NAACL 2022

PennState

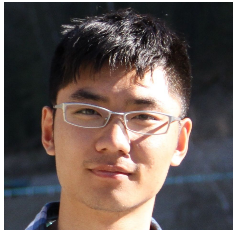
University of Virginia

西湖大學 WESTLAKE UNIVERSITY

Contrastive Data and Learning for Natural Language Processing

NAACL 2022 Tutorial, July 10, 2022

https://contrastive-nlp-tutorial.github.io/



Rui Zhang
Penn State University

Yangfeng Ji
University of Virginia

Yue Zhang
Westlake University

Rebecca J. Passonneau
Penn State University

Tutorial Website

Our tutorial materials are available at https://contrastive-nlp-tutorial.github.io/

Contrastive Data and Learning for Natural Language Processing

Tutorial at NAACL 2022 at Seattle, WA. July 10 - July 15, 2022

Tutorial Time and Location

Tutorial: 2:00pm-5:30pm PDT, July 10, 2022
Zoom Q&A sessions: 6:00pm - 6:45pm PDT, July 10, 2022

Tutorial Materials

1. Tutorial abstract in the conference proceeding [PDF]
2. Tutorial slides [slides]
3. Tutorial video [video]
4. Paper reading list of constrastive learning for NLP [Github]

Paper List

A comprehensive paper list for Contrastive Learning for NLP [github]

README.md

Contrastive Learning for Natural Language Processing

Current NLP models heavily rely on effective representation learning algorithms. Contrastive learning is one such technique to learn an embedding space such that similar data sample pairs have close representations while dissimilar samples stay far apart from each other. It can be used in supervised or unsupervised settings using different loss functions to produce task-specific or general-purpose representations. While it has originally enabled the success for vision tasks, recent years have seen a growing number of publications in contrastive NLP. This first line of works not only delivers promising performance improvements in various NLP tasks, but also provides desired characteristics such as task-agnostic sentence representation, faithful text generation, data-efficient learning in zero-shot and few-shot settings, interpretability and explainability.

Participation + Q&A

Questions are welcomed during the tutorial!

T6: Contrastive Data and Learning for Natural Language Processing

Date: Sunday, July 10, 2022
Time:
Part 1 14:00-15:30 Pacific Daylight Time
Coffee break 15:30-16:00 Pacific Daylight Time
Part 2 16:00-17:30 Pacific Daylight Time

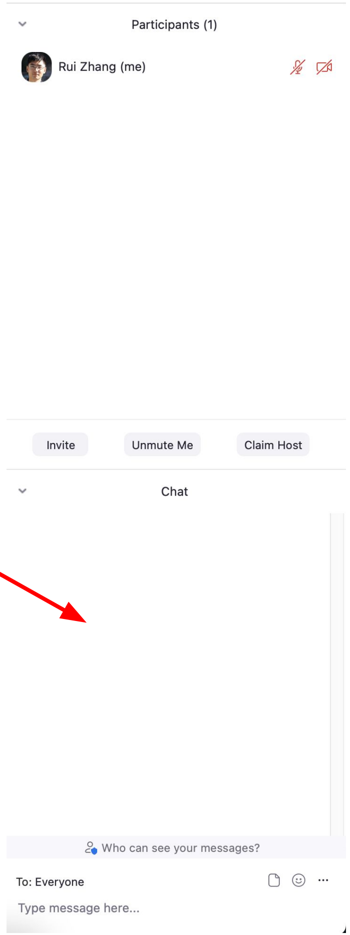
Ask Question in Chat

This is a hybrid session.
This session will take place in **Columbia A** (In-person) and in **Zoom** (Remote). Please click on button to join the session.
Please use the same link for Part 1 and Part 2!

Join Zoom Room

Join Zoom Here

There will be two Tutorial 6 Q'n'A sessions.
Go to Tutorial 6 Q&A session 1
Go to Tutorial 6 Q'n'A session 2



Two Zoom Q&A sessions: 13:30-14:00   18:00-18:45 PDT, July 10, 2022

Contrastive Learning

Learning embeddings such that similar data sample pairs are close while dissimilar sample pairs stay far apart (Chopra et al., 2005)

$$\mathrm{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^+)) \gg \mathrm{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^-))$$

f : encoder, e.g., neural networks

sim : similarity measure, e.g., inner product

$\boldsymbol{x}$: anchor

$\boldsymbol{x}^+$: positive example

$\boldsymbol{x}^-$: negative example

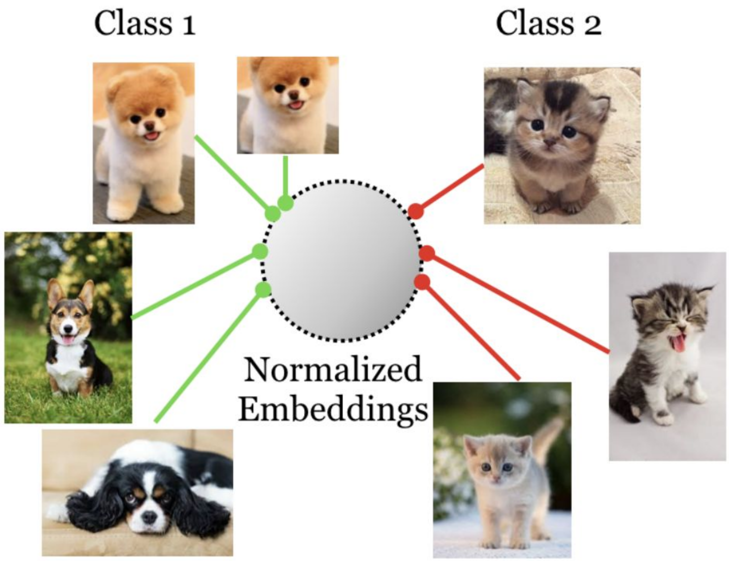
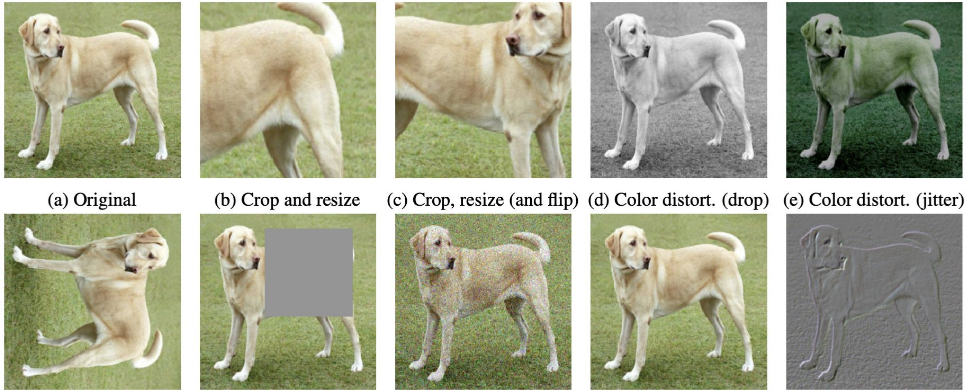


Figure from Khosla et al., 2020

Contrastive Learning in Computer Vision

SimCLR (Chen et al., 2020)



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Contrastive Learning for NLP

(Smith and Eisner, 2005): The first NLP paper introducing “contrastive estimation” as an unsupervised training objective for log-linear models.

Contrastive Estimation: Training Log-Linear Models on Unlabeled Data*

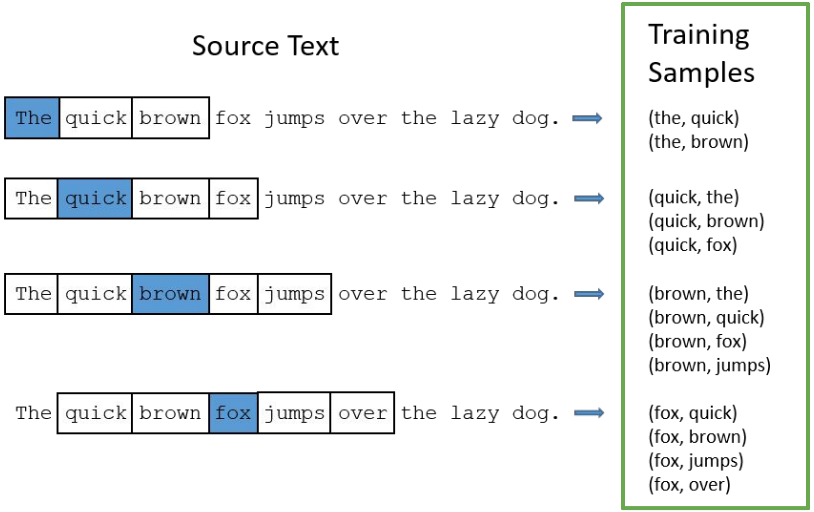
Noah A. Smith and **Jason Eisner**
Department of Computer Science / Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218 USA

$$\prod_i p\left(X_i = x_i \mid X_i \in \boxed{\mathcal{N}(x_i)}, \vec{\theta}\right)$$

“neighborhood” N(xi) is a set of implicit negative examples plus the example xi itself.

Most Successful Example of Contrastive Learning for NLP

word2vec (Mikolov et al., 2013) for word embeddings



word2vec's skip-gram model. Figure from Chris McCormick

$$\boxed{\log \sigma({v'_{w_O}}^\top v_{w_I})} + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \boxed{\left[\log \sigma(-{v'_{w_i}}^\top v_{w_I})\right]}$$

f : word embeddings

sim : inner product

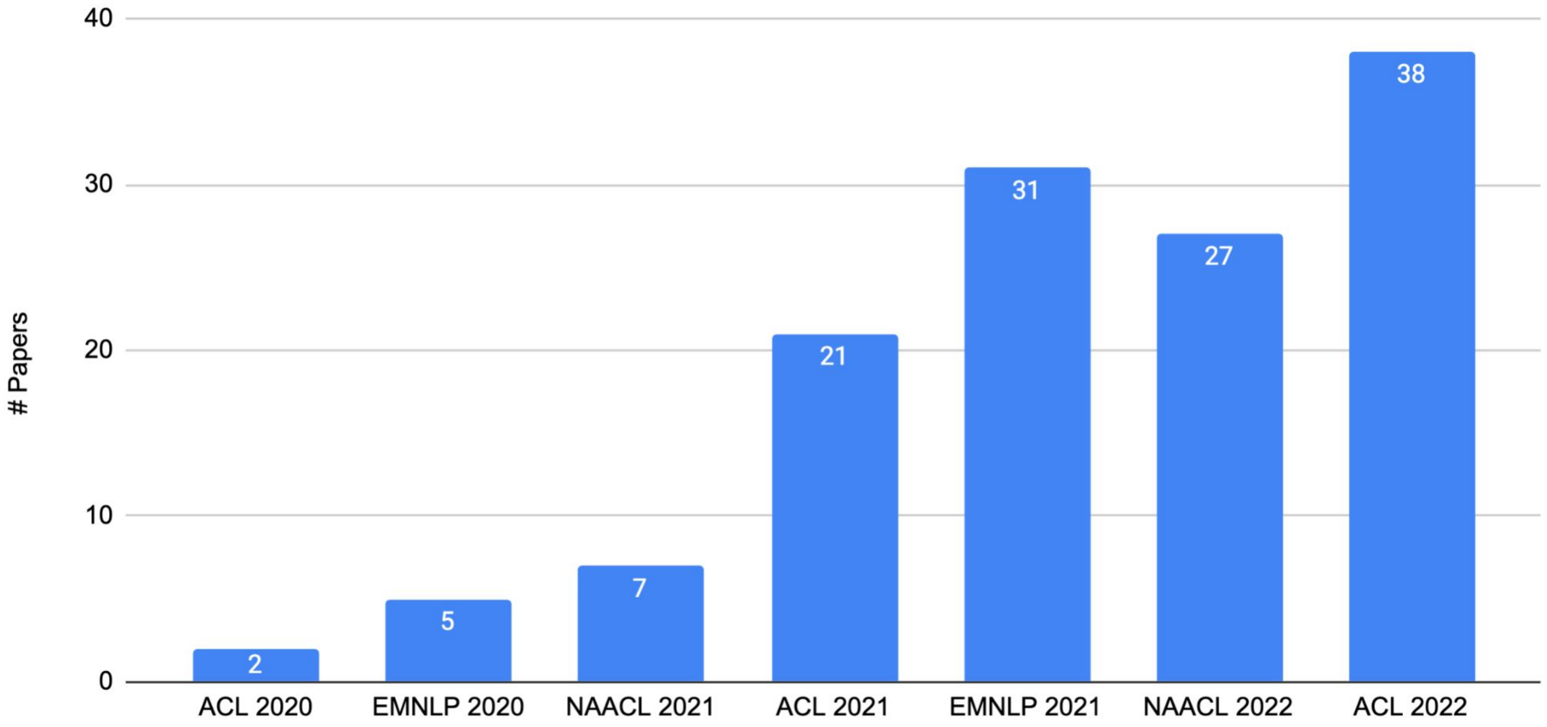
$\boldsymbol{x}$: current word

$\boldsymbol{x}^+$: context word

$\boldsymbol{x}^-$: random word by negative sampling

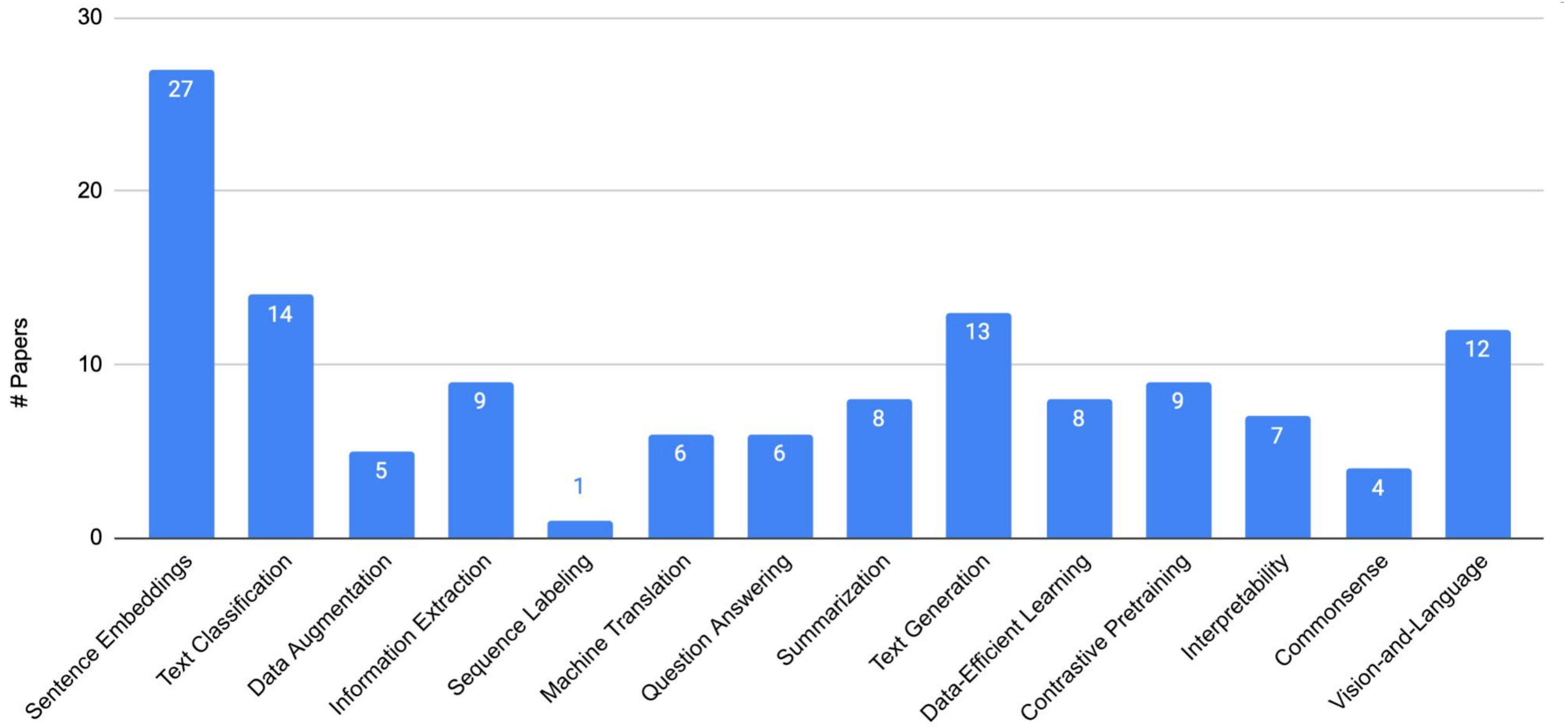
Why give this tutorial today?

Number of papers with titles containing “contrastive learning” in recent NLP conferences



Why give this tutorial today?

Number of papers with titles containing “contrastive learning” in recent NLP conferences



Why give this tutorial today?

- word embeddings ⇨ sentence representations ⇨ various tasks.
  - Classification: Text Classification, Information Extraction
  - Reasoning: Commonsense Reasoning, Question Answering, Fact Verification
  - Generation: Summarization, Machine Translation, Text Generation
  - Multimodal Learning: Vision-and-Language

- performance improvements ⇨ desired characteristics
  - Task-agnostic Sentence Representation
  - Data-efficient Learning in Zero-shot and Few-shot settings
  - Interpretability and Robustness
  - Faithful Text Generation

Agenda

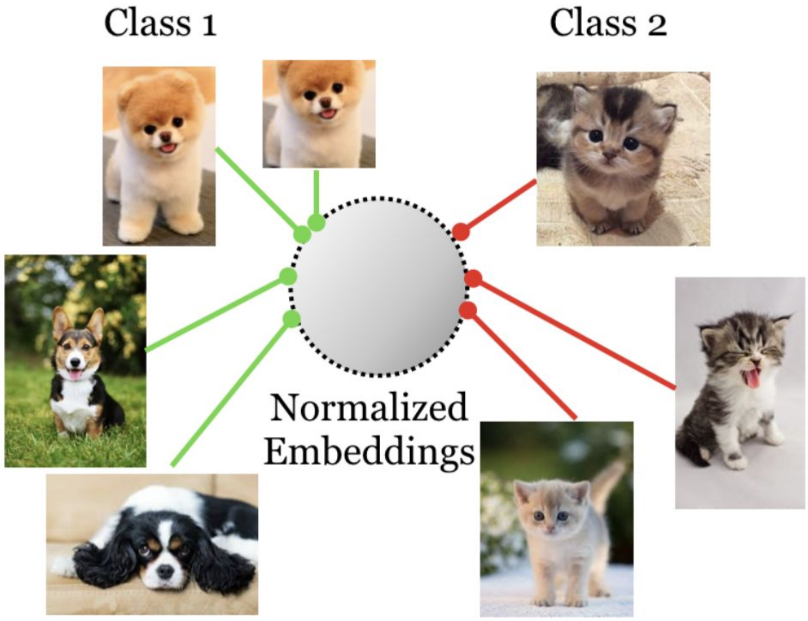
- Part 0: Introduction (Becky, Rui)
- Part 1: Foundations of Contrastive Learning (Rui)
- Part 2: Contrastive Learning for NLP (Yangfeng and Yue)
- Part 3: Summary and Reflection (Yue)

Part 1.

Foundations of Contrastive Learning

What is Contrastive Learning

Learning embeddings such that similar data sample pairs are close while dissimilar sample pairs stay far apart (Chopra et al., 2005)



$$\text{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^+)) \gg \text{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^-))$$

f : encoder, e.g., neural networks

sim : similarity measure, e.g., inner product

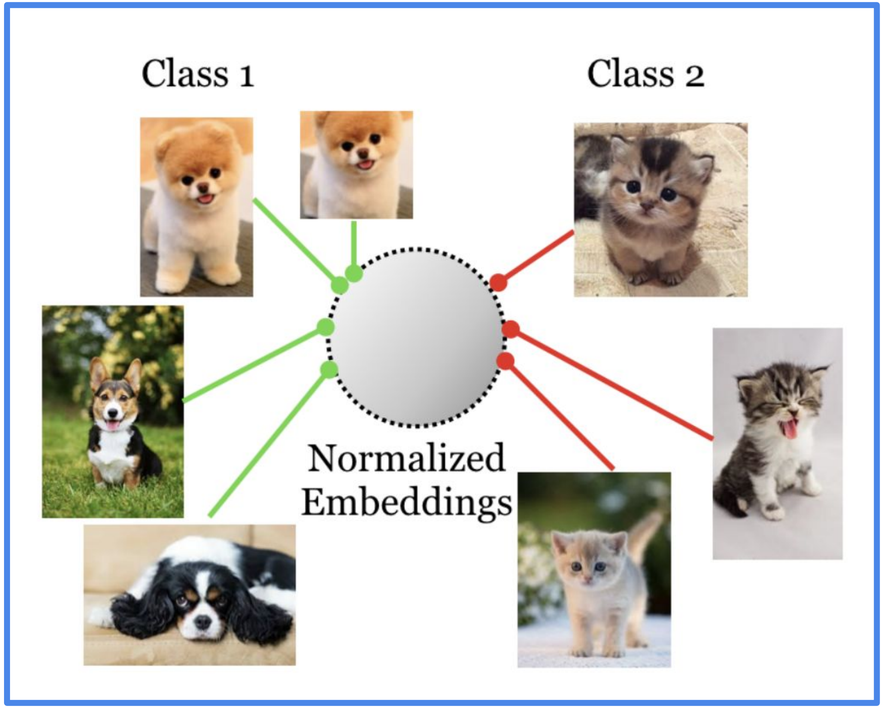
$\boldsymbol{x}$: anchor

$\boldsymbol{x}^+$: positive example

$\boldsymbol{x}^-$: negative example

Two Elements of Contrastive Learning

Contrastive Learning = Contrastive Data Creation + Contrastive Objective Optimization



$$\mathrm{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^+)) \gg \mathrm{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^-))$$

f : encoder, e.g., neural networks

sim : similarity measure, e.g., inner product

$\boldsymbol{x}$: anchor

$\boldsymbol{x}^+$: positive example

$\boldsymbol{x}^-$: negative example

Outline

Part 1. Foundations of Contrastive Learning

- Part 1.1 Contrastive Learning Objectives
- Part 1.2 Contrastive Data Sampling and Augmentation Strategies
- Part 1.3 Analysis of Contrastive Learning

Part 1.1

Contrastive Learning Objectives

Contrastive Learning Objectives

Contrastive Learning = Contrastive Data Creation + **Contrastive Objective Optimization**

$$\boxed{\mathrm{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^{+})) \gg \mathrm{sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^{-}))}$$

- Contrastive Loss
- Triplet Loss
- N-pair Loss
- Noise Contrastive Estimation
  - NCE
  - NEG
  - InfoNCE
  - NT-Xent
- Soft-Nearest Neighbors Loss
- Lifted Structured Loss

Contrastive Loss

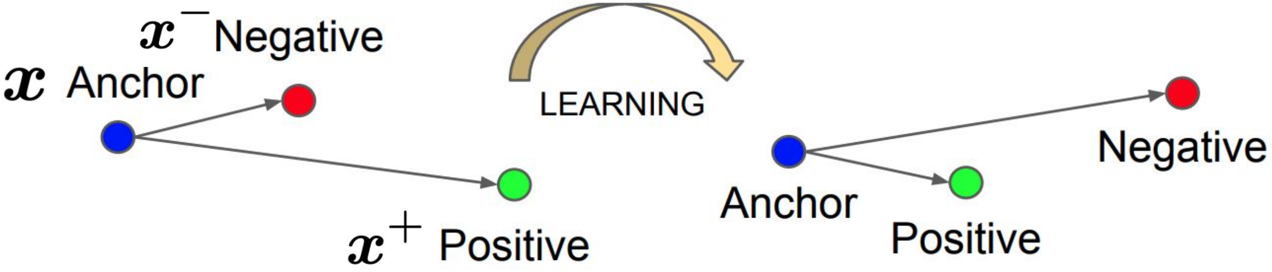
$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boxed{\mathbb{1}[(y_i = y_j)] \| f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \|^2} + \boxed{\mathbb{1}[(y_i \neq y_j)] \max(0, m - \| f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j) \|)^2}$$

minimizes the embedding distance when they are from the same class

maximizes the embedding distance when they are from the same class

[Learning a Similarity Metric Discriminatively, with Application to Face Verification (Chopra et al., 2005)](#)

Triplet Loss

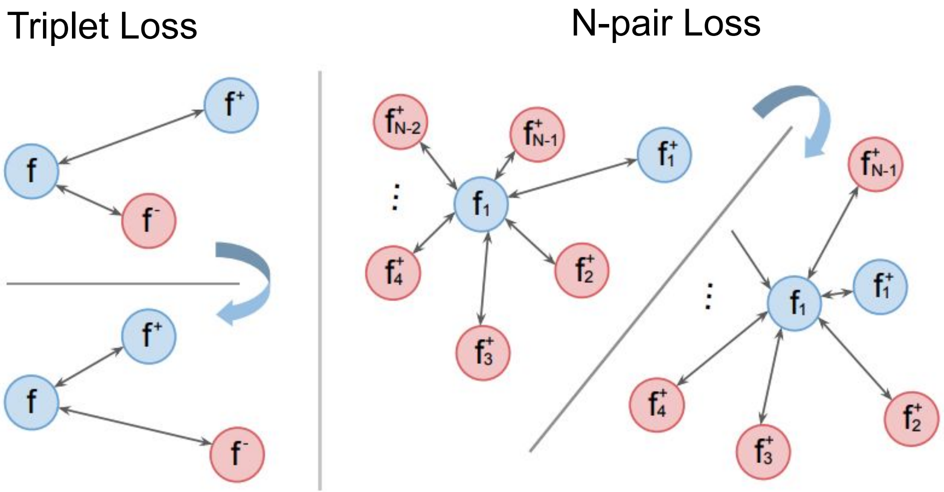


$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{x}^+, \boldsymbol{x}^-) = \max(0, \boxed{m + ||f(\boldsymbol{x}) - f(\boldsymbol{x}^+)||_2^2} - \boxed{||f(\boldsymbol{x}) - f(\boldsymbol{x}^-)||_2^2})$$

We push the the distance between positive and anchor + margin to be smaller than the distance between negative and anchor.

[FaceNet: A Unified Embedding for Face Recognition and Clustering (Schroff et al., 2015)]

N-pair Loss



$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{x}^+, \{\boldsymbol{x}_i^-\}_{i=1}^{N-1}) = \log\left(1 + \sum_{i=1}^{N-1} \exp(f(\boldsymbol{x})^\top f(\boldsymbol{x}_i^-) - f(\boldsymbol{x})^\top f(\boldsymbol{x}^+))\right)$$

- Extend to N-1 negative examples
- Inner product similarity + softmax loss
- Similar to multi-class classification

[Improved Deep Metric Learning with Multi-class N-pair Loss Objective (Sohn, 2016)]

Noise Contrastive Estimation (NCE)

Use Logistic Regression with cross-entropy loss to differentiate positive samples (i.e., target distribution) and negative samples (i.e., noise distribution).

$\ell(\boldsymbol{x})$ Logit function of a sample from the target distribution

$\sigma(\ell(\boldsymbol{x}))$ Probability a sample from the target distribution

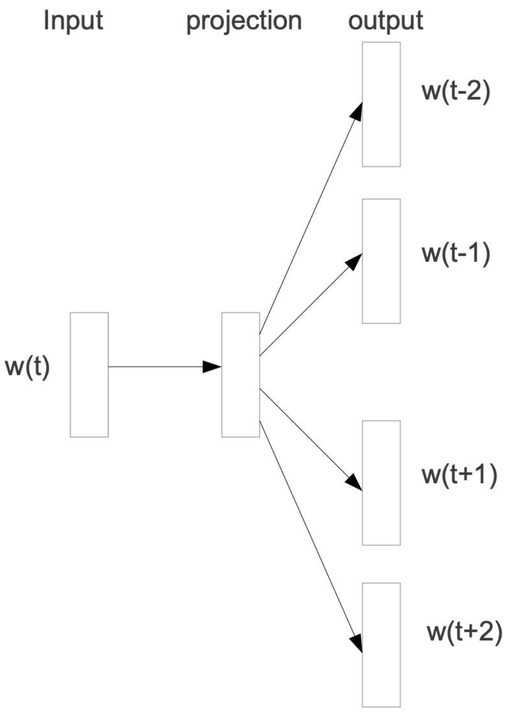
$$
\begin{aligned}
\mathcal{L}(\boldsymbol{x}^+, \boldsymbol{x}^-) &= -\left[\boxed{\log \sigma(\ell(\boldsymbol{x}^+))} + \boxed{\log(1 - \sigma(\ell(\boldsymbol{x}^-)))}\right] \\
&= -\left[\boxed{\log \sigma(\ell(\boldsymbol{x}^+))} + \boxed{\log \sigma(-\ell(\boldsymbol{x}^-))}\right]
\end{aligned}
$$

Noise-contrastive estimation: A new estimation principle for unnormalized statistical models (Gutmann and Hyvärinen, 2010)

Negative Sampling (NEG)

- A variation of NCE used in word2vec
- Logit is inner product of word embeddings
- Random words as negative sampling

$$\underbrace{\log \sigma({v'_{w_O}}^\top v_{w_I})}_{\text{log of probability of positive pairs}} + \underbrace{\sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)}}_{\text{k negative samples}} \underbrace{\left[\log \sigma(-{v'_{w_i}}^\top v_{w_I})\right]}_{\text{log of (1 - probability of negative pairs)}}$$



word2vec's skip-gram model.

Distributed Representations of Words and Phrases and their Compositionality (Mikolov et al., 2013)

InfoNCE

Use softmax loss to differentiate a positive sample from a set of noise examples.

$\boldsymbol{c}$ Context Vector, e.g., anchor point

$X = \{x_1, \ldots, x_N\}$ N samples with 1 positive sample and N-1 negative samples

$$\mathcal{L} = -\log \frac{\boxed{f(\boldsymbol{x}, \boldsymbol{c})}}{\boxed{\sum_{\boldsymbol{x}' \in X} f(\boldsymbol{x}', \boldsymbol{c})}}$$

1 positive sample

1 positive sample + N-1 negative samples

[Representation Learning with Contrastive Predictive Coding (van den Oord et al., 2018)](#)

Normalized Temperature-scaled Cross-Entropy (NT-Xent)

$$\mathcal{L} = -\log \frac{\boxed{\exp(\mathrm{sim}(\boldsymbol{x}, \boldsymbol{x}^{+})/\tau)}}{\boxed{\exp(\mathrm{sim}(\boldsymbol{x}, \boldsymbol{x}^{+})/\tau)} + \boxed{\sum_{j=1}^{N-1} \exp(\mathrm{sim}(\boldsymbol{x}, \boldsymbol{x}_j^{-})/\tau)}}$$

InfoNCE with Cosine Similarity on Normalized Embeddings

Temperature controls the relative importance of the distances between point pairs

- At low temperatures, the loss is dominated by the small distances.
- At high temperatures, the loss is dominated by the large distances.

A Simple Framework for Contrastive Learning of Visual Representations. (Chen et al., 2020)

Soft-Nearest Neighbors Loss

Extend to different numbers of positive (M) and negative examples (N).

$$\mathcal{L}(\boldsymbol{x}, \{\boldsymbol{x}_j^+\}_{j=1}^{M}, \{\boldsymbol{x}_i^-\}_{i=1}^{N}) = -\log\left(\frac{\sum_{j=1}^{M} \exp(-|f(\boldsymbol{x}) - f(\boldsymbol{x}_j^+)|^2/\tau)}{\sum_{l=1}^{M+N} \exp(-|f(\boldsymbol{x}) - f(\boldsymbol{x}_l^+)|^2/\tau)}\right)$$

[Analyzing and Improving Representations with the Soft Nearest Neighbor Loss (Frosst et al., 2019)](#)

Lifted Structured Loss

Lifted Structured Loss explicitly takes into account all pairwise edges within the batch.

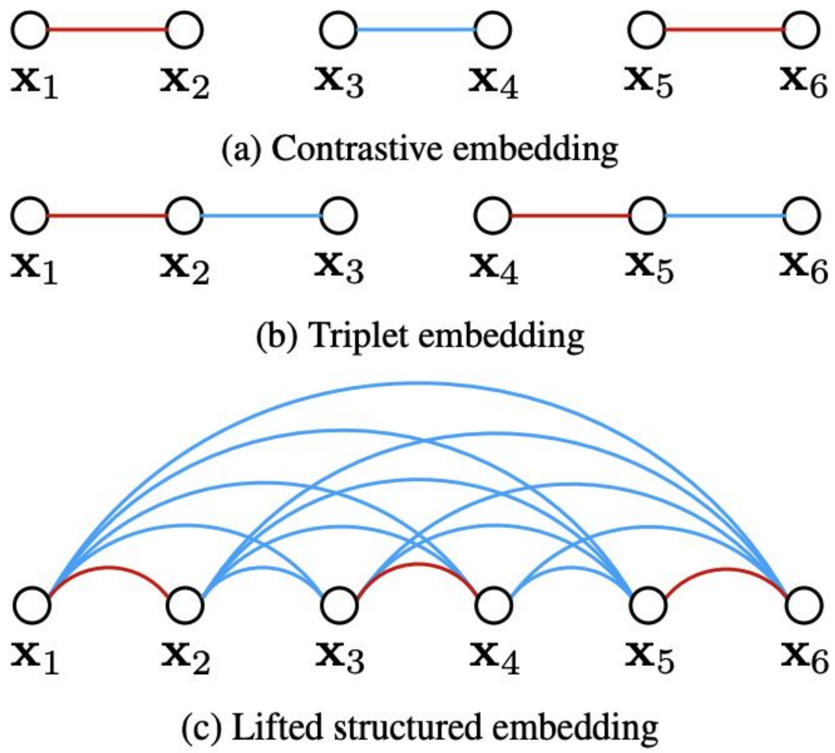
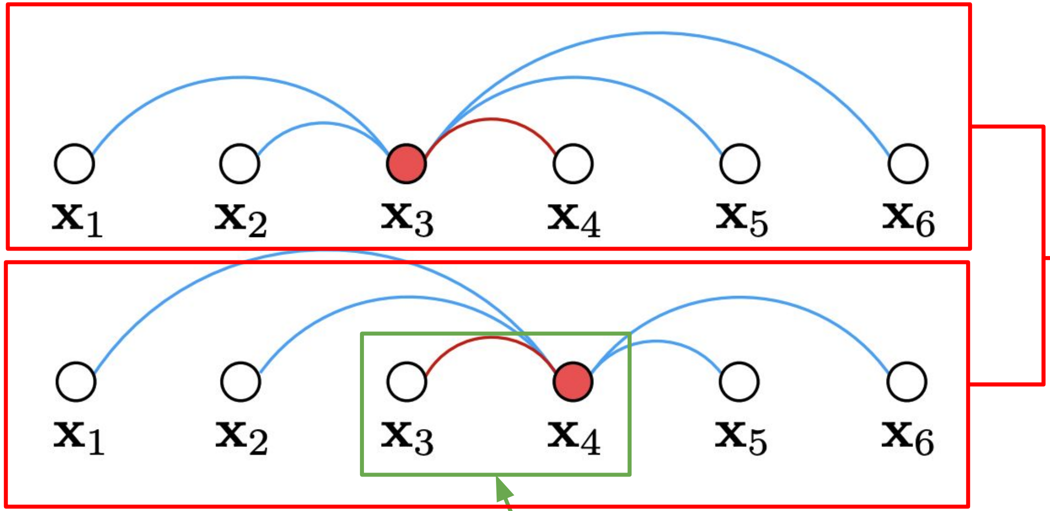


Illustration for a training batch with six examples.
Red edges: similar examples.
Blue edges: dissimilar examples.

[Deep Metric Learning via Lifted Structured Feature Embedding (Song et al., 2016)]

Lifted Structured Loss - Mining the Hardest Negative



Take max to find the hardest negative

$$d_{i,j} = |f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)|$$

$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max\left(0, d_{i,j} + \max\left(\max_{(i,k)} m - d_{i,k}, \max_{(j,l)} m - d_{j,l}\right)\right)^2$$

Lifted Structured Loss - Relaxation

$$d_{i,j} = |f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)|$$

$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max\left(0, d_{i,j} + \max\left(\max_{(i,k)} \; m - d_{i,k}, \; \max_{(j,l)} \; m - d_{j,l}\right)\right)^2$$

$\Downarrow$ Replace the second term with a smooth upper bound to ease optimization

$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max\left(0, d_{i,j} + \log\left(\sum_{(i,k)} \exp(m - d_{i,k}) + \sum_{(j,l)} \exp(m - d_{j,l})\right)\right)^2$$

Summary of Contrastive Learning Objectives

Loss Function	Paper	Contrast Unit			Number of Examples		Used In
		Pair	Triplet	Set	# of positive	# of negative	
Contrastive Loss	(Chopra et al., 2005)	✓			0/1	0/1	
Triplet Loss	(Schroff et al., 2015)		✓		1	1	
N-pair Loss	(Sohn, 2016)			✓	1	$N-1$	
NCE	(Gutmann and Hyvärinen, 2010)	✓			0/1	0/1	
Negative Sampling	(Mikolov et al., 2013)			✓	1	$N-1$	word2vec
InfoNCE	(van den Oord et al., 2018)			✓	1	$N-1$	
NT-Xent	(Chen et al., 2020)			✓	1	$N-1$	simCLR,simCSE,CLIP
Soft-Nearest Neighbors Loss	(Frosst et al., 2019)			✓	M	N	
Lifted Structured Loss	(Oh Song et al., 2016)			✓	M	N	

Part 1.2

Contrastive Data Sampling and Augmentation Strategies

Self-Supervised Contrastive Learning

Positive: Data Augmentation

Negative: Random, e.g., In-batch Negatives

The Biggest Advantage: No label is required!

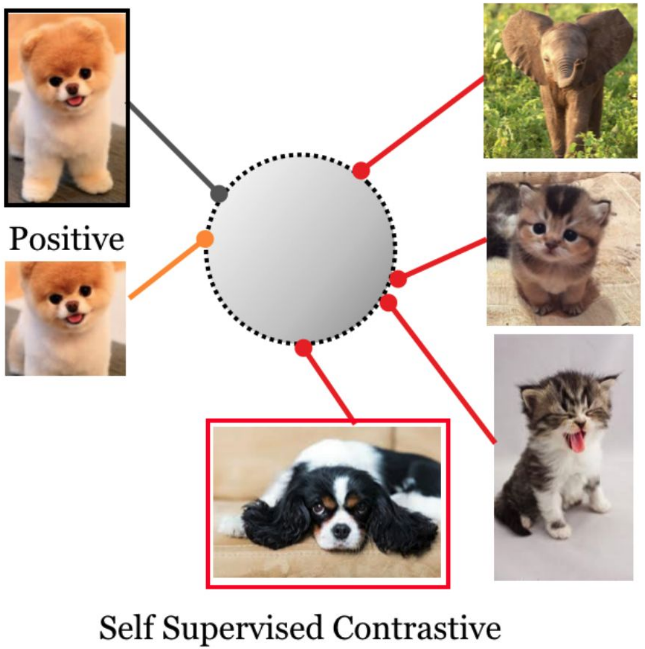


Figure from (Khosla et al., 2020)

Four Challenges of Self-Supervised Contrastive Learning

1. Non-trivial Data Augmentation
2. Risk of “Sampling Bias” (i.e., False Negative)
3. Hard Negative Mining
4. Large Batch Size

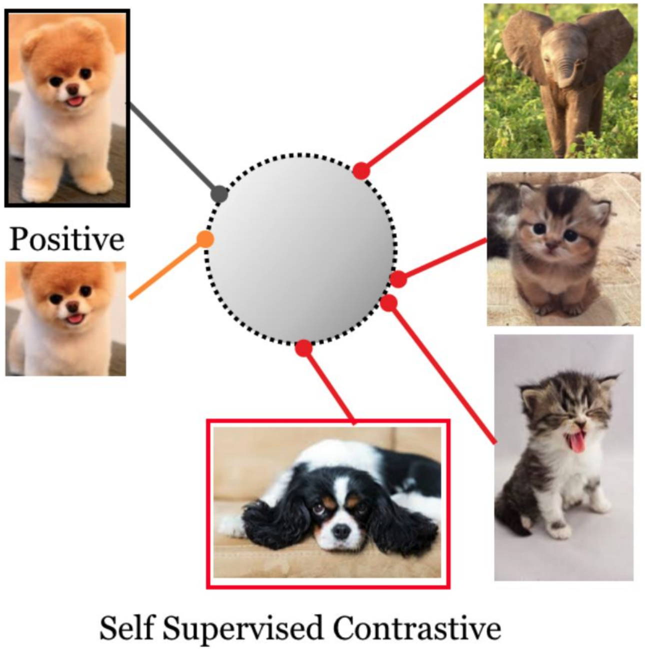
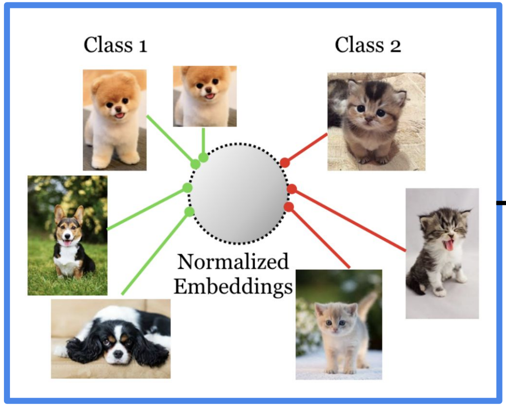


Figure from (Khosla et al., 2020)

Contrastive Data Sampling and Augmentation Strategies

Contrastive Learning = **Contrastive Data Creation** + Contrastive Objective Optimization



- Self-Supervised Contrastive Learning
  - Data Augmentation
  - Sampling Bias
  - Hard Negative Mining
  - Large Batch Size
- Supervised Contrastive Learning

Data Augmentation for Text

Text Space

- Lexical Editing (token-level)
- Back-Translation (sentence-level)

Embedding Space

- Dropout
- Cutoff
- Mixup

Manual

Lexical Editing

Synonym Replacement

Random Insertion

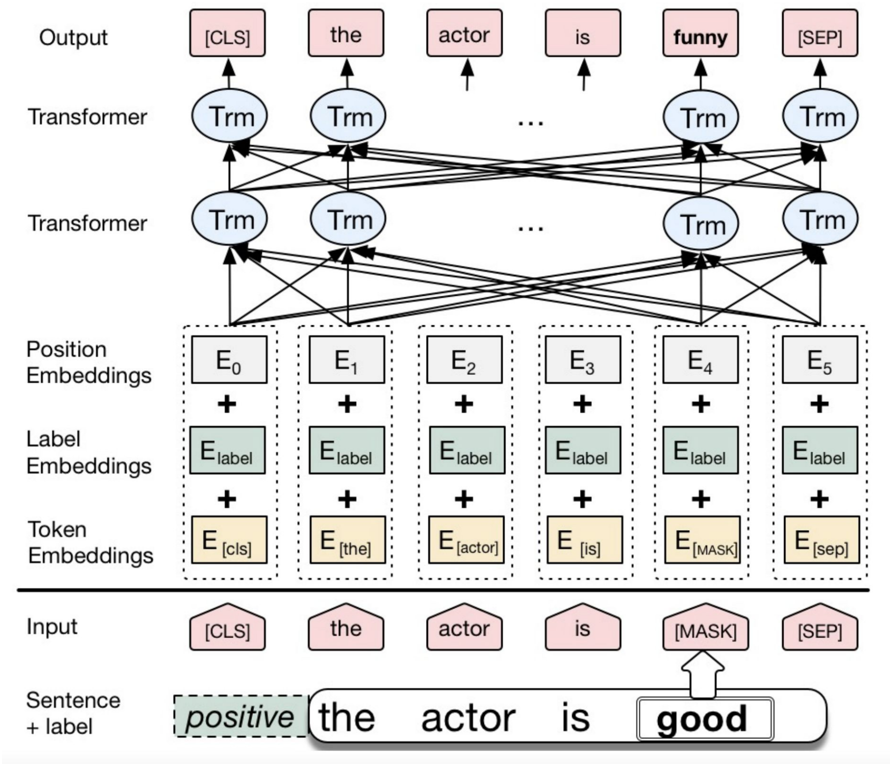
Random Swap

Random Deletion

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.

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. (Wei and Zhou, 2019)

Word Replacement with c-BERT

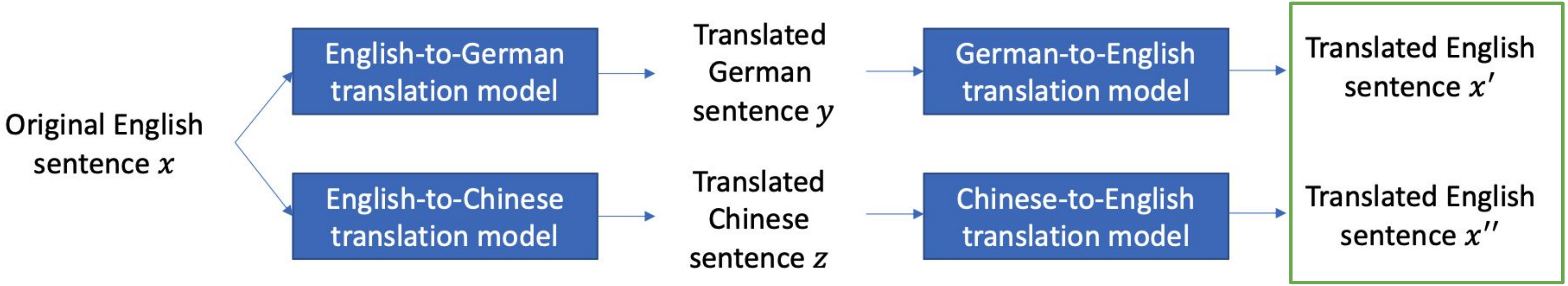


[Conditional BERT Contextual Augmentation (Wu et al., 2018)]

Back-Translation

Create paraphrases of the sentence using back-translation.

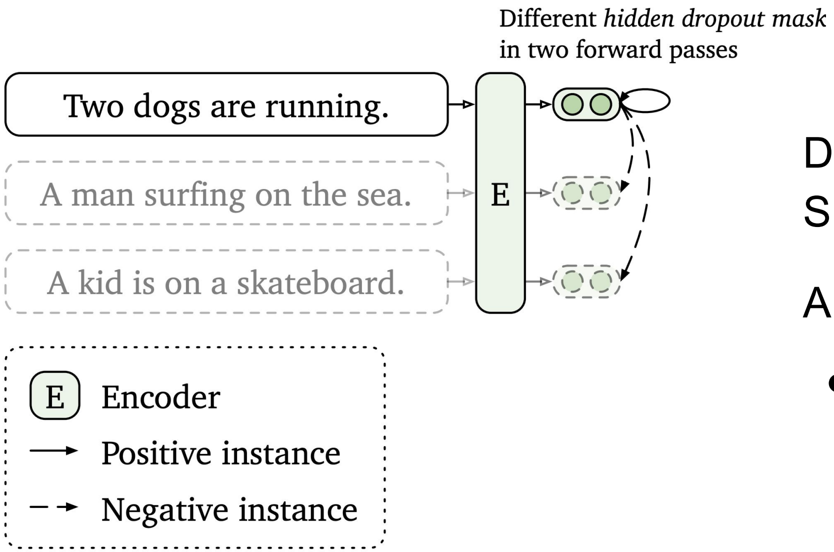
- Positive: translated sentences from the same sentences.
- Negative: translated sentences from different sentences.



[Improving Neural Machine Translation Models With Monolingual Data (Sennrich et al., 2016)]

[CERT: Contrastive Self-supervised Learning for Language Understanding (Fang et al., 2020)]

Dropout



SimCSE (Unsupervised Version)

Dropout for Data Augmentation in Embedding Space

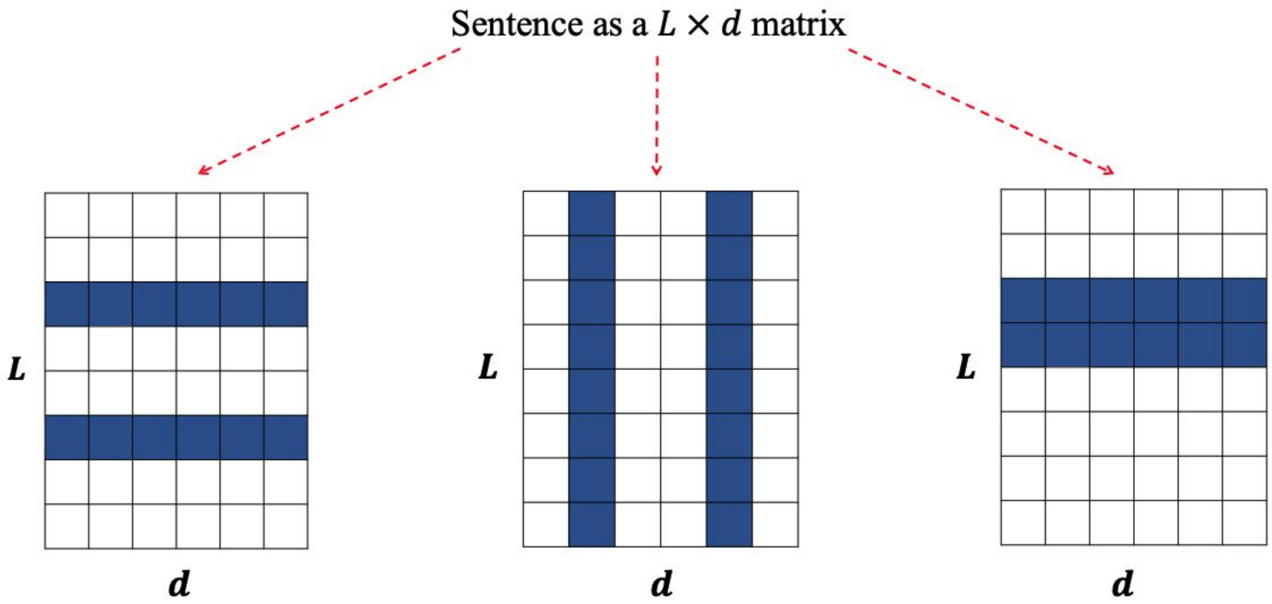
Apply dropout to sentence encoder outputs.

- Positive: Two different dropout masks create two different embeddings for the same sentence as a “positive pair”.
- Negative: in-batch negatives.

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

Cutoff

A structured version of dropout.



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

Blue area are “cutoff” to be zero.

[A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation. (Shen et al., 2020)]

mixup

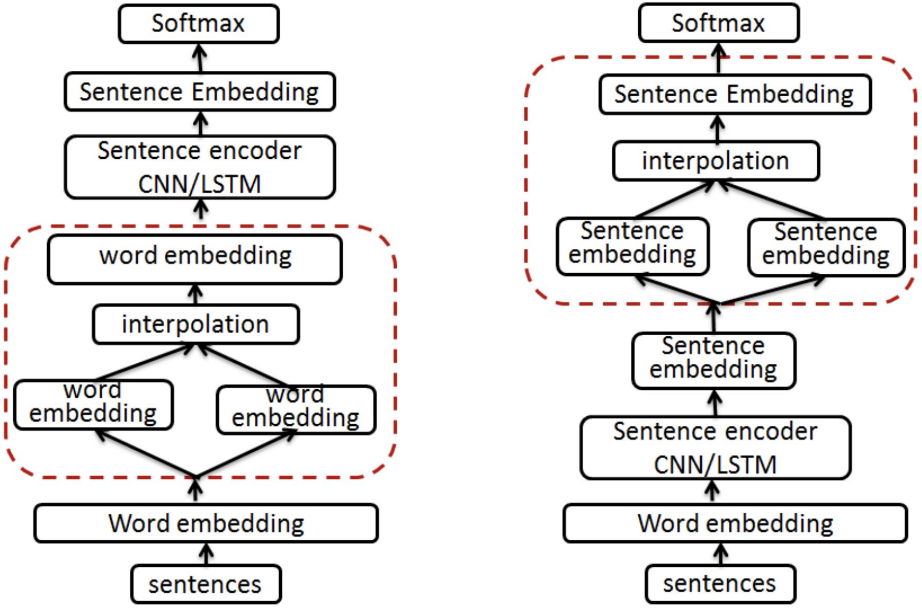
linear interpolation over a pair of samples.

$$(B^i; y^i) \text{ and } (B^j; y^j)$$

$$\widetilde{B}_t^{ij} = \lambda B_t^i + (1 - \lambda) B_t^j$$

$$\widetilde{y}^{ij} = \lambda y^i + (1 - \lambda) y^j$$

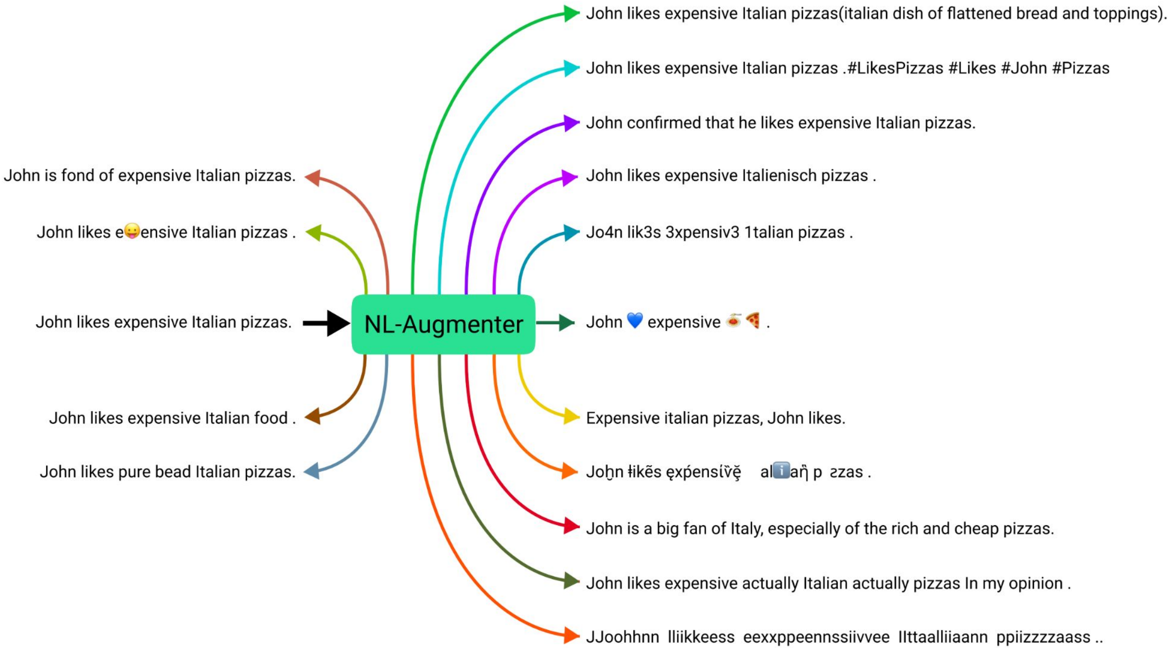
Figure 1: Illustration of wordMixup (left) and senMixup (right), where the added part to the standard sentence classification model is in red rectangle.



[mixup: Beyond Empirical Risk Minimization. (Zhang et al., 2017)](#)
[Augmenting Data with Mixup for Sentence Classification: An Empirical Study. (Guo et al., 2019)](#)
[MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. (Chen et al., 2020)](#)

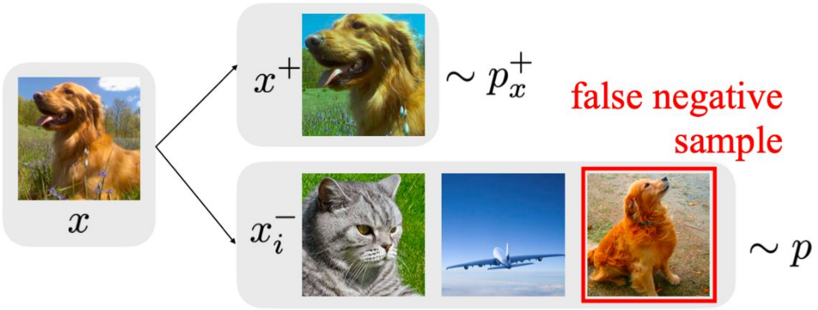
NL-Augmenter: Manual Data Augmentation

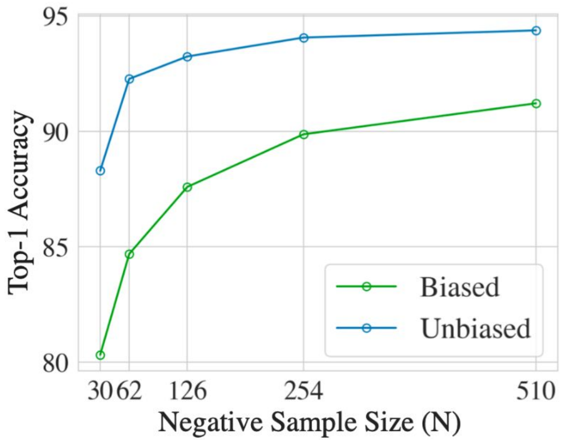
- Crowdsource Wisdom-of-Researchers
- 117 ways of doing Data Augmentation
- Use or Contribute at https://github.com/GEM-benchmark/NL-Augmenter



[NL-Augmenter A Framework for Task-Sensitive Natural Language Augmentation (Dhole et al., 2021)](#)

Sampling Bias





Problem: Because we don't know the label, we may accidentally create false negative by sampling examples from the same class.

Debiased Contrastive Learning (Chuang et al., 2020)

Debiased Contrastive Learning

Key Idea: Assume a prior probability between positive and negative, then approximate the distribution of negative examples to debias the loss.

$$p(x') = \tau^+ p_x^+(x') + \tau^- p_x^-(x')$$

Then samples N samples (may contain positive and negative) and M positive samples

$$\text{replace } p_x^- \text{ in } L^N_{\text{Unbiased}} \text{ with } p_x^-(x') = (p(x') - \tau^+ p_x^+(x'))/\tau^-$$

$$-\log \frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + N g\left(x, \{u_i\}_{i=1}^N, \{v_i\}_{i=1}^M\right)}$$

[Debiased Contrastive Learning (Chuang et al., 2020)](#)

Hard Negative Mining

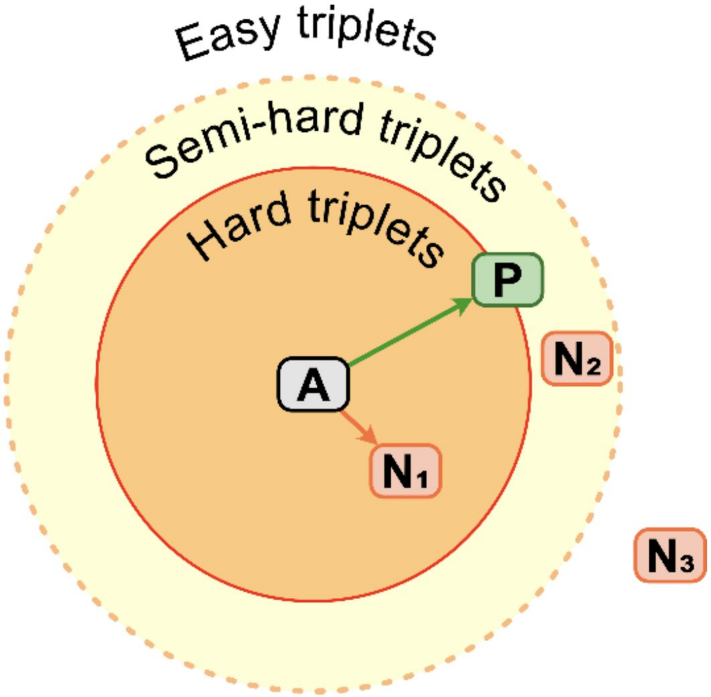


Figure from Kurowski et al., 2021

A: Anchor.   P: Positive.   N: Negative

We want to AN is greater than AP, at least by the margin.

Hard Negative Mining: Find hard negatives

Hard Negative Mining by Importance Sampling

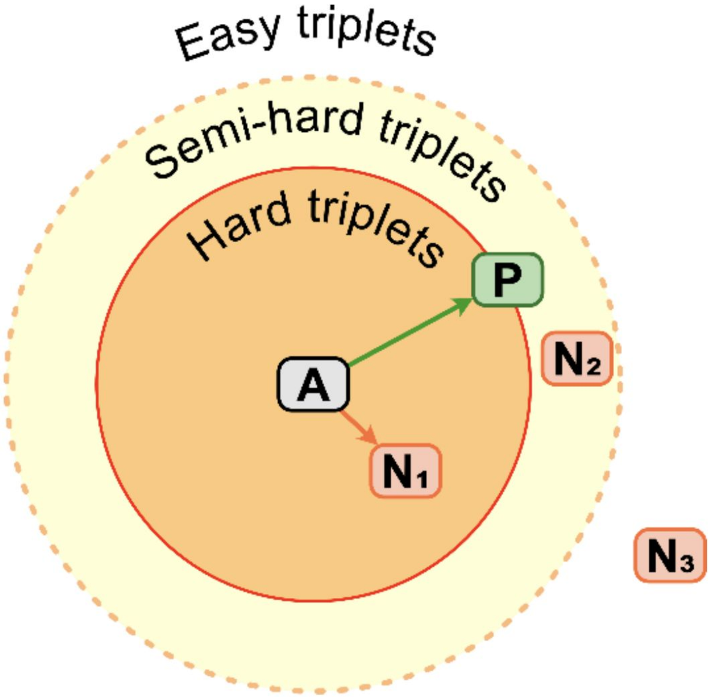


Figure from Kurowski et al., 2021

$$\boxed{q_\beta(x^-)} \propto \boxed{e^{\beta f(x)^\top f(x^-)}} \cdot \boxed{p(x^-)}$$

new sampling probability

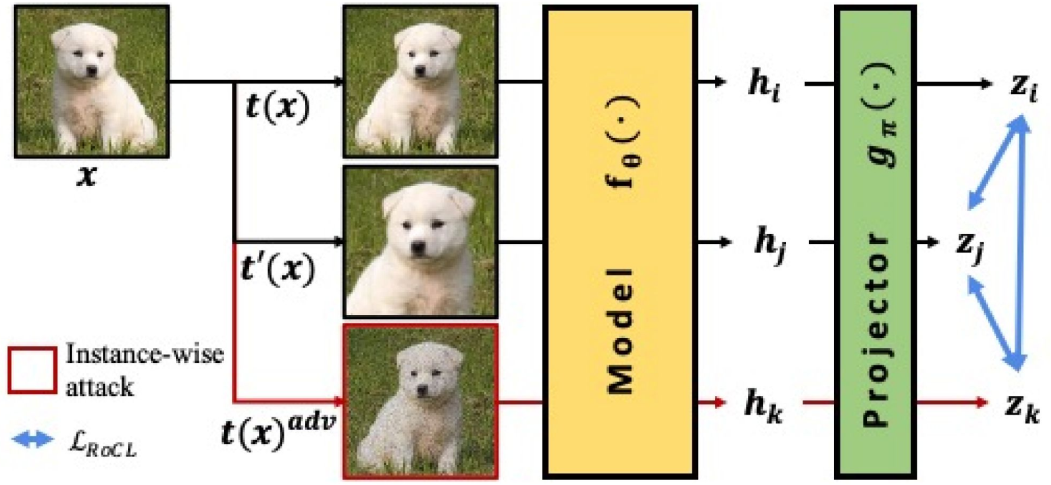
similarity

original sampling probability

Key Idea: If this negative sample is close to the anchor sample, then we up-weight its probability of being selected.

Contrastive Learning with Hard Negative Samples (Robinson et al., 2021)

Hard Positive Mining by Adversarial Examples



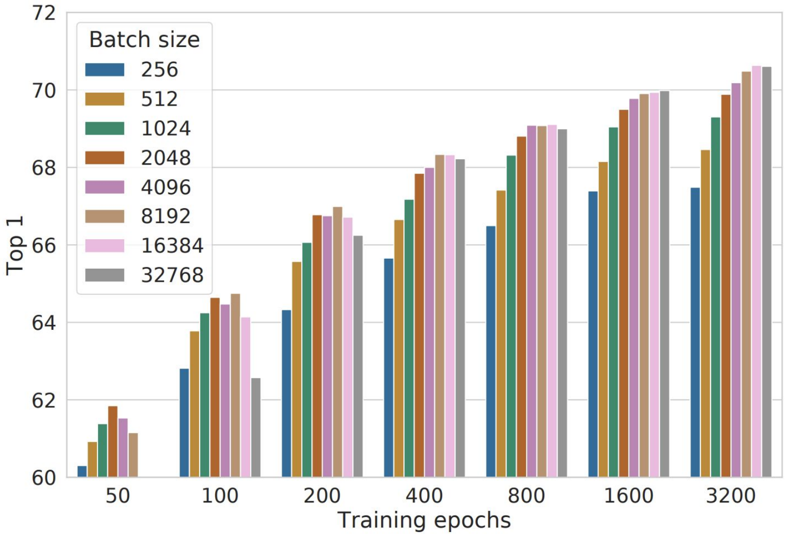
(a) Robust contrastive learning training

Create adversarial examples that are positive but confuses the model.

Use Contrastive Learning to train with “Hard Positive” examples for robustness.

[Adversarial Self-Supervised Contrastive Learning (Kim et al., 2020)]

Large Batch Size



SimCLR of ResNet-50 trained with different batch sizes and epochs.

“*We train with larger batch size (up to 32K) and longer (up to 3200 epochs).*”

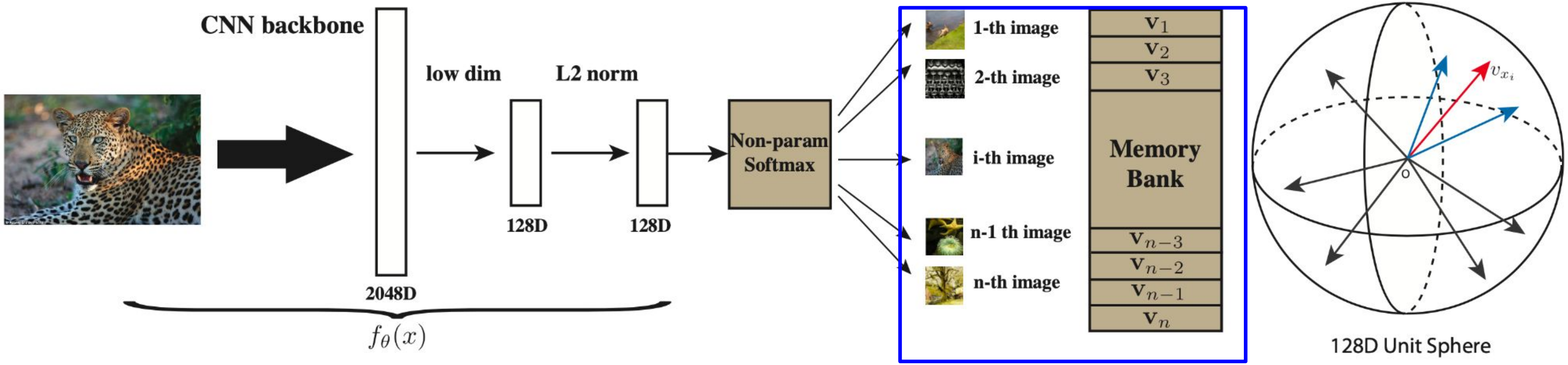
— Chen et al., SimCLR

“*We use a very large minibatch size of 32,768.*”

— Radford et al., CLIP

Memory Bank to Reduce Computation

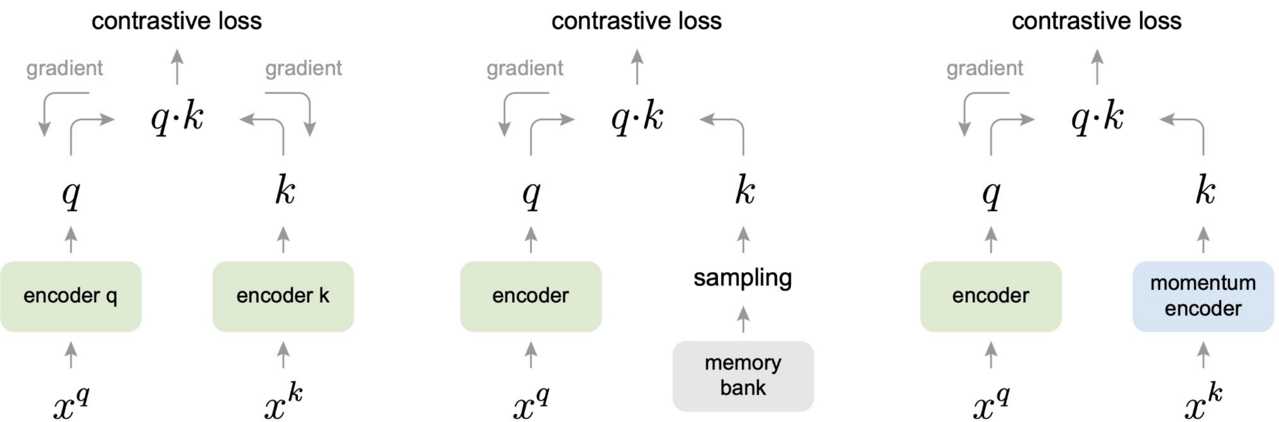
Memory Bank: Compute and store the representations in advances, instead of computing embeddings for all examples in a batch.



Instance-level discrimination uses contrastive learning to maximally scatter the features of training samples over the 128-dimensional unit sphere. Embeddings are stored in a Memory Bank.

[Unsupervised Feature Learning via Non-Parametric Instance Discrimination. (Wu et al., 2018)]

Momentum Contrast (MoCo) to Scale the Number of Negative Examples

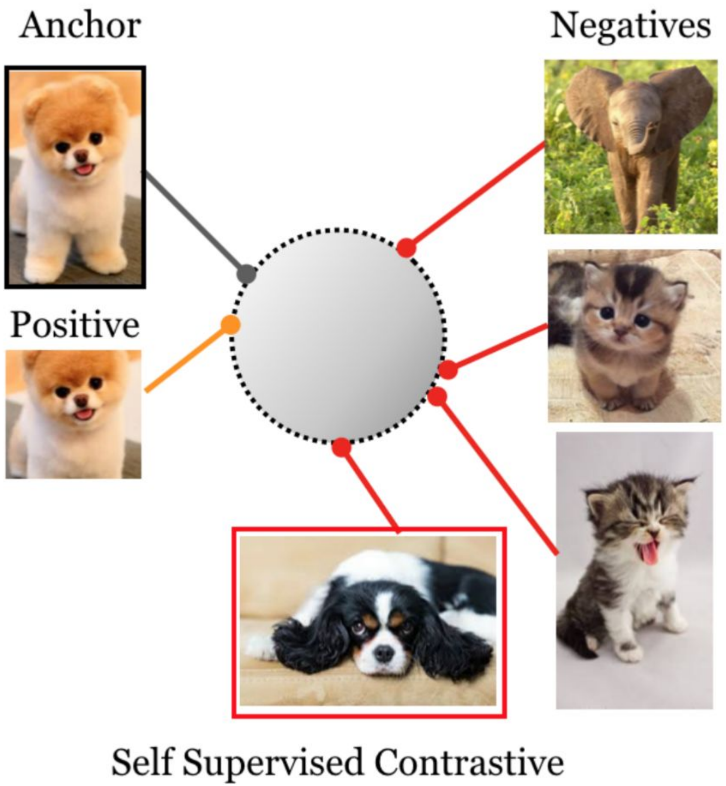


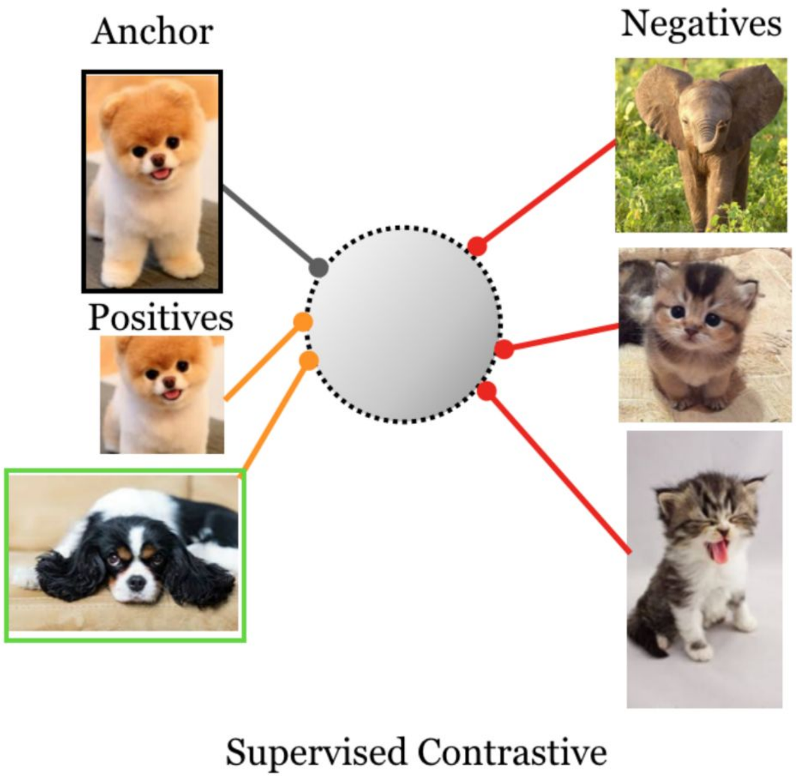
(a) end-to-end (b) memory bank (c) MoCo

- Traditional: a encoder for query and a decoder for key. The number of negative samples is restricted to the size of the mini-batch.
- Momentum Contrast
  - Scale the number of negative sample by maintaining a queue.
  - The key encoder is updated using momentum.
  - A large and consistent dictionary for stable training.

[Momentum Contrast for Unsupervised Visual Representation Learning (He et al., 2020)]

From Self-Supervised to Supervised Contrastive Learning





[Supervised Contrastive Learning (Khosla, et al., 2020)]

Supervised Contrastive Learning

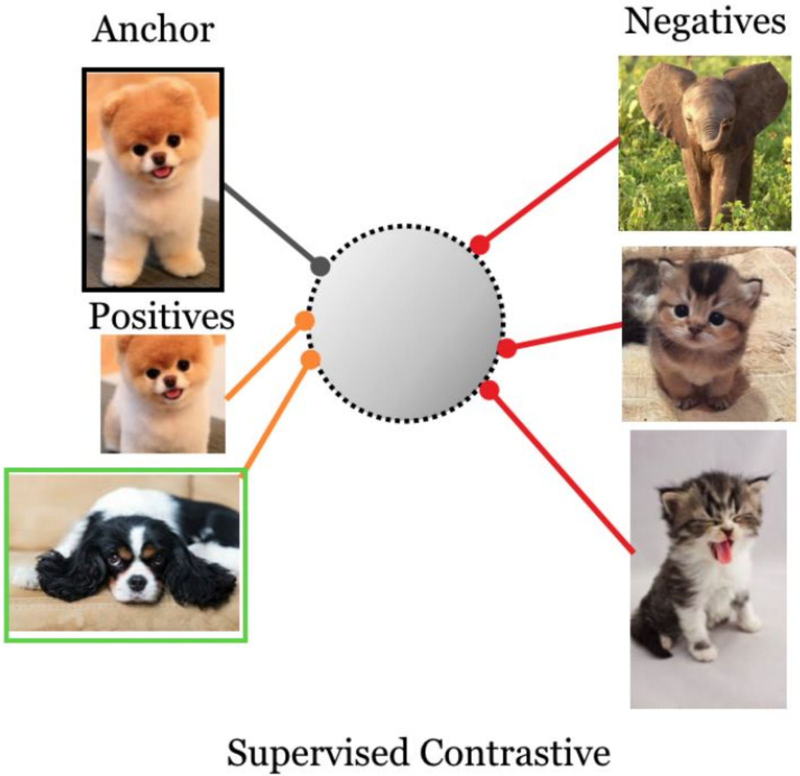
Positive: Same Class

Negative: Different Class

Pros
- No Need for Data Augmentation
- No Risk of “False Negative”
- No Need for Large Batch Size

Cons
- Need Label

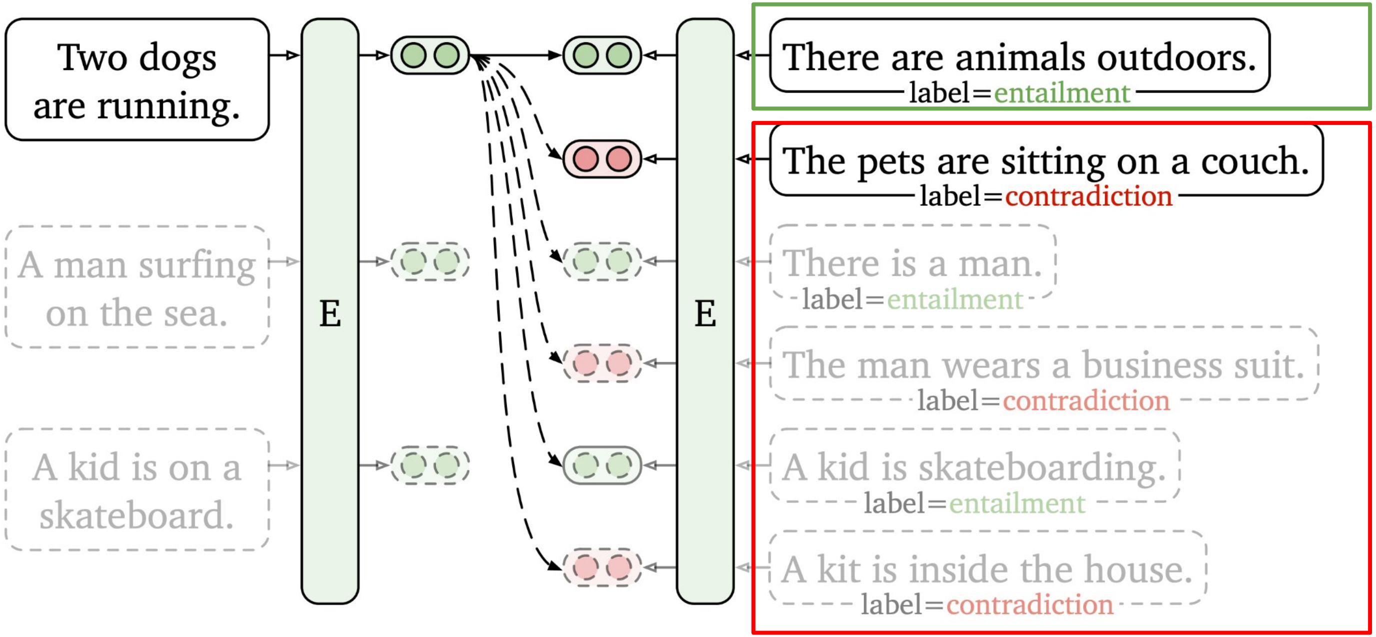
Sentence-BERT, SimCSE, DPR, CLIP



Supervised Contrastive Learning (Khosla, et al., 2020)

SimCSE (Supervised Version)

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

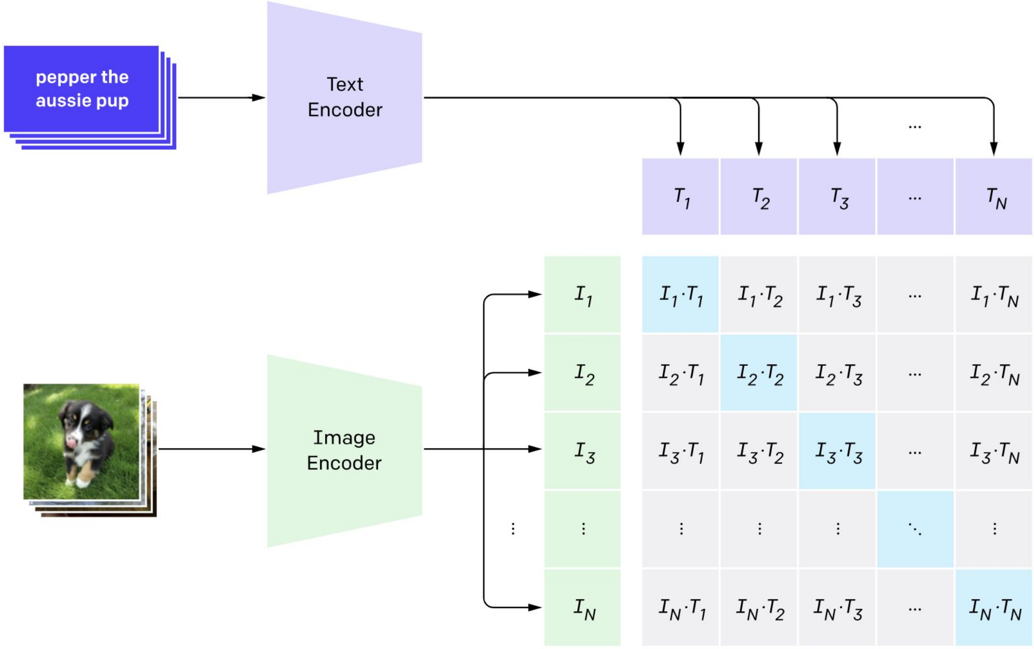


[SimCSE: Simple Contrastive Learning of Sentence Embeddings. (Gao et al., 2021)]

CLIP

Supervision: Image Captions

- Positive: N correct image-caption pairs
- Negative: N(N-1) in-batch negative



[Learning Transferable Visual Models From Natural Language Supervision. (Radford et al., 2021)]

Summary of Contrastive Data Strategies

- Self-Supervised Contrastive Learning
  - Data Augmentation
    - Lexical Edits
    - Back-translation
    - Dropout, Cutoff, mixup
    - Manual
  - Sampling Bias
    - Debiased Contrastive Learning
  - Hard Negative Mining
    - Importance Sampling
    - Adversarial Examples
  - Large Batch Size
    - Memory Bank
    - Momentum Encoder
- Supervised Contrastive Learning

Part 1.3

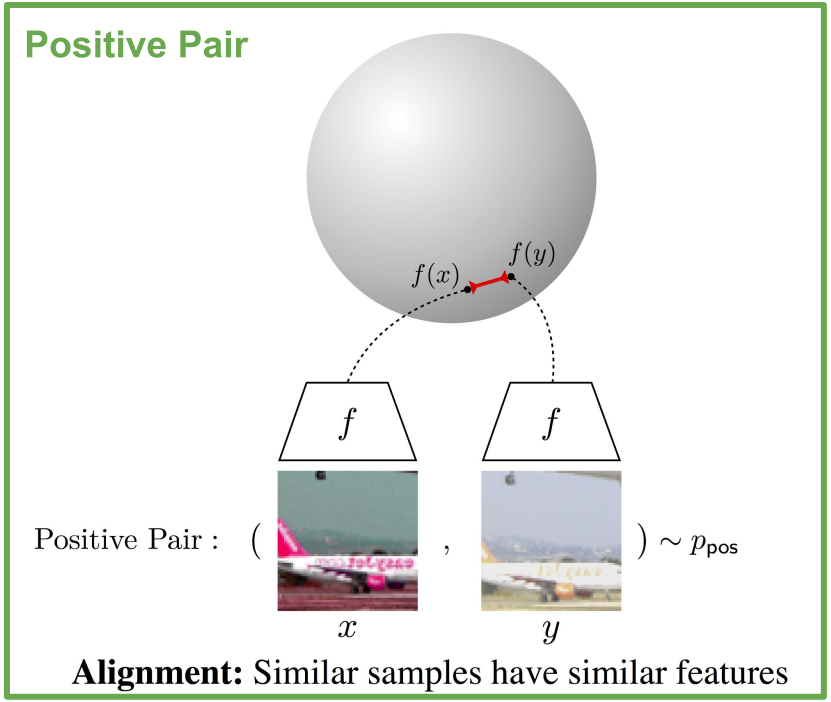
Analysis of Contrastive Learning

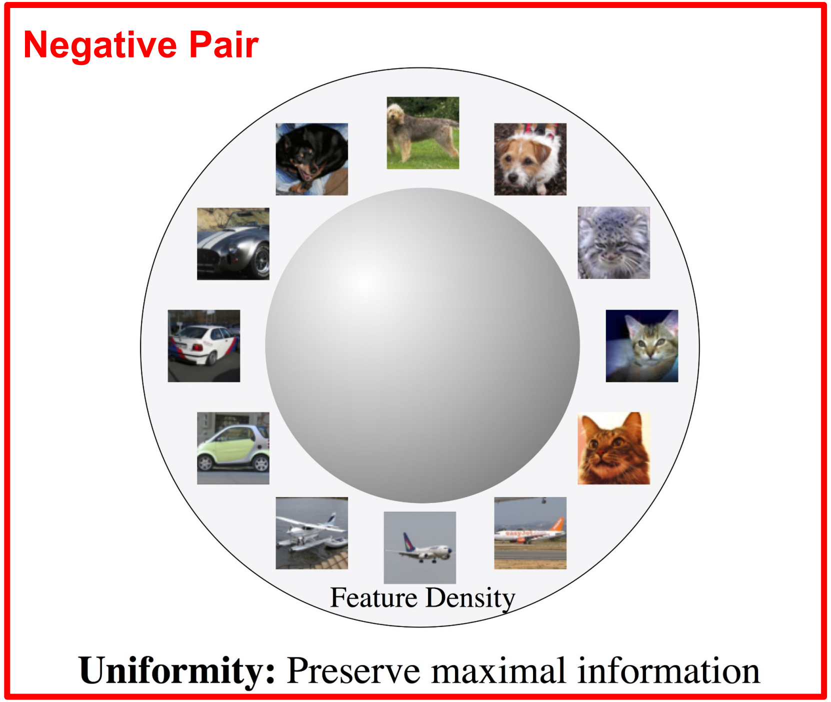
Analysis of Contrastive Learning

- Geometric Interpretation
- Connection to Mutual Information
- Theoretical Analysis
- Robustness and Security

Geometric Interpretation of Contrastive Learning

Two geometric forces on the hypersphere (the n-dimensional sphere, i.e. when embeddings are normalized).

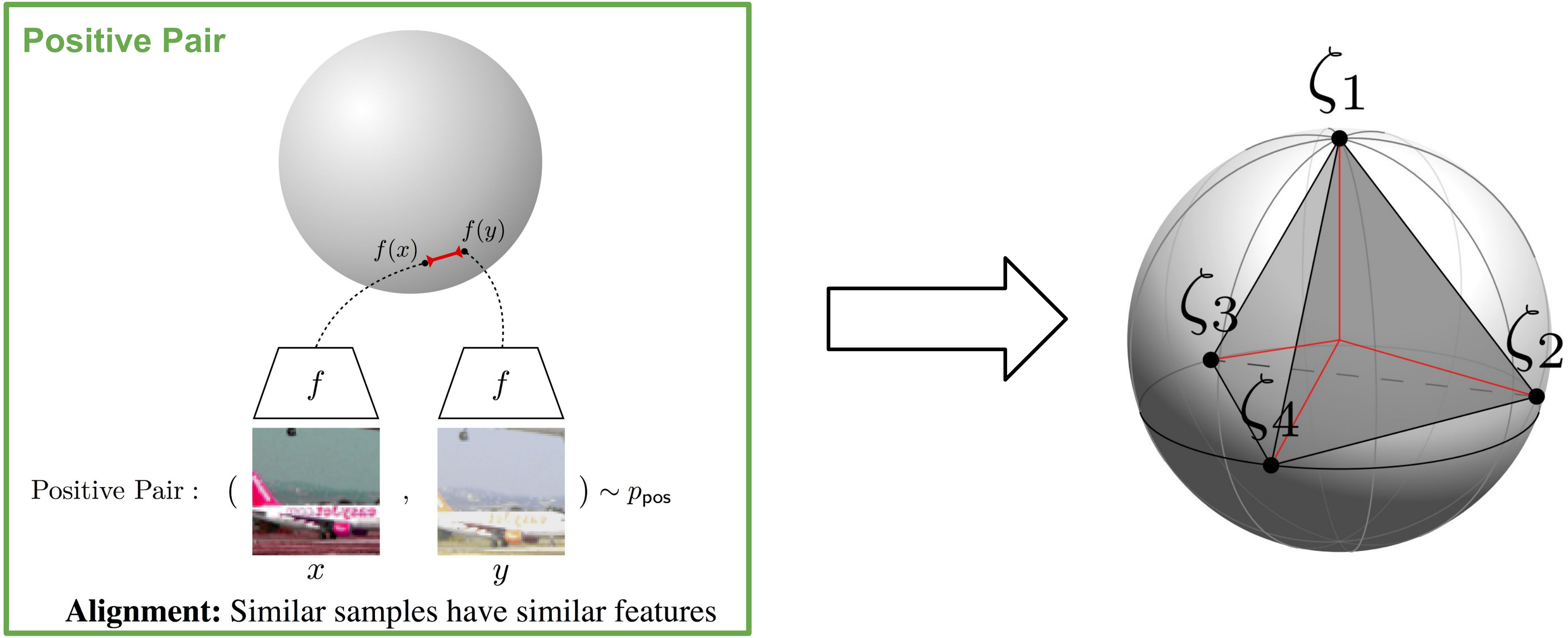




[Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. (Wang and Isola, 2020)]

Geometric Interpretation of Supervised Contrastive Learning

When the class label is used, then supervised contrastive learning will converge to **class collapse** to a regular simplex.



[Dissecting Supervised Contrastive Learning (Graf et al., 2021)]
[Perfectly Balanced: Improving Transfer and Robustness of Supervised Contrastive Learning (Chen et al., 2022)]

Mutual Information

The Mutual Information (MI) between two random variables is a measure of how dependent they are on one another.

- If two random variables are independent, MI will be zero.
- Maximize Mutual Information: make them as dependent as possible.
- Minimize Mutual Information: make them as independent as possible.

$$I(\boldsymbol{x}; \boldsymbol{x}^+) = D_{\mathrm{KL}}(p(\boldsymbol{x}, \boldsymbol{x}^+) \| p(\boldsymbol{x})p(\boldsymbol{x}^+)) = \sum_{(\boldsymbol{x}, \boldsymbol{x}^+)} p(\boldsymbol{x}, \boldsymbol{x}^+) \log \frac{p(\boldsymbol{x}|\boldsymbol{x}^+)}{p(\boldsymbol{x})}$$

InfoNCE

Use softmax loss to differentiate a positive sample from a set of noise examples.

$\boldsymbol{c}$ Context Vector, e.g., anchor

$X = \{x_1, \ldots, x_N\}$ N samples with 1 positive sample and N-1 negative samples

The probability mass that x_i is the positive example, and every other is negative.

$$p(d = i|X, \boldsymbol{c}) = \frac{\boxed{p(\boldsymbol{x}_i|\boldsymbol{c}) \prod_{l \neq i} p(\boldsymbol{x}_l)}}{\boxed{\sum_{j=1}^{k} p(\boldsymbol{x}_j|\boldsymbol{c}) \prod_{l \neq j} p(\boldsymbol{x}_l)}}$$

The probability mass for all possible cases.

InfoNCE as Maximizing Mutual Information

$$p(d=i|X,\boldsymbol{c}) = \frac{p(\boldsymbol{x}_i|\boldsymbol{c})\prod_{l\neq i} p(\boldsymbol{x}_l)}{\sum_{j=1}^{k} p(\boldsymbol{x}_j|\boldsymbol{c})\prod_{l\neq j} p(\boldsymbol{x}_l)} = \frac{\frac{p(\boldsymbol{x}_i|\boldsymbol{c})}{p(\boldsymbol{x}_i)}}{\sum_{j=1}^{k} \frac{p(\boldsymbol{x}_j|\boldsymbol{c})}{p(\boldsymbol{x}_j)}}$$

$$\boxed{f(\boldsymbol{x},\boldsymbol{c}) \propto \frac{p(\boldsymbol{x}|\boldsymbol{c})}{p(\boldsymbol{x})}}$$

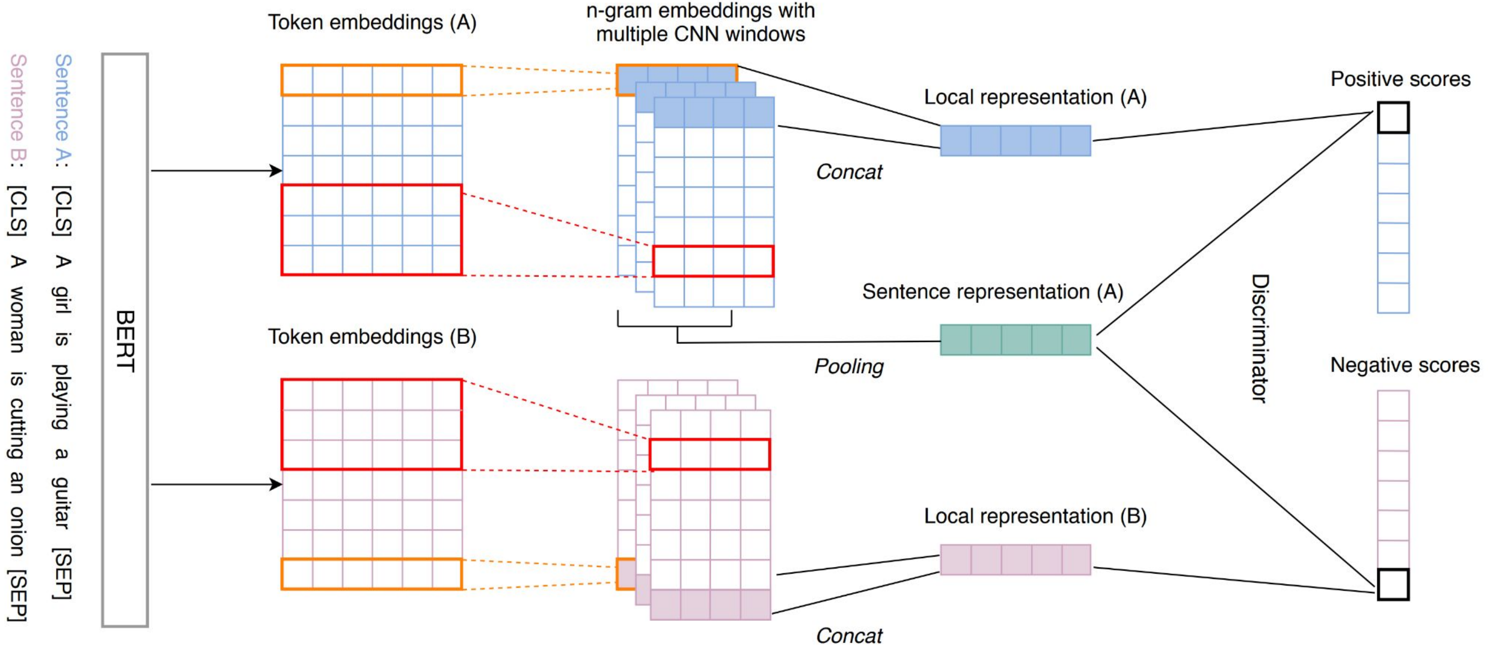
The scoring function we want to learn, and it is proportional to mutual information between x and c

$$\mathcal{L} = -\log \frac{f(\boldsymbol{x},\boldsymbol{c})}{\sum_{\boldsymbol{x}'\in X} f(\boldsymbol{x}',\boldsymbol{c})}$$

Representation Learning with Contrastive Predictive Coding (van den Oord et al., 2018)
Learning deep representations by mutual information estimation and maximization (Hjelm et al., 2018)
Learning Representations by Maximizing Mutual Information Across Views (Bachman et al., 2019)
On Variational Bounds of Mutual Information (Poole et al., 2019)

Maximizing Mutual Information for Sentence Embeddings

Info-Sentence-BERT: Maximizing Mutual Information of global representation and local representation of the same sentence.



[An Unsupervised Sentence Embedding Method by Mutual Information Maximization. (Zhang et al., 2020)]

Contrastive Data Selection by Minimizing Mutual Information

If we don't have annotated labels available, how shall we select the views to which the representations should be invariant?

The InfoMin Hypothesis: The views that yield the best results should discard as much information in the input as possible except for the task relevant information (e.g., object labels).

An illustration of three regimes of information captured during contrastive multiview learning. Views should not share too much information (**left**) or too little information (**right**), but should find an optimal mix (the "sweet spot", **middle**) that maximizes the downstream performance.

[What Makes for Good Views for Contrastive Learning? (Tian et al., 2020)](#)

Figure from [Google Blog](#)

Theoretical Analysis for Contrastive Learning

- **Framework** connecting unlabeled data with downstream supervised tasks.
- **Provable guarantees**: Unsupervised loss is surrogate for average supervised loss

Framework

Semantic similarity $\approx$ membership in **same latent class**.

Connection

$\mathcal{X}$: Set of inputs, $\mathcal{C}$: Set of classes, ρ: Distribution over $\mathcal{C}$

$\mathcal{D}_c$: Universal distribution over $\mathcal{X}$ conditioned on class c.

$\updownarrow$ $\updownarrow$

Unlabeled Data

Similarity data: $(x, x^+) \sim \mathcal{D}_{sim}$

$$c^+ \sim \rho$$
$$(x, x^+) \sim \mathcal{D}_{c^+}^2$$

Negative samples: $x^- \sim \mathcal{D}_{neg}$

$$c^- \sim \rho$$
$$x^- \sim \mathcal{D}_{c^-}$$

Supervised Tasks

Task: Subset of latent classes

$$\mathcal{T} = \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$$

Labeled samples: $(x, c) \sim \mathcal{D}_{\mathcal{T}}$

$$c \sim \mathcal{T}$$
$$x \sim \mathcal{D}_c$$

Unsupervised Loss Bounds Supervised Loss

$\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^d,\ \|f(\cdot)\| \leq R\}$: Function class of interest.

τ: Probability that two classes sampled from ρ are the same.

$\widehat{f}$: Minimizer from $\mathcal{F}$ of **empirical unsupervised loss**.

Theorem 2: Generalization Bound

With probability at least $1 - \delta$,

$$\boldsymbol{L_{sup}(\widehat{f})} \leq \frac{1}{1-\tau}\left[\min_{f \in \mathcal{F}} \boldsymbol{L_{un}(f)} - \tau + Gen_M\right]$$

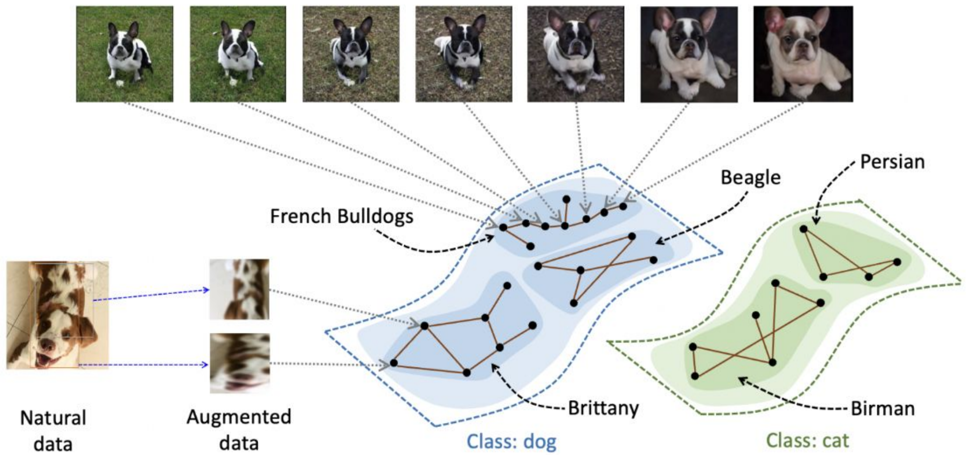
where

$$Gen_M = O\left(R\frac{\mathcal{R}_{\mathcal{S}}(\mathcal{F})}{M} + R^2\sqrt{\frac{\log\frac{1}{\delta}}{M}}\right)$$

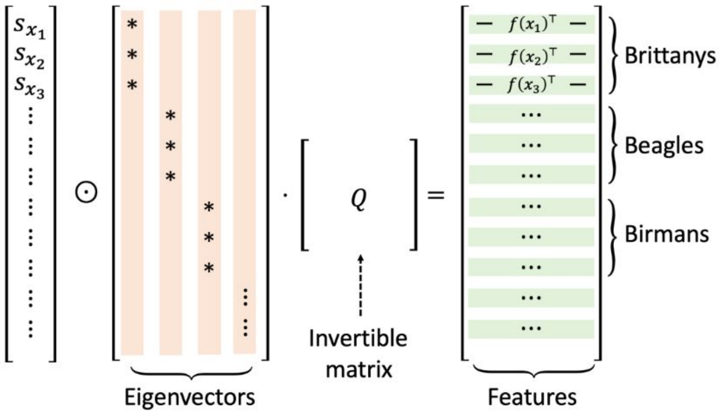
[A Theoretical Analysis of Contrastive Unsupervised Representation Learning. (Arora et al., 2019)](#) [[poster]](#)

Theoretical Analysis for Contrastive Learning

- **Population Augmentation Graph**: Two augmented data are connected if they are views of the same natural datapoint. Subgraphs for subclasses.
- **Spectral Contrastive Loss**: Connect contrastive learning to spectral decomposition on the adjacency matrix of the graph.



Left: The population augmentation graph.

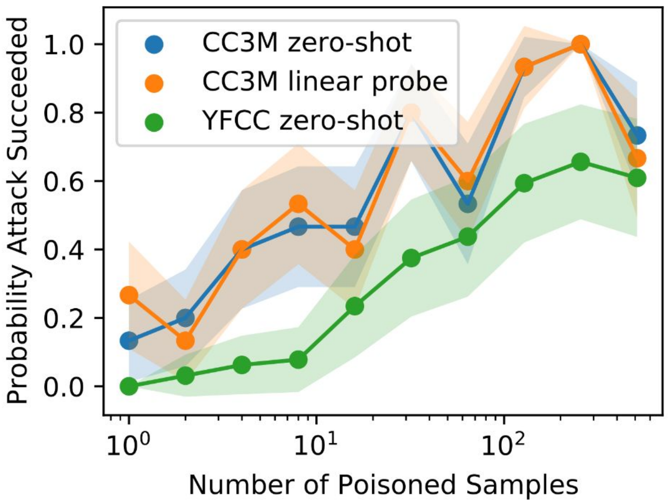


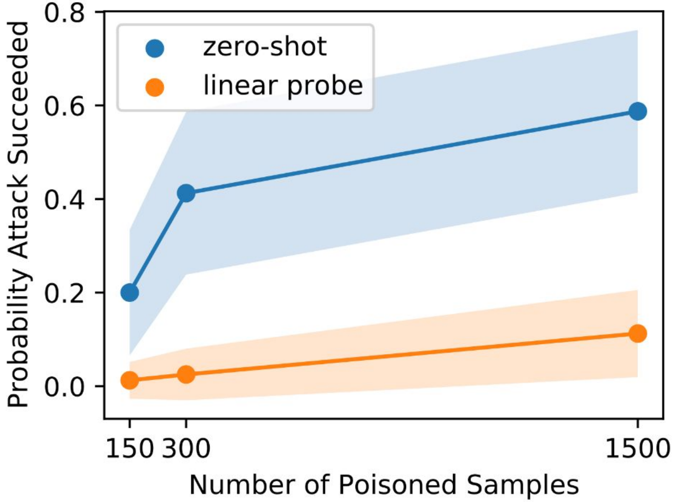
Right: Decomposition of the learned representations.

[Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. (HaoChen et al., 2021)]

Robustness and Security of Contrastive Learning Models

Can we trust contrastive learning models like CLIP trained on noisy and uncurated training datasets?





Targeted Poisoning: Misclassify a particular test input to an target label. Poisoning 0.0001% of a dataset (3 out of the 3 million images).

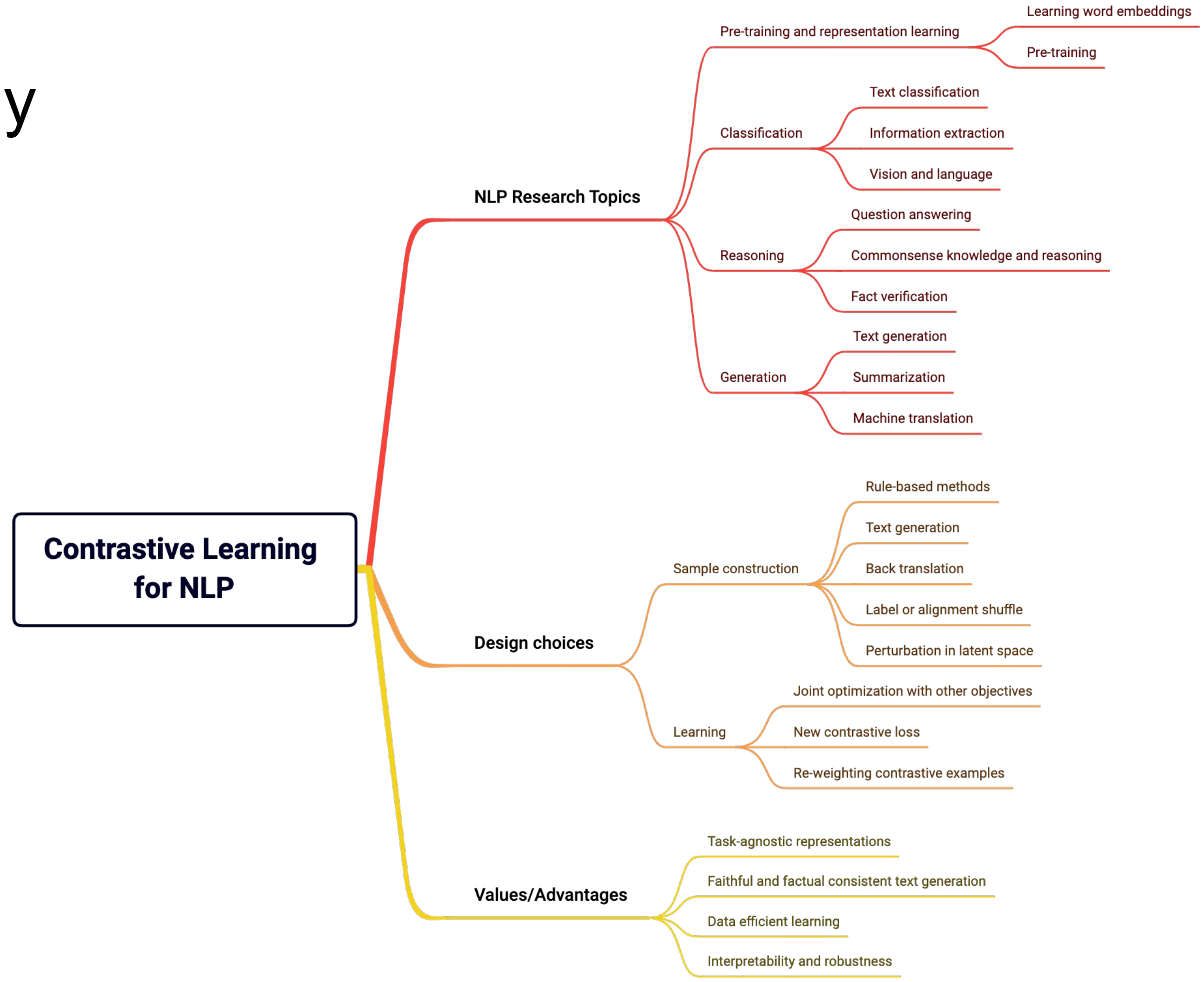
Backdoor Attack: Misclassify *any* image by overlaying a small patch. Poisoning 0.01% of a dataset (300 images of the 3 million-example).

[Poisoning and Backdooring Contrastive Learning. (Carlini and Terzis, 2021)](#)

Part 2.

Contrastive Learning for NLP

Taxonomy


Contrastive Learning for NLP
NLP Research Topics
Pre-training and representation learning
Learning word embeddings
Pre-training
Classification
Text classification
Information extraction
Vision and language
Reasoning
Question answering
Commonsense knowledge and reasoning
Fact verification
Generation
Text generation
Summarization
Machine translation
Design choices
Sample construction
Rule-based methods
Text generation
Back translation
Label or alignment shuffle
Perturbation in latent space
Learning
Joint optimization with other objectives
New contrastive loss
Re-weighting contrastive examples
Values/Advantages
Task-agnostic representations
Faithful and factual consistent text generation
Data efficient learning
Interpretability and robustness

NLP Applications

Overall, there are four categories of NLP applications that contrastive learning can help:

1. Classification, e.g.,
   - **Text classification**
   - Information extraction
   - Vision and language
2. Reasoning, e.g.,
   - Commonsense knowledge and reasoning
   - **Question answering**
   - Fact verification
3. Generation, e.g.,
   - **Text generation**
   - Summarization
   - Machine translation
4. Pre-training and representation learning
   - **Pre-training**
   - Word embeddings

Key Questions in NLP Applications

Regardless the type of application, two common questions needed to be answered:

- How to construct contrastive examples? E.g.,
  - Rule-based methods
  - Text generation
  - Back translation
  - Label or alignment shuffle
  - Perturbation in latent space
- How to use contrastive examples? E.g.,
  - Joint optimization with other training objectives
  - New contrastive loss
  - Re-weighting contrastive examples

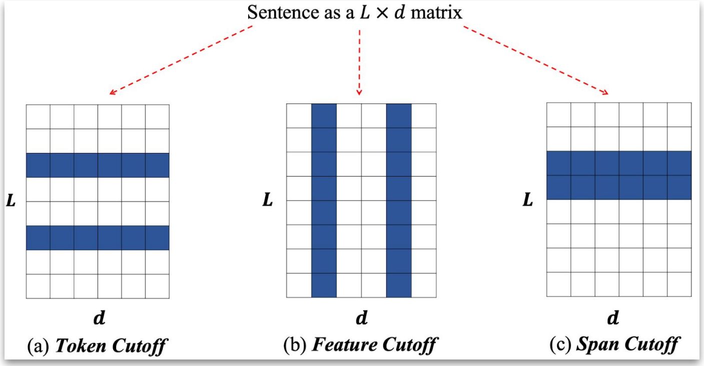
Meta Comments

- Each work can be categorized by
  - NLP research topics
  - Design choices
  - Values/advantages
- Presentation format
  - Present each selected work based on its research topic
    - Then, discuss its design choice
  - Summarize the values/advantages in a separate subsection
- About taxonomy
  - Comments are welcomed for comprehensiveness and better organization

Classification: Cutoff

- Inspired by multi-view learning (e.g., Blum and Mitchell, 1998)

- Sample construction
  - Cut off dense representations to construct some similar examples
  - Three cutoff operations:
    - Token cutoff
    - Feature cutoff
    - Span cutoff
- Training
  - Minimize the cross entropy loss of cutoff samples
  - Minimize the JS divergence of all cutoff samples, regarding one original example



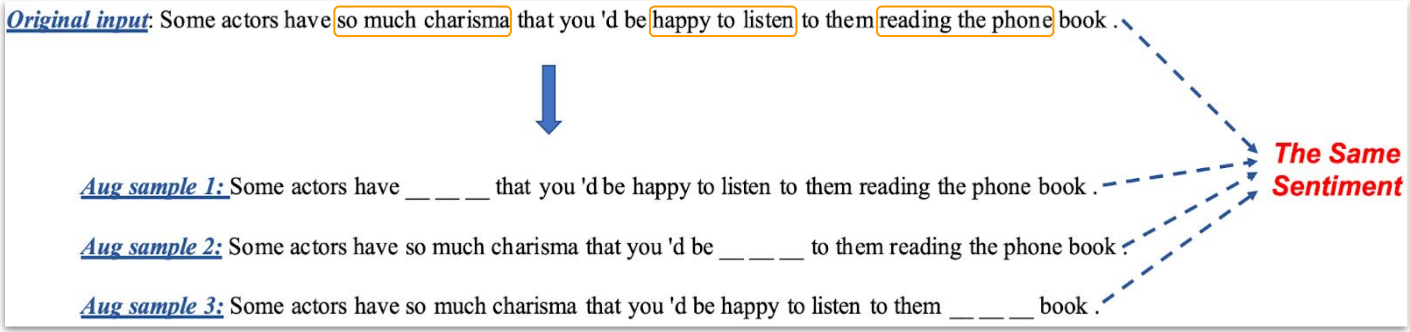
$$\mathcal{L} = \mathcal{L}_{\text{ce}}(x, y) + \boxed{\alpha \sum_{i=1}^{N} \mathcal{L}_{\text{ce}}(x^i_{\text{cutoff}}, y)}$$
$$+ \boxed{\beta \mathcal{L}_{\text{divergence}}(x, x^1_{\text{cutoff}}, x^2_{\text{cutoff}}, \ldots, x^N_{\text{cutoff}}, y)}$$

$$\mathcal{L}_{\text{divergence}} = \frac{1}{N+1} \sum_{i=0}^{N} \text{KL}[p(y|x^i_{\text{cutoff}})||p_{avg}]$$

[A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation (Shen et al., 2020)]
[Combining Labeled and Unlabeled Data with Co-Training (Blum and Mitchell, 1998)]

Classification: Cutoff (Cont.)

- *Span cutoff* is arguably the most popular one and has been adopted by other works in contrastive learning and data augmentation (e.g., Ye et al., 2021).
- Examples of the span cutoff effect (Shen et al., 2020)



- More structural than Dropout (Srivastava et al., 2014)

A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation (Shen et al., 2020)
Dropout: A Simple Way to Prevent Neural Networks from Overfitting (Srivastava et al., 2014)
Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning (Ye et al., 2021)

Classification: CERT

- Contrastive self-supervised Encoder Representation from Transformers (CERT)
- Auxiliary task
  - Predict whether two augmented data are from the same original sample
- Sample construction
  - Based on the texts in target task
  - Back-translation
  - Using English-German and English-Chinese translation systems
- Training
  - Momentum Contrastive Learning (He et al., 2020; MoCo)

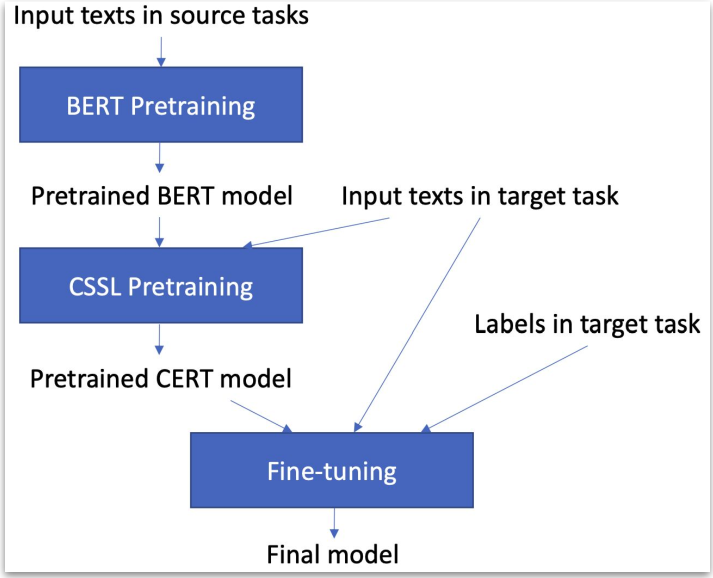
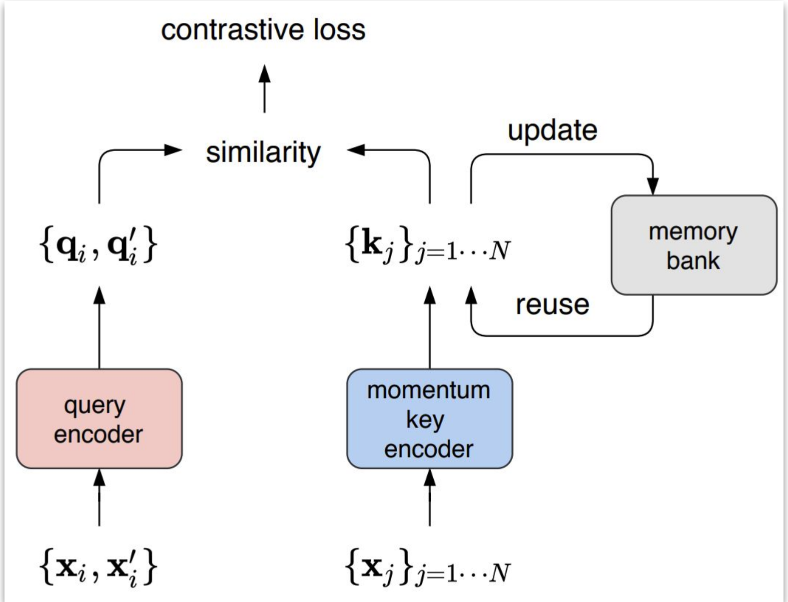


Figure: Use contrastive learning for pre-training on the target task

[CERT: Contrastive Self-supervised Learning for Language Understanding (Fang et al., 2020)]
[Momentum Contrast for Unsupervised Visual Representation Learning (He et al., 2020)]

Classification: CoDA

- Contrast-enhanced and Diversity-promoting Data Augmentation (CoDA)
- Sample construction
  - Stack different data augmentation methods together
  - Use *five different label-preserving operations*, including cutoff and back-translation
- Training
  - The model should encourage the augmented sample x'_i to be closer to x_i, than x_j
  - With a momentum encoder + memory bank



CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding (Qu et al., 2020)

Classification: CoDA (Cont.)

Comparison on singla transformation and multiple translations (Qu et al., 2020)

- MNLI-m development set
- Methods
  - Original data (ori)
  - c-BERT
  - Back-translation (back)
  - Cutoff (cut)
  - Mixup (mix)
  - Adversarial (adv)

Method	MNLI-m (Acc)	MMD
RoBERTa-base	87.6	-
Single Transformation		
+ back-translation	88.5	0.63
+ c-BERT	88.0	0.01
+ cutoff	88.4	0.02
+ mixup (ori, ori)	88.2	0.06
+ adversarial	88.5	0.65
Multiple Transformations		
+ random (back, cut, adv)	88.4	-
+ mix (ori, back)	88.4	0.11
+ mix (back, adv)	88.6	0.81
+ stack (back, cut)	88.5	0.62
+ stack (back, adv)	**88.8**	1.14
+ stack (back, cut, adv)	88.5	1.14
+ stack (back, adv, cut)	88.4	1.14

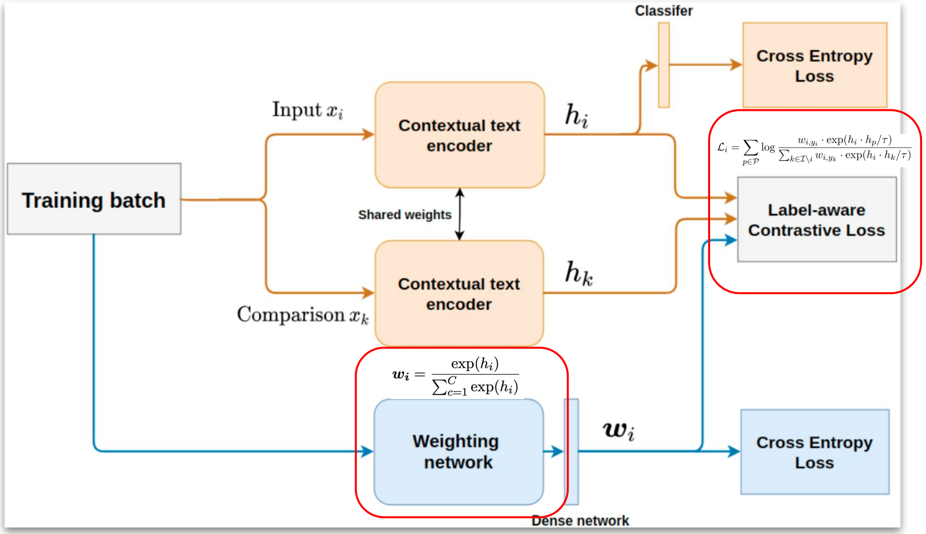
CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding (Qu et al., 2020)

Classification: Not all negatives are equal

- Sample construction
  - Use existing positive/negative samples
  - Positive samples: samples in the same minibatch with the same labels
  - Negative samples: samples in the same minibatch with different labels
- Training
  - A weighted label-aware contrastive loss

$$\mathcal{L}_i = \sum_{p \in \mathcal{P}} \log \frac{w_{i,y_i} \cdot \exp(h_i \cdot h_p/\tau)}{\sum_{k \in \mathcal{I} \setminus i} w_{i,y_k} \cdot \exp(h_i \cdot h_k/\tau)}$$

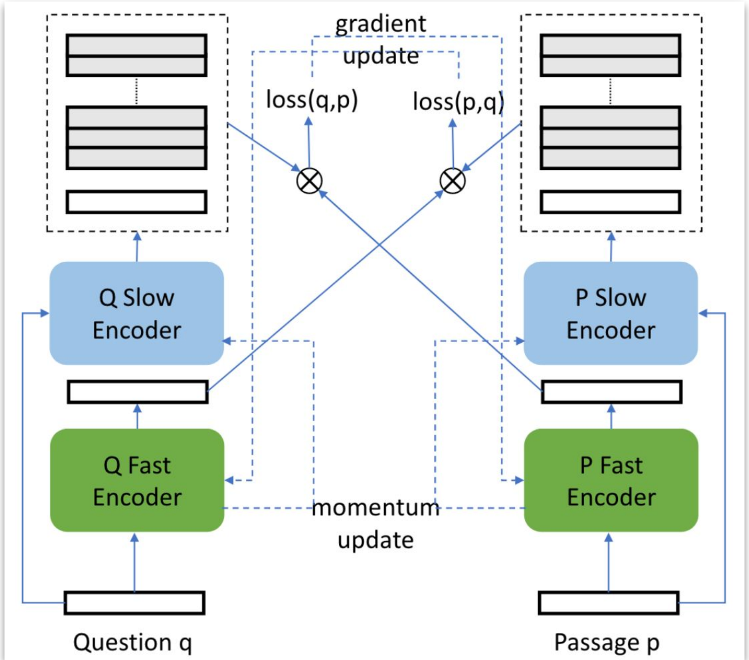
$$\boldsymbol{w_i} = \frac{\exp(h_i)}{\sum_{c=1}^{C} \exp(h_i)}$$



[Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification (Suresh and Ong, 2021)]

Question Answering: xMoCo

- Sample construction
  - Use passages (without the corresponding questions) from the original training data as *negative* examples
- Training
  - Address the asymmetric issue in question-passage pairs for the momentum contrastive learning
  - Employ two sets of fast/slow encoders and jointly optimize the question-passage matching task
    - Fast encoders: trained with gradients
    - Slow encoders: trained with momentum updates



[xMoCo: Cross Momentum Contrastive Learning for Open-Domain Question Answering (Yang et al., 2021)](#)

Question Answering: xMoCo (Cont.)

The same model architecture can be used in any other scenarios with *asymmetric* input pairs

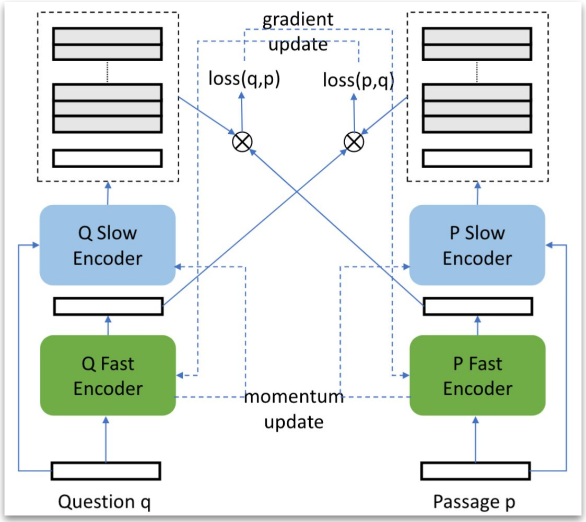


Figure: xMoCo

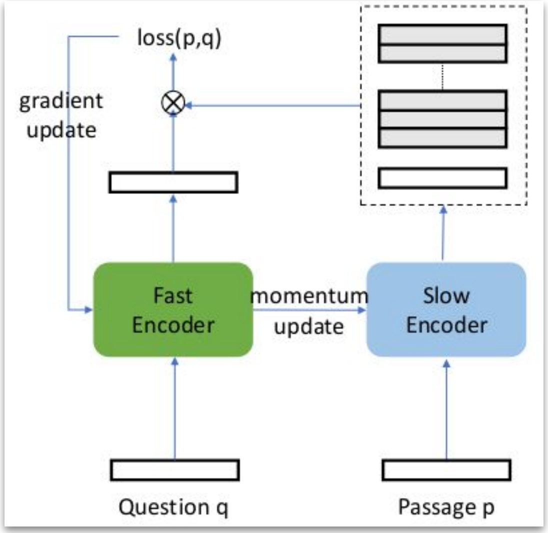
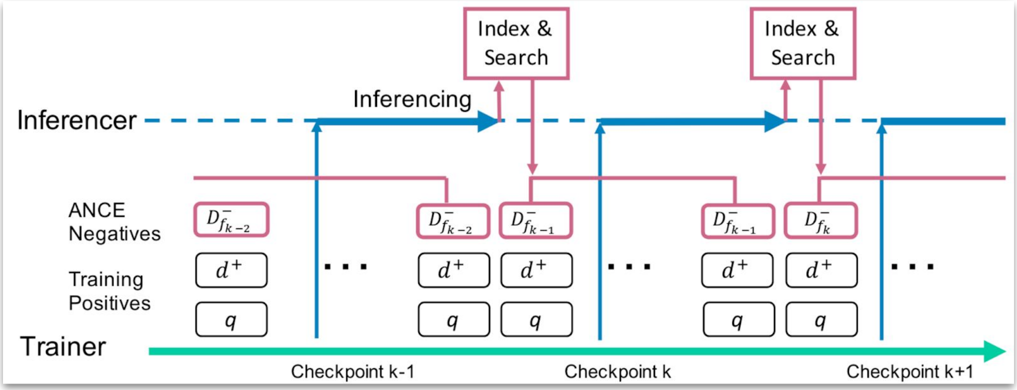


Figure: MoCo

[xMoCo: Cross Momentum Contrastive Learning for Open-Domain Question Answering (Yang et al., 2021)]

Question Answering: ANCE

- Goal:
  - The bottleneck of dense retrieval is the domination of uninformative negatives sampled in mini-batch training
- Sample construction
  - Obtain negative samples from the top retrieved documents
  - By definition, they are the hardest negatives for the current model
- Training
  - Asynchronous Index Refresh (Guu et al., 2020)

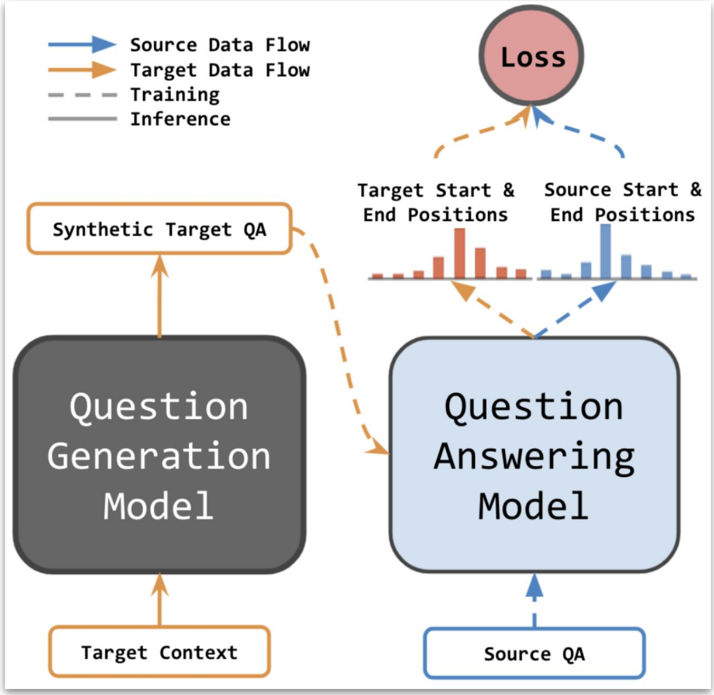


[Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval (Xiong et al., 2021; ANCE)]
[REALM: Retrieval-Augmented Language Model Pre-Training (Guu et al., 2020)]

Question Answering: CAQA

- Problem definition
  - Source domain: (context, question, answer)
  - Target domain: context only
- Sample construction
  - Generate question-answer pairs from target context using QAGen-T5
- Training
  - Define the contrastive adaptation loss on a mini-batch with samples from both source and target domain

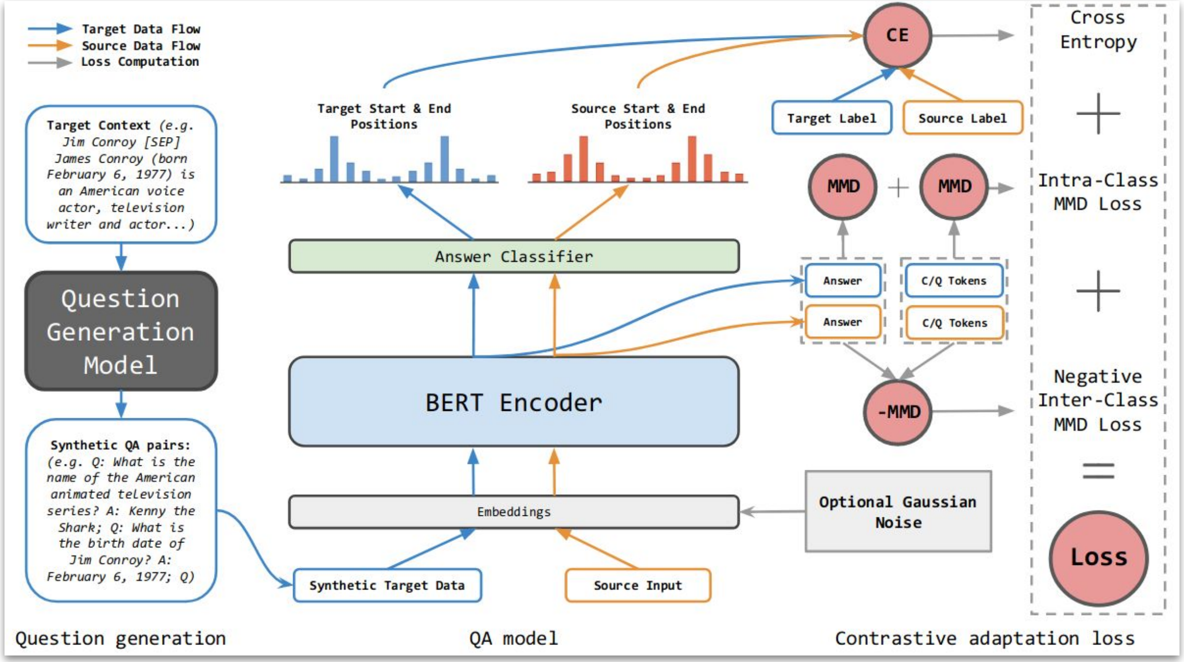
$$
\begin{aligned}
\mathcal{L}_{\text{con}}(\boldsymbol{X}) = & \frac{1}{|\boldsymbol{X}|^2} \sum_{i=1}^{|\boldsymbol{X}|} \sum_{j=1}^{|\boldsymbol{X}|} k(\phi(\boldsymbol{x}_{\text{a}}^{(i)}), \phi(\boldsymbol{x}_{\text{a}}^{(j)})) \\
& + \frac{1}{|\boldsymbol{X}|^2} \sum_{i=1}^{|\boldsymbol{X}|} \sum_{j=1}^{|\boldsymbol{X}|} k(\phi(\boldsymbol{x}_{\text{cq}}^{(i)}), \phi(\boldsymbol{x}_{\text{cq}}^{(j)})) \\
& - \frac{1}{|\boldsymbol{X}|^2} \sum_{i=1}^{|\boldsymbol{X}|} \sum_{j=1}^{|\boldsymbol{X}|} k(\phi(\boldsymbol{x}_{\text{a}}^{(i)}), \phi(\boldsymbol{x}_{\text{cq}}^{(j)}))
\end{aligned}
$$



Contrastive Domain Adaptation for Question Answering using Limited Text Corpora (Yue et al., 2021)

Question Answering: CAQA (Cont.)

The whole training pipeline with three components in the loss function



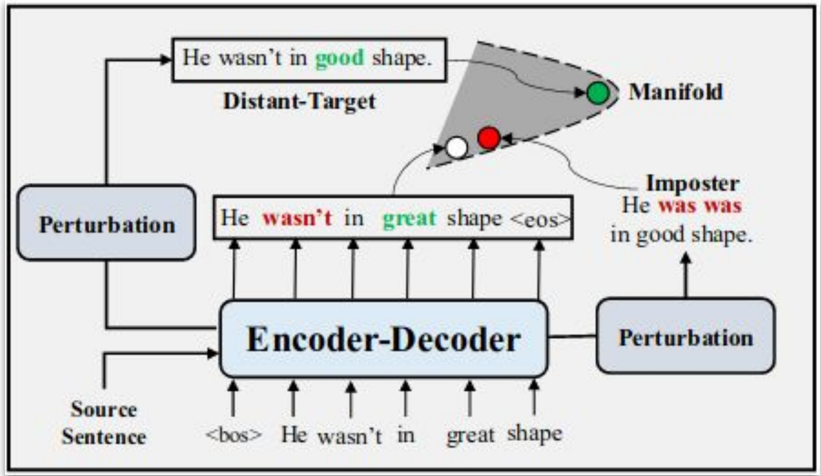
[Contrastive Domain Adaptation for Question Answering using Limited Text Corpora (Yue et al., 2021)]

Text Generation: CLAPS

In conditional text generation

- Sample construction
  - Negative: adding *small perturbation* to minimize the conditional likelihood
  - Positive: adding *large perturbation* while enforcing a high conditional likelihood
- Training
  - Maximize the similarity between the positive pairs and minimize the similarity between negative pairs

$$\mathcal{L}_{cont}(\theta) = \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_\mathbf{x}^{(i)}, \mathbf{z}_\mathbf{y}^{(i)})/\tau)}{\sum_{\mathbf{z}_\mathbf{y}^{(j)} \in S} \exp(\text{sim}(\mathbf{z}_\mathbf{x}^{(i)}, \mathbf{z}_\mathbf{y}^{(j)})/\tau)}$$



[Contrastive Learning with Adversarial Perturbations for Conditional Text Generation (Lee et al., 2021)]

Text Generation: Counter-contrastive learning

- Sample construction
  - Positive: feed the original sentence into the discriminator twice with *different dropout masks*
  - Negative: generate *random sentences* from the pre-trained generator
- Training
  - The counter-contrastive learning objective

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

Algorithm 1 Adversarial Training of CCL.

```
Require: generator G_θ; discriminator D_φ; samples of real data S; generator training step g; discriminator training step k; the generator pretraining epochs l.
Pretrain G_θ using MLE on S for l epochs
repeat
    for g steps do
        Sample a minibatch from real data S
        Generate a minibatch from G_θ
        Construct positive pairs by feeding the real samples to D_φ twice with different dropout masks, and negative samples from x_i^- ~ G_θ.
        Update G_θ via Eq. (1)
        Update G_θ via Eq. (3) (CCL training)
    end for
    for k steps do
        Sample a minibatch from real data S
        Sample a minibatch from the generated data
        Train the discriminator D_φ by Eq. (1)
    end for
until convergence
```

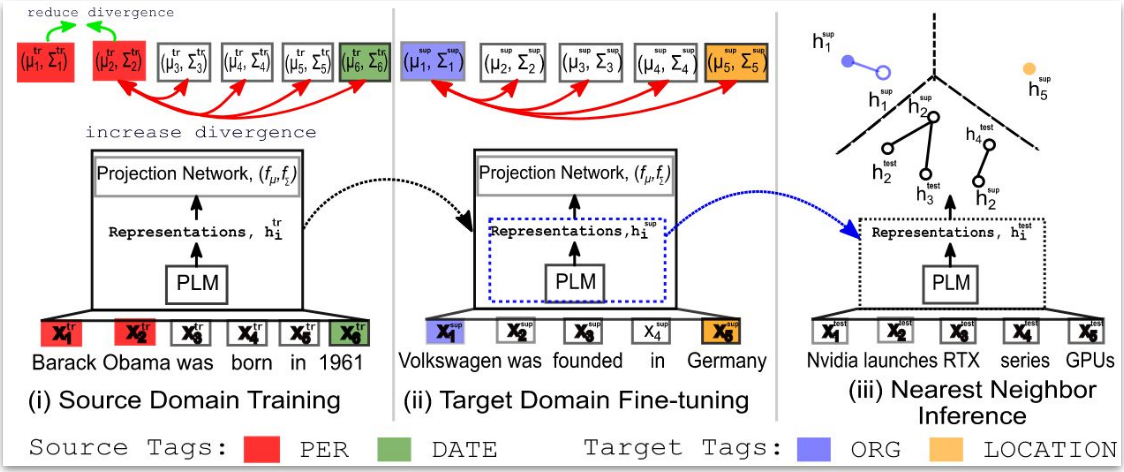
[Counter-Contrastive Learning for Language GANs (Chai et al., 2021)](#)

Named Entity Recognition

- Words in the same NER category should have similar embeddings
- Representing words with Gaussian embeddings, instead of single vectors
- Sample construction: positive - words in the same category
- Training

$$\ell(p) = -\log \frac{\sum_{(x_q,y_q)\in\mathcal{X}_p} \exp(-d(p,q))/|\mathcal{X}_p|}{\sum_{(x_q,y_q)\in\mathcal{X},p\neq q} \exp(-d(p,q))}$$

$$\mathcal{X}_p = \{(x_q, y_q) \in \mathcal{X} \mid y_p = y_q, p \neq q\}$$



[CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning (Das et al., 2022)]

Advantages of Contrastive Learning for NLP

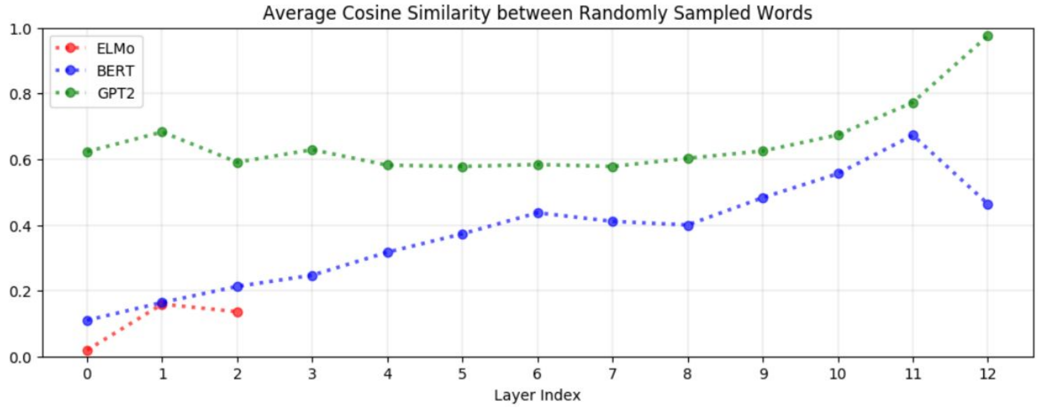
In addition to the performance benefits, we can also summarize the advantages of contrastive learning for NLP from the following four aspects:

- Task-agnostic representations
- Faithful text generation
- Data-efficient learning
- Interpretability and explainability
  - *Discussed in the next section*

On Sentence Representations

Some open questions in learning generic sentence representations

- Benefit downstream applications
  - E.g., Classification on the GLUE benchmark
- Avoid anisotropy (Ethayarajh, 2019)
  - Anisotropic embedding space indices poor semantic similarity (Li et al., 2020)



[How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings (Ethayarajh, 2019)]

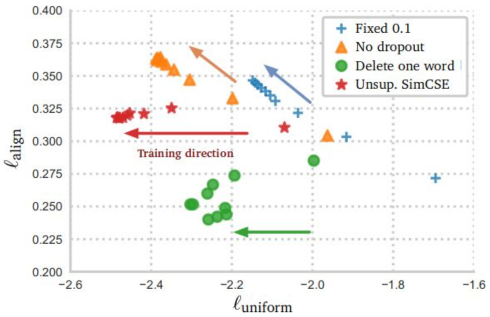
Task-agnostic Sentence Representations

Contrastive learning can help learning task-agnostic representations

- For further pre-training (e.g., CERT; Fang et al., 2020)
  - Additional pre-training with contrastive loss
  - Use back-translation for sample construction
- Avoiding representation collapse (e.g., SimCSE; Gao et al., 2021)
  - Use dropout to create positive samples *implicitly*
  - Make representations from similar examples stay closer, and in general representations are uniformly distributed in the space (measured by *alignment* and *uniformity*)

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

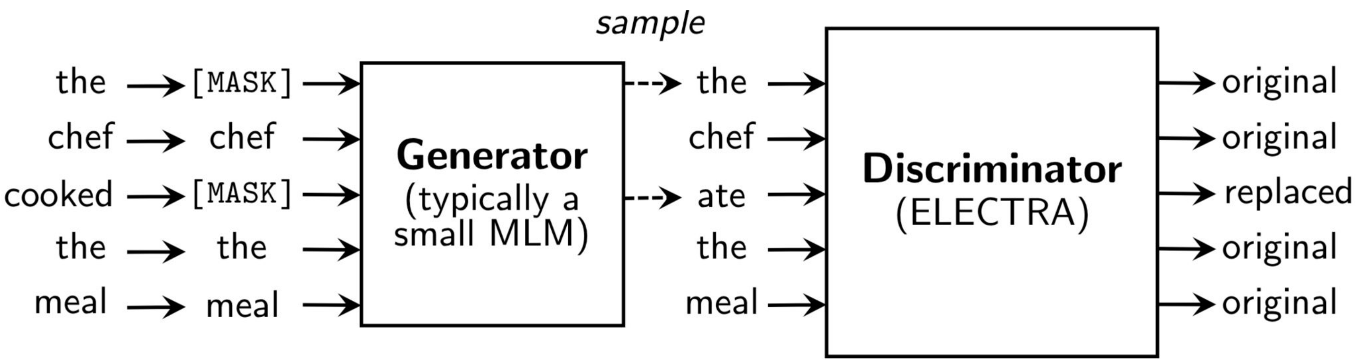
$$\ell_{\text{uniform}} \triangleq \log \mathop{\mathbb{E}}_{x,y \overset{i.i.d.}{\sim} p_{\text{data}}} e^{-2\|f(x)-f(y)\|^2}$$



[CERT: Contrastive Self-supervised Learning for Language Understanding (Fang et al., 2020)]
[SimCSE: Simple Contrastive Learning of Sentence Embeddings (Gao et al., 2021)]

For Pre-training: ELECTRA

- Instead of predicted masked tokens, ELECTRA takes a corrupted sentence and predict whether each token is *original/replaced*
- This prediction task is similar to Noise-Contrastive Estimation (NCE)

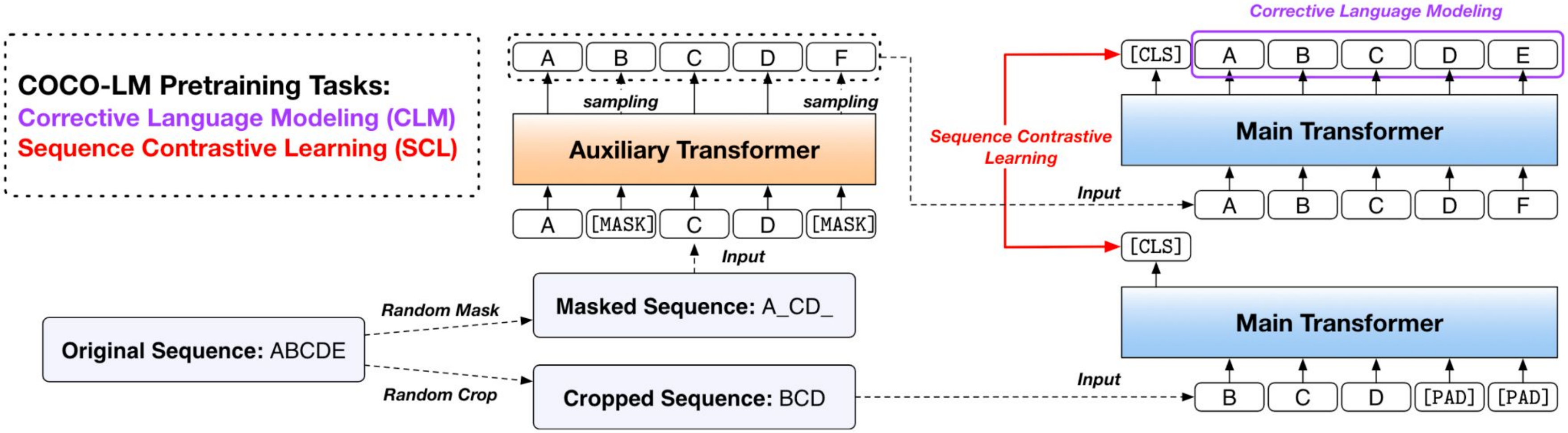


[ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators (Clark et al., 2020)](#)
[Noise-contrastive estimation: A new estimation principle for unnormalized statistical models (Gutmann and Hyvarinen, 2010)](#)

Pre-training: COCO-LM

COCO-LM proposes two training methods

- CLM: train the model to recover the original tokens, conceptually similar to ELECTRA
- SCL: align different views of the same input (created by data augmentation), against with unrelated negative instances



[COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining (Ment et al., 2021)](#)

On Text Generation: Hallucination

There are some common issues on current text generation approaches

- Hallucination (Maynez et al., 2020)
  - *Intrinsic hallucinations*: manipulating the information present in the input
  - *Extrinsic hallucinations*: adding information not directly inferable from the input

PTGEN	UKIP leader Nigel Goldsmith has been elected as the new mayor of London to elect a new Conservative MP.
TCONVS2S	Former London mayoral candidate Zac Goldsmith has been chosen to stand in the London mayoral election.
TRANS2S	Former London mayor Sadiq Khan has been chosen as the candidate to be the next mayor of London.
GPT-TUNED	Conservative MP Zac Goldwin's bid to become Labour's candidate in the 2016 London mayoral election.
BERTS2S	Zac Goldsmith has been chosen to contest the London mayoral election.

On Faithfulness and Factuality in Abstractive Summarization (Maynez et al., 2020)

Faithful and Factual-consistent Text Generation

Strategies in contrastive learning for faithful and factual-consistent text generation

- Leverage automatically generated texts as negative examples
  - E.g., for text summarization (Cao and Wang, 2021; CLIFF)

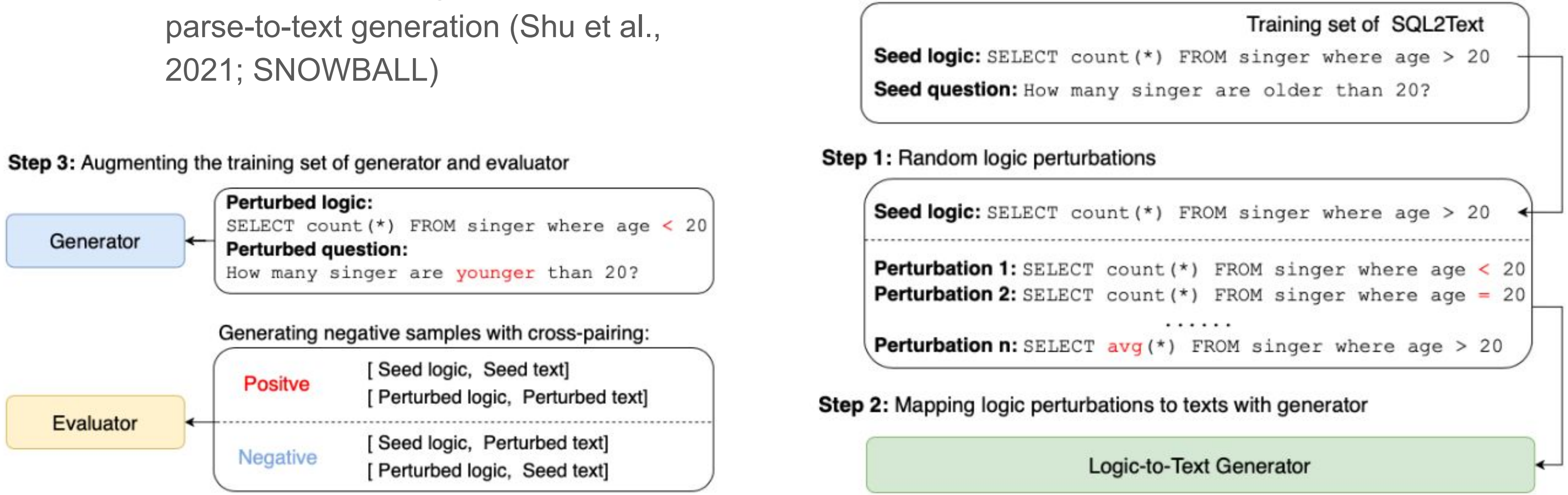
$$l_{cl}^{x} = -\frac{1}{\binom{|P|}{2}} \sum_{\substack{y_i, y_j \in P \\ y_j \neq y_i}} \log \frac{\exp(\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_j)/\tau)}{\sum_{\substack{y_k \in P \cup N \\ y_k \neq y_i}} \exp(\mathrm{sim}(\boldsymbol{h}_i, \boldsymbol{h}_k)/\tau)}$$

- P: a set of reference summaries
- U: a set of erroneous summaries

CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization (Cao and Wang, 2021)

Faithful and Factual-consistent Text Generation

- Alter inputs based on certain rules
  - Perturb the input logic forms in parse-to-text generation (Shu et al., 2021; SNOWBALL)

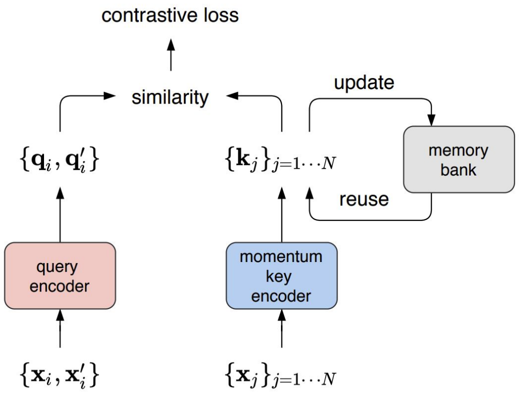


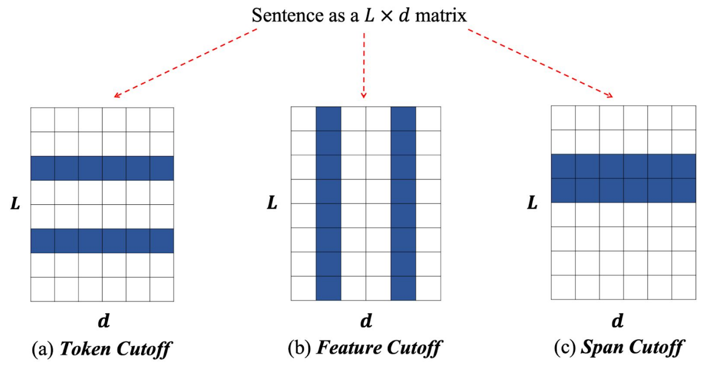
Logic-Consistency Text Generation from Semantic Parses (Shu et al., 2021)

On Data Efficiency

Contrastive learning can help address the data scarcity issues via many ways

- Via data augmentation and self-supervision
  - Synthesize contrast-enhanced and diverse examples (Qu et al., 2021; CoDA)
  - Augment training data with latent representation modification (Shen et al., 2020; Cutoff)

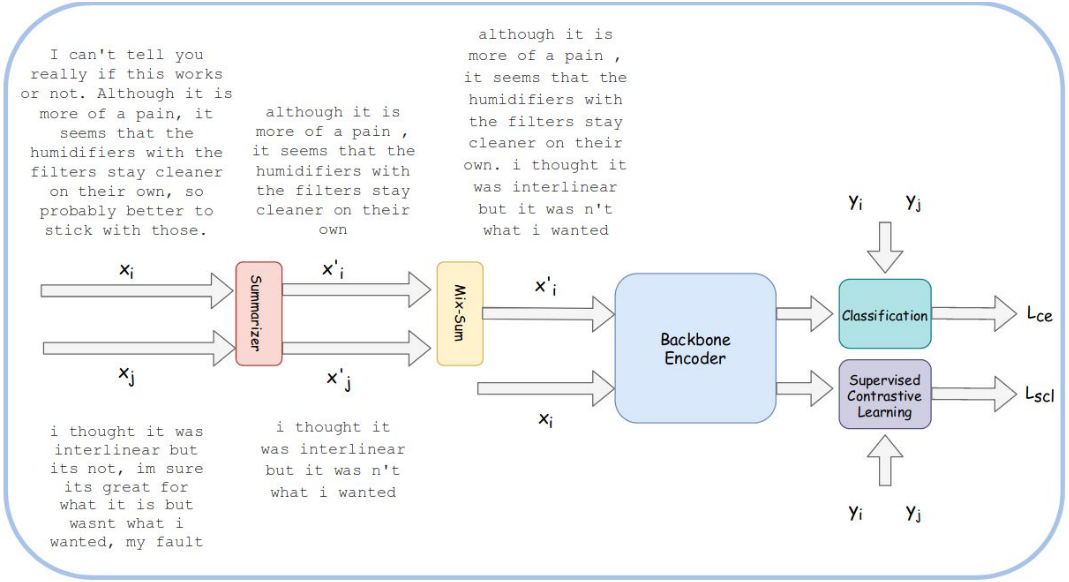




[CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding (Qu et al., 2020)]
[A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation (Shen et al., 2020)]

On Data Efficiency (Cont.)

- Via domain adaptation
  - Concatenate the text from the summary of a text and the residual words from another text regarding its summary (Du et al., 2021; mixsum)



[Constructing contrastive samples via summarization for text classification with limited annotations (Du et al., 2021)](#)

On Data Efficiency (Cont.)

- Selecting contrastive examples works better than traditional sample selection strategies in active learning

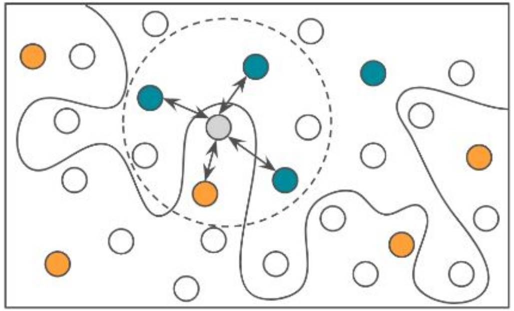


Figure 1: Illustrative example of our proposed method CAL. The solid line (model decision boundary) separates data points from two different classes (blue and orange), the coloured data points represent the labeled data and the rest are the unlabeled data of the pool.

for x_p *in* $\mathcal{D}_{pool}$ **do**

$\{(x_l^{(i)}, y_l^{(i)})\}, i = 1, ..., k \leftarrow \text{KNN}(\Phi(x_p), \Phi(\mathcal{D}_{\textbf{lab}}), k)$

$p(y|x_l^{(i)}) \leftarrow \mathcal{M}(x_l^{(i)}), i = 1, ..., k$

$p(y|x_p) \leftarrow \mathcal{M}(x_p)$

$\text{KL}(p(y|x_l^{(i)})||p(y|x_p)), i = 1, ..., k$

$s_{x_p} = \frac{1}{k} \sum_{i=1}^{k} \text{KL}(p(y|x_l^{(i)})||p(y|x_p))$

end

$Q = \underset{x_p \in \mathcal{D}_{\textbf{pool}}}{\text{argmax}} s_{x_p}, |Q| = b$

Output: Q

[Active Learning by Acquiring Contrastive Examples (Margatina et al., 2021)]

Interpretability and Robustness

Contrastive Explanation

- Why P

Q: Why did you rob a bank?

References:
[1] Miller, Tim. “Contrastive Explanation: a Structural-Model Approach.” The Knowledge Engineering Review 36 (2021): e14. doi:10.1017/S0269888921000102.

Contrastive Explanation

- Why P

Q: Why did you rob a bank?
A: Because that is where the money is.

References:
[1] Miller, Tim. “Contrastive Explanation: a Structural-Model Approach.” The Knowledge Engineering Review 36 (2021): e14. doi:10.1017/S0269888921000102.

Contrastive Explanation

- Why P
- Why P rather than Q

Q: Why dog?	Q: Why dog rather than cat?
Features: head, tail, run, head shape, tail shape, fur, bark	Features: ~~head~~, tail, ~~run~~, head shape, tail shape, ~~fur~~, bark
Shares features	

References:
[1] Miller, Tim. "Contrastive Explanation: a Structural-Model Approach." The Knowledge Engineering Review 36 (2021): e14. doi:10.1017/S0269888921000102.

Contrastive Explanation

- Why P
- Why P rather than Q
    - Social scientists show that explanations are contrastive.
    - Contrastive explanations facilitate modeling

References:
[1] Miller, Tim. "Contrastive Explanation: a Structural-Model Approach." The Knowledge Engineering Review 36 (2021): e14. doi:10.1017/S0269888921000102.

Contrastive Explanation for Commonsense Reasoning

i) I picked up a bag of peanuts and raisins for a snack. I wanted a sweeter snack out so I ate the __ for now. *Contrastive Expl. - Peanuts are salty while raisins tend to be sweet.*
ii) The geese prefer to nest in the fields rather than the forests because in the __ predators are more hidden. *Contrastive Expl. - Forests are denser than fields*

Table 1: Examples of Winograd Schema Instances where the correct and incorrect answer choices are highlighted in blue and red respectively. Choices are *contrasted* along attributes like taste (for i) and density of vegetation (for ii) by humans to explain why they prefer some answer choice.

References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Baseline
  - Input $\longrightarrow$ Output
- Self-talk
  - Input $\longrightarrow$ Why P? $\longrightarrow$ Output
- Contrastive Explanation
  - Input $\longrightarrow$ Why P rather than Q? $\longrightarrow$ Output

Contrastive Explanation for Commonsense Reasoning

- Source of Prompt
  - Human labeling of reasoning:
    - 64%--76% use contrast.
  - Select templates with ≥10 instances
  - Use the templates to prompt a pretrained language model

References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Templates

Complete list of Contrastive Prompt Templates	Commonsense Task/Instance Type
Temporal:	PIQA (Consists of events)
OPT1 happened before/after OPT2	
OPT1 takes longer than OPT2	
OPT1 takes longer to _ than OPT2	
OPT1 happened for a longer time than OPT2	

References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Templates

Complete list of Contrastive Prompt Templates	Commonsense Task/Instance Type
Personal Characteristics: OPT1 likes _ while OPT2 likes _ OPT1 likes _ while OPT2 does not like _ OPT1 likes to _ while OPT2 likes to _ OPT1 likes to _ while OPT2 does not like to _ OPT1 prefers _ while OPT2 prefers _ OPT1 prefers _ while OPT2 does not prefer _ OPT1 prefers to _ while OPT2 prefers to _ OPT1 prefers to _ while OPT2 does not prefer to _ OPT1 thinks _ while OPT2 thinks _ OPT1 thinks _ while OPT2 does not thinks _	WSC (if PERSON entity tag is detected)

References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Templates

Complete list of Contrastive Prompt Templates	Commonsense Task/Instance Type
Object Characteristic: OPT1 is/are smaller than OPT2 OPT1 is/are larger than OPT2 OPT1 is/are slower than OPT2 OPT1 is/are faster than OPT2 OPT1 is _ than OPT2 OPT1 are _ than OPT2 OPT1 is _ while OPT2 is _ OPT1 is _ but OPT2 is _ OPT1 is _ however OPT2 is _ OPT1 are _ while OPT2 are _ OPT1 are _ but OPT2 are _ OPT1 are _ however OPT2 are _ OPT1 has _ while/but/however OPT2 has/does not have _ OPT1 have _ while/but/however OPT2 have/do not have _ OPT1 is made of/to _ however OPT2 is made of/to _ OPT1 is made of/to _ while OPT2 is made of/to _	WSC and PIQA

References:

[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Templates

Complete list of Contrastive Prompt Templates	Commonsense Task/Instance Type
Spatial: OPT1 is above OPT2 OPT1 is below OPT2 OPT1 is to the right of OPT2 OPT1 is to the left of OPT2 OPT1 is inside OPT2 OPT1 is outside OPT2 _ is closer to OPT1 and father away from OPT2 OPT1 is closer to _ while OPT2 is father away from _	WSC and PIQA

References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Templates

Complete list of Contrastive Prompt Templates	Commonsense Task/Instance Type
Usecase: OPT1 can _ while OPT2 can/cannot _ OPT1 is/can be used for OPT2 OPT1 is/can be used to do OPT2 OPT1 is/can be used for _ but OPT2 cannot OPT1 is/can be used for _ while OPT2 is used for _ OPT1 is/can be s used for _ but OPT2 is used for _ OPT1 is/can be used to _ while OPT2 is used to _ OPT1 is/can be used to _ but OPT2 is used to _	WSC(No PERSON entity) and PIQA

References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Templates

Complete list of Contrastive Prompt Templates	Commonsense Task/Instance Type
Causes: OPT1 has _ because _ while OPT2 is _ because _ OPT1 can cause _ while OPT2 causes/results in _ Since _ it can OPT1 but not OPT2 Since _ it can OPT1 but because it is not _ it can't OPT2	WSC (No PERSON entity) and PIQA

References:

[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

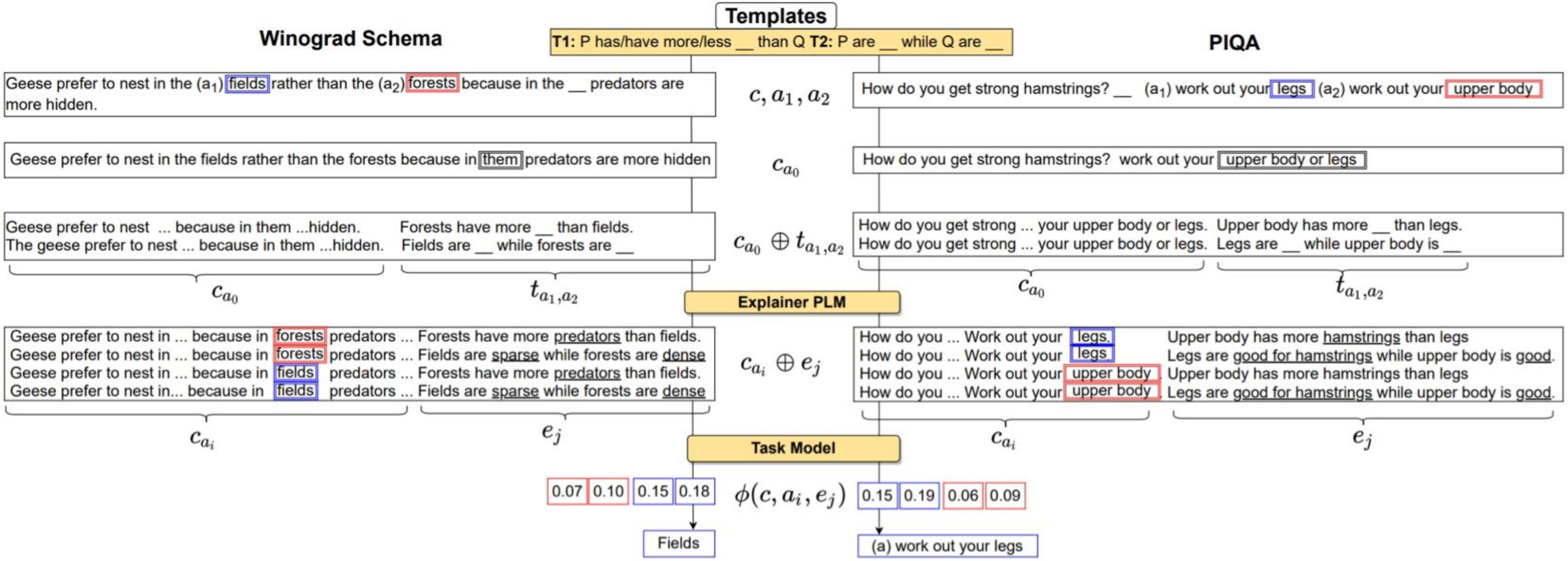
- Templates

Complete list of Contrastive Prompt Templates	Commonsense Task/Instance Type
Miscellaneous: _ can be OPT1 but cannot be OPT2 OPT1 means to _ while OPT2 means to _ OPT1 is defined as _ while OPT2 is defined as _ _ OPT1 _ OPT2 _ OPT1 but not OPT2 OPT1 exists while an OPT2 doesn't	WSC (No PERSON entity) and PIQA

References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Commonsense Reasoning

- Model



References:

[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

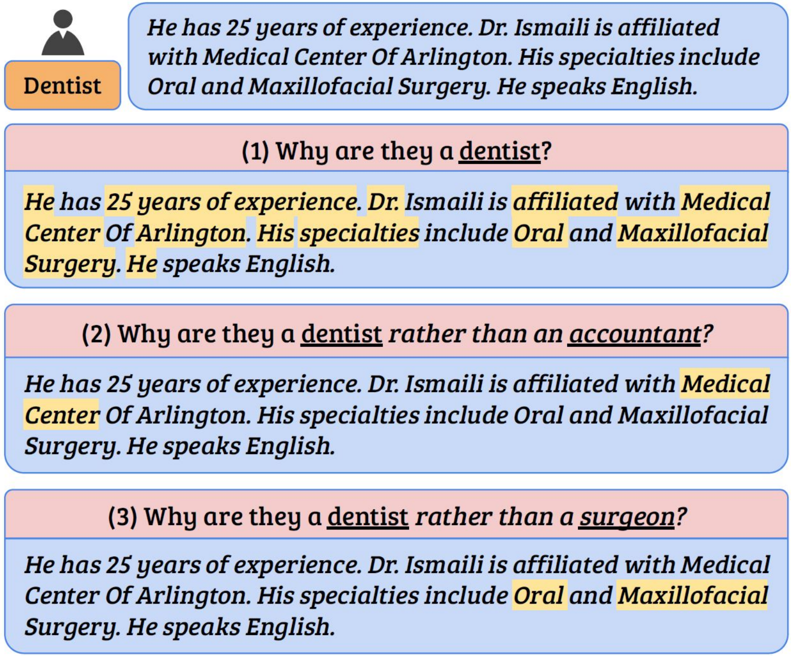
Contrastive Explanation for Commonsense Reasoning

- Results

	Explainer PLM (# Params)	Task model	WGRD ZS	FT	PIQA ZS	FT	WSC ZS	WGND ZS
1. Context-only	GPT2-XL (1.5B)	GPT2-XL	54.8	77.9	62.6	80.1	61.5	60.0
2. Unconstrained	GPT2-XL		54.9	77.8	63.9	80.7	61.4	60.0
3. Self-Talk	GPT2-XL		55.1	78.4	69.5	82.3	62.0	61.3
4. Contrastive	BART-Large(680M)		56.8	78.9	71.8	82.8	63.2	62.9
5. (Ours)	T5-Large (770M)		59.2	79.1	72.5	83.5	63.5	63.2
6.	T5-11B(11B)		**60.3**	**79.6**	**73.4**	**83.9**	**64.1**	**63.5**

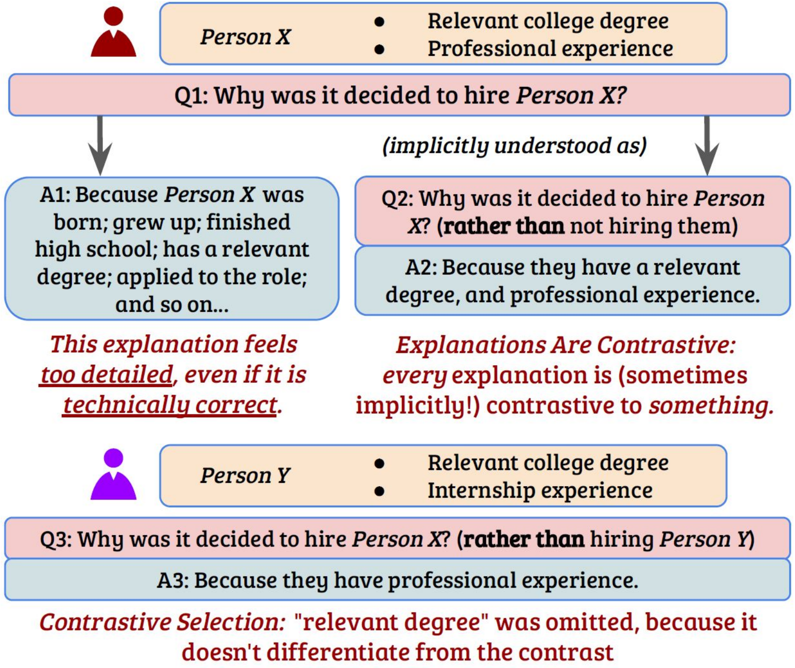
References:
[1] Paranjape, Bhargavi, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Prompting Contrastive Explanations for Commonsense Reasoning Tasks." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4179-4192. 2021.

Contrastive Explanation for Model Interpretability



References:
[1] Jacovi, Alon, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi and Yoav Goldberg. "Contrastive Explanations for Model Interpretability." EMNLP (2021).

Contrastive Explanation for Model Interpretability



References:
[1] Jacovi, Alon, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi and Yoav Goldberg. "Contrastive Explanations for Model Interpretability." EMNLP (2021).

Contrastive Explanation for Model Interpretability

- Find a latent contrastive representation in the input space

- Project input representation into a spece that minimally separates two decisions
  - Fact
  - Foil

- Measure Contrastiveness by computing behavior change before and after projection

References:
[1] Jacovi, Alon, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi and Yoav Goldberg. "Contrastive Explanations for Model Interpretability." EMNLP (2021).

Contrastive Explanation for Model Interpretability

- Results on NLI

Concept	Tot.	Gold			Predicted		
		%E	%C	%N	%E	%C	%N
Overlap	5.7	63.7	29.6	6.7	64.4	29.2	6.3
Hypothesis	52.3	33.9	33.1	32.9	49.1	24.7	26.2
Hyp-Neg	14.8	21.4	61.0	17.6	21.6	60.7	17.7

concept	fact	foil		
		E	C	N
Overlap	E	-	0.006	0.433
Hypothesis (MultiNLI)	E	-	-0.005	-0.031
Hyp-Negation	C	0.195	-	0.051

References:
[1] Jacovi, Alon, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi and Yoav Goldberg. "Contrastive Explanations for Model Interpretability." EMNLP (2021).

Contrastive Explanation for Model Interpretability

- Results on NLI

Fact	Foil (**gold**)	Input with Highlights
entailment		P: A nun uses her camera to take a photo of an interesting site.
	none	H: A nun taking photos of a interesting site outside.
	contradict.	H: A nun taking photos of a interesting site outside.
	neutral	H: A nun taking photos of a interesting site outside.
neutral		P: A couple bows their head as a man in a decorative robe reads from a scroll in Asia with a black late model station wagon in the background.
	none	H: A light black late model station wagon is in the background.
	entailment	H: A light black late model station wagon is in the background.
	contradict.	H: A light black late model station wagon is in the background.
neutral		P: Girl plays with colorful letters on the floor.
	none	H: The girl is having fun learning her letters.
	entailment	H: H: The girl is having fun learning her letters.
	contradict.	H: The girl is having fun learning her letters.
neutral		P: Three men with blue jerseys try to score a goal in soccer against the other team in white jerseys and their goalie in green.
	none	H: Some men with jerseys are in a bar, watching a soccer match.
	entailment	H: Some men with jerseys are in a bar, watching a soccer match.
	contradict.	H: Some men with jerseys are in a bar, watching a soccer match.

References:
[1] Jacovi, Alon, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi and Yoav Goldberg. "Contrastive Explanations for Model Interpretability." EMNLP (2021).

Contrastive Explanation for Model Interpretability

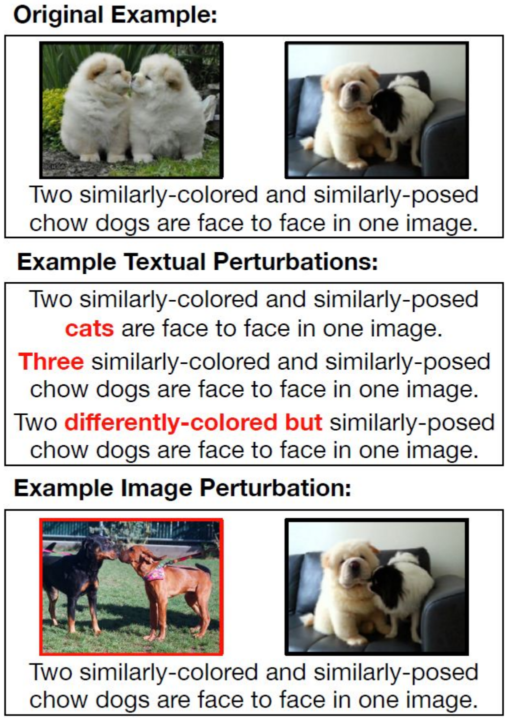
- Results on NLI

Biography / Profession / Gender
She also works as a Restitution Specialist while being the liaison to the Victim Compensation Board. Ms. Azevedo was named an OVSRS Outstanding Partner due to her dedication to providing critical information to staff so victims can obtain their court-ordered restitution while offenders can be held accountable. / **paralegal** / **F**
Peter also has substantial experience representing clients in government investigations, including criminal and regulatory investigations, and internal investigations conducted on behalf of clients. / **attorney** / **M**

References:
[1] Jacovi, Alon, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi and Yoav Goldberg. "Contrastive Explanations for Model Interpretability." EMNLP (2021).

Contrast and Robustness

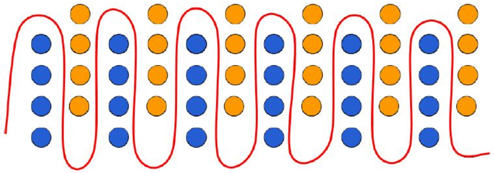
- Make small change in the data to alter the output



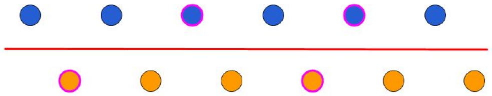
References:
[1] Gardner, Matt, Yoav Artzi, Jonathan Berant, Ben Bogin, Sihao Chen, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Eric Wallace, Ally Zhang and Ben Zhou. "Evaluating Models' Local Decision Boundaries via Contrast Sets." FINDINGS (2020).

Contrast and Robustness

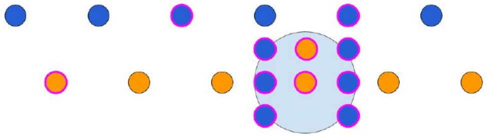
- Robustness issues and the idea for testing



(a) A two-dimensional dataset that requires a complex decision boundary to achieve high accuracy.



(b) If the same data distribution is instead sampled with systematic gaps (e.g., due to annotator bias), a simple decision boundary *can perform well on i.i.d. test data* (shown outlined in pink).



(c) Since filling in all gaps in the distribution is infeasible, a *contrast set* instead fills in a local ball around a test instance to evaluate the model's decision boundary.

References:

[1] Gardner, Matt, Yoav Artzi, Jonathan Berant, Ben Bogin, Sihao Chen, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Eric Wallace, Ally Zhang and Ben Zhou. "Evaluating Models' Local Decision Boundaries via Contrast Sets." FINDINGS (2020).

Contrast and Robustness

Dataset	Original Instance	Contrastive Instance (color = edit)
IMDb	Hardly one to be faulted for his ambition or his vision, it is genuinely unexpected, then, to see all Park's effort add up to so very little. . . . The premise is promising, gags are copious and offbeat humour abounds but it all fails miserably to create any meaningful connection with the audience. *(Label: Negative)*	Hardly one to be faulted for his ambition or his vision, **here we see all Park's effort come to fruition.** . . . The premise is **perfect**, gags are **hilarious** and offbeat humour abounds, **and it creates a deep** connection with the audience. *(Label: Positive)*

- Making minimum changes to differentiate data.

References:
[1] Gardner, Matt, Yoav Artzi, Jonathan Berant, Ben Bogin, Sihao Chen, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Eric Wallace, Ally Zhang and Ben Zhou. "Evaluating Models' Local Decision Boundaries via Contrast Sets." FINDINGS (2020).

Contrast and Robustness

- 10 Sets
- No model in the loop
- Human performance remains stable

Dataset	Original Test	Contrast Set	
IMDb	94.3	93.9	(–0.4)
PERSPECTRUM	91.5	90.3	(–1.2)
QUOREF	95.2	88.4	(–6.8)
ROPES	76.0	73.0	(–3.0)

References:
[1] Gardner, Matt, Yoav Artzi, Jonathan Berant, Ben Bogin, Sihao Chen, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Eric Wallace, Ally Zhang and Ben Zhou. "Evaluating Models' Local Decision Boundaries via Contrast Sets." FINDINGS (2020).

Contrast and Robustness

- Model performance decreases significantly

Dataset	# Examples	# Sets	Model	Original Test	Contrast		Consistency
NLVR2	994	479	LXMERT	76.4	61.1	(–15.3)	30.1
IMDb	488	488	BERT	93.8	84.2	(–9.6)	77.8
MATRES	401	239	CogCompTime2.0	73.2	63.3	(–9.9)	40.6
UD English	150	150	Biaffine + ELMo	64.7	46.0	(–18.7)	17.3
PERSPECTRUM	217	217	RoBERTa	90.3	85.7	(–4.6)	78.8
DROP	947	623	MTMSN	79.9	54.2	(–25.7)	39.0
QUOREF	700	415	XLNet-QA	70.5	55.4	(–15.1)	29.9
ROPES	974	974	RoBERTa	47.7	32.5	(–15.2)	17.6
BoolQ	339	70	RoBERTa	86.1	71.1	(–15.0)	59.0
MC-TACO	646	646	RoBERTa	38.0	14.0	(–24.0)	8.0

References:
[1] Gardner, Matt, Yoav Artzi, Jonathan Berant, Ben Bogin, Sihao Chen, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Eric Wallace, Ally Zhang and Ben Zhou. "Evaluating Models' Local Decision Boundaries via Contrast Sets." FINDINGS (2020).

Contrast and Robustness

- On Sentiment
- Counterfactual data labeling

Types of Revisions	Examples
Recasting *fact* as *hoped for*	The world of Atlantis, hidden beneath the earth's core, is fantastic The world of Atlantis, hidden beneath the earth's core is **supposed** to be fantastic
Suggesting sarcasm	thoroughly captivating **thriller-drama, taking a deep and realistic** view thoroughly mind numbing **"thriller-drama", taking a "deep" and "realistic" (who are they kidding?)** view
Inserting modifiers	The presentation of simply Atlantis' landscape and setting The presentation of Atlantis' **predictable** landscape and setting
Replacing modifiers	"Election" is a highly fascinating and thoroughly **captivating** thriller-drama "Election" is a highly expected and thoroughly **mind numbing** "thriller-drama"
Inserting phrases	Although there's hardly any action, the ending is still shocking. Although there's hardly any action **(or reason to continue watching past 10 minutes)**, the ending is still shocking.
Diminishing via qualifiers	which, while usually containing some reminder of harshness, become **more and more intriguing**. which, usually containing some reminder of harshness, became **only slightly more intriguing**.
Differing perspectives	Granted, **not all of the story makes full sense**, but the film doesn't feature any amazing new computer-generated visual effects. Granted, **some of the story makes sense**, but the film doesn't feature any amazing new computer-generated visual effects.
Changing ratings	one of the worst ever scenes in a sports movie. **3 stars out of 10**. one of the wildest ever scenes in a sports movie. **8 stars out of 10**.

References:
[1] Kaushik, Divyansh, Eduard H. Hovy and Zachary Chase Lipton. "Learning the Difference that Makes a Difference with Counterfactually-Augmented Data." ArXiv abs/1909.12434 (2020): n. pag.

Contrast and Robustness

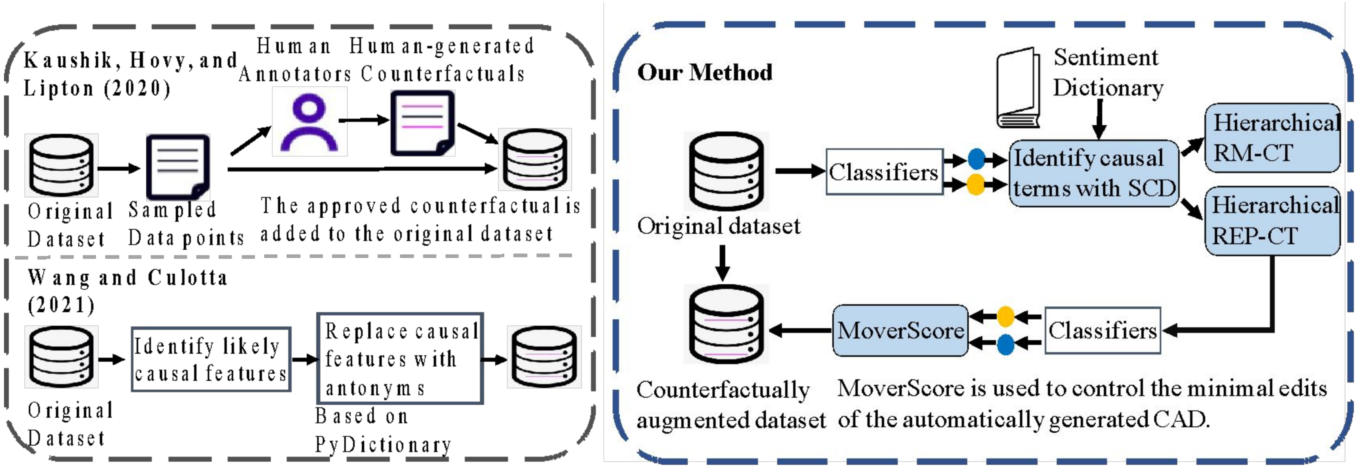
- Results

Training data	SVM		NB		ELMo		Bi-LSTM		BERT	
	O	R	O	R	O	R	O	R	O	R
Orig. (1.7k)	**80.0**	51.0	**74.9**	47.3	**81.9**	66.7	**79.3**	55.7	**87.4**	82.2
Rev. (1.7k)	58.3	**91.2**	50.9	**88.7**	63.8	**82.0**	62.5	**89.1**	80.4	**90.8**
Orig. – Edited	57.8	–	59.1	–	50.3	–	60.2	–	49.2	–
Orig. & Rev. (3.4k)	83.7	**87.3**	**86.1**	**91.2**	**85.0**	**92.0**	**81.5**	**92.0**	88.5	**95.1**
Orig. (3.4k)	**85.1**	54.3	82.4	48.2	82.4	61.1	80.4	59.6	**90.2**	86.1
Orig. (19k)	**87.8**	60.9	84.3	42.8	86.5	64.3	86.3	68.0	93.2	88.3
Orig. (19k) & Rev.	**87.8**	**76.2**	**85.2**	**48.4**	**88.3**	**84.6**	**88.7**	**79.5**	**93.2**	**93.9**

References:
[1] Kaushik, Divyansh, Eduard H. Hovy and Zachary Chase Lipton. “Learning the Difference that Makes a Difference with Counterfactually-Augmented Data.” ArXiv abs/1909.12434 (2020): n. pag.

Contrast and Robustness

- Approach



Overview of previous CAD methods are shown on the left side, while the pipeline of our method is shown on the right. Hierarchical RM-CT and Hierarchical REP-CT (are our methods for automatically generating CAD, respectively. SCD denotes sampling and sensitivity of contextual decomposition. Sentiment Dictionary refers to the opinion lexicon published by Hu and Liu. [2]

References:

[1] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. “Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis.” ArXiv abs/2106.15231 (2021): n. pag.

[2] Hu, Minqing and Bing Liu. “Mining and summarizing customer reviews.” Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (2004): n. pag.

Contrast and Robustness

- Results

Models	Parameter	Training / Testing data				AC: (Our method)			
		O/O	CF/O	CF/CF	O/CF	C/O	AC/O	C/CF	AC/CF
SVM(TF-IDF)	-	80.0	58.3	91.2	51.0	83.7	**84.8**	87.3	86.1
Bi-LSTM	0.2M	79.3	62.5	89.1	55.7	81.5	**82.2**	92.0	88.5
Transformer-based Models									
BERT [ICLR,2021]	110M	87.4	80.4	90.8	82.2	88.5	**90.6**	95.1	92.2
WWM-BERT-Large	335M	91.2	86.9	96.9	93.0	91.0	**91.8**	95.3	94.1
XLNet-Large	340M	**95.3**	90.8	98.0	93.9	93.9	**94.9**	96.9	95.5
RoBERTa-Large	355M	93.4	91.6	96.9	93.0	93.6	**94.1**	96.7	94.3

References:
[1] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. “Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis.” ArXiv abs/2106.15231 (2021): n. pag.

Contrast and Robustness

- Two different robustness issues

Sentence	Label	Predict
creepy but ultimately unsatisfying thriller	Negative	Negative
creepy but lastly unsat-isfying thriller	Negative	Positive
creepy but ultimately satisfying thriller	Positive	Negative

References:
[1] Wang, Dong, Ning Ding, Pijian Li and Haitao Zheng. "CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding." ArXiv abs/2107.00440 (2021): n. pag.

Contrast and Robustness

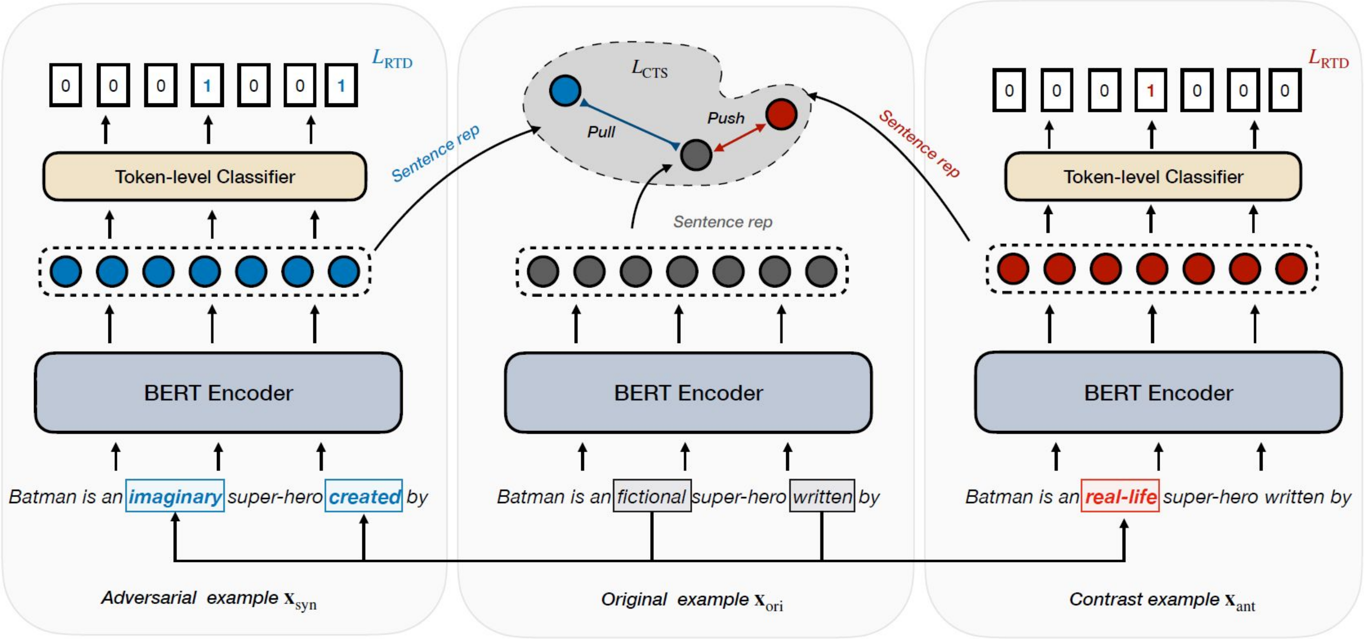
- Adversarial and contrastive examples are different

Model	Method	IMDB Adv	IMDB Rev	SNLI Adv	SNLI Rev
BERT-base	Vanilla	88.7	89.8	48.6	73.0
	FreeLB	91.9 (+3.2)	87.7 (−2.1)	56.1 (+7.5)	71.4 (−1.6)
RoBERTa-base	Vanilla	93.9	93.0	55.1	75.2
	FreeLB	95.2 (+1.3)	92.6 (−0.4)	58.1 (+3.0)	74.6 (−0.6)

References:
[1] Wang, Dong, Ning Ding, Pijian Li and Haitao Zheng. “CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding.” ArXiv abs/2107.00440 (2021): n. pag.

Contrast and Robustness

- Trained on Book Corpus and Wikipedia.



References:

[1] Wang, Dong, Ning Ding, Pijian Li and Haitao Zheng. "CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding." ArXiv abs/2107.00440 (2021): n. pag.

Contrast and Robustness

Rev – adversarial data

Con – contrastive data

Model	IMDB			PERSPECTRUM			BoolQ			SNLI		
	Ori	Rev	Con	Ori	Rev	Con	Ori	Rev	Con	Ori	Rev	Con
BERT	92.2	89.8	82.4	74.7	72.8	57.6	60.9	57.6	36.1	89.8	73.0	65.1
RoBERTa	93.6	93.0	87.1	80.6	78.8	65.0	69.6	60.6	43.9	90.8	75.2	67.8
CLINE	**94.5**	**93.9**	**88.5**	**81.6**	**80.2**	**72.2**	**73.9**	**63.9**	**47.8**	**91.3**	**76.0**	**69.2**

References:
[1] Wang, Dong, Ning Ding, Pijian Li and Haitao Zheng. “CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding.” ArXiv abs/2107.00440 (2021): n. pag.

Contrast and Robustness

Counterfactual Data Augmentation

- Generating manual counterfactuals [1]:-- expensive and time-consuming
- Fully automatic generation [2]:-- task-specific; dictionary-dependent
- An Example of Spurious Patterns in Sentiment Analysis:

Raw: "**Nolan's films** always **shock** people, thanks to his **superb** directing skills" **-- POS**

Artifacts: "**Martin**'s **movies** always shock people, thanks to his superb directing skills" **-- NEG**

References:
[1] Kaushik, Divyansh, Amrith Rajagopal Setlur, Eduard H. Hovy and Zachary Chase Lipton. "Explaining The Efficacy of Counterfactually-Augmented Data." ArXiv abs/2010.02114 (2021): n. pag.
[2] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. "Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis." ArXiv abs/2106.15231 (2021): n. pag.
[3] Lu, Jinghui, Linyi Yang, Brian Mac Namee and Yue Zhang. "A Rationale-Centric Framework for Human-in-the-loop Machine Learning." ACL (2022).

Contrast and Robustness

Semi-fact Data Augmentation + Dynamic Human Intervention

- Efficient
- Robust
- Model-agnostic

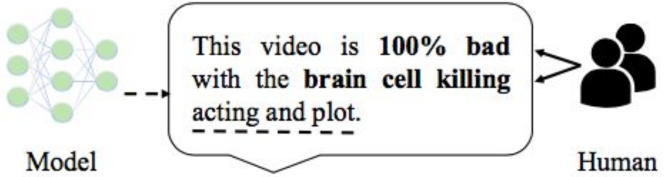


Figure 1: A negative movie review with human annotated causal terms (bold text) and spurious patterns recognised by the model (underlined text).

Rationales: ***"100% bad"*** and ***"brain cell killing"***
Spurious Patterns: "<u>acting and plot</u>"

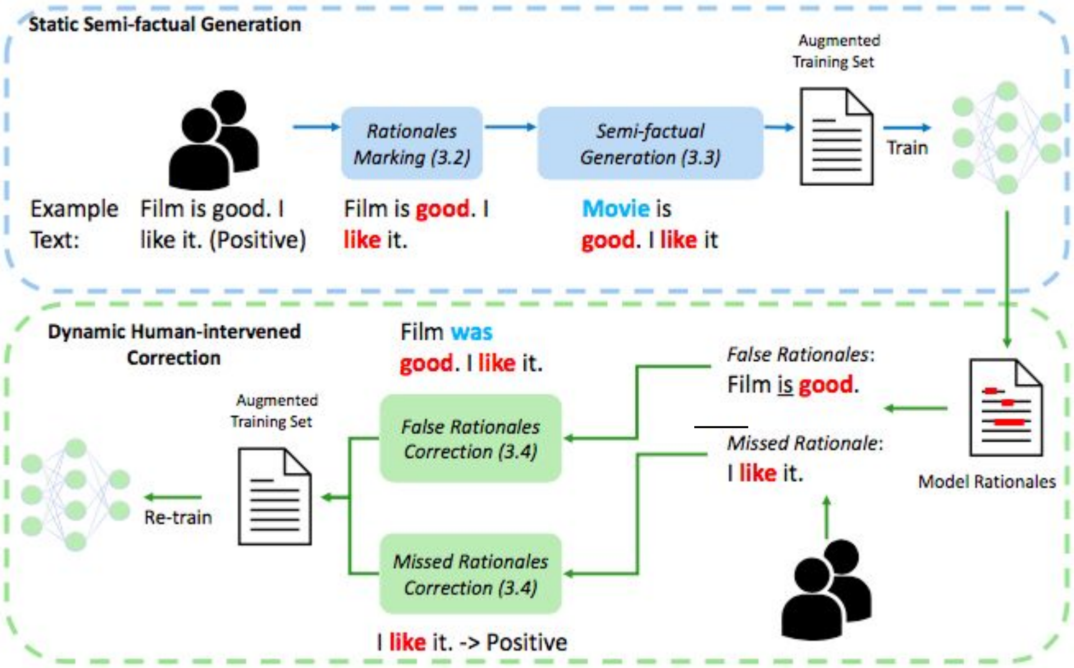
References:
[1] Kaushik, Divyansh, Amrith Rajagopal Setlur, Eduard H. Hovy and Zachary Chase Lipton. "Explaining The Efficacy of Counterfactually-Augmented Data." ArXiv abs/2010.02114 (2021): n. pag.
[2] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. "Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis." ArXiv abs/2106.15231 (2021): n. pag.
[3] Lu, Jinghui, Linyi Yang, Brian Mac Namee and Yue Zhang. "A Rationale-Centric Framework for Human-in-the-loop Machine Learning." ACL (2022).

Contrast and Robustness

Red text highlights rationales identified by human annotators.

Blue text indicates words replaced in raw text.

<u>Underlined text </u>shows spurious patterns identified by the model.



References:

[1] Kaushik, Divyansh, Amrith Rajagopal Setlur, Eduard H. Hovy and Zachary Chase Lipton. "Explaining The Efficacy of Counterfactually-Augmented Data." ArXiv abs/2010.02114 (2021): n. pag.

[2] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. "Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis." ArXiv abs/2106.15231 (2021): n. pag.

[3] Lu, Jinghui, Linyi Yang, Brian Mac Namee and Yue Zhang. "A Rationale-Centric Framework for Human-in-the-loop Machine Learning." ACL (2022).

Contrast and Robustness

Sentiment	Examples
Negative	Origin: The attempt at a "lesbian scene" was **sad**. Augment 1: The hint at a "lesbian scene" was **sad** . Augment 2: The attempt at a "kiss scene" was **sad** .
Positive	Origin: I **recommended** this film a lot, specially in this difficult times for the planet . Augment 1: I **recommended** you film a lot, specially in this difficult times for the planet . Augment 2: I **recommended** this movie a lot, specially in this difficult times for the planet .

Blue spans were synonyms used as replacements and **bold font** were rationales identified by human annotators.

Average time to identify rationales in a review: 183.68 seconds (OUR METHOD)

Average time to generate a counterfactual review: average 300 seconds

Given the fact that our approach using 100 labelled examples can outperform manual CAD [1] using the entire training set of 1,707 examples.

Our approach is **27.88** times more efficient than manually generated CAD.

References:

[1] Kaushik, Divyansh, Amrith Rajagopal Setlur, Eduard H. Hovy and Zachary Chase Lipton. "Explaining The Efficacy of Counterfactually-Augmented Data." ArXiv abs/2010.02114 (2021): n. pag.

[2] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. "Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis." ArXiv abs/2106.15231 (2021): n. pag.

[3] Lu, Jinghui, Linyi Yang, Brian Mac Namee and Yue Zhang. "A Rationale-Centric Framework for Human-in-the-loop Machine Learning." ACL (2022).

Contrast and Robustness

Sentiment	Examples
Negative	Origin: but this is **pathetic**! Micawber was nothing more than a mid-nineteenth century Kramer. SCD: but this is **pathetic**! <u>Micawber was</u> nothing more than a mid-nineteenth century Kramer. Augment 1: but this is **pathetic**! Perkins became nothing more than a mid-nineteenth century Kramer. Augment 2: but this is **pathetic**! It had nothing more than a mid-nineteenth century Kramer.
Positive	Origin: Soylent Green is a wild movie that I **enjoyed** very much . SCD: <u>Soylent Green</u> is a wild movie that I **enjoyed** very much . Augment 1: Gang Orange is a wild movie that I **enjoyed** very much . Augment 2: Village Spring is a wild movie that I **enjoyed** very much .

<u>Underlined spans </u>were false rationales given by the model through SCD. Blue spans were synonyms used as replacements, and **bold font** were rationales identified by human annotators.

- SCD: sampling and sensitivity of contextual decomposition – A post-hoc method to detect the model’s attention.

References:
[1] Kaushik, Divyansh, Amrith Rajagopal Setlur, Eduard H. Hovy and Zachary Chase Lipton. “Explaining The Efficacy of Counterfactually-Augmented Data.” ArXiv abs/2010.02114 (2021): n. pag.
[2] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. “Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis.” ArXiv abs/2106.15231 (2021): n. pag.
[3] Lu, Jinghui, Linyi Yang, Brian Mac Namee and Yue Zhang. “A Rationale-Centric Framework for Human-in-the-loop Machine Learning.” ACL (2022).

Contrast and Robustness

Baseline Methods	In-domain	SemEval-2017	SST-2	Yelp	Amazon
Static (50 gold)	88.60±1.11	77.28±9.11	79.29±5.14	91.53±2.06	89.63±1.65
Static + 350 auto (400)	90.16±0.85	80.54±2.81	81.26±1.97	93.03±1.08	90.09±1.79
AL (100 gold)	88.64±1.75	78.61±5.90	80.50±3.37	92.47±0.68	89.80±1.91
CAD-based Methods					
Manual CAD (3,414 gold)	**92.70±0.53**	69.98±3.99	80.30±2.03	91.87±1.09	90.48±1.09
Automatics CAD (1,707 gold+1,707 auto)	91.82±0.74	79.39±5.37	80.60±3.10	91.92±0.97	90.46±1.08
Our Dynamic Methods					
Dynamic (100 gold + 700 auto)	90.84±0.99	80.32±4.31	**82.40±2.14**	**93.19±1.24**	**90.51±2.17**
Dynamic-MR (100 gold + 700 auto)	91.06±1.21	79.04±4.92	82.24±2.59	93.03±1.92	90.22±2.74
Dynamic-FR (100 gold + 700 auto)	89.85±1.38	**82.39±1.88**	81.59±1.82	92.98±0.91	90.12±2.42

Average results from 10 times experiments. Results on in-distribution and OOD data. Values in brackets are the training set size. AL: Active Learning. Manual CAD [1], Automatic CAD [2]. Our methods are Dynamic-MR: Missing Rationale Correction, Dynamic-FR: False Rationale Correction, Dynamic: Dynamic Human-intervened Correction.

References:
[1] Kaushik, Divyansh, Amrith Rajagopal Setlur, Eduard H. Hovy and Zachary Chase Lipton. "Explaining The Efficacy of Counterfactually-Augmented Data." ArXiv abs/2010.02114 (2021): n. pag.
[2] Yang, Linyi, Jiazheng Li, P'adraig Cunningham, Yue Zhang, Barry Smyth and Ruihai Dong. "Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis." ArXiv abs/2106.15231 (2021): n. pag.
[3] Lu, Jinghui, Linyi Yang, Brian Mac Namee and Yue Zhang. "A Rationale-Centric Framework for Human-in-the-loop Machine Learning." ACL (2022).

Part 3.

Summary and Reflection

Summary

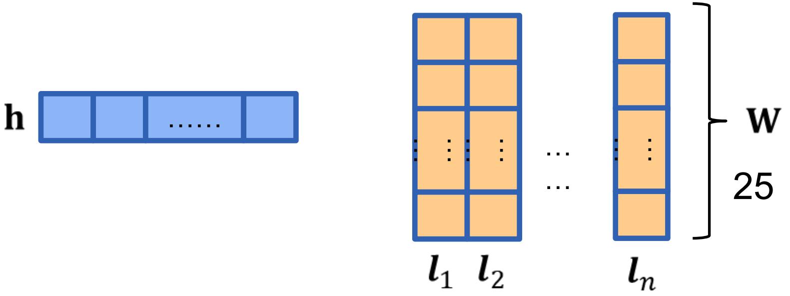
Contrast is a Broad Concept

- Has social scientific motivation
- Useful for model training, pre-training evaluation and interpretation
- Traditional training methods are also contrastive to some extent

Contrastive Learning VS Predictive Learning

- Predictive Learning

Using SoftMax as a typical example



$$\mathrm{SoftMax}(\mathbf{W} \cdot \mathbf{h})$$

$$\mathrm{dot}(s_i) = \mathbf{h} \cdot \boldsymbol{l}_i$$

$$\mathcal{L} = -\log \frac{e^{dot(s_g)}}{\sum_i e^{dot(s_i)}}$$

$\boldsymbol{l}_g$: gold label

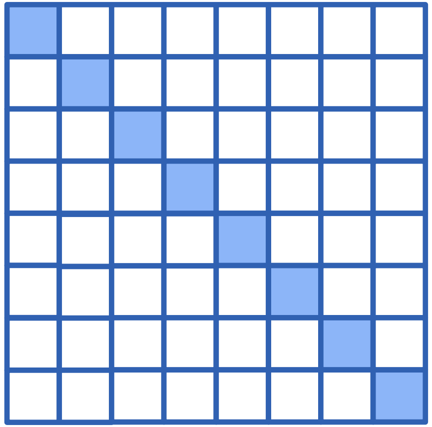
Contrastive Learning VS Predictive Learning

- Contrastive Learning

  Using SimCLR as a typical example

$$\cos_i = \frac{\mathbf{h}_a \cdot \mathbf{h}_i}{|\mathbf{h}_a| \cdot |\mathbf{h}_i|}$$

$$\mathcal{L} = -\log \frac{e^{\cos_+}}{\sum_i e^{\cos_i}}$$



Contrastive Learning VS Predictive Learning

$$\mathcal{L} = -\log \frac{e^{dot(s_g)}}{\sum_i e^{dot(s_i)}} \qquad \mathcal{L} = -\log \frac{e^{\cos_+}}{\sum_i e^{\cos_i}} \qquad \mathrm{dot} = \boldsymbol{a} \cdot \boldsymbol{b} \qquad \cos_i = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\mathbf{a}| \cdot |\mathbf{b}|}$$

- Both can be in the form of SoftMax/InfoMax.
- Both compute vector similarities with predictive learning focusing on similarities between hidden vectors and label embeddings.
- Contrastive learning uses normalization, calculating cosine. Some work [1] investigates it for predictive learning.
- Contrastive learning takes negative samples from a batch, while predictive learning takes all incorrect labels.

References:
[1] Wang, H., Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jin Zhou and Wenyu Liu. “CosFace: Large Margin Cosine Loss for Deep Face Recognition.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 5265-5274.

Contrastive Learning VS Predictive Learning

- Key elements include
    - What to contrast
    - How to make contrast
    - The goal
- These elements are correlated

References:
[1] Wang, H., Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jin Zhou and Wenyu Liu. "CosFace: Large Margin Cosine Loss for Deep Face Recognition." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 5265-5274.

What To Contrast

- Positive Samples
  - Perturbation
    - Back Translation [1]
    - Deletion [2]
    - Truncating [2]
    - Synonym Replacement [2]
    - Dropout [3]

References:
[1] Qu, Yanru, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han and Weizhu Chen. “CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding.” ArXiv abs/2010.08670 (2021): n. pag.
[2] Wu, Zhuofeng, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun and Hao Ma. “CLEAR: Contrastive Learning for Sentence Representation.” ArXiv abs/2012.15466 (2020): n. pag.
[3] Gao, Tianyu, Xingcheng Yao and Danqi Chen. “SimCSE: Simple Contrastive Learning of Sentence Embeddings.” ArXiv abs/2104.08821 (2021): n. pag.

What To Contrast

- Positive Samples
  - Perturbation
  - Matching Pairs
    - Image & Text [1]
    - Query & Doc [2]
    - Cross-lingual Tokens, Segments and Sentences [3]

References:
[1] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision." ICML (2021).
[2] Yang, Nan, Furu Wei, Binxing Jiao, Daxin Jiang and Linjun Yang. "xMoCo: Cross Momentum Contrastive Learning for Open-Domain Question Answering." ACL (2021).
[3] Li S, Yang P, Luo F, et al. Multi-Granularity Contrasting for Cross-Lingual Pre-Training[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1708-1717.

What To Contrast

- Positive Samples
  - Perturbation
  - Matching Pairs
  - Gold Labels [1] [2]

References:
[1] Gunel B, Du J, Conneau A, et al. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning[C]//International Conference on Learning Representations. 2020.
[2] Li L, Song D, Ma R, et al. KNN-BERT: Fine-Tuning Pre-Trained Models with KNN Classifier[J]. arXiv preprint arXiv:2110.02523, 2021.

What To Contrast

- Negative Samples
  - Different Instances in Batch [1] [2]
  - Influence of Batch Size [3] [4] [5]

References:
[1] Sohn, Kihyuk. “Improved Deep Metric Learning with Multi-class N-pair Loss Objective.” NIPS (2016).
[2] Chen, Ting, Simon Kornblith, Mohammad Norouzi and Geoffrey E. Hinton. “A Simple Framework for Contrastive Learning of Visual Representations.” ArXiv abs/2002.05709 (2020): n. pag.
[3] He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie and Ross B. Girshick. “Momentum Contrast for Unsupervised Visual Representation Learning.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 9726-9735.
[4] Yeh, Chun-Hsiao, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen and Yann LeCun. “Decoupled Contrastive Learning.” ArXiv abs/2110.06848 (2021): n. pag.
[5] Gao, Luyu, Yunyi Zhang, Jiawei Han and Jamie Callan. “Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup.” REPL4NLP (2021).

What To Contrast

- Negative Samples
  - Different Instances in Batch
  - Sampled negative instances by similarity

References:
[1] Schroff, Florian, Dmitry Kalenichenko and James Philbin. "FaceNet: A unified embedding for face recognition and clustering." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 815-823.
[2] Cui, Yin, Feng Zhou, Yuanqing Lin and Serge J. Belongie. "Fine-Grained Categorization and Dataset Bootstrapping Using Deep Metric Learning with Humans in the Loop." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 1153-1162.
[3] Li L, Song D, Ma R, et al. KNN-BERT: Fine-Tuning Pre-Trained Models with KNN Classifier[J]. arXiv preprint arXiv:2110.02523, 2021.

What To Contrast

- Negative Samples
  - Different Instances in Batch
  - Sampled negative instances by similarity
  - Hard Negative Samples [1] [2] [3]

References:
[1] Schroff, Florian, Dmitry Kalenichenko and James Philbin. “FaceNet: A unified embedding for face recognition and clustering.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 815-823.
[2] Cui, Yin, Feng Zhou, Yuanqing Lin and Serge J. Belongie. “Fine-Grained Categorization and Dataset Bootstrapping Using Deep Metric Learning with Humans in the Loop.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 1153-1162.
[3] Xia, Jun, Lirong Wu, Ge Wang, Jintao Chen and Stan Z.Li. “ProGCL: Rethinking Hard Negative Mining in Graph Contrastive Learning.” (2021).
[4] Li L, Song D, Ma R, et al. KNN-BERT: Fine-Tuning Pre-Trained Models with KNN Classifier[J]. arXiv preprint arXiv:2110.02523, 2021.

How To Contrast

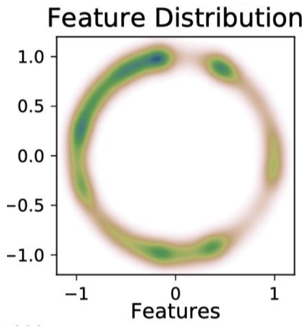
- Normalize Vectors
- Pairwise Similarity Score
- Pairwise Loss [1]
  - Max Margin [2]
  - Log-likelihood [3]

Two Standard forms of losses in NLP [4]

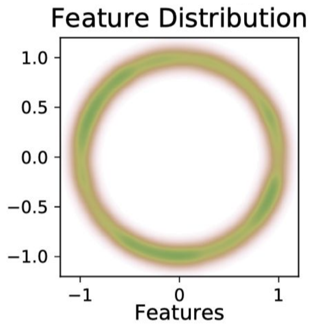
References:
[1] Boudiaf, Malik, Jérôme Rony, Imtiaz Masud Ziko, Éric Granger, Marco Pedersoli, Pablo Piantanida and Ismail Ben Ayed. "A Unifying Mutual Information View of Metric Learning: Cross-Entropy vs. Pairwise Losses." ECCV (2020).
[2] Schroff, Florian, Dmitry Kalenichenko and James Philbin. "FaceNet: A unified embedding for face recognition and clustering." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 815-823.
[3] Goldberger, Jacob, Sam T. Roweis, Geoffrey E. Hinton and Ruslan Salakhutdinov. "Neighbourhood Components Analysis." NIPS (2004).
[4] Zhang, Yue, and Zhiyang Teng. Natural language processing: a machine learning perspective. Cambridge University Press, 2021.

The Goal

- Obtain nice vector representations
  - Uniformity[1]
  - Multi-lingual[2]



Supervised cross entropy



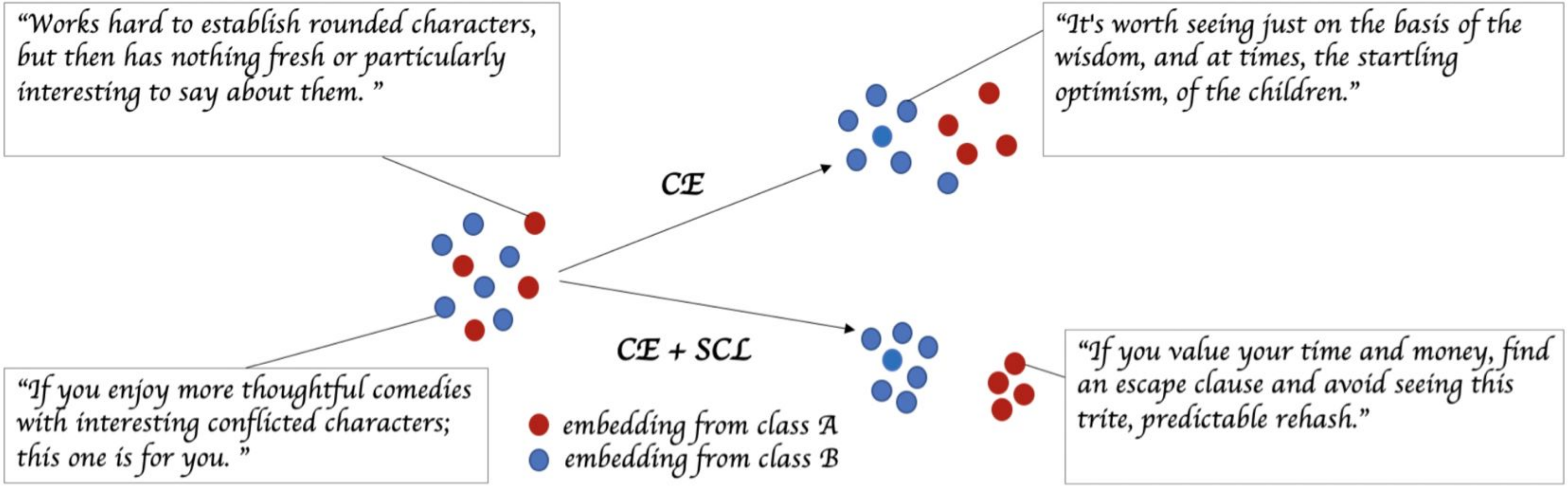
Unsupervised Contrastive Learning

Normalized feature distribution on a unit sphere of R^2 [1]

References:
[1] Wang, Tongzhou and Phillip Isola. "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere." ICML (2020).
[2] Li, Shicheng, Pengcheng Yang, Fuli Luo and Jun Xie. "Multi-Granularity Contrasting for Cross-Lingual Pre-Training." FINDINGS (2021).

The Goal

- Obtain nice vector representations
- Improve supervised learning



References:
[1] Gunel B, Du J, Conneau A, et al. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning[C]//International Conference on Learning Representations. 2020.
[2] P. Khosla, P. Teterwak, et. al. . Supervised contrastive learning. NeurIPS, 2020.
[3] Li L, Song D, Ma R, et al. KNN-BERT: Fine-Tuning Pre-Trained Models with KNN Classifier[J]. arXiv preprint arXiv:2110.02523, 2021.

The Goal

- Obtain nice vector representations
- Improve supervised learning
- Facilitate retrieval

References:
[1] Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih. "Dense Passage Retrieval for Open-Domain Question Answering." ArXiv abs/2004.04906 (2020): n. pag.

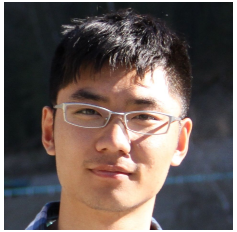
The Goal

- Obtain nice vector representations
- Improve supervised learning
- Facilitate retrieval
- Rank candidates

References:
[1] Liu, Yixin and Peng Liu. “SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization.” ACL (2021).

Thank You! Any Questions?

Slides and Video at https://contrastive-nlp-tutorial.github.io/



Rui Zhang
Penn State University

Yangfeng Ji
University of Virginia

Yue Zhang
Westlake University

Rebecca J. Passonneau
Penn State University