Contrastive learning

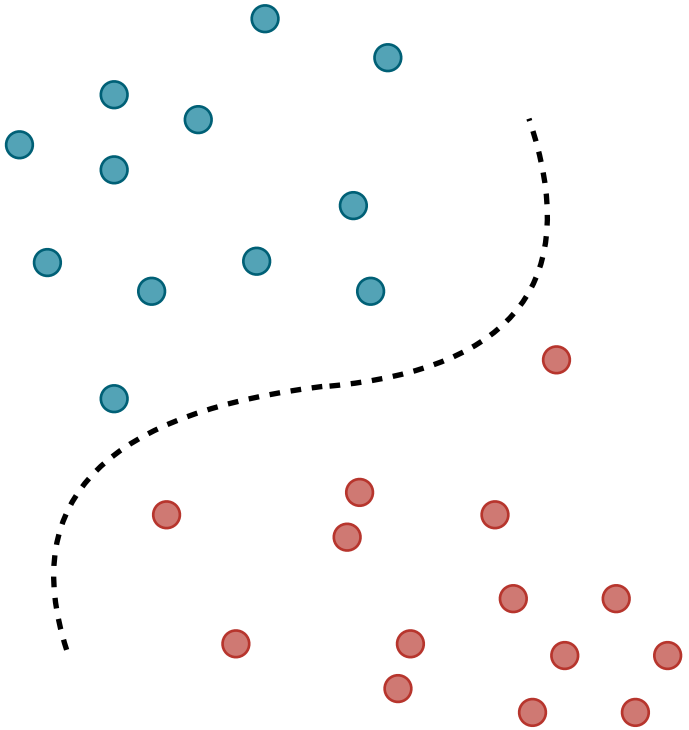
Giacomo Camposampiero

Seminar in Deep Neural Networks, 03.05.2022

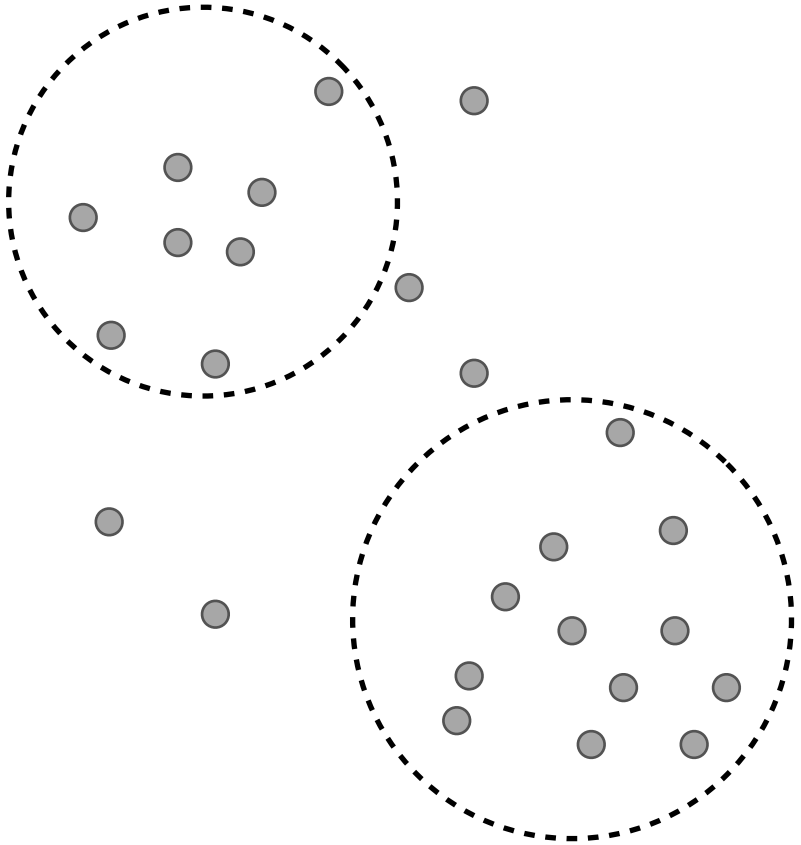
ETH zürich


Introduction

Learning Approaches



Supervised Learning

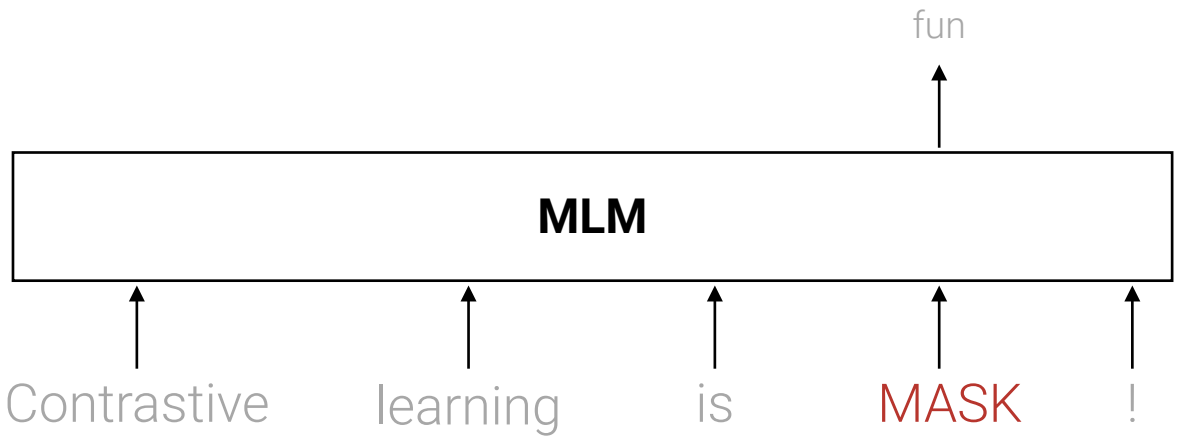


Unsupervised Learning

Self-supervised Learning

A middle-ground between supervised and unsupervised learning

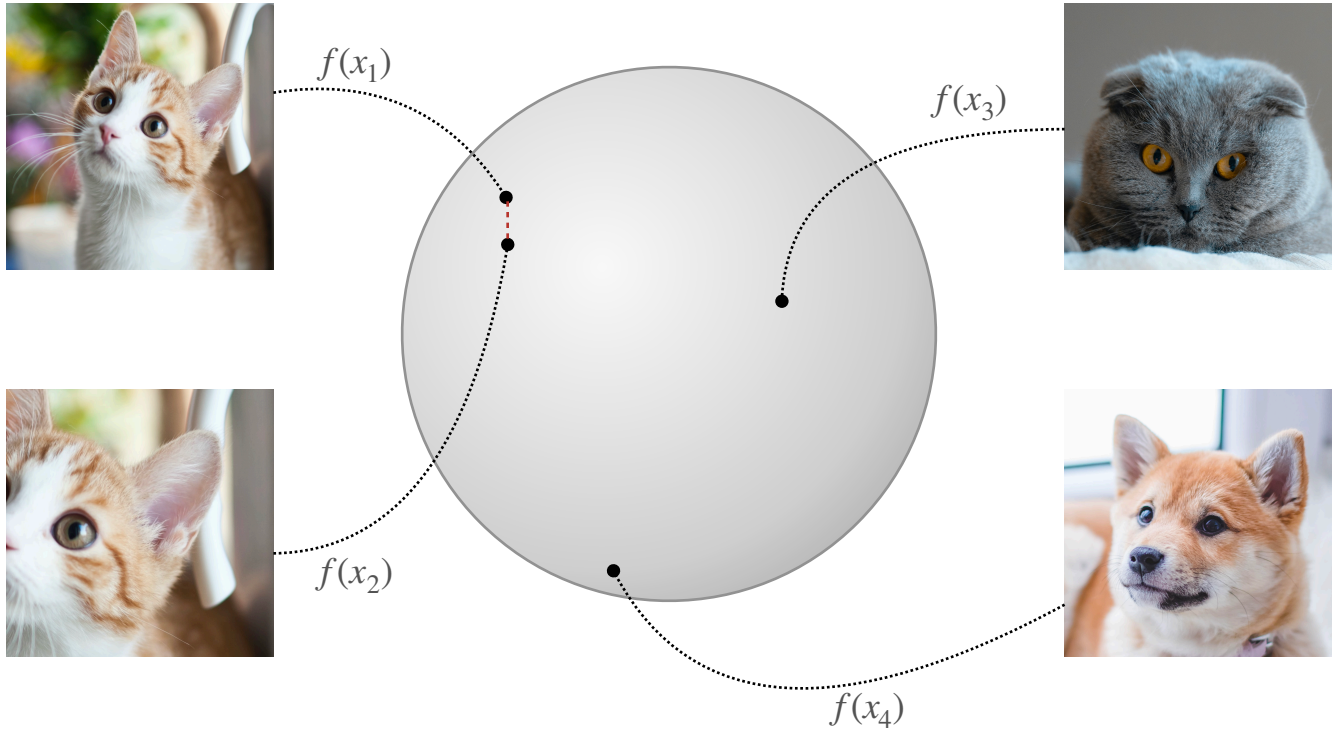
- unlabeled dataset
- supervised training task (**auxiliary task**)



Masked Language Models (MLM) training objective is an example of self-supervised learning task.

Contrastive Learning

In the latent space, similar pairs stay together while dissimilar are pushed away



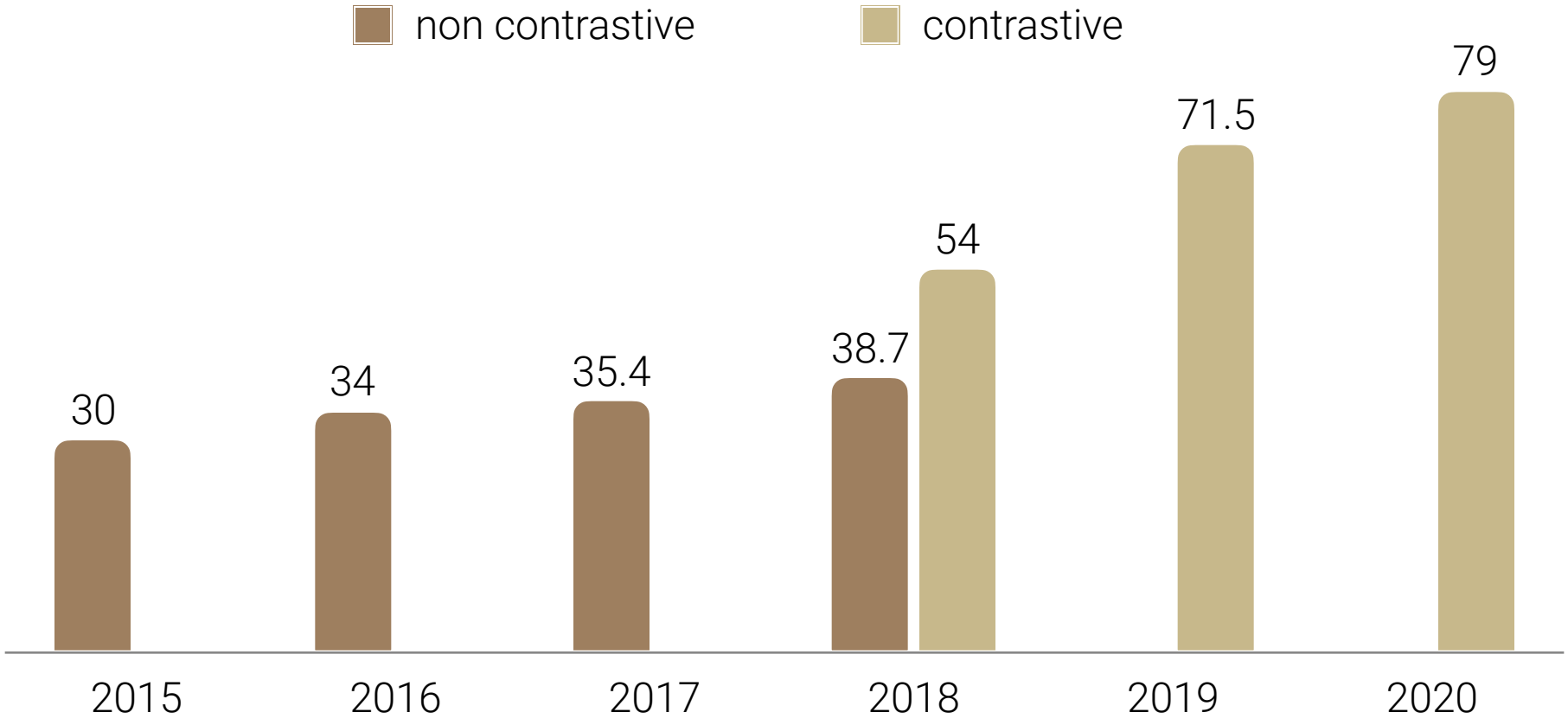
⚠️ in the unsupervised setting, similar pairs are augmentations of the same sample

Intuition behind Contrastive Learning

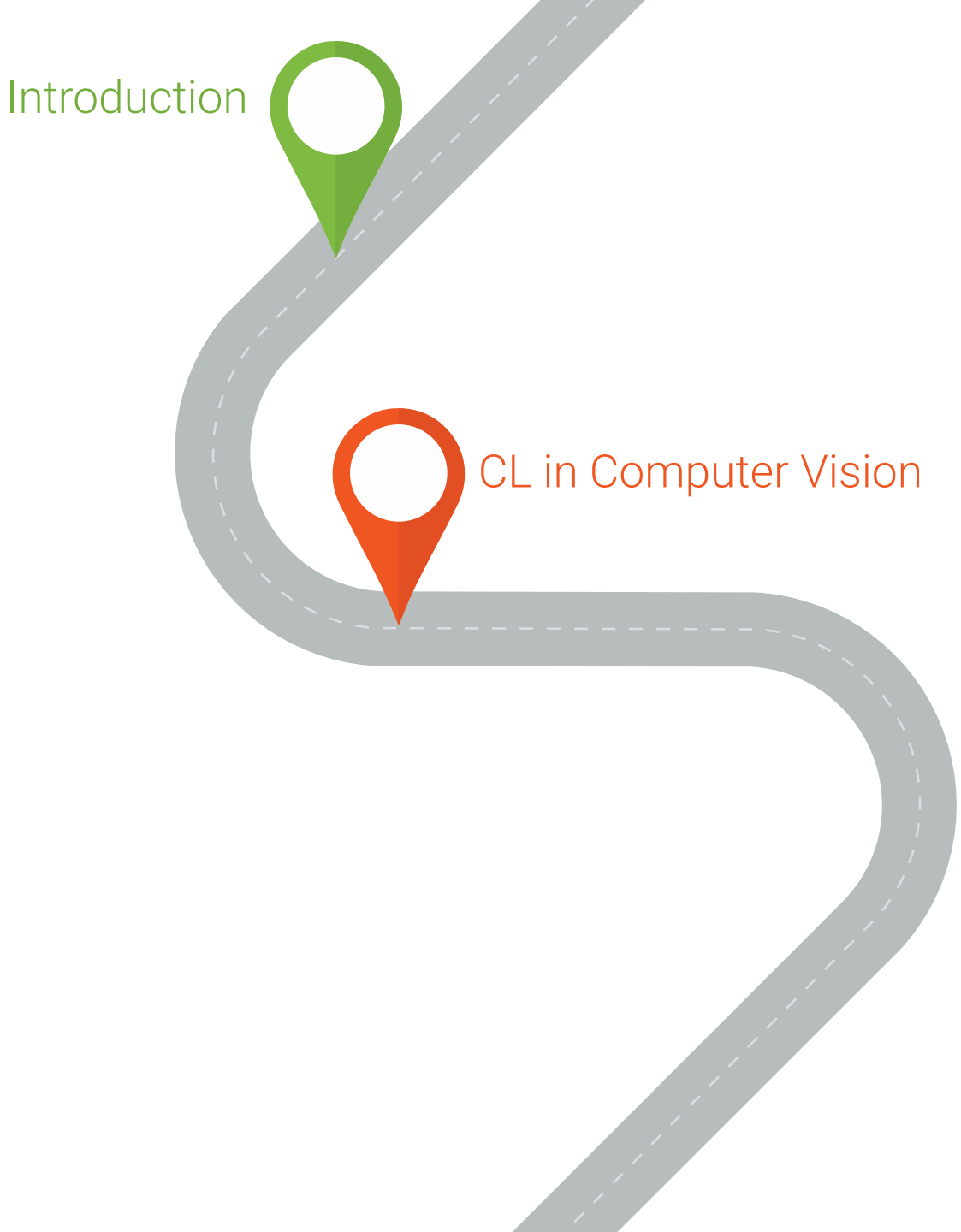
The encoder learn to extract most shared informations between positive pairs (preserving semantic information), while remaining invariant to other noise factors




Does it work?

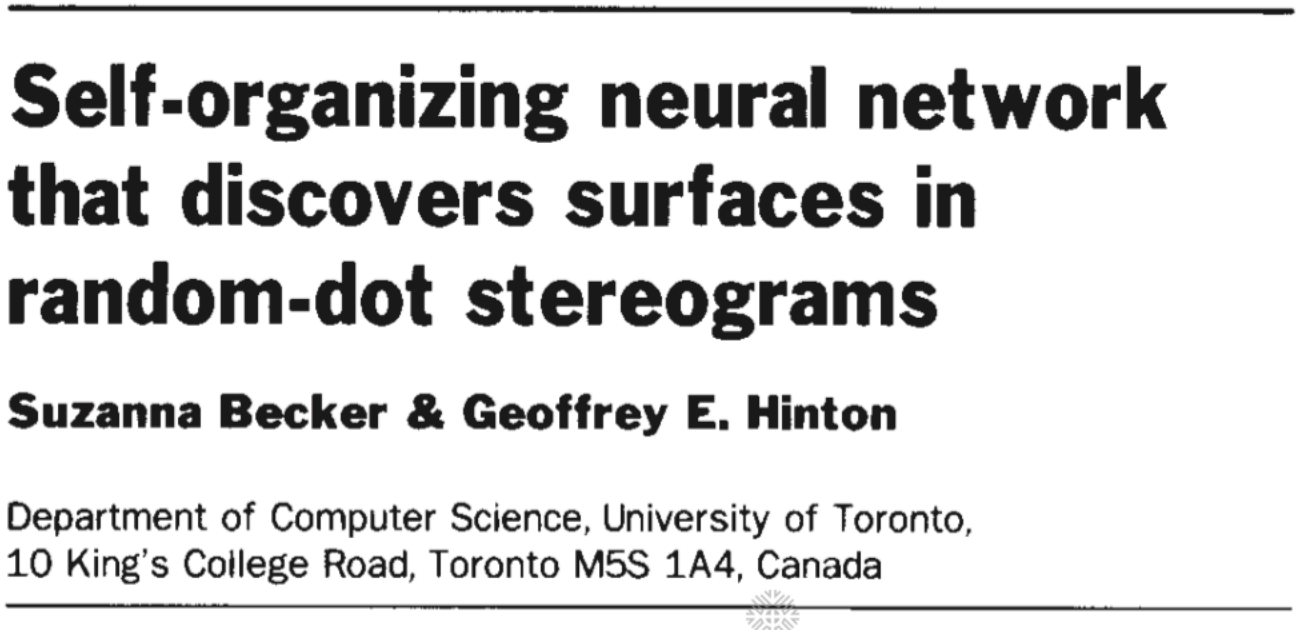


Best-scoring self-supervised approaches on the ImageNet Linear Benchmark.
Courtesy of Yonglong Tian, Contrastive Learning: A General Self-supervised Learning Approach


Introduction
CL in Computer Vision

Contrastive learning in Computer Vision

Contrastive Learning approaches for self-supervised learning originated in the Computer Vision field; first contribution dates back to 1992 (Becker & Hinton)



Self-organizing neural network that discovers surfaces in random-dot stereograms

Suzanna Becker & Geoffrey E. Hinton

Department of Computer Science, University of Toronto,
10 King's College Road, Toronto M5S 1A4, Canada

Contrastive learning in Computer Vision

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

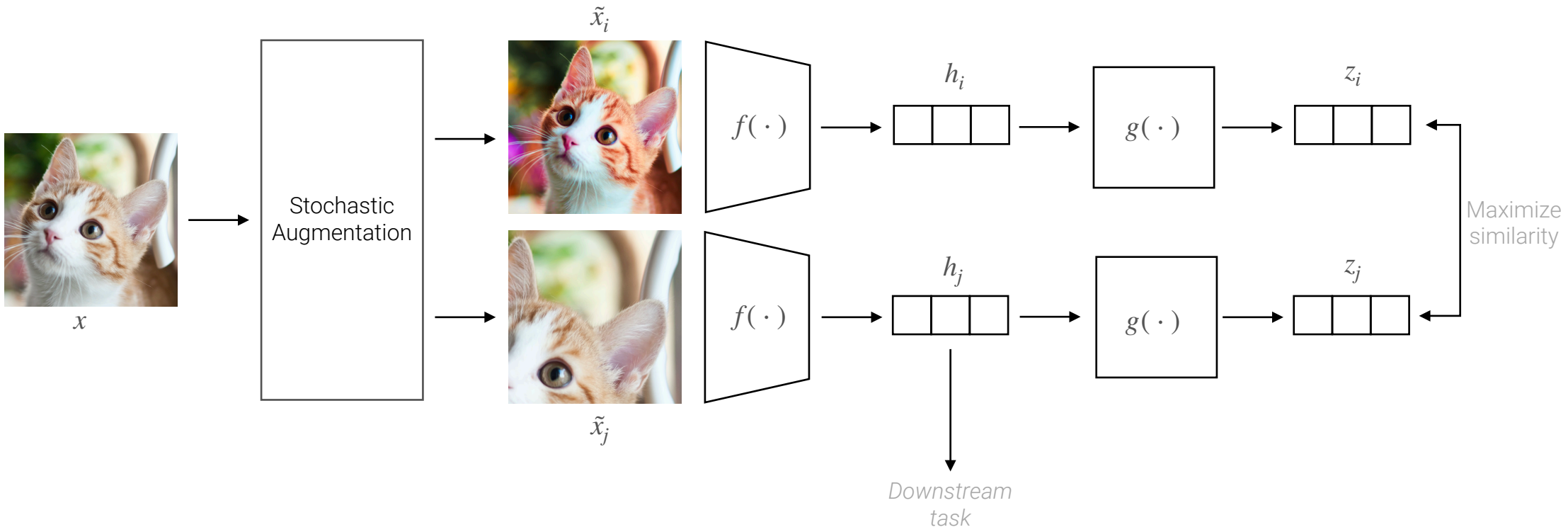
- random cropping
- random color distortions
- random Gaussian blur

Many relevant works out there

- MoCo (He et al. 2019)
- **SimCLR** (Chen et al. 2020)
- BYOL (Grill et al. 2020)
- Barlow Twins (Zbontar et al. 2021)

SimCLR (Chen et al. 2020)

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



SimCLR (Chen et al. 2020)

Training procedure

1. randomly sample a mini-batch of N examples; each sample is augmented using two different transformations, resulting in $2N$ samples in total

2. given one positive pair, the other $2(N-1)$ samples are considered as negatives; a representation for each sample is obtained using the encoder $f(\cdot)$

3. compute $z = g(\cdot)$ applying an additional non-linear projection to the embedding

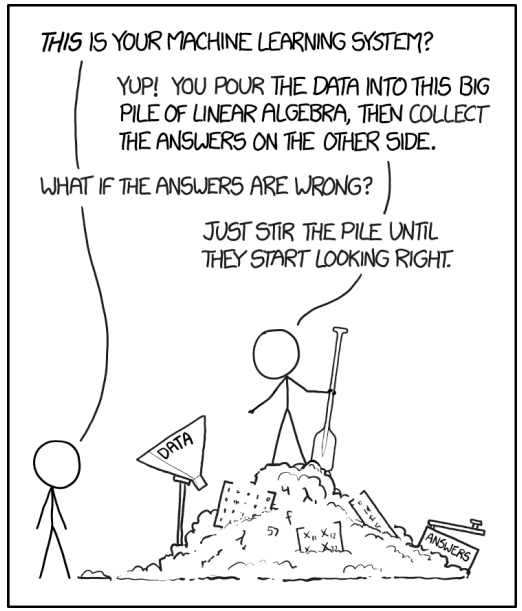
4. compute the loss across all positive pairs in a mini-batch using cosine similarity $sim(\cdot, \cdot)$

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

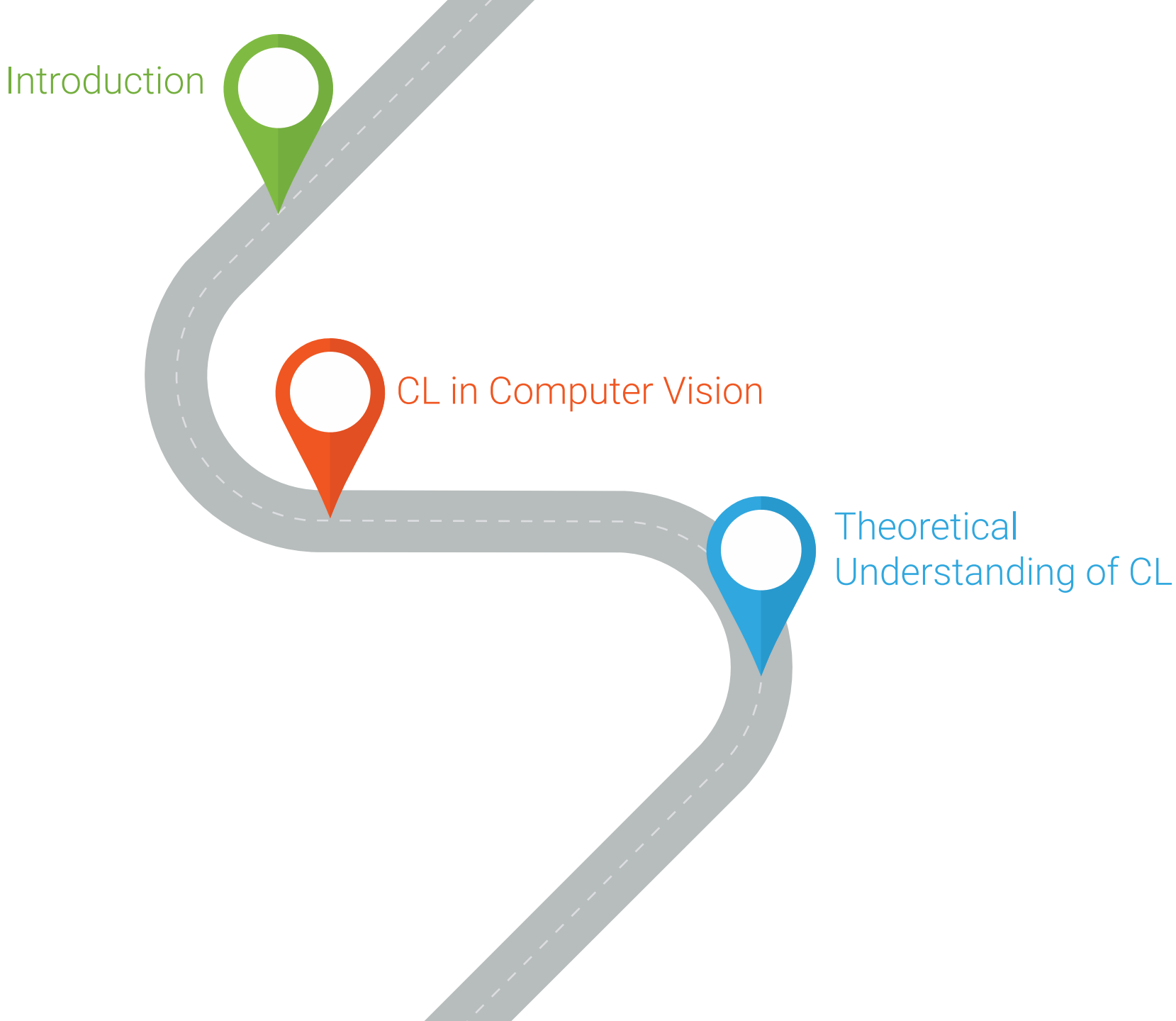
SimCLR (Chen et al. 2020)

Key findings of the paper

- contrastive learning benefits more from stronger data augmentation than supervised learning
- composition of multiple data augmentation operations is necessary to yield good embeddings
- contrastive learning benefits from larger batch sizes and longer trainings (more negative samples)



https://xkcd.com/1838


Introduction
CL in Computer Vision
Theoretical
Understanding of CL

Theoretical Understanding of Contrastive Learning

In the last few years, many efforts went into a theoretical understanding of the Contrastive framework and the reasons why it produces such effective representations.

MIT CSAIL Lab, Massachusetts (USA)

Contrastive learning formalization (Wang & Isola 2020)

Wang and Isola in their work define a general and unified objective called *contrastive loss*

- $p_{data} \rightarrow$ *data distribution over* $\mathbb{R}^n$
- $p_{pos}(\cdot, \cdot) \rightarrow$ *distribution of positive pairs over* $\mathbb{R}^n \times \mathbb{R}^n$.
- $f : \mathbb{R}^n \rightarrow S^{m-1}$ *encoder that maps data to* ℓ_2 *normalized embeddings of dimension* m
- $\tau > 0 \rightarrow$ *scalar temperature hyper-parameter (controls entropy)*
- $M \in \mathbb{Z}_+ \rightarrow$ *fixed number of negative samples*

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

InfoMax principle

Many empirical works are motivated by the InfoMax principle of maximizing $I(f(x); f(y))$ for $(x, y) \sim p_{pos}$ where the general contrastive loss is usually interpreted as a lower bound on the mutual information.

For example, Tian et al. 2019 show that

$$I(z_i; z_j) \geq \log(k) - \mathcal{L}_{contrastive}$$

However, this motivation has proven to have issues both from a theoretical (McAllester & Statos 2020) and empirical (Tschannen et al. 2019) perspective.

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

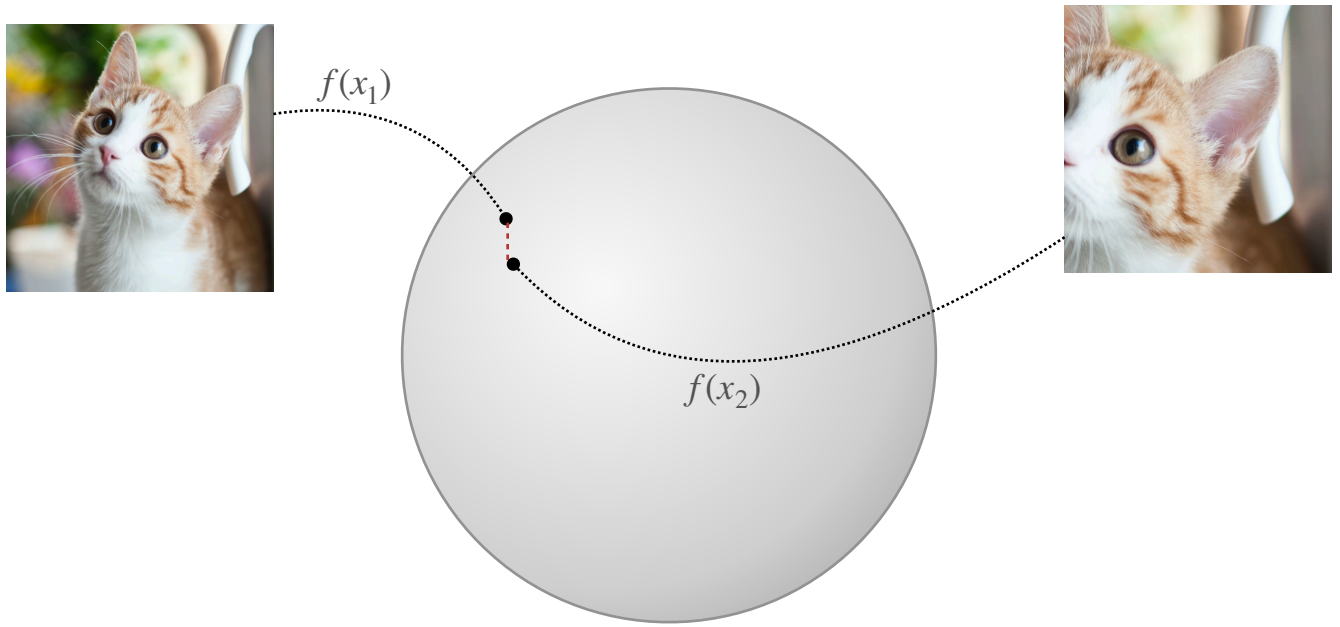
A more recent work from Wang and Isola backed contrastive learning with a different theoretical explanation. In their work, they identified two main properties of the embeddings that might be the origin of the good performances of contrastive learning frameworks:

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

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

Alignment: measures the noise-invariance property, straightforwardly defined as

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



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

Uniformity: measures how uniformly distributed are the feature vectors in the hypersphere, defined as

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

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

Theorem 1 (*asymptotic of the contrastive loss*)

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

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

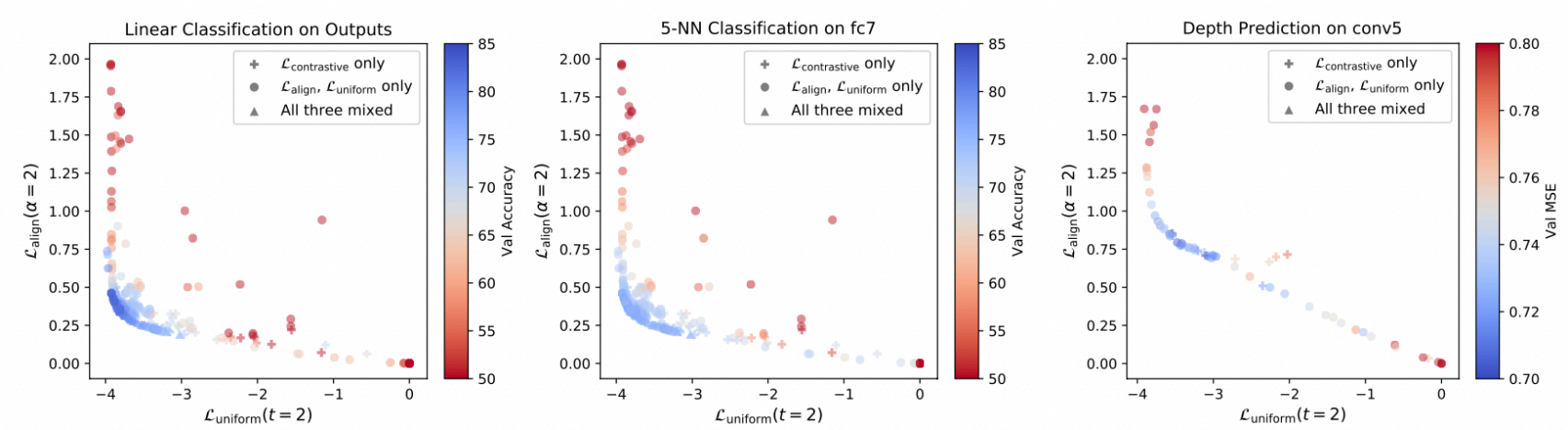
We have the following results:

1. the first term is minimized iff is perfectly aligned
2. if perfectly uniform encoders exist, they form the exact minimizers of the second term
3. for the convergence, the absolute deviation from the limit decays in $\mathcal{O}(M^{-1/2})$

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

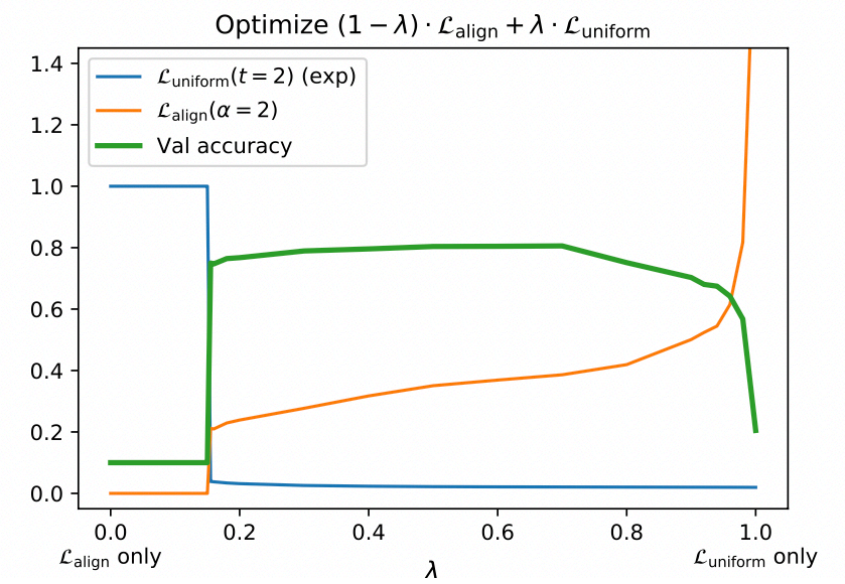


Understanding Contrastive Representation Learning through Alignment and Uniformity on the hypersphere, Wang & Isola 2020

- alignment and uniformity are valid properties across many representation learning variants (in their paper authors experimented image and text)

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

- both alignment and uniformity are necessary for good representations

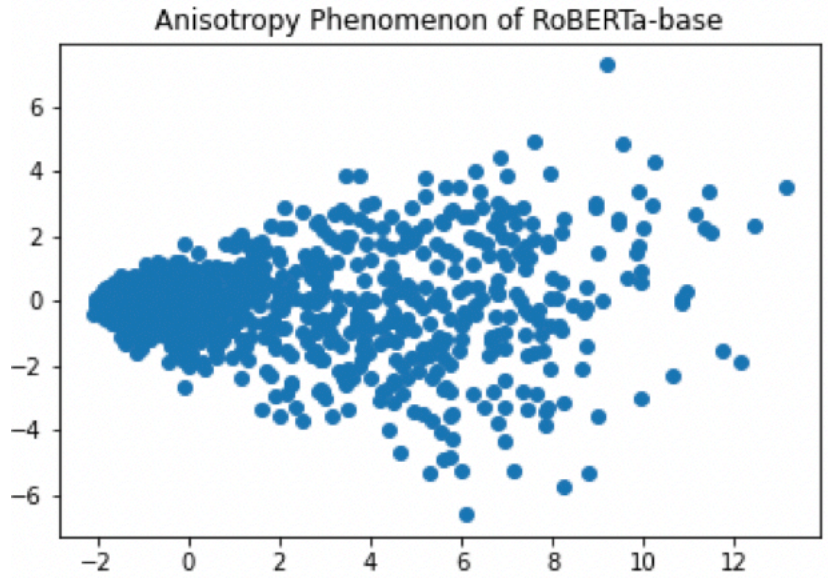


Understanding Contrastive Representation Learning through Alignment and Uniformity on the hypersphere, Wang & Isola 2020

- directly optimizing alignment and uniformity losses at the same time can lead to better results w.r.t. contrastive loss for limited number of negative samples

Anisotropy

Recent works (Ethayarajh 2019; Li et al. 2020) have identified anisotropy to be a problem in language representations: learned embeddings occupy only a narrow cone in the vector space.



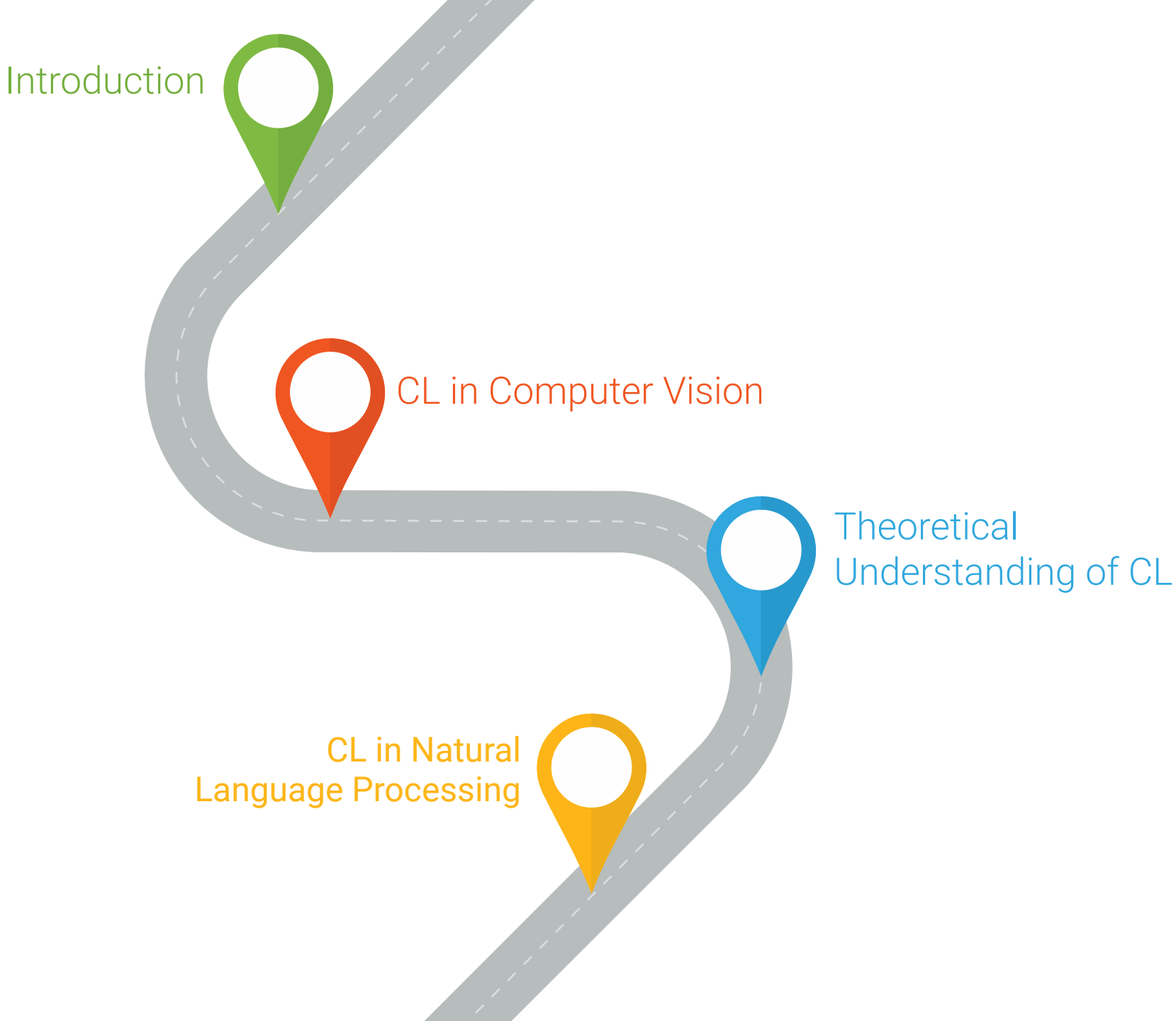
Analyzing the Anisotropy Phenomenon in Transformer-based Masked Language Models, Luo 2021

Anisotropy

Wang et al. (2020) further show that the anisotropy problem is related to the drastic decay on the singular values of word embedding matrices in language models.

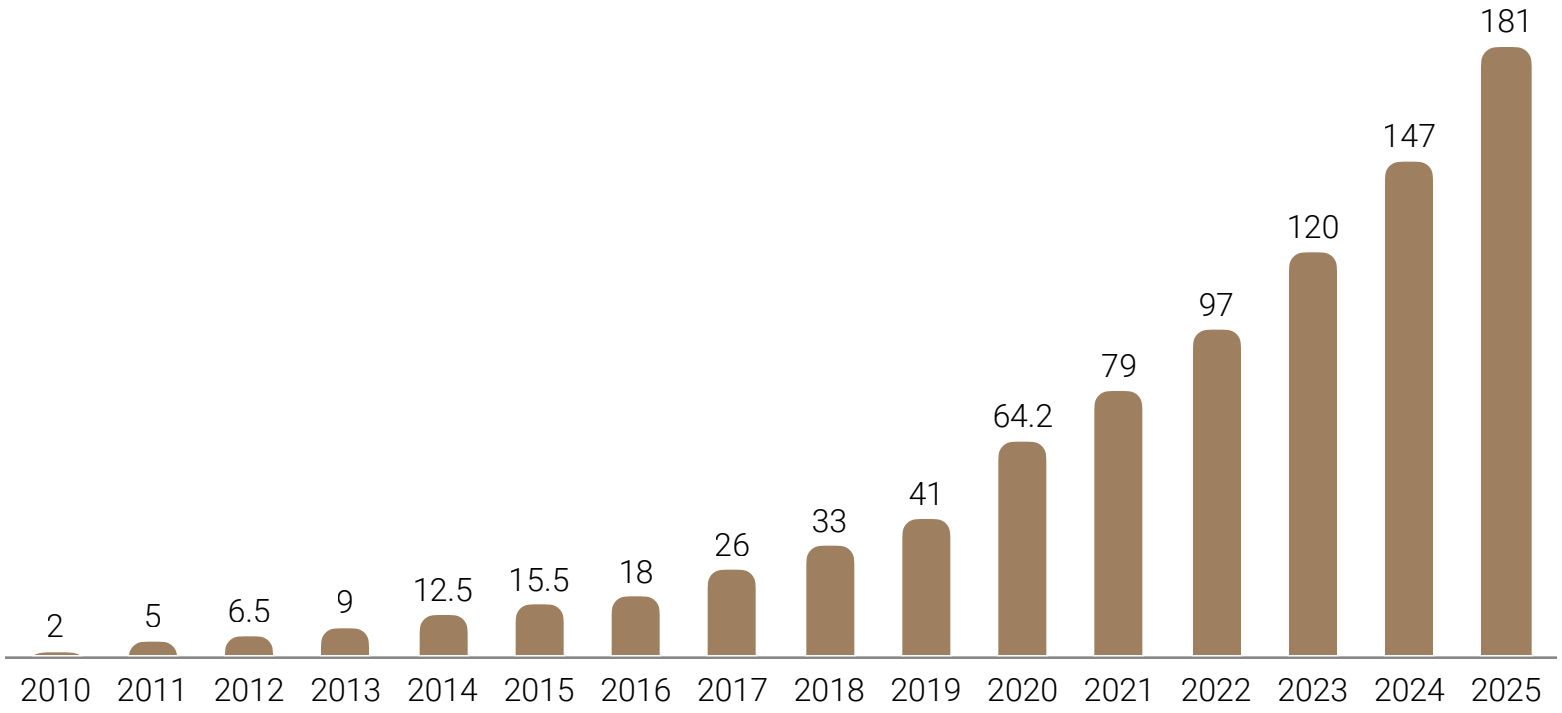
This problem is intuitively correlated to the uniformity property identified by Wang and Isola, since they’re both concerned with the distribution of the embeddings in the latent space.

The authors derive that, when the contrastive loss is optimized, the top eigenvalues of the embedding matrix are reduced, flattening the singular spectrum of the embedding space. As a result, contrastive learning is expected to alleviate the anisotropy problem and improve the uniformity of the embeddings distribution in the latent space.


Introduction
CL in Computer Vision
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 (e.g. language models have existed since the ‘90s) 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
- sentence order prediction

What about **Contrastive Learning**?

Contrastive Learning in Natural Language Processing

Still not very established, but recent success of Contrastive Learning in other fields (e.g. image representation in Computer Vision) stimulated research about it.

Main problem to tackle: data augmentation is challenging in NLP, and preserving the semantic meaning of the sample after a transformation is often a non-trivial task.

Main approaches used for augmentation

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

SimCSE (Gao et al. 2021)

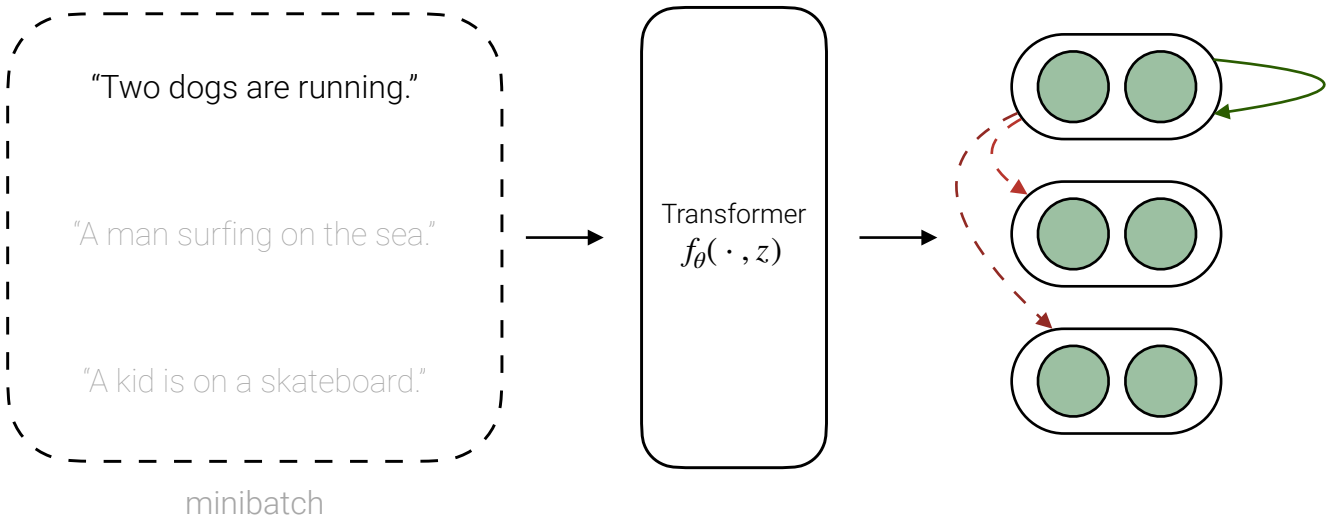
Simple contrastive learning framework to learn sentence embeddings, using **dropout** as augmentation technique during training.

Four main contributions:

- self-supervised contrastive framework for sentence embedding
- supervised contrastive framework for NLI sentence embedding
- contrastive loss as a solution to anisotropy
- evaluation of produced embeddings

SimCSE (Gao et al. 2021)

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



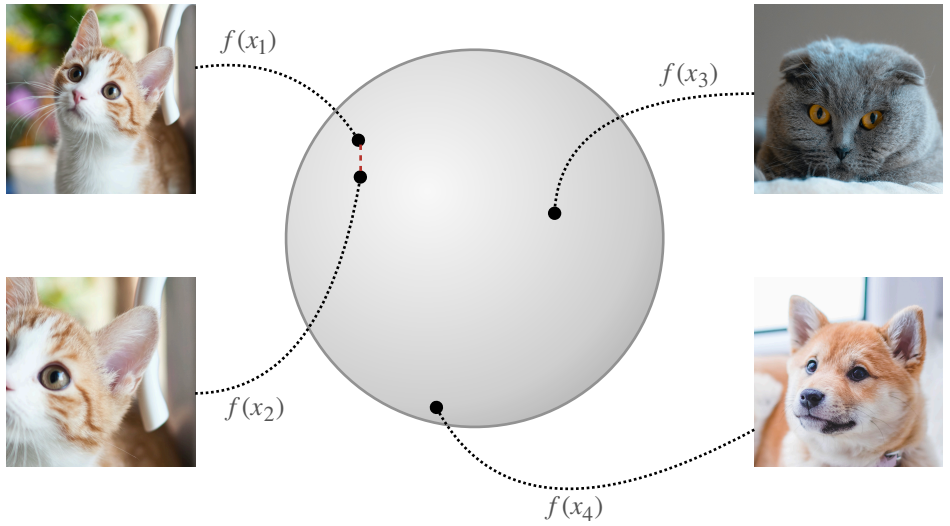
The loss used to train the model is a classic contrastive loss

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

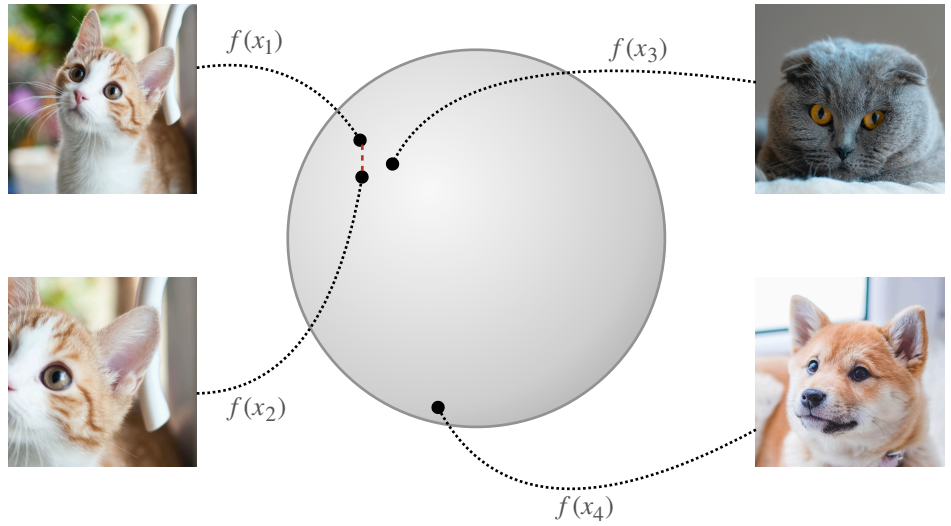
Brief detour: Supervised Contrastive Learning

So far, we have only dealt with self-supervised contrastive learning. However, other formulations of contrastive losses are possible.

A supervised approach to Contrastive Learning can be used when we want to fine-tune embeddings for a specific downstream task.



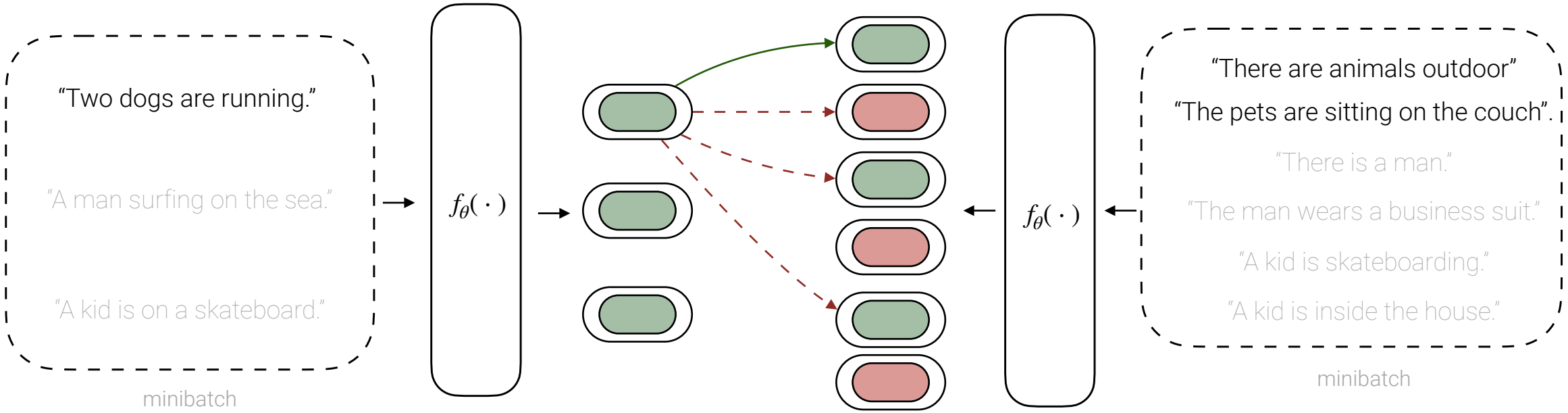
Unsupervised setting



Supervised setting

SimCSE (Gao et al. 2021)

The authors also experiment using a **supervised learning** framework using NLI datasets for training, considering triplets (x_i, x_i^+, x_i^-) where the positive and negative samples are the labeled entailed and contradicting hypothesis respectively.



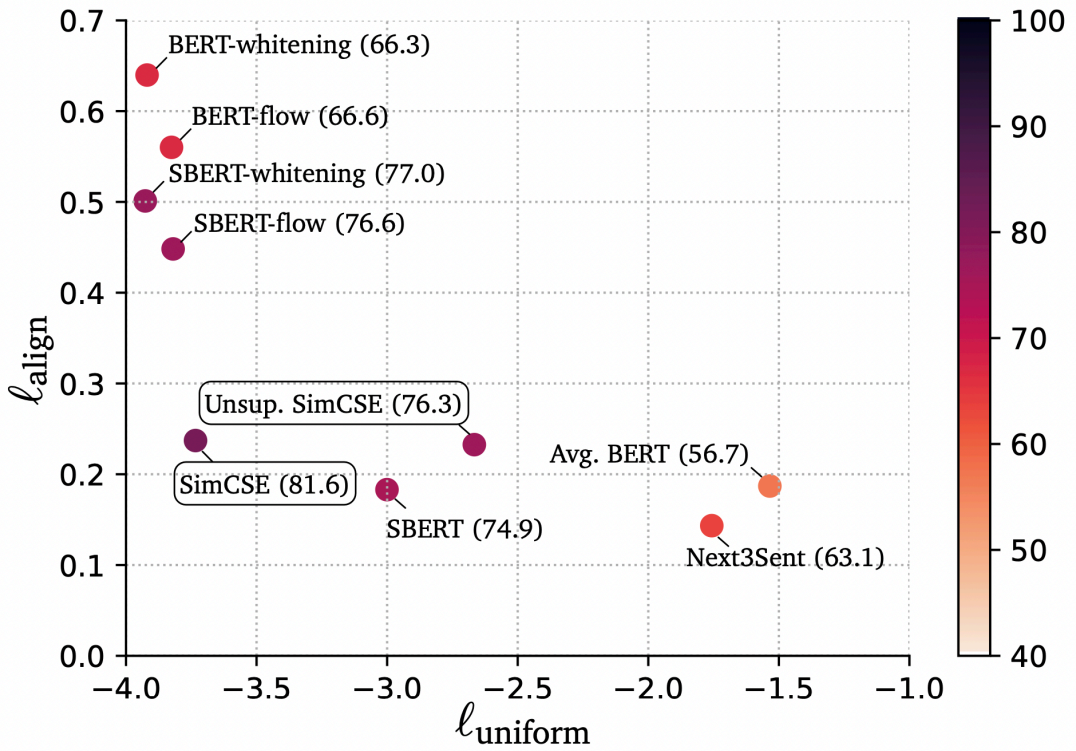
SimCSE (Gao et al. 2021)

The loss used in this case to train the model is quite similar to the self-supervised loss and is defined as

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

SimCSE (Gao et al. 2021)

Finally, the authors **evaluate** the quality of learned embeddings. They compare the proposed models (supervised and unsupervised) to other sentence embedders, using as comparison the alignment and uniformity metrics defined by Wang and Isola



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

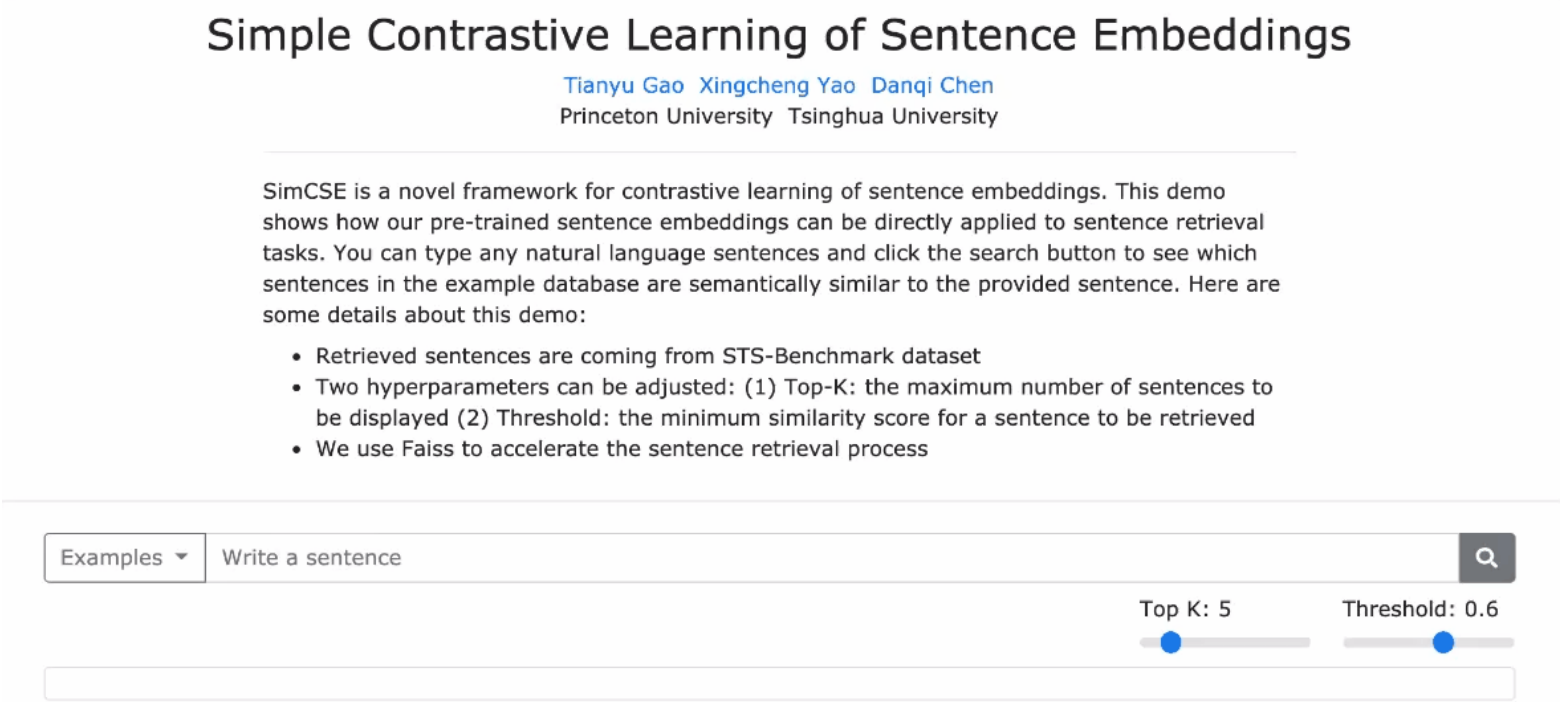
SimCSE (Gao et al. 2021)

Finally, the authors evaluate both the unsupervised and the supervised frameworks on intrinsic and extrinsic tasks, achieving state-of-the-art results in most of the benchmarks.

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

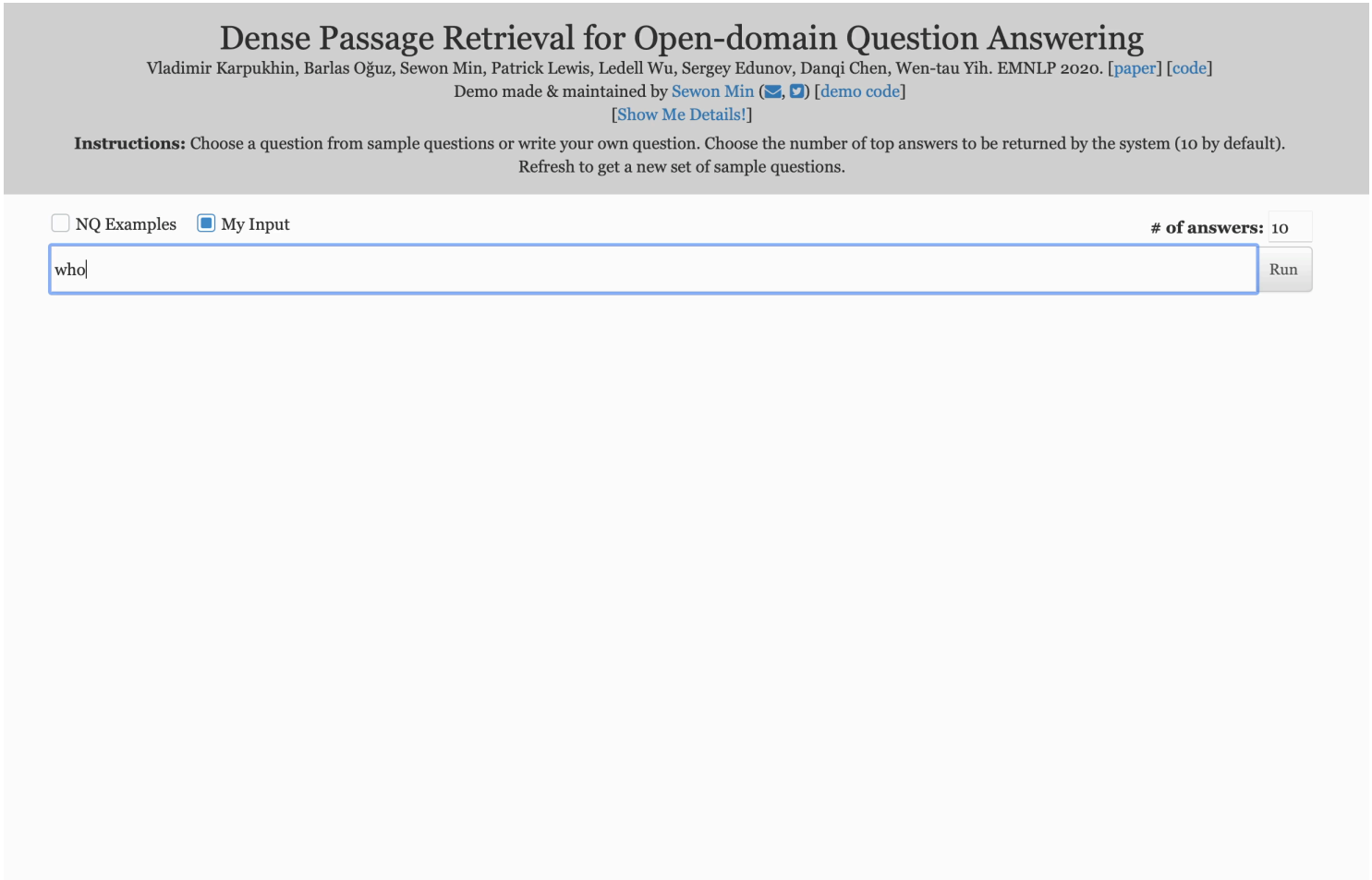
Extrinsic evaluation: 7 transfer tasks (Movie Review, Custom Review, Subjectivity Summarization and others), training a logistic classifier on top of (frozen) sentence embeddings

SimCSE (Gao et al. 2021)



Interactive demo from Gao et al., Source: https://github.com/princeton-nlp/SimCSE/blob/main/figure/demo.gif

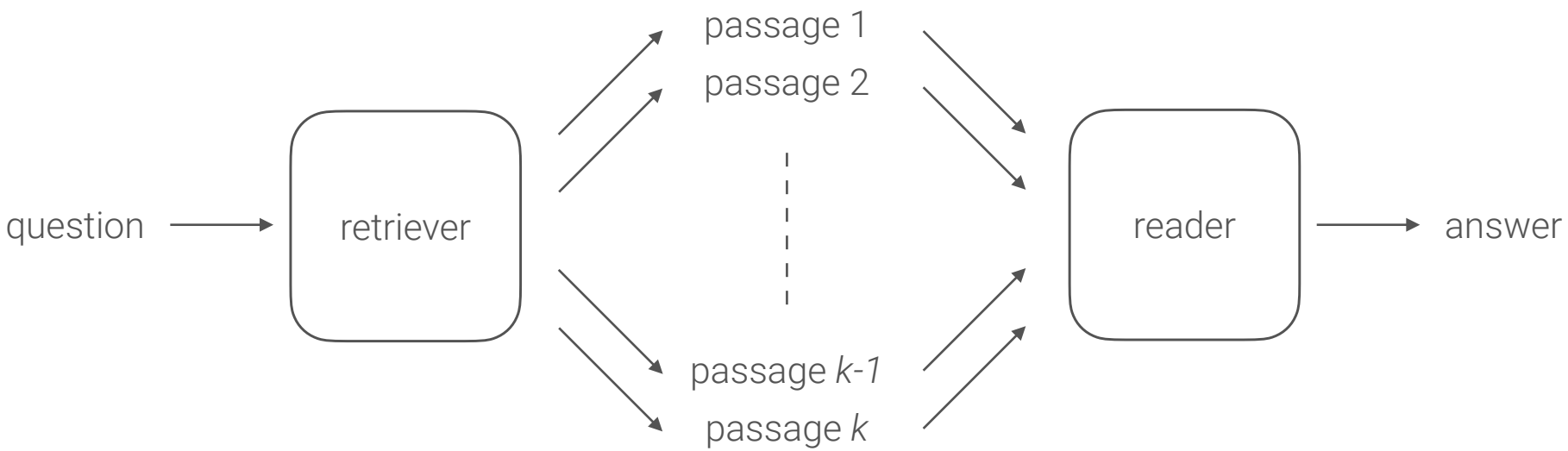
DPR (Karpukhin et al. 2020)



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

DPR (Karpukhin et al. 2020)

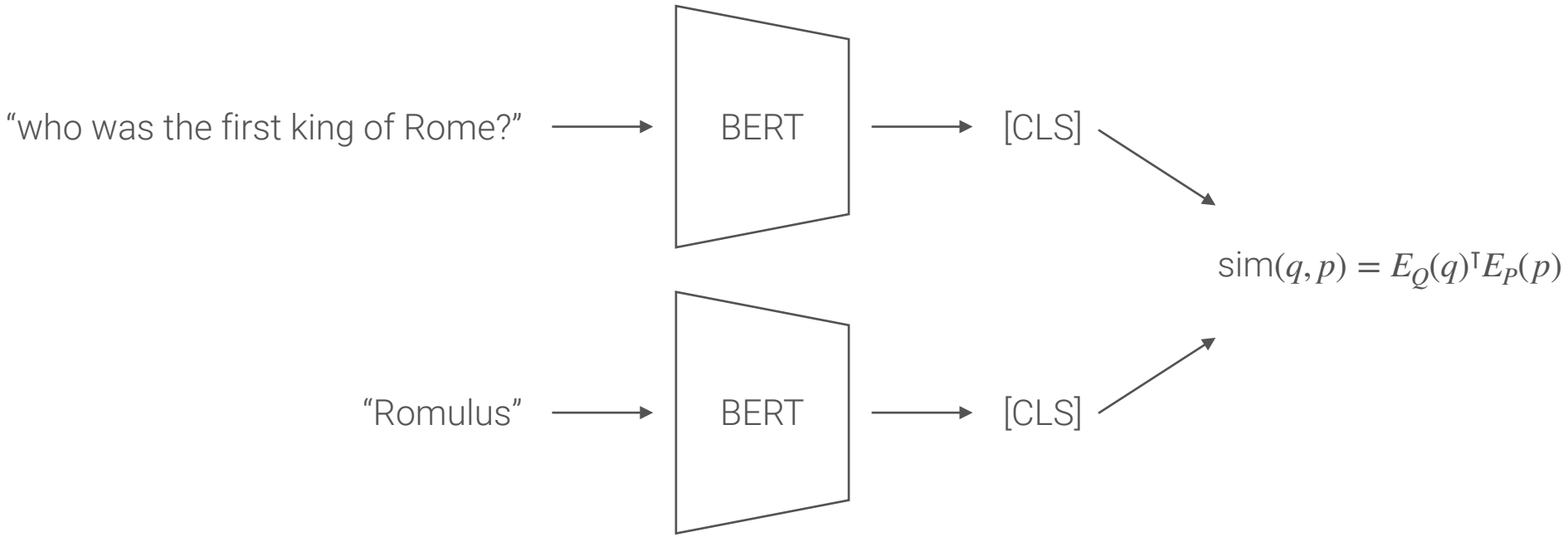
Open-domain question answering (QA) is a task that answers factoid questions using knowledge learnt from a large collection of documents.



DPR (Karpukhin et al. 2020)

Dense Passage Retriever (DPR) is a passage retriever based on contrastive learning. In this setting, positive pairs are no longer augmented version of the same entity, but *(question, answer)* pairs

$$\mathcal{L}(q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,n}^-) = -\log \frac{\exp\left(\text{sim}\left(q_i, p_i^+\right)\right)}{\exp\left(\text{sim}\left(q_i, p_i^+\right)\right) + \sum_{j=1}^{n} \exp\left(\text{sim}\left(q_i, p_{i,j}^-\right)\right)}$$



DPR (Karpukhin et al. 2020)

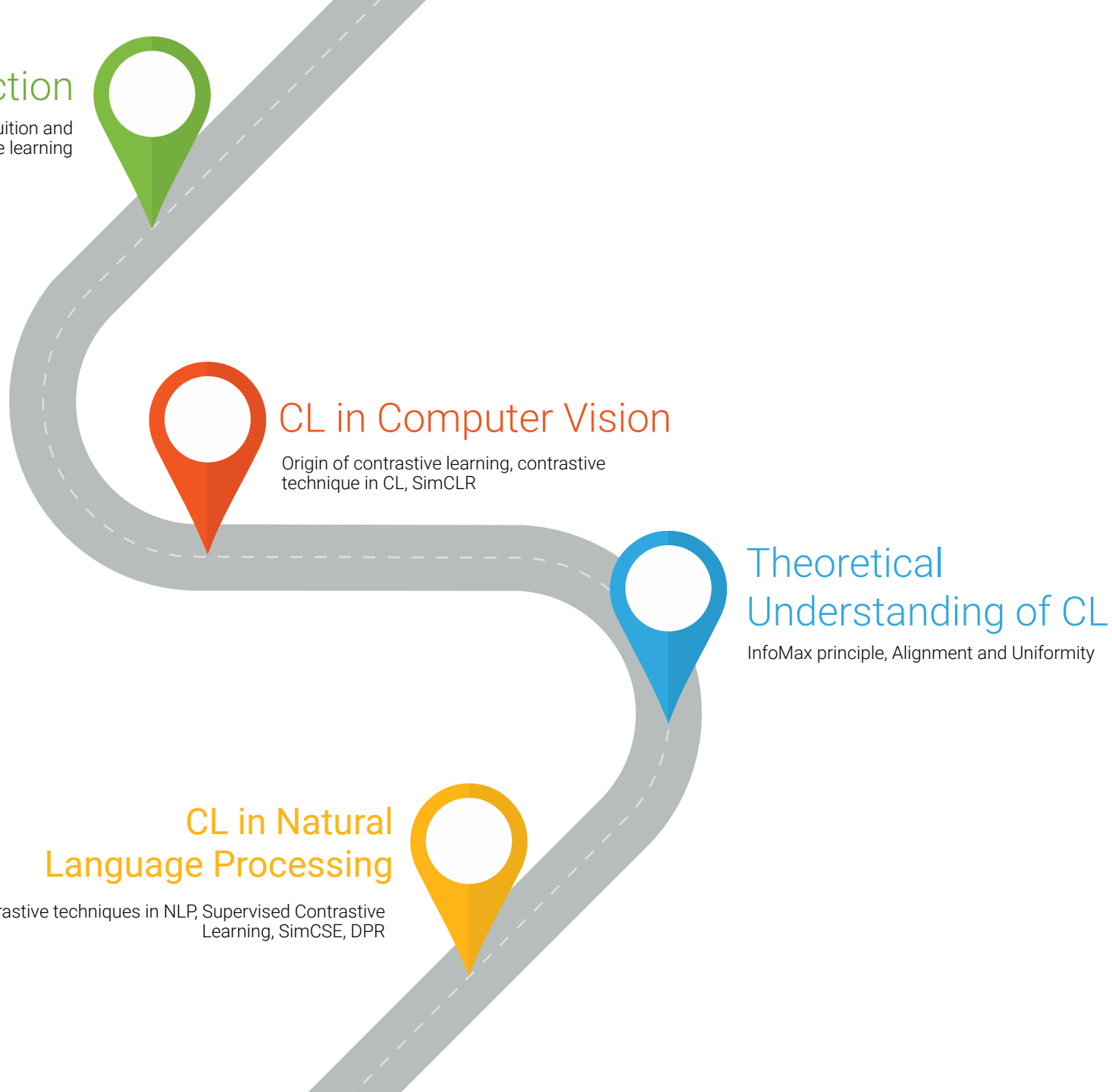
Different techniques to choose negative samples explored

- standard 1-to-N training setting
- in-batch negative sampling (memory efficient, allows more negative samples)
- in-batch negative sampling + additional hard negative for BM25

DPR (Karpukhin et al. 2020)

DPR outperform state-of-the-art BM25 on both top-*k* accuracy (intrinsic evaluation) and leads to improvements for the downstream task of end-to-end exact match (extrinsic evaluation) on 4 out of 5 chosen training sets.

Also, it's worth to mention that with the help of FAISS in-memory index DPR can be made incredibly efficient during inference time, more than 4 times faster than BM25. However, sparse representations such that BM25 are way less time expensive to train w.r.t. dense representations.


Introduction
Self-supervised learning, intuition and motivation about contrastive learning
CL in Computer Vision
Origin of contrastive learning, contrastive technique in CL, SimCLR
Theoretical Understanding of CL
InfoMax principle, Alignment and Uniformity
CL in Natural Language Processing
Contrastive techniques in NLP, Supervised Contrastive Learning, SimCSE, DPR

Thanks for your attention!

References

- Lilian Weng's blog (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

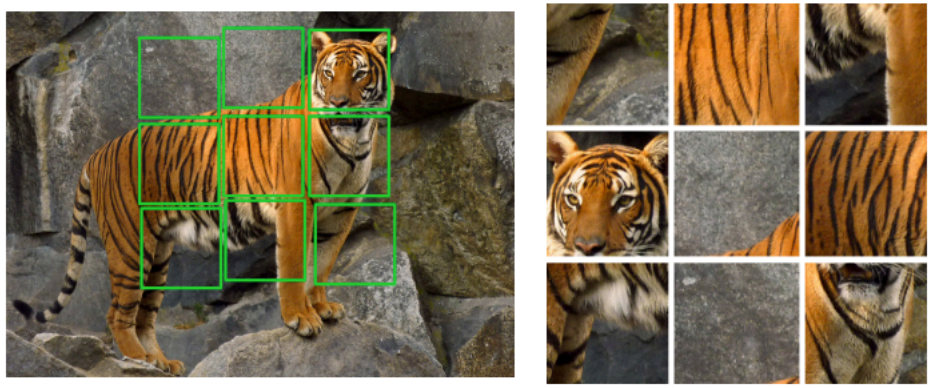
- Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu and Weizhu Chen, # A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation, 2020
- Jason Wei and Kai Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, 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

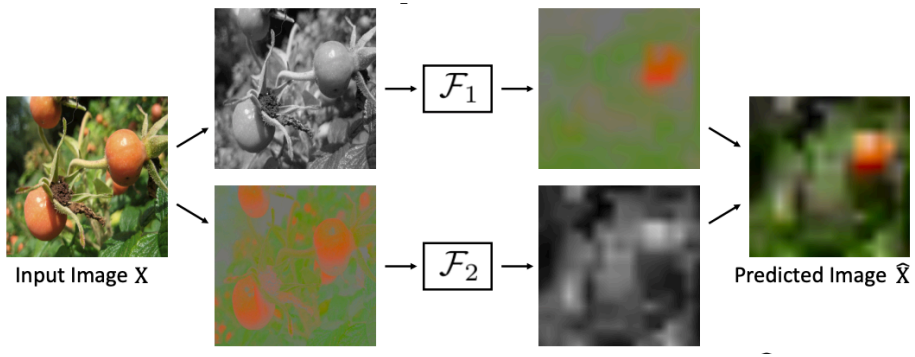
Previous Self-Supervised Learning approaches (courtesy of Yonglong Tian)

Context (Doersch et al. 2015)

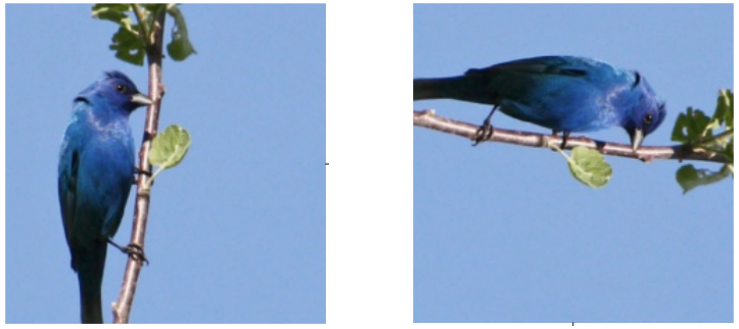
JigSaw Puzzle (Noroozi et al. 2016)



Colorization (Zhang et al. 2016, 2017)



Rotation Prediction (Gidaris et al. 2018)

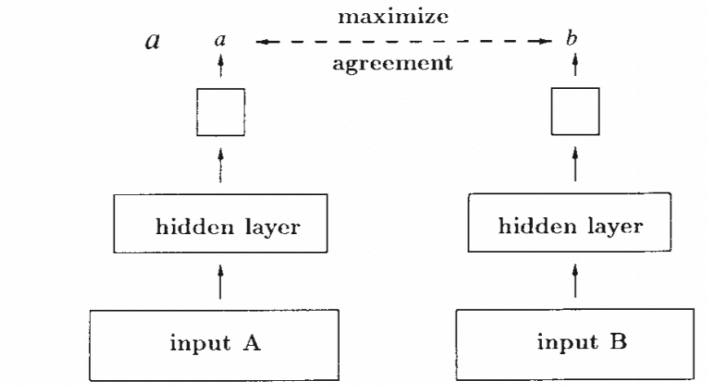


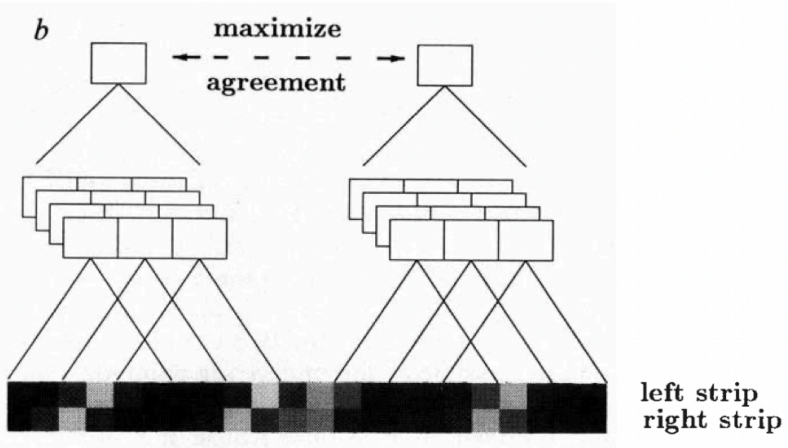
Self-organizing neural network that discovers surfaces in random-dot stereograms (Becker & Hinton 1992)

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.

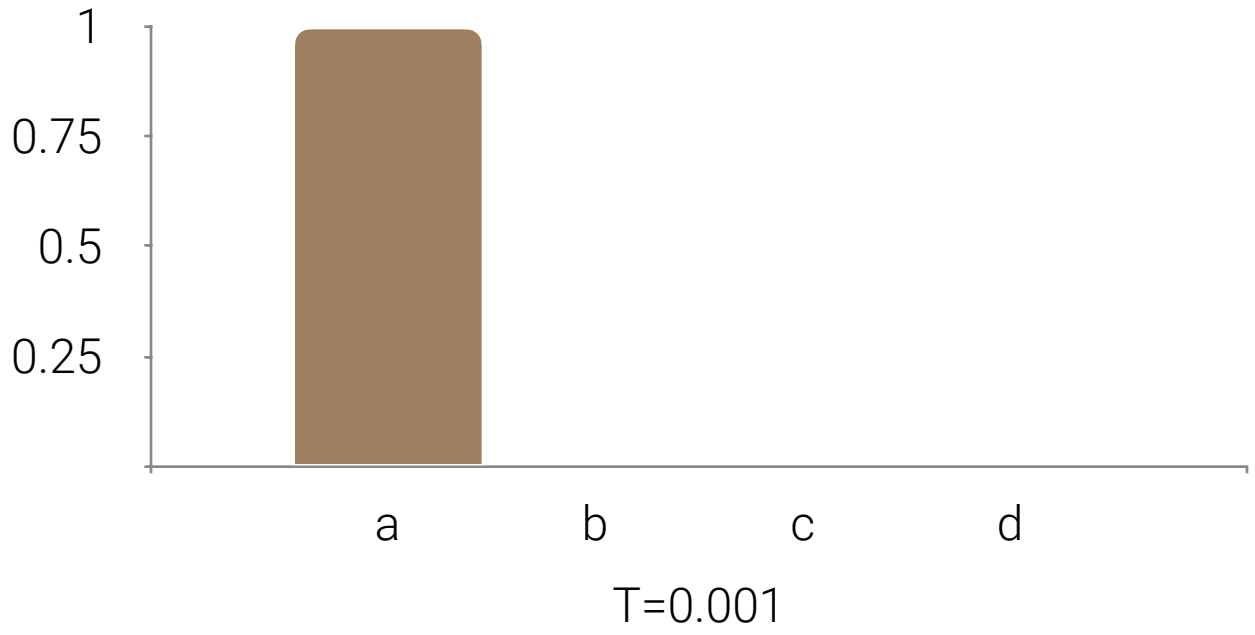




Softmax Temperature

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



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.

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 $WW^\intercal$.