# Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap

ICLR 2022

Yifei Wang*, Qi Zhang*, Yisen Wang, Jiansheng Yang, Zhouchen Lin

Peking University

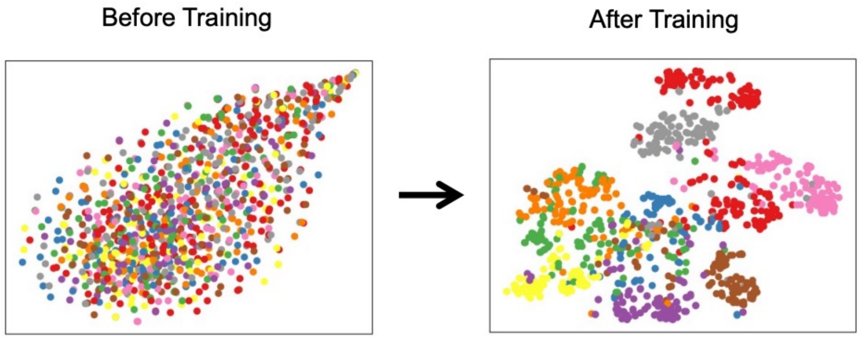
# Background: Contrastive Learning Learns Clustered Representations

PEKING UNIVERSITY

- Contrastive Learning (CL)
  - arguably the SOTA method for Self–Supervised Learning (SSL)
- Simple Learning Paradigm (e.g., InfoNCE)
  - Pull close positive samples $x^+$: random augmentations of the same samples
  - Push away negative samples $x^-$: augmented samples of independent samples

$$\mathcal{L}_{\mathrm{NCE}}(f) = \mathbb{E}_{p(x,x^+)}\mathbb{E}_{\{p(x_i^-)\}}\left[-\log \frac{\exp(f(x)^\top f(x^+))}{\sum_{i=1}^{M}\exp(f(x)^\top f(x_i^-))}\right]$$

- Empirically, CL can successfully cluster samples together



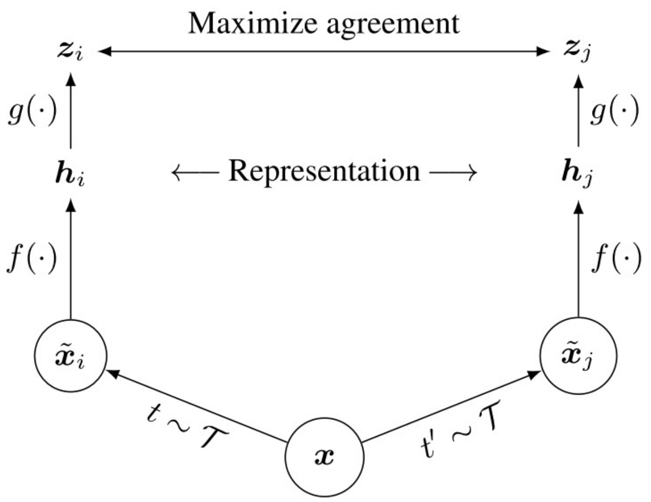




Figure from Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020.

**But Why?**

# Previous Theoretical Understandings and Their Limitations

北京大学 PEKING UNIVERSITY

- Arora et al. (2019)
  - Establish upper and lower bounds between contrastive loss and downstream loss
  - Assume that positive samples (x, x’) are conditionally independent given y
  - –> too strong assumption, which hardly holds in practice
- Wang & Isola (2020)
  - Propose the perspective of **alignment** and **uniformity**
  - However, we prove that these two properties alone are not enough!
  - Prop 3.1: there exists cases when a random encoder also minimizes the InfoNCE loss

learn class inseparable features even with perfect aligned positive samples and uniform negative samples. Colors denote classes.

**Proposition 3.1** (Class-uniform Features Also Minimize the InfoNCE Loss). *For $N$ training examples of $K$ classes, consider the case when features $\{f(x_i)\}_{i=1}^N$ are randomly distributed in $\mathbb{S}^{m-1}$ with maximal uniformity (i.e., , minimizing the 2nd term of Eq. 1) while also satisfying $\forall x_i, x_i^+ \sim p(x, x^+), f(x_i) = f(x_i^+)$. Because we have these two properties, the InfoNCE loss achieves its minimum. However, the downstream classification accuracy is at most $1/K + \varepsilon$ and $\varepsilon$ is nearly zero when $N$ is large enough.*

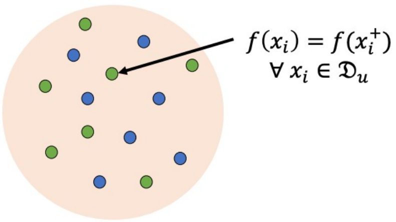



Figure 2: An illustration of the case when contrastive learning fails to learn class-separated features even if the features are uniformly distributed and positive samples are perfect aligned. Each color denotes a class.

# Previous Theoretical Understandings and Limitations

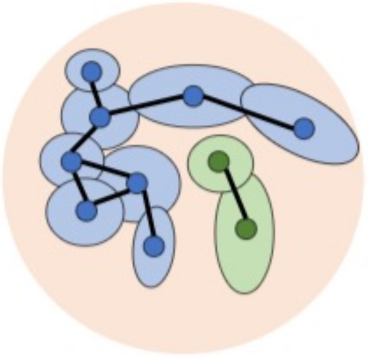
北京大學 PEKING UNIVERSITY

- Arora et al. (2019)
  - Establish upper and lower bounds on downstream loss using contrastive loss
  - Assume that positive samples (x, x') are conditionally independent given y
  - –> too strong assumption, which hardly holds in practice ☹
- Wang & Isola (2020)
  - Propose the perspective of alignment and uniformity
  - However, we prove that these two properties alone are not enough!
  - Prop 3.1: there exists cases when a random encoder also minimizes the InfoNCE loss ☹

How to establish **guarantees** on downstream performance **with minimal and practical assumptions**?

# A New *Augmentation Overlap* Theory for Contrastive Learning

- Augmentation Graph $\mathcal{G}$
  - Nodes: natural samples $x_i$
  - Edge: $e_{ij}$ exists if the two have augmentation overlap
- Three Practical Assumptions (informal)
  - Augmentations do not change the labels
  - The intra-class augmentation subgraph is connected
  - Good alignment of positive samples (encoder capacity)



Augmentation Graph (input space)

(b) Intra-class samples are more alike via augmented views.

- Obtained Guarantees on Downstream Performance (measured by CE loss)

$$\mathcal{L}_{\mathrm{NCE}}(f^\star) - \mathcal{O}\left(M^{-1/2}\right) \leq \mathcal{L}^{\mu}_{\mathrm{CE}}(f^\star) + \log(M/K) \leq \mathcal{L}_{\mathrm{NCE}}(f^\star) + \mathcal{O}\left(M^{-1/2}\right).$$

  - For the optimal encoder f*, contrastive learning is almost as good as supervised learning (with asymptotically closed upper and lower bounds)

# Measuring Augmentation Overlap

- An unsupervised evaluation metric ARC (Average Relative Confusion)
  - Designed based on our augmentation overlap theory
  - Aligns well with downstream accuracy across different augmentation strength
  - Can be used for unsupervised model selection!!

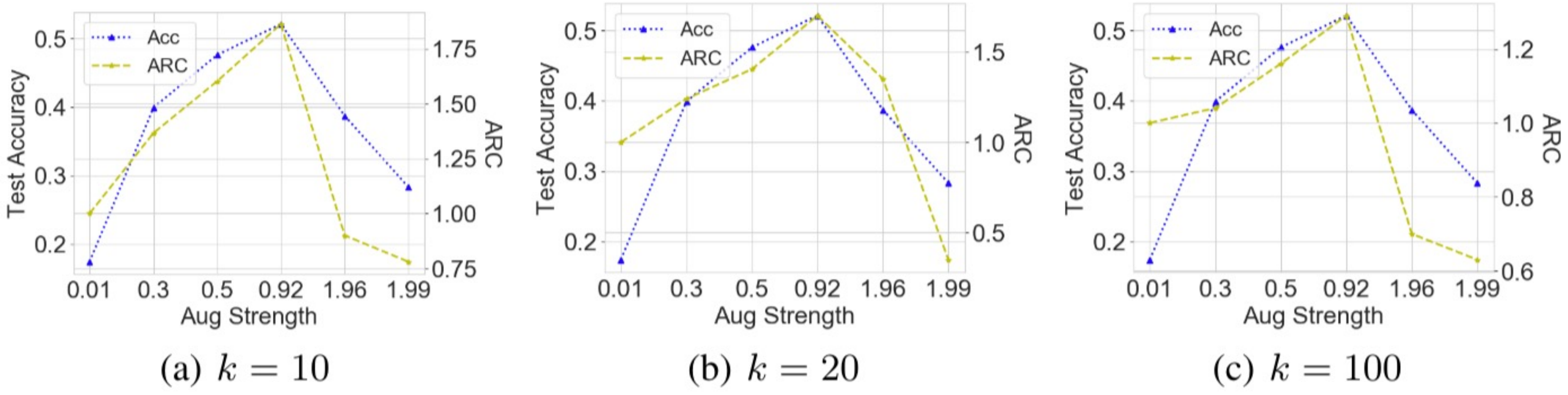



Figure 7: Average Relative Confusion (ARC) and downstream accuracy *v.s.* different augmentation strength on CIFAR-10 (SimCLR) with different number of nearest neighbors $k$.

The code for computing ARC is available at [https://github.com/zhangq327/ARC](https://github.com/zhangq327/ARC)

# Takeaways

- Contributions
  - We characterize the failure of previous analysis of contrastive learning.
  - Theoretically, we establish a new augmentation overlap theory with guarantees on downstream performance using more practical assumptions.
  - Empirically, we show the theoretically inspired ARC metric is a good indicator for unsupervised evaluation of contrastive learning.
- **Key insight:** an alternative understanding of contrastive learning
  - the role of aligning positive samples is more like a surrogate task than an ultimate goal
  - the overlapped augmented views (i.e., the chaos) create a ladder for contrastive learning to gradually learn class-separated representations.

*Chaos isn't a pit. Chaos is a ladder.*

-- "Littlefinger" Petyr Baelish
*Game of Thrones*

北京大學
PEKING UNIVERSITY

# Thanks for Listening