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)

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

InfoNCE loss used in SimCLR

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

\[
\begin{align*}
\text{Alignment of positive samples} &= \text{graph convolution on the augmentation graph } A \\
\text{Uniformity of negative samples} &= \text{reversed graph convolution on the estimated graph } A'
\end{align*}
\]
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

\[ \begin{align*}
\text{Alignment of positive samples} &= \text{graph convolution on the augmentation graph } A \\
\text{Uniformity of negative samples} &= \text{reversed graph convolution on the estimated graph } A'
\end{align*} \]

Equilibrium is attained when the estimated graph = the augmentation graph
A unified view of learning and inference via graphs

A formal correspondence between two domains in two scenarios

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

an optimization view of MPNN (known)

NN learning (backpropagation) ➔
NN inference (forward propagation)

a graph propagation view of NN learning (this paper)

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 ↔ $\ell_2$ 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

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Method</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>ImageNet-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18</td>
<td>SimCLR</td>
<td>84.5</td>
<td>56.1</td>
<td>62.3</td>
</tr>
<tr>
<td></td>
<td>SimCLR-Attn</td>
<td><strong>85.4</strong></td>
<td><strong>56.9</strong></td>
<td><strong>63.1</strong></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>SimCLR</td>
<td>88.2</td>
<td>59.8</td>
<td>66.0</td>
</tr>
<tr>
<td></td>
<td>SimCLR-Attn</td>
<td><strong>89.4</strong></td>
<td><strong>60.7</strong></td>
<td><strong>66.7</strong></td>
</tr>
</tbody>
</table>

• SimSiam with multi-stage aggregation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Top-1 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td>SimSiam</td>
<td>83.82</td>
</tr>
<tr>
<td></td>
<td>SimSiam-MultiStage</td>
<td><strong>84.75</strong></td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>SimSiam</td>
<td>56.34</td>
</tr>
<tr>
<td></td>
<td>SimSiam-MultiStage</td>
<td><strong>58.87</strong></td>
</tr>
<tr>
<td>ImageNet-100</td>
<td>SimSiam</td>
<td>68.76</td>
</tr>
<tr>
<td></td>
<td>SimSiam-MultiStage</td>
<td><strong>70.52</strong></td>
</tr>
</tbody>
</table>

(a) multi-stage aggregation
Takeaways

• 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

Thanks for Listening!

Yifei Wang (Peking University)