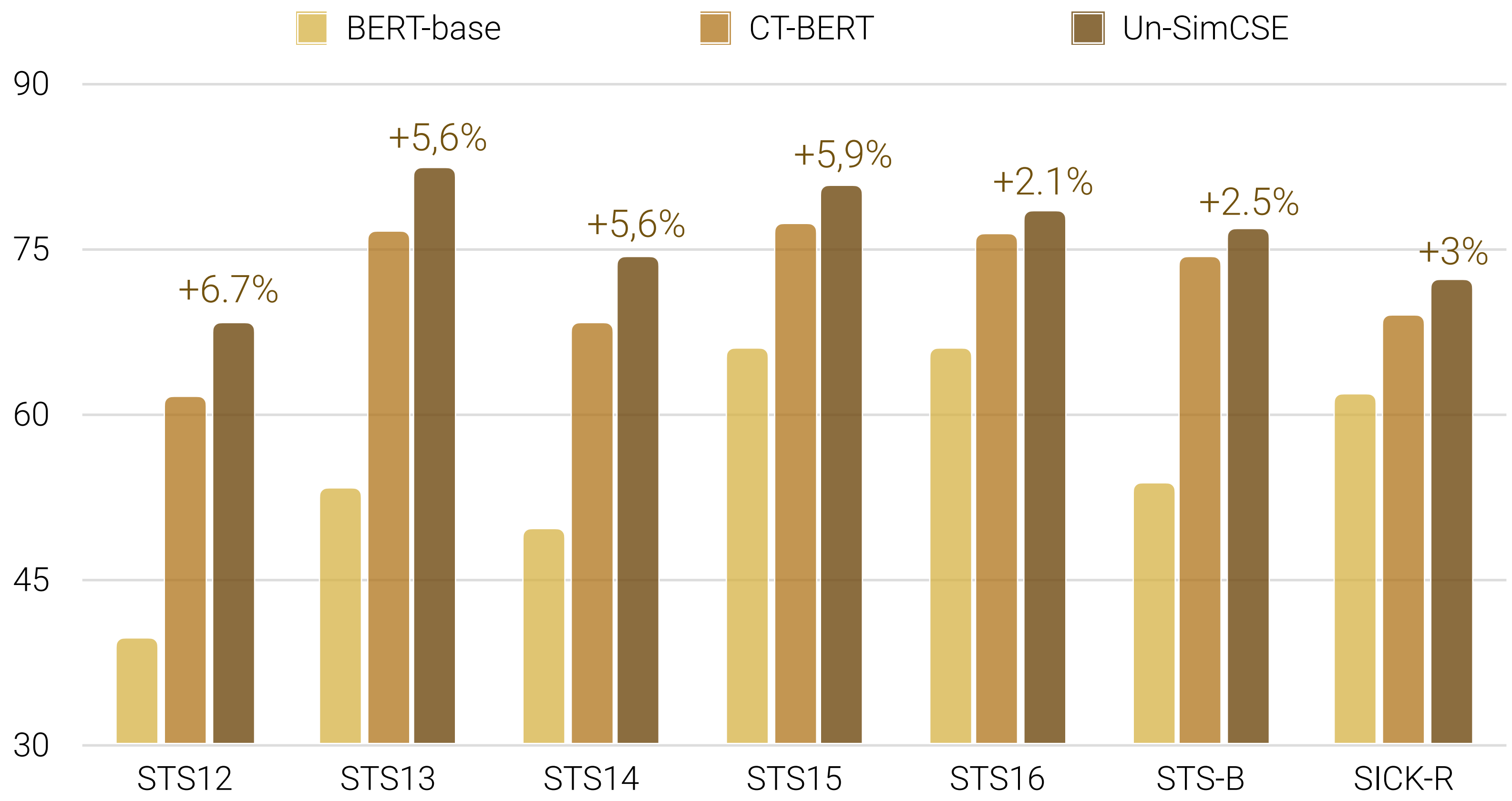
Finally, the authors **evaluate** the quality of learned embeddings.



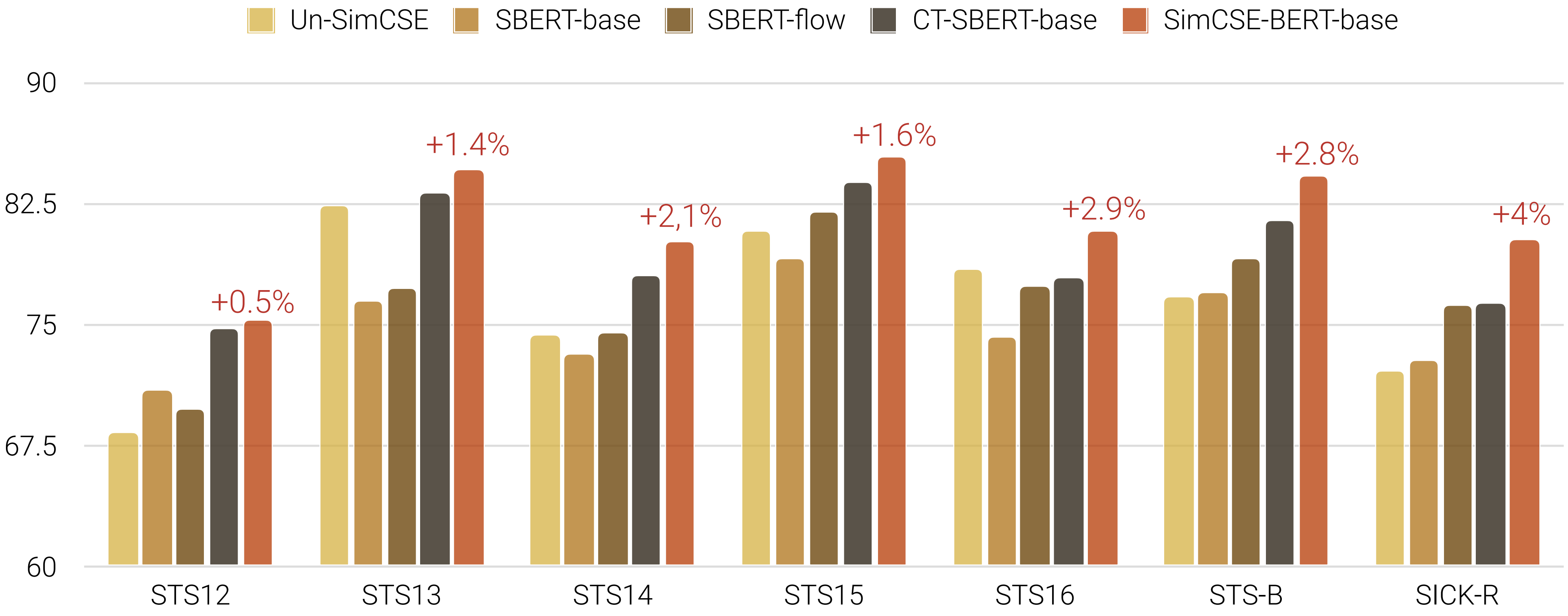
Alignment and Uniformity of difference sentence embedders, Gao et al. 2021

SimCSE (Gao et al. 2021)

Intrinsic evaluation: 7 semantic textual similarity tasks, using Spearman’s correlation index



SimCSE (Gao et al. 2021)



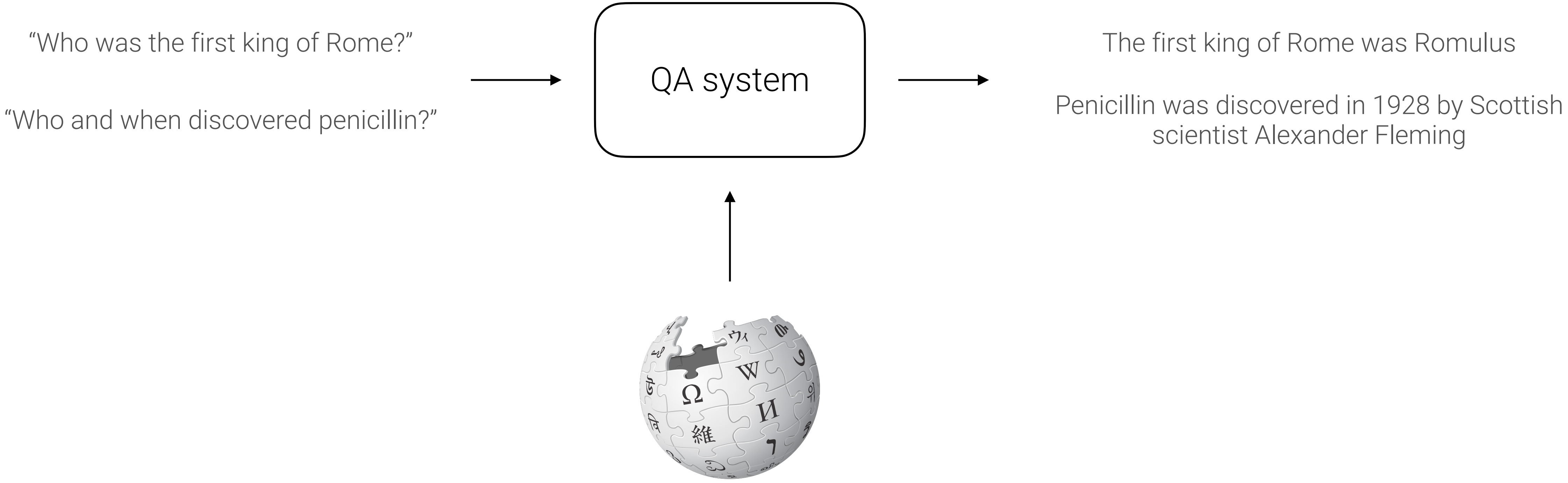
SimCSE (Gao et al. 2021)

Extrinsic evaluation: 7 transfer learning tasks, using a logistic classifier on top of (frozen) embeddings

- Unsup-SimCSE achieves SoTA in most of the datasets
- Sup-SImCSE outperformed by S-BERT

Open-domain question answering

Open-domain question answering (QA) is a task that answers questions using knowledge learnt from a large collection of documents



“Who was the first king of Rome?”

“Who and when discovered penicillin?”

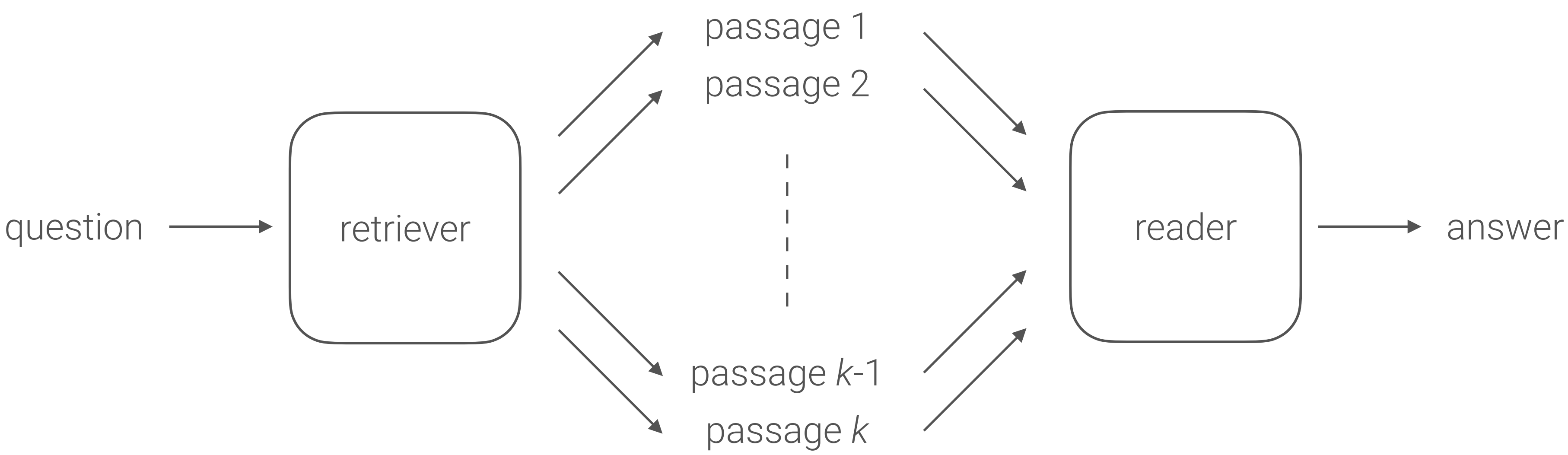
QA system

The first king of Rome was Romulus

Penicillin was discovered in 1928 by Scottish scientist Alexander Fleming

QA system structure

Modern QA systems have usually two main components: a retriever and a reader



Passage retriever

Two approaches for the implementation:

- TF-IDF or BM25 retrievers (inverted index)
- dense retrievers

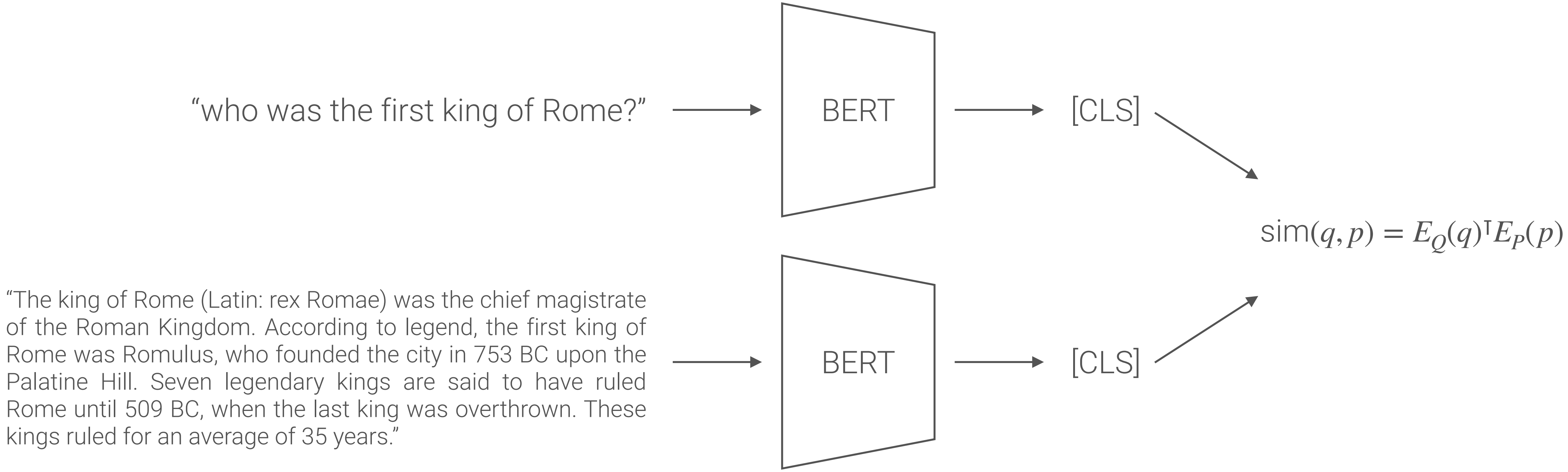
DPR (Karpukhin et al. 2020)

Dense Passage Retriever (DPR) is a passage retriever based on contrastive learning

$$+ \quad -\log \frac{\exp\left(\mathrm{sim}(h_i^{z_i}, h_i^{z'_i})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}(h_i^{z_i}, h_j^{z'_j})/\tau\right)} \quad = \text{SoTA}$$

DPR (Karpukhin et al. 2020)

The framework is based on two encoders



"who was the first king of Rome?"

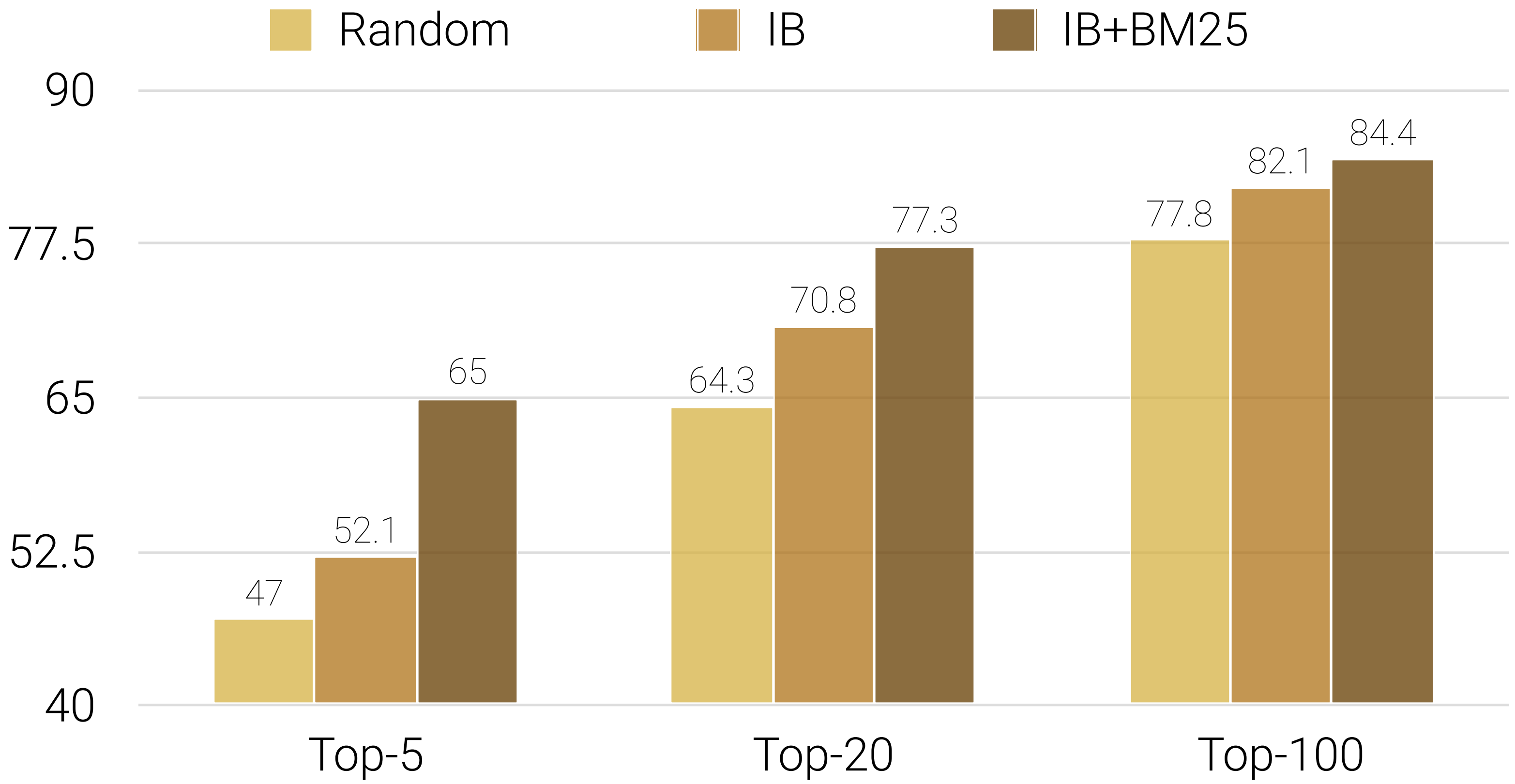
"The king of Rome (Latin: rex Romae) was the chief magistrate of the Roman Kingdom. According to legend, the first king of Rome was Romulus, who founded the city in 753 BC upon the Palatine Hill. Seven legendary kings are said to have ruled Rome until 509 BC, when the last king was overthrown. These kings ruled for an average of 35 years."

$$\text{sim}(q, p) = E_Q(q)^\intercal E_P(p)$$

Inference time?

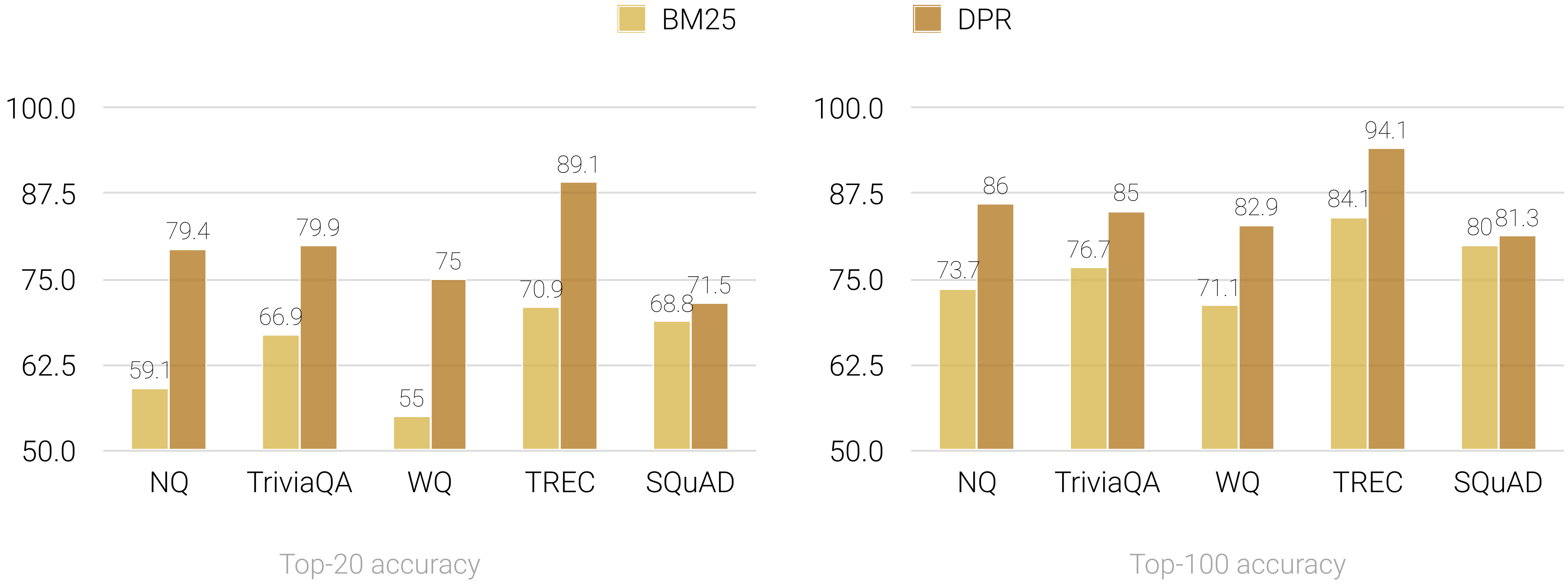
DPR (Karpukhin et al. 2020)

Different techniques to choose negative samples explored



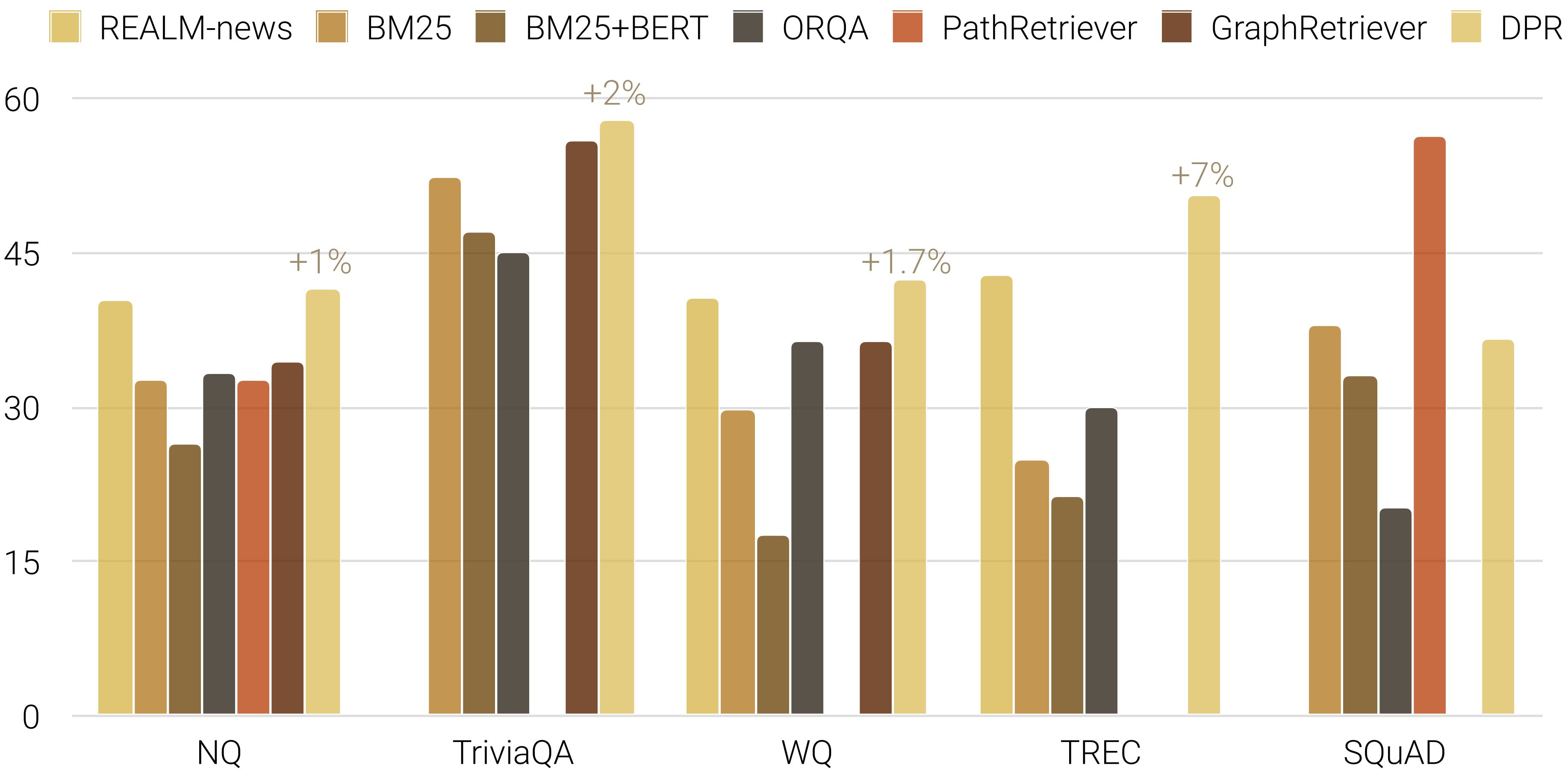
DPR (Karpukhin et al. 2020)

Top-k accuracies of the retriever



DPR (Karpukhin et al. 2020)

End-to-end QA accuracies

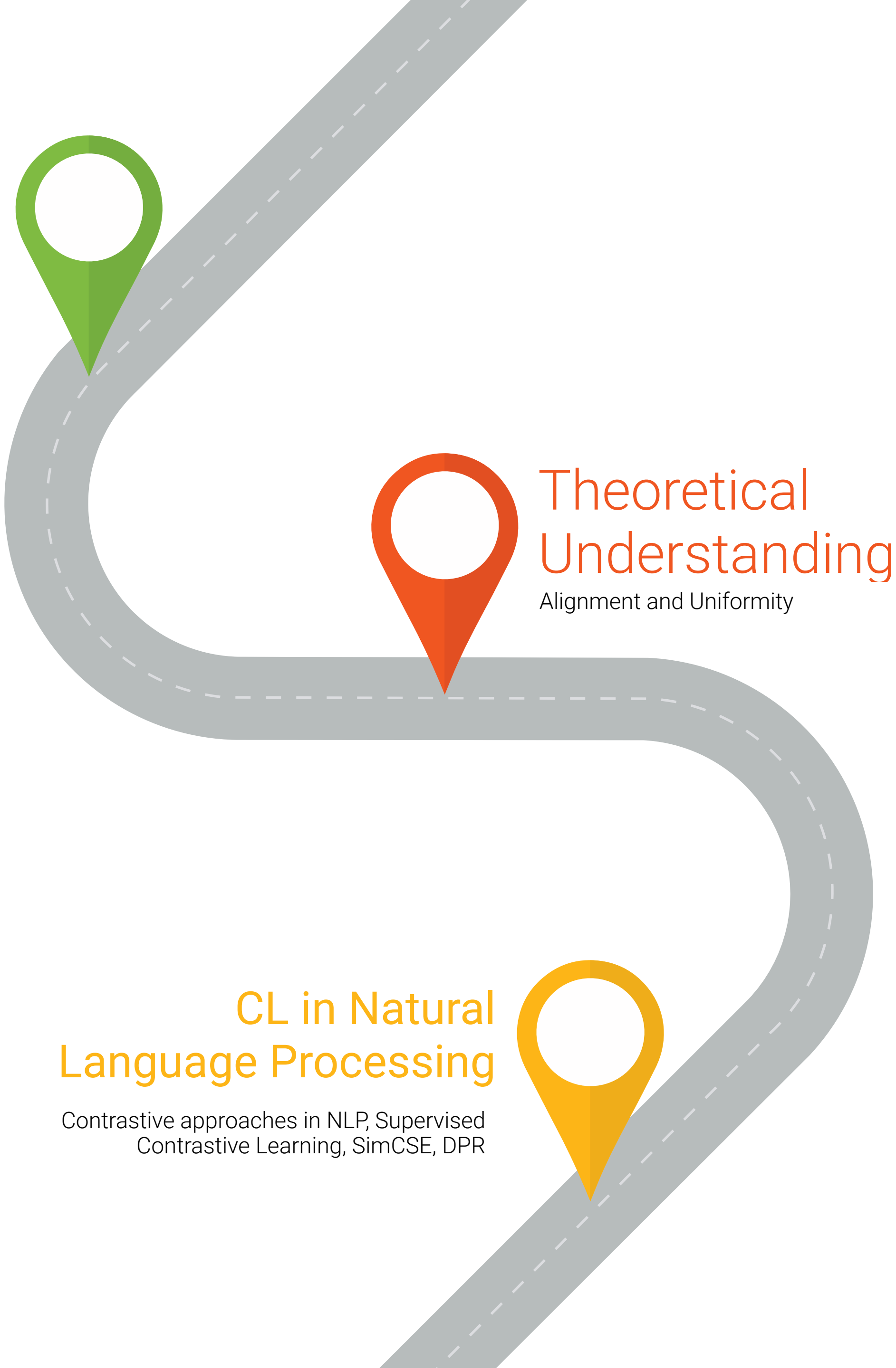


Conclusion

Contrastive learning in NLP is stil a very dynamic research area

- DiffCSE (sequel of SimCSE) released only few days ago (credits to Till Aczél for recommending it 💡)

Main open challenge remains data augmentation



Introduction

Self-supervised learning, intuition and formalization of contrastive learning, SimCLR

Theoretical Understanding of CL

Alignment and Uniformity

CL in Natural Language Processing

Contrastive approaches in NLP, Supervised Contrastive Learning, SimCSE, DPR

Thanks for your attention!

References

- Lilian Weng's blog (link)
- Tianyu Gao, SimCSE: Simple Contrastive Learning of Sentence Embeddings @ EMNLP 2021 (link)
- Yonglong Tian, Contrastive Learning: A General Self-supervised Learning Approach (link)
- Suzanna Becker and Geoffrey E. Hinton, Self-organizing neural network that discovers surfaces in random-dot stereograms, 1992
- Ting Chen, Simon Kornblith, Mohammad Norouzi and Geoffrey Hinton, A Simple Framework for Contrastive Learning of Visual Representations, 2020
- Tongzhou Wang and Phillip Isola, Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, 2020
- Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly and Mario Lucic, On Mutual Information Maximization for Representation Learning, 2020
- David McAllester, Karl Stratos, Formal Limitations on the Measurement of Mutual Information, 2018
- Yonglong Tian, Dilip Krishnan and Phillip Isola, Contrastive Multiview Coding, 2019
- Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding and Pengtao Xie, CERT: Contrastive Self-supervised Learning for Language Understanding, 2020

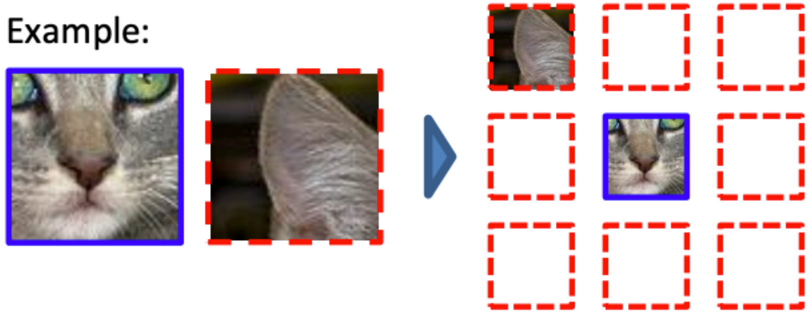
References

- Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu and Weizhu Chen, # A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation, 2020
- Jason Wei and Kai Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, 2019
- Tianyu Gao, Xingcheng Yao and Danqi Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, 2021
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu and Dilip Krishnan, Supervised Contrastive Learning, 2020
- Yannic Kilcher, Supervised Contrastive Learning
- Kawin Ethayarajh, How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings, 2019
- Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li, On the sentence embeddings from pre-trained language models, 2020
- Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu, Improving neural language generation with spectrum control, 2020
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih, Dense Passage Retrieval for Open-Domain Question Answering, 2020

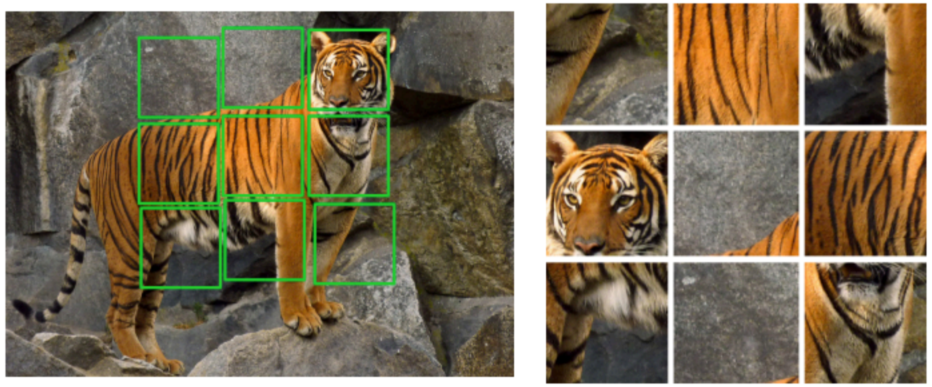
Supplementary material

Previous Self-Supervised Learning approaches (courtesy of Yonglong Tian)

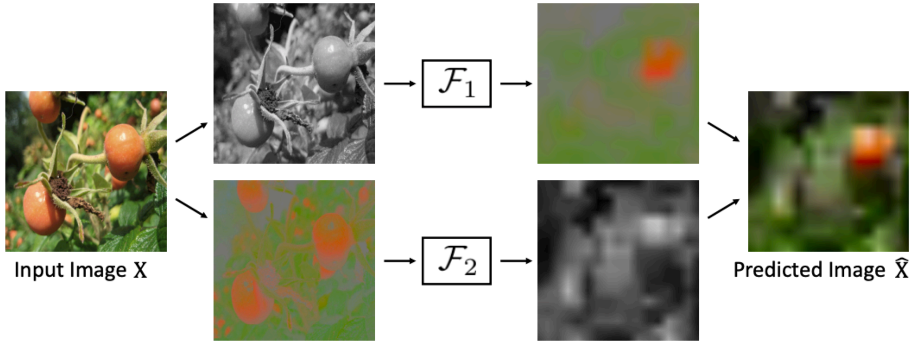
Context (Doersch et al. 2015)



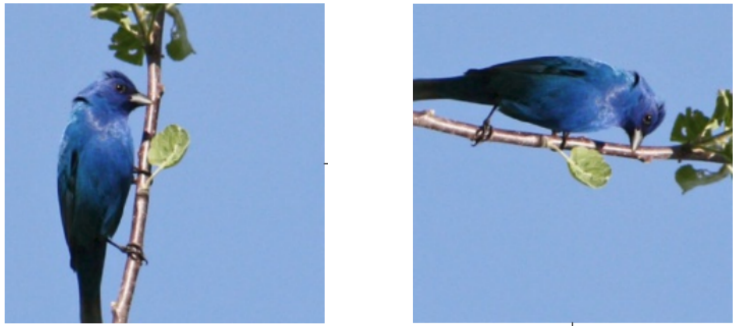
JigSaw Puzzle (Noroozi et al. 2016)



Colorization (Zhang et al. 2016, 2017)



Rotation Prediction (Gidaris et al. 2018)

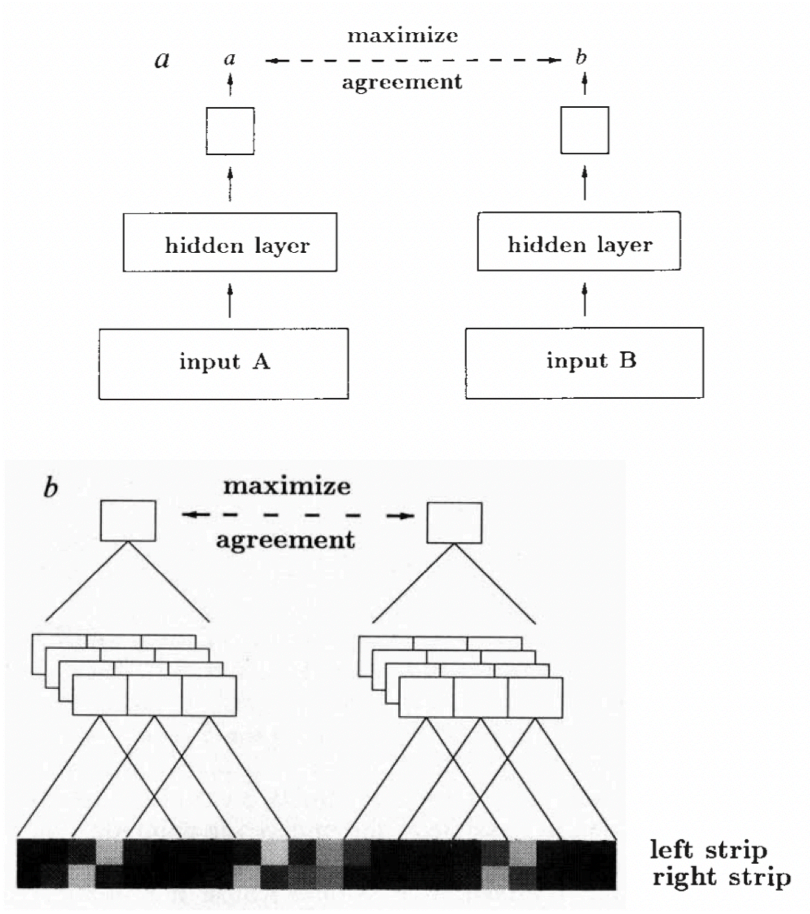


Self-organizing neural network that discovers surfaces in random-dot stereograms (Becker & Hinton 1992)

In their work, Becker and Hinton train multiple separate modules that look at separate but correlated part of the input and attempt to produce the same output.

In their paper, they refer to the correlation between this strategy and the maximization of the mutual information between the underlying signal and the average of the the two encoders.

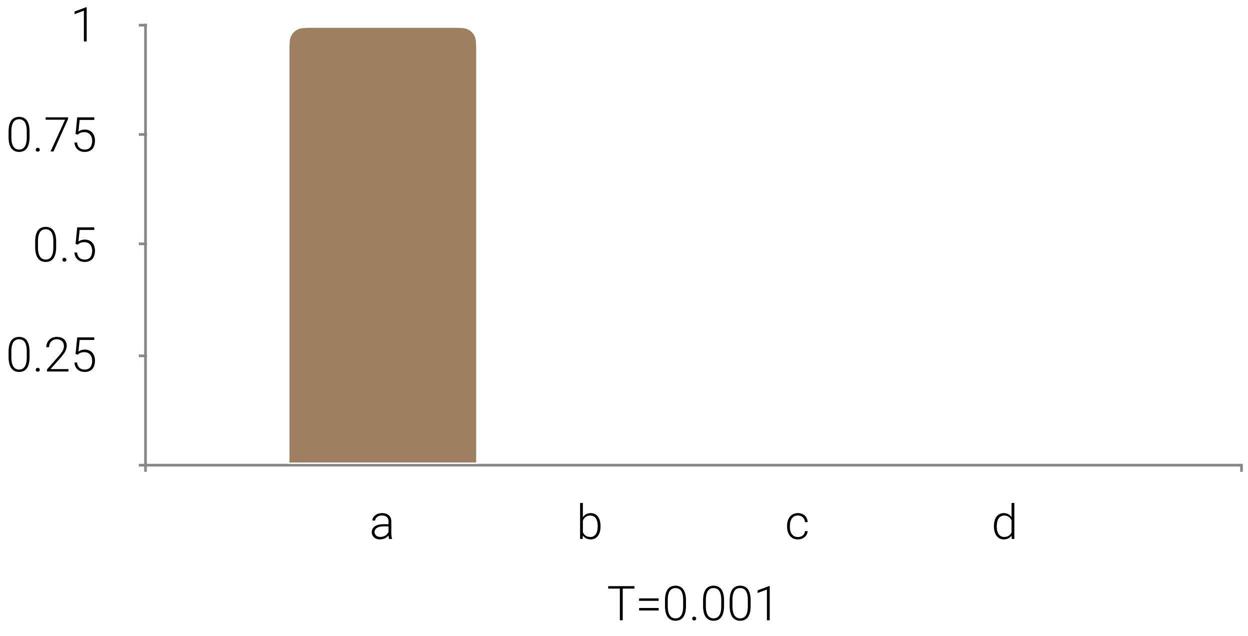
Therefore, maximizing the mutual information between the latent representations the two network should learn to extract a “pure” version of the underlying common signal, which is attended by each network together with an independent random gaussian noise that corrupts the true distribution.



Softmax Temperature

In the softmax, $\tau > 0$ is the so called “temperature” parameter, that allows to control the entropy of a distribution, while preserving the relative rank of each event.

As $T \rightarrow \infty$, we approach the uniform distribution (maximum entropy). As $T \rightarrow 0$, all the mass of the distribution tends to be placed on the same element.



Alignment and Uniformity formalization

Alignment loss definition

$$\mathcal{L}_{align}(f;\alpha) = \mathop{\mathbb{E}}_{(x,y)\sim p_{pos}} \left[||f(x) - f(y)||_2^{\alpha}\right], \qquad \alpha > 0$$

Uniformity loss definition

$$\mathcal{L}_{uniform}(f;t) = log \mathop{\mathbb{E}}_{(x,y)\overset{\text{i.i.d.}}{\sim} p_{data}} \left[\exp\left(-t\,||f(x) - f(y)||_2^2\right)\right], \qquad t > 0$$

More on Alignment and Uniformity

Uniformity metric is not as straightforward as alignment to define. This metric must be both **asymptotically correct** (the optimization of this metric should converge to the uniform distribution) and **empirically reasonable with a finite number of points**.

The authors therefore decide to consider the Radial Basis Function kernel (**RBF**, also know as Gaussian potential kernel) and define the uniformity loss as the logarithm of the average pairwise RBF (since it’s nicely tied to the uniform distribution on the unit sphere and can be used to represent a general class of kernels, including Riesz *s*-potentials).

In their work, they show that the uniform distribution is the unique minimizer for uniform metric and that this convergence is *weak*.

In the Appendix, the authors also discuss further properties of the uniformity loss and characterize its optimal values and range .

Perfect Alignment and Perfect Uniformity

In their work, Wang and Isola also define the notion of optimal encoder for each of the two proposed metric.

Perfect Alignment. We say an encoder f is perfectly aligned if $f(x) = f(y)$ a.s. $(x, y) \sim p_{pos}$

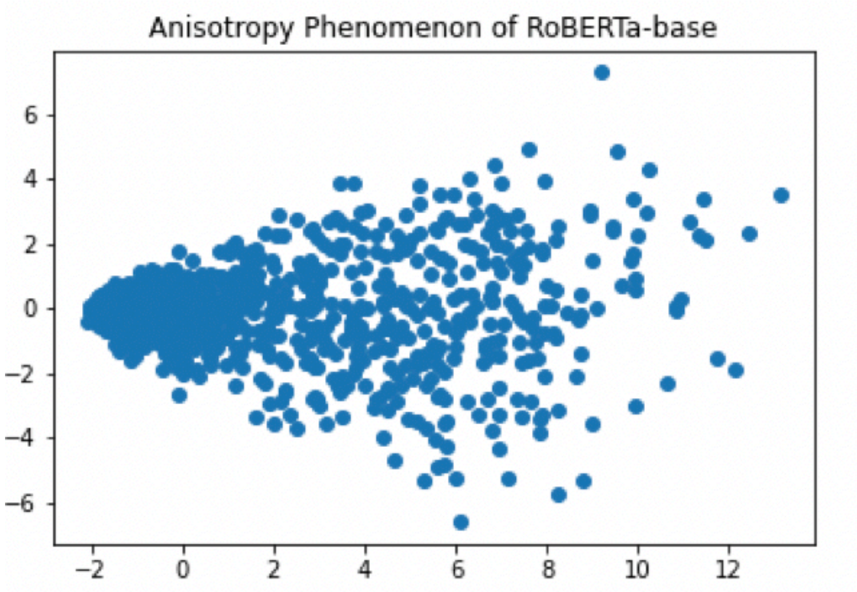
Perfect Uniformity. We say an encoder f is perfectly uniform if the distribution $f(x)$ for $x \sim p_{data}$ is the uniform distribution σ_{m-1} on $\mathcal{S}^{m-1}$ (the hypersphere)

The authors also note that is not always possible to achieve perfect uniformity, e.g. when the data manifold in $\mathbb{R}^n$ is lower dimensional than the feature space in $\mathcal{S}^{m-1}$. Moreover, they also highlight that when p_{data} and p_{pos} are formed from sampling augmented sample from a finite dataset, there cannot be an encoder which achieves both perfect alignment and perfect uniformity. This is because perfect alignment implies that all augmentations from a single element have the same feature vector (and therefore they cannot be uniformly distributed in the latent space).

However, perfectly uniform encoder functions exists under the conditions that $n \geq m - 1$ and p_{data} has bounded density.

Anisotropy

Recent works (Ethayarajh 2019; Li et al. 2020) have identified anisotropy to be a problem in language representations.

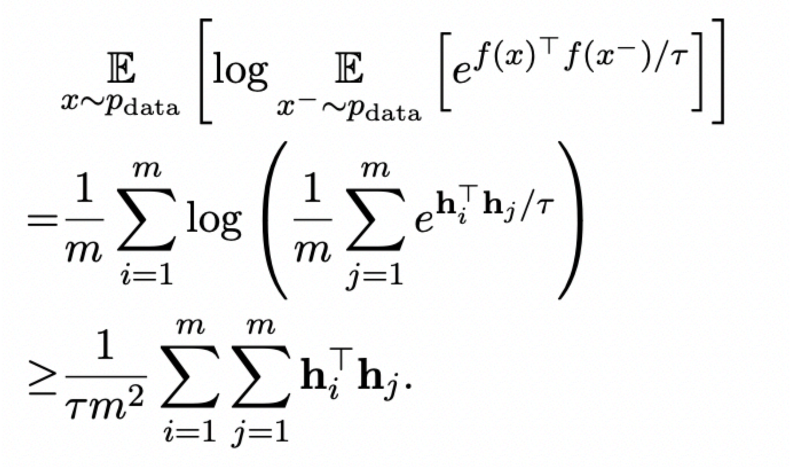


Analyzing the Anisotropy Phenomenon in Transformer-based Masked Language Models, Luo 2021

The problem of anisotropy is intuitively correlated to the uniformity metric defined by (Wang and Isola 2020). In fact, (Gao et al. 2021) prove the optimization of the contrastive loss can flatten the eigenvalues spectrum (one of the believed causes of anisotropy) and eventually alleviate the problem of anisotropy.

Proof for Eigenvalues Limit (Gao et al. 2021)

Let the finite set of samples $\{x_i\}_{i=1}^{m}$ be uniformly distributed. We can then derive the following formula using Jensen inequality



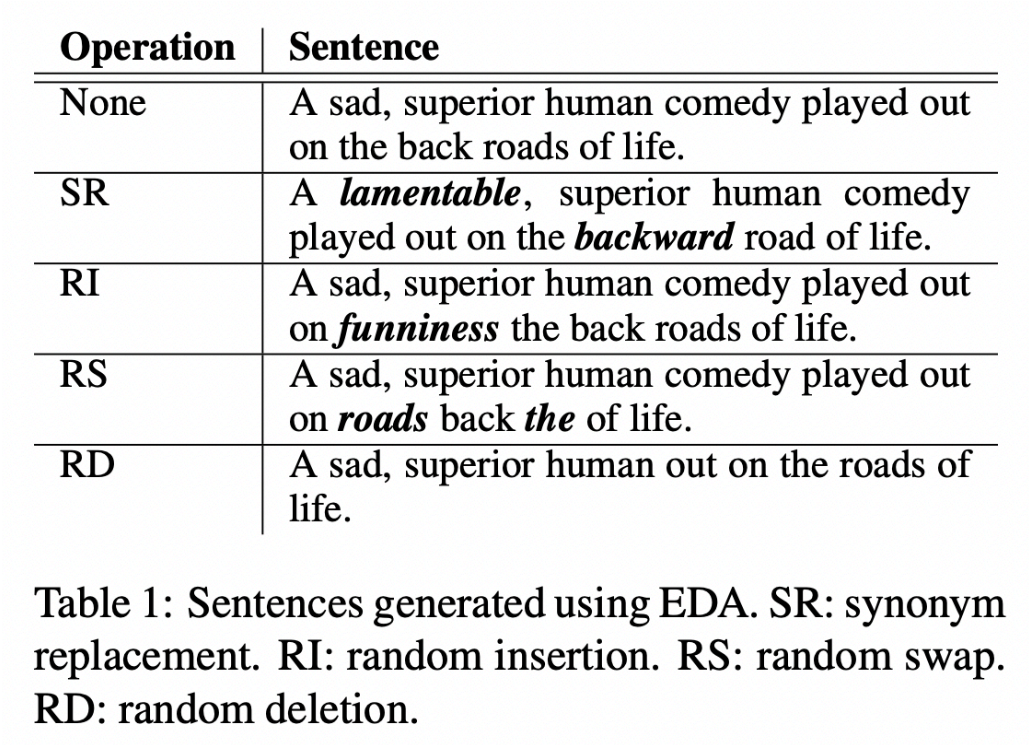
$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim p_{\text{data}}}\left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}}\left[e^{f(x)^\top f(x^-)/\tau}\right]\right] \\
=&\frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{1}{m}\sum_{j=1}^{m}e^{\mathbf{h}_i^\top \mathbf{h}_j/\tau}\right) \\
\geq&\frac{1}{\tau m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{h}_i^\top \mathbf{h}_j.
\end{aligned}$$

Let W be the sentence embedding matrix, where the *i*-th row is the embedding $h_i = f(x_i)$. Optimizing the second term of the contrastive loss essentially minimizes an upper bound of the summation of all elements in $WW^\intercal$, i.e. $\text{Sum}(WW^\intercal) = \sum_{i=1}^{m}\sum_{j=1}^{m} h_i^\intercal h_j$.

Since the embeddings are normalized, all the element of the diagonal of this matrix are 1, and hence $\text{tr}(WW^\intercal)$ is constant. According to (Merikosky 1984), if all the elements of $WW^\intercal$ are positive, then $\text{Sum}(WW^\intercal)$ is an upper bound for the largest eigenvalue of $WW^\intercal$. Hence, minimizing the second term of the contrastive loss minimizes the magnitude of the eigenvalues of $WW^\intercal$.

Augmentation techniques in NLP

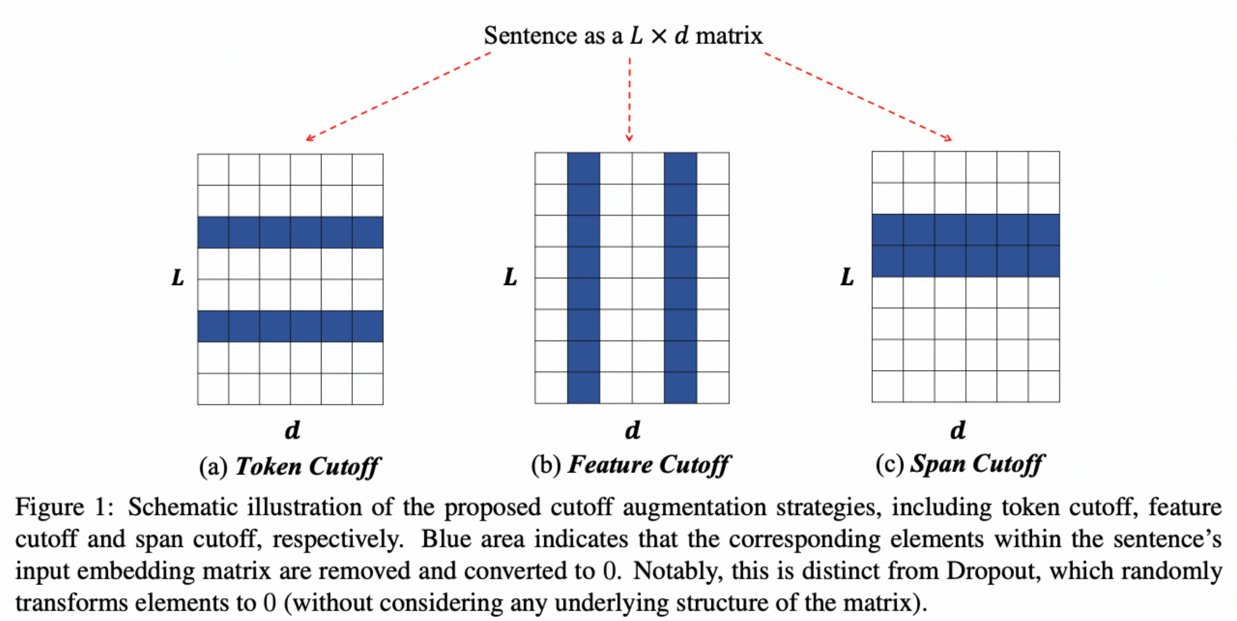
Lexical edits (Wei and Zhou, 2019)



Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

Table 1: Sentences generated using EDA. SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

Cutoff (Shen et al. 2020)



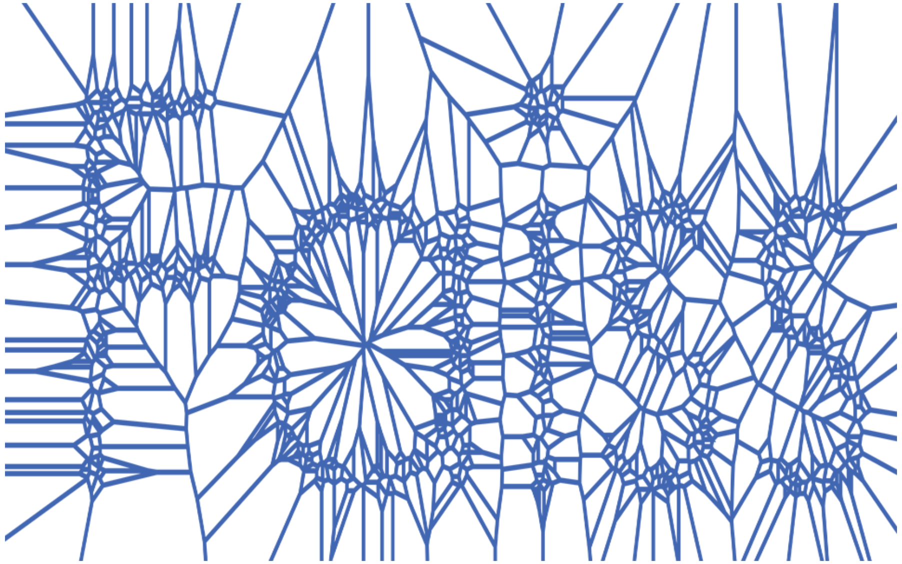
(a) ***Token Cutoff*** (b) ***Feature Cutoff*** (c) ***Span Cutoff***

Figure 1: Schematic illustration of the proposed cutoff augmentation strategies, including token cutoff, feature cutoff and span cutoff, respectively. Blue area indicates that the corresponding elements within the sentence's input embedding matrix are removed and converted to 0. Notably, this is distinct from Dropout, which randomly transforms elements to 0 (without considering any underlying structure of the matrix).

FAISS indexing

FAISS (Facebook AI Similarity Search) is an extremely efficient, open-source library for similarity search and clustering of dense vectors, which can easily be applied to billions of vectors and that is based on nearest-neighbor search.

FAISS is used to efficiently index all the embedded passages in memory and to retrieve the *k* most relevant of them using a query based on vector similarity.



DPR (Karpukhin et al. 2020)

DPR demo: http://qa.cs.washington.edu:2020