Theory of Contrastive Learning

UC San Diego

Weili Cao

Definition of Contrastive Learning

Pull together representations of similar data points and push apart representations of dissimilar data points in an embedding space

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

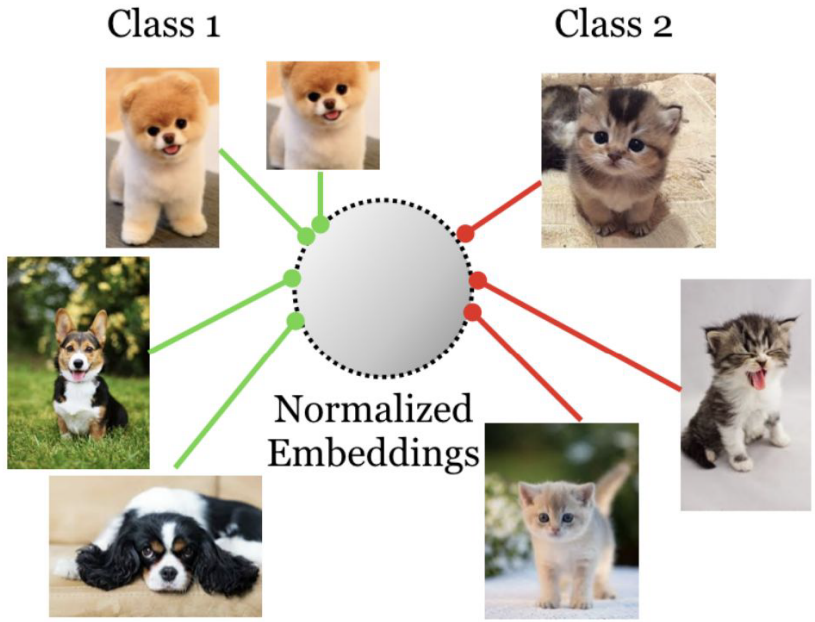
f : encoder, e.g., neural networks

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

$\boldsymbol{x}$: anchor

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

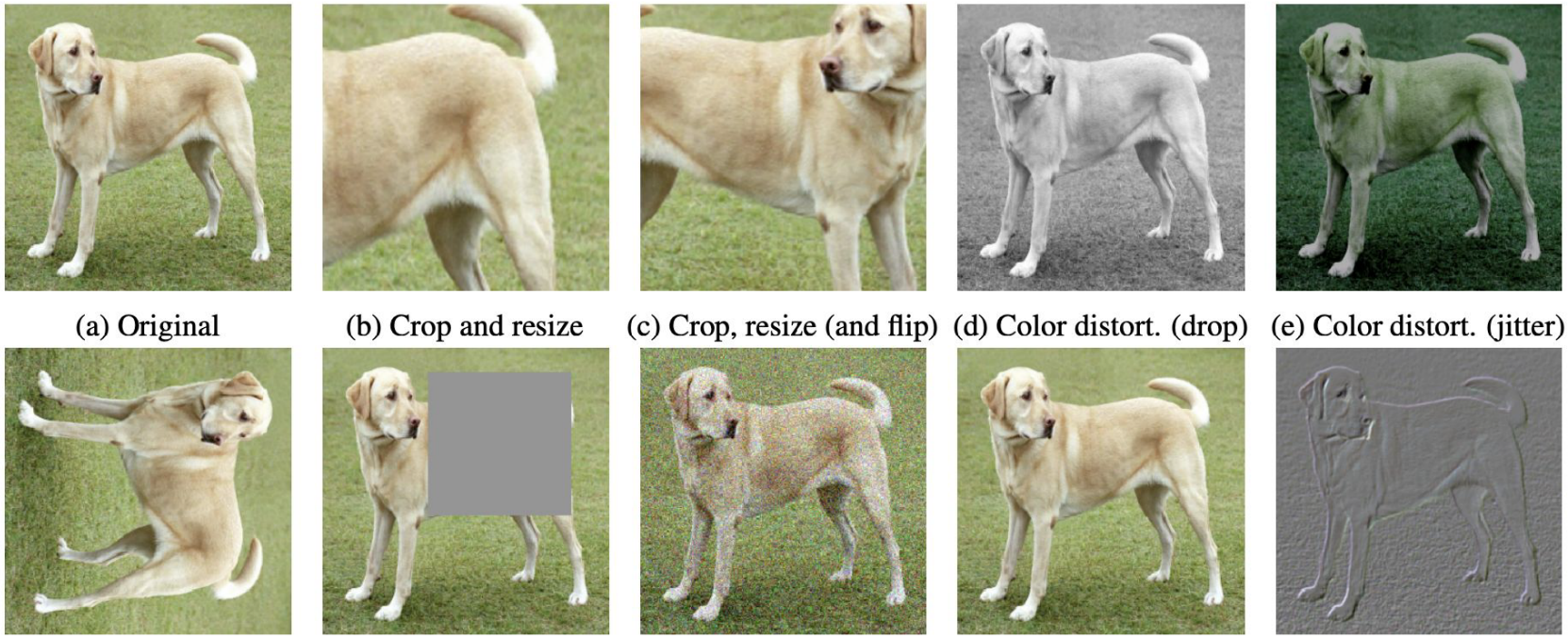
$\boldsymbol{x}^-$: negative example



Supervised Contrastive

Khosla et al. 2020

Contrastive Learning in Computer Vision

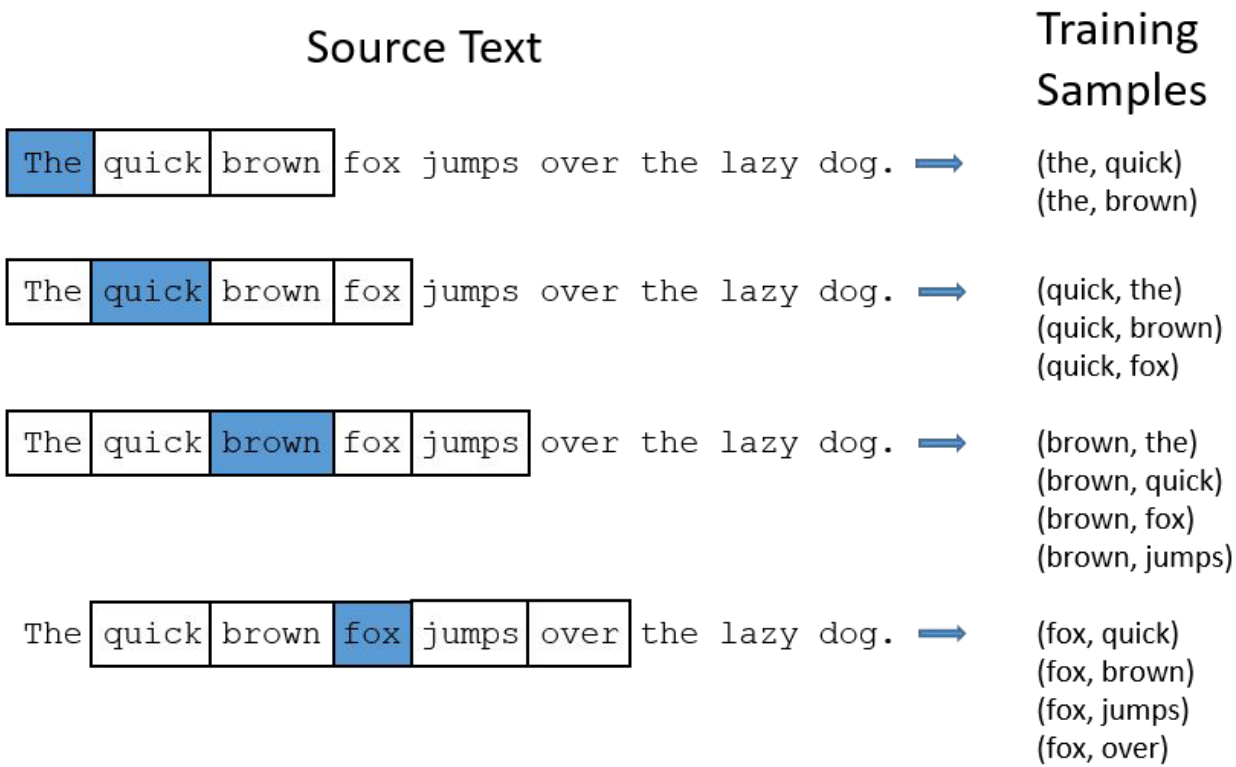


(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

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

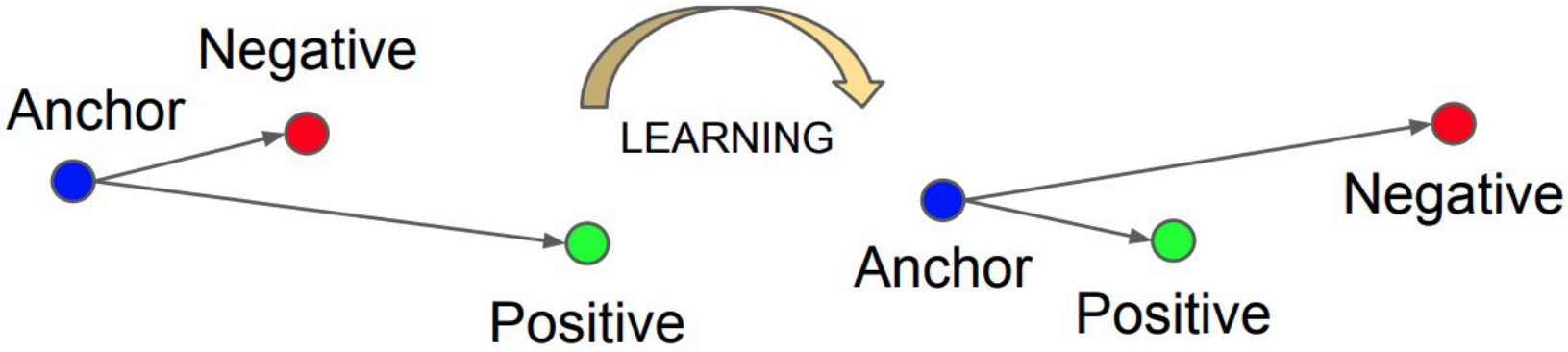
example of image augmentation Chen et al. 2020

Contrastive Learning in Natural Language Processing



this is an example of Word2Vec Mikolov et al. 2013, use words in a sentence as positive pairs and random word as negative sample

Triplet Loss



The idea is to push similar instances (anchor and positive) closer in the embedding space and pull dissimilar instances (anchor and negative) further apart.

$$\text{Loss}(\textit{Anchor}, \textit{Positive}, \textit{Negative}) =$$

$$\max(0, \text{sim}(\textit{Anchor}, \textit{Negative}) - \text{sim}(\textit{Anchor}, \textit{Positive}) + \text{margin})$$

NT-Xent Loss

- Normalized Temperature-scaled Cross Entropy loss
- This is the most popular loss function in contrastive learning
- We use cosine similarity
- τ is the temperature parameter that can scale the similarity, M is the number of negative samples

$$L(x,y) = \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}, x_i \overset{\text{iid}}{\sim} p_{\text{data}}} \left[-\log \frac{e^{\langle f(x), f(y)\rangle/\tau}}{e^{\langle f(x), f(y)\rangle/\tau} + \sum_{i=1}^{M} e^{\langle f(x), f(x_i^-)\rangle/\tau}} \right]$$

NT-Xent Loss

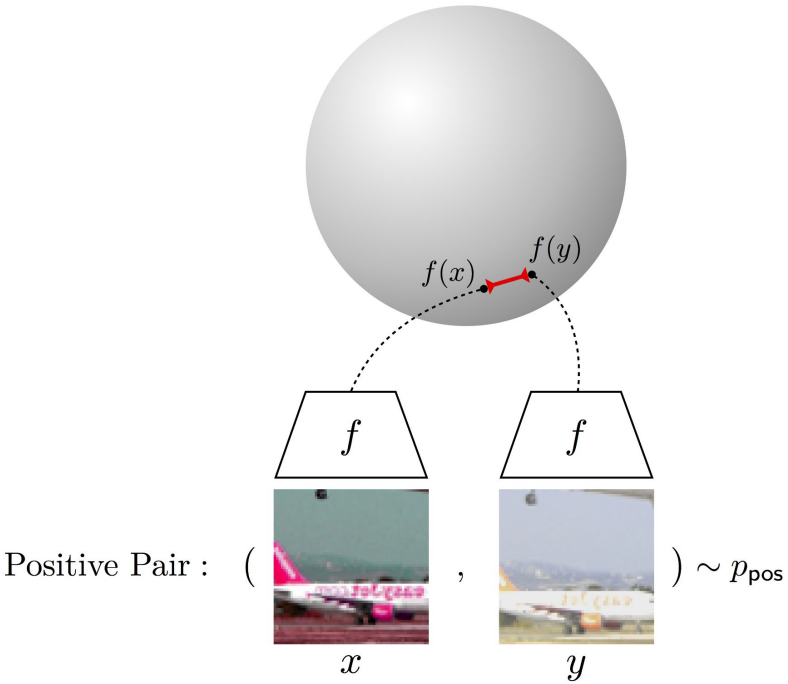
- Contrastive learning with NT-Xent Loss has gained empirical success in various downstream tasks
- But people do not have a good understand of why it works
- Next, I will introduce two papers giving insight on why contrastive learning with NT-Xent Loss works

Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere

Wang and Isola 2020

- Introduces two terms, Alignment and Uniformity
- Propose quantifiable metrics for Alignment and Uniformity as two measures of representation quality
- Shows that by minimizing NT-Xent Loss, we are actually optimizing Alignment and Uniformity
- Finds strong agreement between both metrics and downstream task performance.

Alignment and Uniformity



Alignment: Similar samples have similar features

Uniformity: Preserve maximal information

Defining Alignment and Uniformity

- Alignment:

$$L_{\text{align}}(f;\alpha) \triangleq \underset{(x,y)\sim p_{\text{pos}}}{\mathbb{E}} \left[||f(x)-f(y)||^{\alpha}\right]; \quad \alpha > 0$$

- We add Euclidean distances of all positive pairs.

- Uniformity:

$$L_{\text{uniform}}(f;t) \triangleq \log \underset{(x,y)\overset{\text{iid}}{\sim} p_{\text{data}}}{\mathbb{E}} \left[e^{-t||f(x)-f(y)||^2}\right]$$

- We chose Gaussian potential kernel to measure uniformity.

Perfect Alignment and Uniformity

Definition (Perfect Alignment) an encoder f is *perfectly aligned* if $f(x) = f(y)$ a.s. over $(x; y) \sim p_{\text{pos}}$.

Definition (Perfect Uniformity) an encoder f is *perfectly uniform* if the distribution of $f(x)$ for $x \sim p_{\text{data}}$ is the uniform distribution σ_{m-1} on S^{m-1}.

Asymptotic of the Loss

- Recall the NT-Xent Loss: $L_{\text{contrastive}}(f;\tau,M) =$

$$\mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}},x_i \overset{\text{iid}}{\sim} p_{\text{data}}} \left[\log \frac{e^{f(x)^T f(y)/\tau}}{e^{f(x)^T f(y)/\tau} + \sum_{i=1}^{M} e^{f(x_i)^T f(y)/\tau}}\right]$$

- By some math tricks: $\lim_{M\to\infty} L_{\text{contrastive}}(f;\tau,M) - \log M =$

$$-\frac{1}{\tau}\mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}}\left[f(x)^T f(y)\right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}}\left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}}\left[e^{f(x^-)^T f(x)/\tau}\right]\right]$$

- Note that for a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with $\mathbb{E}[X_i] = \mu$, we have $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$ as $n \to \infty$

Asymtotic of the Loss

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

- The first term is minimized iff f is perfectly aligned
- If perfectly uniform encoders exist, they form the exact minimizers of the second term.
- By minimizing the NT-Xent Loss, we are actually optimizing Uniformity and Alignment!

Experimental result

	Loss Formula	Validation Set Accuracy ↑	
		top1	top5
Best $L_{\text{contrastive}}$ only	$L_{\text{contrastive}}(\tau=0.07)$	72.80%	91.64%
Best L_{align} and L_{uniform} only	$3 \cdot L_{\text{align}}(a=2) + L_{\text{uniform}}(t=3)$	**74.60%**	**92.74%**
Best among all encoders	$3 \cdot L_{\text{align}}(a=2) + L_{\text{uniform}}(t=3)$	**74.60%**	**92.74%**

Figure: IMAGENET-100 encoder evaluations. Numbers show validation set accuracies of linear classifiers trained on encoder penultimate layer activations.

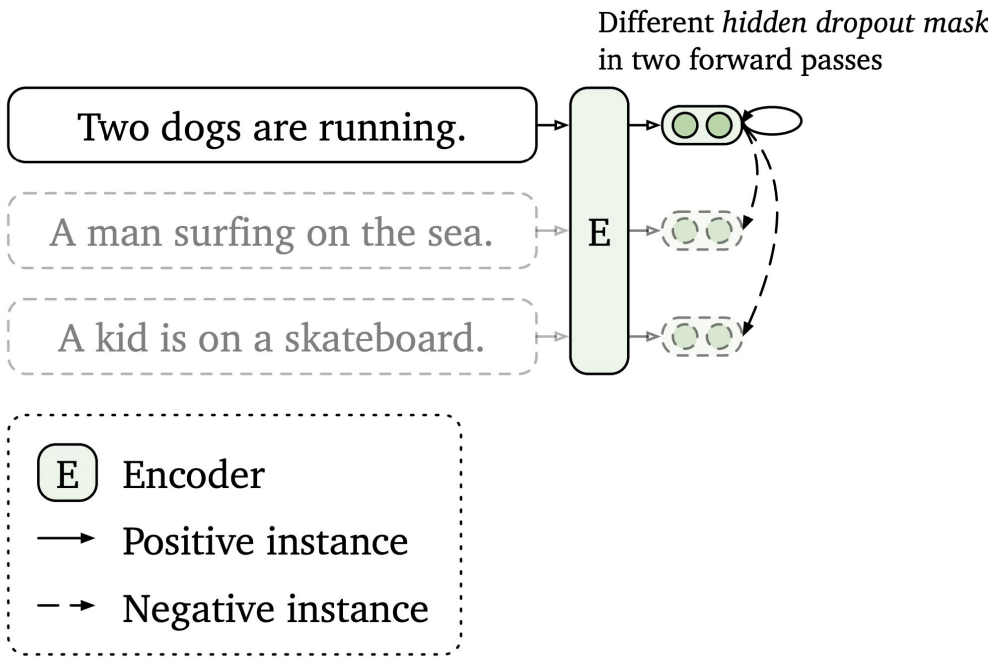
SimCSE: Simple Contrastive Learning of Sentence Embeddings Gao et al. 2021

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.

Figure: These are old data augmentation techniques: Synonym Replacement, Random Insertion, Random Swap, Random Deletion

SimCSE: Simple Contrastive Learning of Sentence Embeddings

In SimCSE, positive pair is an embeddings of the same sentence with different standard dropout masks, which is applied to the token-level embedding.



Analysis of SimCSE

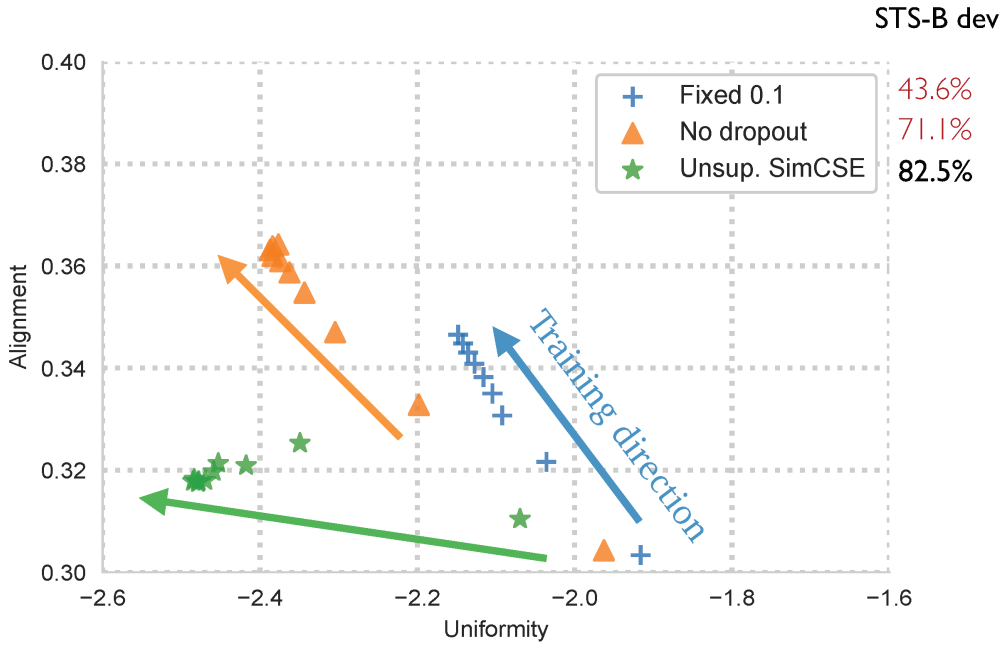


Figure: Fixed 0.1 means dropout rate is 0.1 but same dropout mask is applied

SimCSE improves uniformity while keeping good alignment!

A Theoretical Analysis of Contrastive Unsupervised Representation Learning Arora et al. 2019

- Introduces latent classes
- Assumes semantically similar points are sampled from the same latent class
- Shows provable guarantees on the performance of optimizing NT-Xent loss

Notion Setup

- χ: the set of all possible data points
- (x, x^+): positive pair that comes from a distribution $\mathcal{D}_{\text{sim}}$
- $x_1^-, x_2^-, x_3^-, \ldots$: i.i.d. negative samples from $\mathcal{D}_{neg}$
- Learning is done over $\mathcal{F}$, a class of representation functions $f : \chi \to \mathbb{R}^d$, such that $||f(\cdot)|| \leq R$ for some $R > 0$

Latent Class

Let C denote the set of all latent classes. Associated with each class $c \in C$ is a probability distribution $\mathcal{D}_c$ over χ.

- $\mathcal{D}_c(x)$ captures how relevant x is to class c
- For example, χ could represent all natural images and c could represent the class "dog". So $\mathcal{D}_c$, the distribution associated with class c, assigns high probabilities to images containing doges and low/zero probabilities to other images.
- Classes can overlap arbitrarily

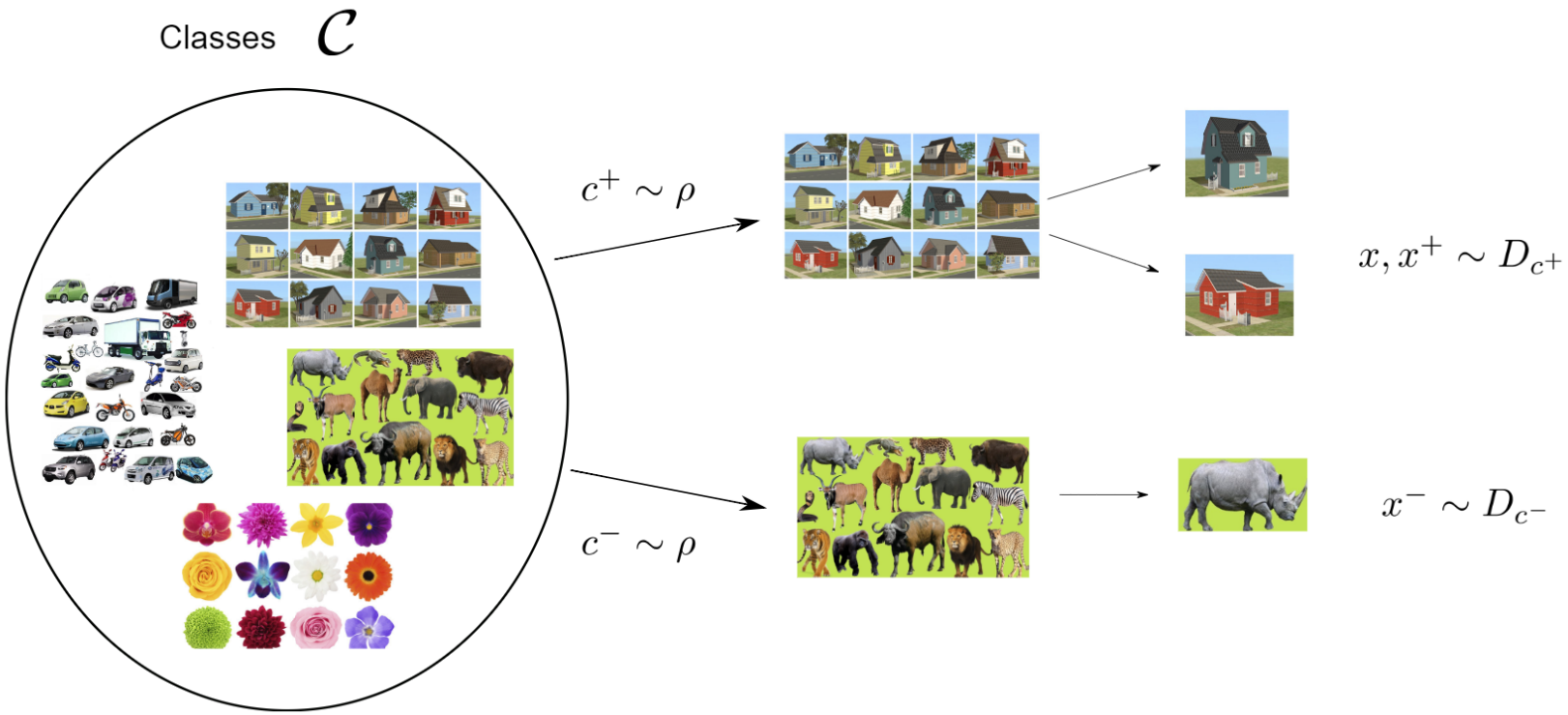
Distribution

- We assume a distribution ρ over the classes that characterizes how these classes naturally occurs in the unlabelled data.
- Distributions of training data is different from that of testing data.
- Distributions of points of unlabelled data includes positive sample as a result of data augmentation. (There are synthetic points in the distribution)
- Distributions of testing data is how the data points naturally distributed in real life.

Similarity

- Assume similar data points x, x^+ are i.i.d. draws from the same class distribution $\mathcal{D}_c$ for some class c picked randomly according to the measure ρ
- $\mathcal{D}_{\text{sim}}(x, x^+) = \mathop{\mathbb{E}}_{c\sim\rho} \mathcal{D}_c(x)\mathcal{D}_c(x^+)$
- Negative samples are drawn from margin of $\mathcal{D}_{\text{sim}}$
- $\mathcal{D}_{\text{neg}}(x^-) = \mathop{\mathbb{E}}_{c\sim\rho} \mathcal{D}_c(x^-)$

Example of similar and dissimilar pair



Learning a Representation Function

- $L_{un}(f) = \mathbb{E}\left[\ell(\{f(x)^T(f(x^+) - f(x^-))\})\right]$
- f is a representation function from $\mathcal{F}$, $f(x)$ maps the original data point x into a new representation which captures certain features or properties of the data. (often we call this "embeddings")
- We find $\hat{f}$ by minimizing the empirical loss function
- This $\hat{f}$ will be subsequently used for downstream tasks
- Empirically, $\hat{f}$ achieves good performance on downstream tasks, but we want to mathematically show this!

Supervised Task

- Task: subset of latent classes $\mathcal{T} = \{c1, c2, ..., ck\} \in C$
- A label $c \in \{c1, c2, ..., ck\}$ is picked according to a distribution $\mathcal{D}_{\mathcal{T}}$. Then, a sample x is drawn from $\mathcal{D}_c$. Together they form a labeled pair (x,c) with distribution

$$\mathcal{D}_{\mathcal{T}}(x, c) = \mathcal{D}_c(x)\mathcal{D}_{\mathcal{T}}(c)$$

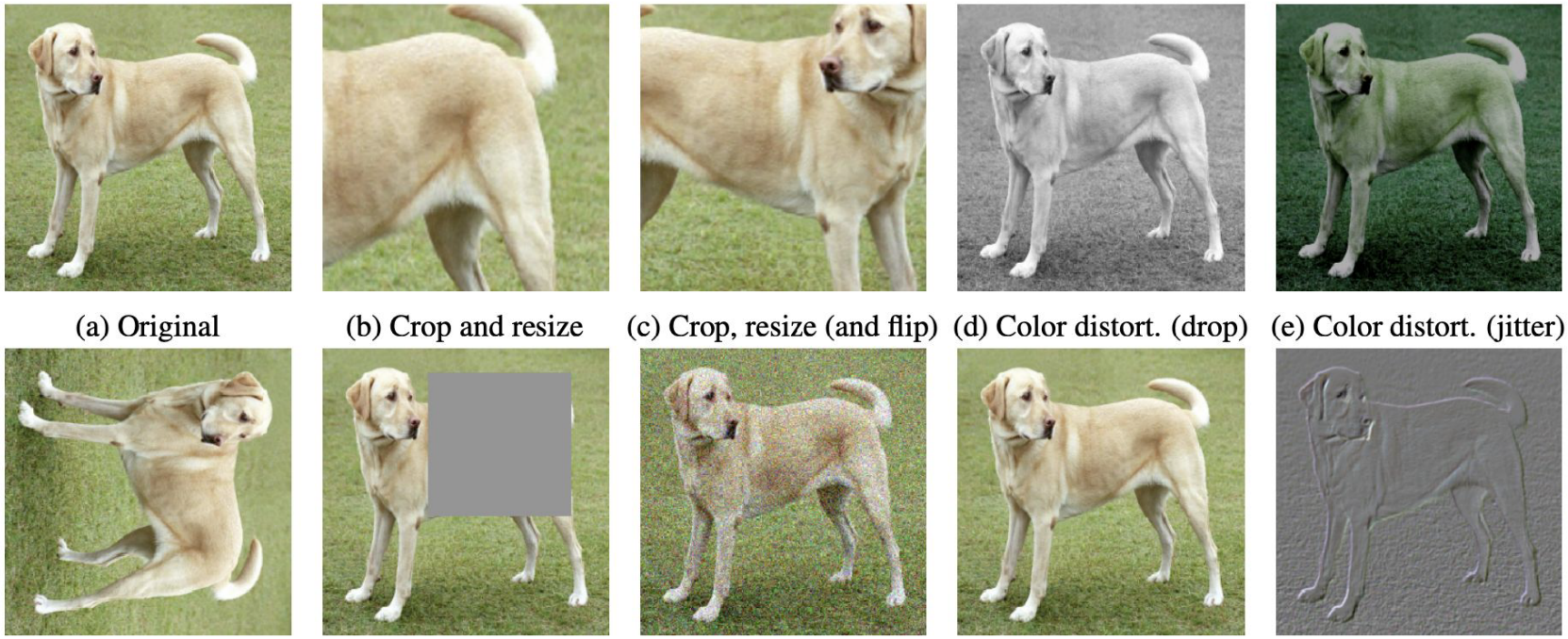
Provable Guarantee

- The paper is able to show that $L_{sup}(\hat{f}) \leq \alpha L_{un}(f) + \eta \text{Gen}_M + \beta s(f), \forall f \in \mathcal{F}$
- $L_{sup}(\hat{f})$ means the performance of $\hat{f}$ on downstream task
- It is bounded by $\alpha L_{un}(f)$, the loss function that we are minimizing
- This provably shows that by minimizing the loss functions of unsupervised learning, we can gain improvement on the performance of downstream tasks.

Key Take Away

- The paper gives an explanation of the success of contrastive learning: the classes in downstream tasks and their associated data distributions $\mathcal{D}_c$ are the same as in the unlabeled data.
- Data augmentation in unsupervised learning preserves semantic meaning. (See next slide for an example)
- This provides a path to formalizing how capturing similarity in unlabeled data can lead to quantitative guarantees on downstream tasks.

Data augmentation preserves semantic meaning



(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

Conclusion

- NT-Xent contrastive loss optimizes uniformity and alignment
- Under the concept of latent class and several assumptions, we can show provable guarantee of NT-Xent loss on downstream tasks.

Future Work of Alignment and Uniformity

- We lack a thorough understanding regarding the reason why we need to optimize uniformity and alignment.
- The paper only shows that, empirically, most recent successful techniques in contrastive learning achieves good results on uniformity and alignment.
- But are they in fact related to the representation learning methods? Do they actually agree with the representation quality (measured by downstream task performance)?

Future Work of second paper

- The concept of latent class seems to make sense, but there are a lot of assumptions in the setting.
- Similar pairs are assumed to be drawn from the same latent class.
- The authors assume a distribution ρ over unlabeled data and that the classes in downstream tasks and their associated data distributions $\mathcal{D}_c$ are the same as those in the unlabeled data.

S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

T. Gao, X. Yao, and D. Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Q&A

Thank you for listening!