ICLR

北京大学
PEKING UNIVERSITY

A Message Passing Perspective on Learning Dynamics of Contrastive Learning

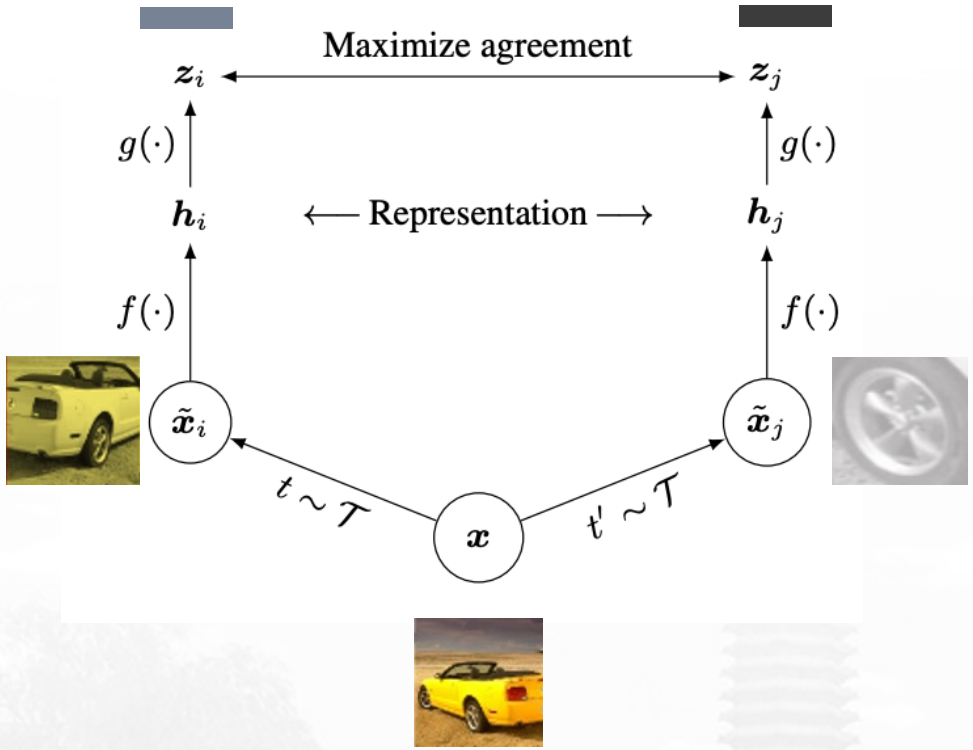
Presenter: Yifei Wang (yifei_wang@pku.edu.cn)

Joint work with Qi Zhang, Tianqi Du, Jiansheng Yang, Zhouchen Lin, Yisen Wang

Peking University

Background: contrastive learning and theoretical understandings

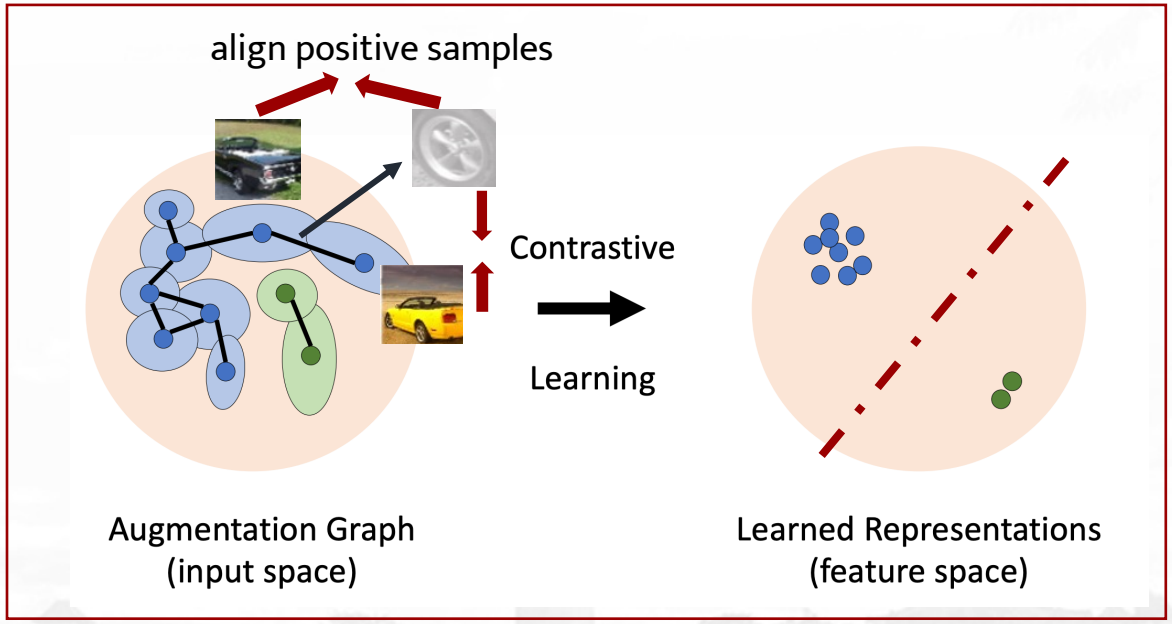
Example: SimCLR (Chen et al., 2020)



InfoNCE loss used in SimCLR

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

Generalization analysis based on augmentation graph

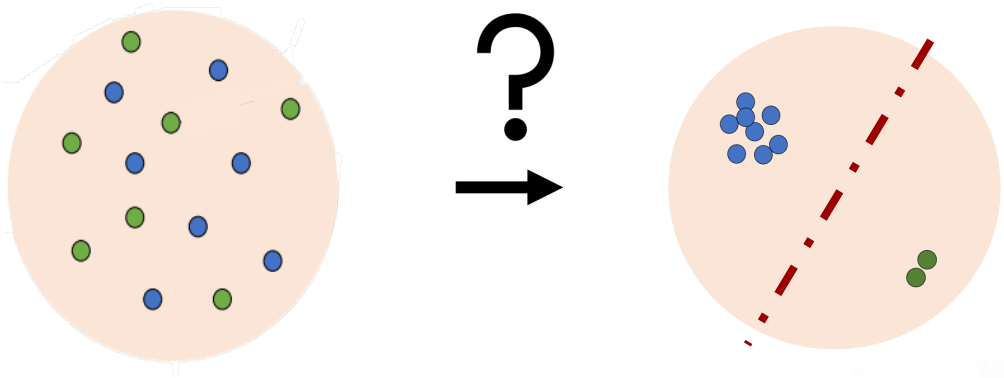


A diagram of the theory of Haochen et al. (2021); Wang et al. (2022)
Taken from Wang et al. (2022)

Motivation: how contrastive learning reaches the optimal solution?

Known: downstream generalization of optimal classifier

Unknown: how CL learns features along training?

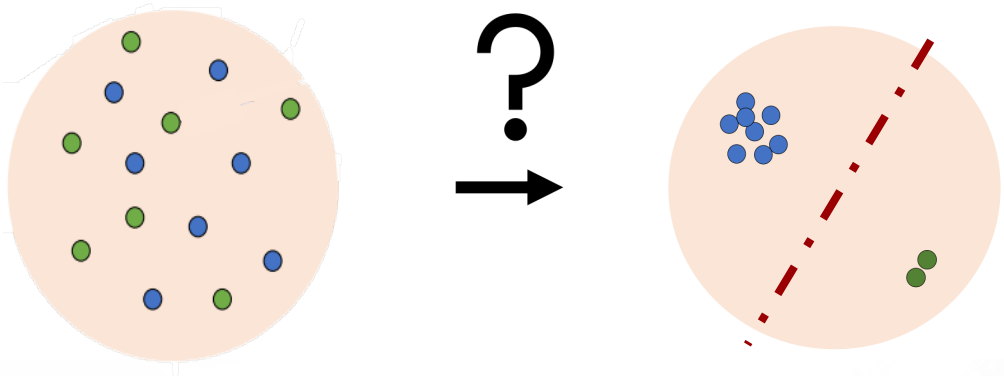


obstacle: NN training dynamics is hard to analyze

Motivation: how contrastive learning reaches the optimal solution?

Known: downstream generalization of optimal classifier

Unknown: how CL learns features along training?



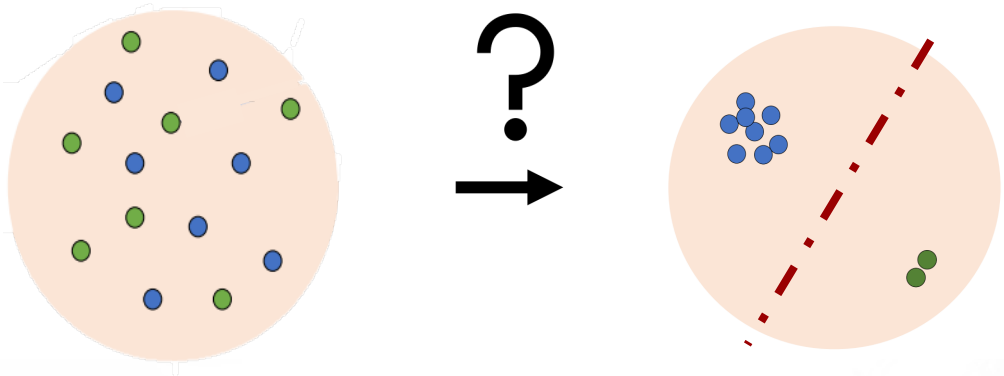
obstacle: NN training dynamics is hard to analyze

our choice: focus on the update of output features F (the unconstrained feature setting)

Summary of main results

Known: downstream generalization of optimal classifier

Unknown: how CL learns features along training?



obstacle: NN training dynamics is hard to analyze

our choice: focus on the update of output features F (i.e., the unconstrained feature setting)

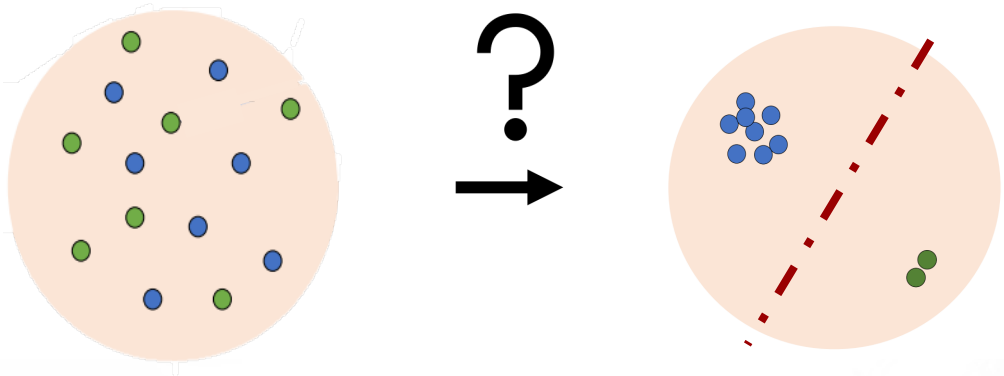
Our theoretical results: contrastive learning amounts to message passing on graphs

- Alignment of positive samples = graph convolution on the augmentation graph A
- Uniformity of negative samples = reversed graph convolution on the estimated graph A'

Summary of main results

Known: downstream generalization of optimal classifier

Unknown: how CL learns features along training?



obstacle: NN training dynamics is hard to analyze

our choice: focus on the update of output features F (i.e., the unconstrained feature setting)

Our theoretical results: contrastive learning amounts to message passing on graphs

- Alignment of positive samples = graph convolution on the augmentation graph A
- Uniformity of negative samples = reversed graph convolution on the estimated graph A'

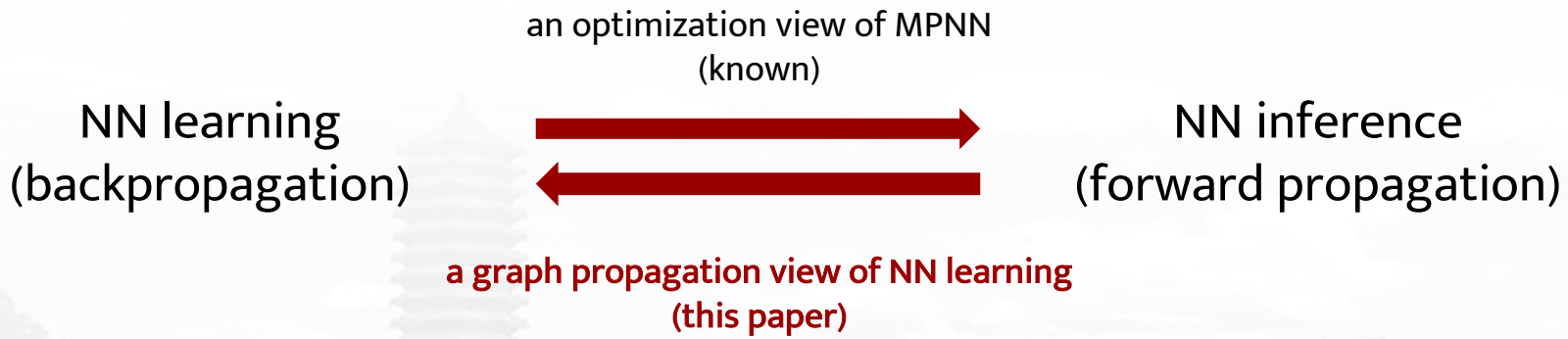
➡ ***Equilibrium is attained when the estimated graph = the augmentation graph***

A unified view of learning and inference via graphs

北京大学 PEKING UNIVERSITY

A formal correspondence between two domains in two scenarios

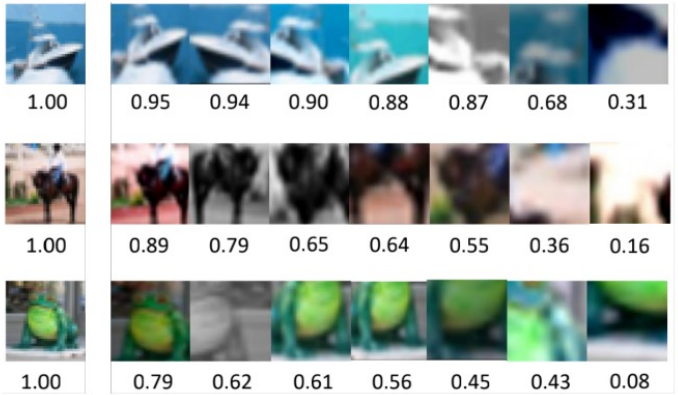
Key components	Contrastive learning	MPNN inference
Adjacency matrix	Between training samples	Between nodes in an input graph
Initial features	Random	Inputs (given, sometimes also random)
MP updates	During training steps	During layerwise propagation
Equilibrium	Of training	Of forward propagation



Complete the missing piece in the unified view!

Connections, Analogies, and New Designs

- This unified perspective allows us to view techniques in the two domains interchangeably
  - Connections
    - Graph Convolution ⟺ Alignment update
    - Oversmoothing ⟺ Feature collapse
  - Existing techniques
    - NodeNorm / LayerNorm ⟺ ℓ_2 normalization of features (SimCLR)
    - PairNorm ⟺ centering and l2 normalization of features (DINO)



- New Designs (transferring advanced GNN techniques for SSL)
  - GraphAttention (GAT, Transformer) → Attentive alignment loss (adaptively aggregate positive samples)

$$\mathcal{L}_{\text{attn-align}}(\theta) = \frac{1}{2}\mathbb{E}_{x,x^+}\alpha(x,x^+)\|f_\theta(x) - f_\theta(x^+)\|^2.$$

  - Multi-stage Aggregation (JKNet, APPNP, SIGN) → Multi-stage alignment loss helps prevent collapse

$$\mathcal{L}_{\text{multi-align}}(\theta) = -\mathbb{E}_{\bar{x}}\mathbb{E}_{x|\bar{x}}(f_\theta(x)^\top z_{\bar{x}}).$$

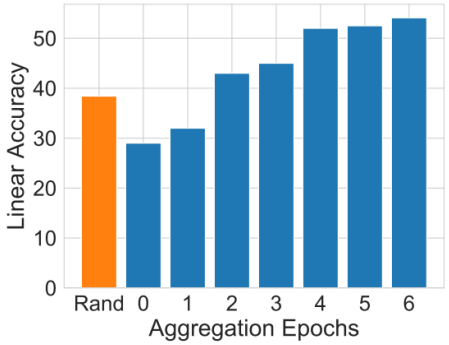
$$z_{\bar{x}} = \frac{1}{r}\sum_{i=1}^{r} z_{\bar{x}}^{(t-i)}$$

Results

- SimCLR with attentive alignment loss

Backbone	Method	CIFAR-10	CIFAR-100	ImageNet-100
ResNet-18	SimCLR	84.5	56.1	62.3
	SimCLR-Attn	**85.4**	**56.9**	**63.1**
ResNet-50	SimCLR	88.2	59.8	66.0
	SimCLR-Attn	**89.4**	**60.7**	**66.7**

- SimSiam with multi-stage aggregation



(a) multi-stage aggregation

Dataset	Method	Top-1 Acc (%)
CIFAR-10	SimSiam	83.82
	SimSiam-MultiStage	**84.75**
CIFAR-100	SimSiam	56.34
	SimSiam-MultiStage	**58.87**
ImageNet-100	SimSiam	68.76
	SimSiam-MultiStage	**70.52**

Takeaways

北京大学 PEKING UNIVERSITY

- **Contrastive learning (implicitly) performs message passing on graphs during training**
  - alignment = graph convolution on the augmentation graph
  - uniformity = reversed graph convolution on the estimated graph

- **A unified view of contrastive learning and message passing neural networks**
  - every learning problem defines a message passing scheme on the graph
  - an optimization step implicitly performs a feature propagation step

- **Inherent connections between existing techniques in two domains, and inspired new ones**
  - graph attention
  - multi-stage aggregation

Yifei Wang*, Qi Zhang*, Tianqi Du, Jiansheng Yang, Zhouchen Lin, Yisen Wang. A Message Passing Perspective on Learning Dynamics of Contrastive Learning. In ICLR 2023.

Thanks for Listening!

Yifei Wang (Peking University)