Identical and Fraternal Twins:
Fine-Grained Semantic
Contrastive Learning of Sentence Representations

Qingfa Xiao\textsuperscript{1,3}, Shuangyin Li\textsuperscript{1,\textsuperscript{48}} and Lei Chen\textsuperscript{2,3}

\textsuperscript{1}South China Normal University
\textsuperscript{2}The Hong Kong University of Science and Technology
\textsuperscript{3}The Hong Kong University of Science and Technology (Guangzhou)
qingfaxiao@m.scnu.edu.cn, shuangyinli@scnu.edu.cn, leichen@cse.ust.hk
Learning universal representations of sentences has wide applications in NLP

- Semantic matching
- Sentence clustering
- Information retrieval
- ...
Background: Contrastive Learning

Contrastive learning is an intuitive and effective training objective that aims to create desired semantic representations by

- bringing semantically positive instances closer together
- pushing away those that are not semantically negative.
Motivation:
The lack of fine-grained semantic discrimination ability via contrastive learning.
Motivation:
The lack of fine-grained semantic discrimination ability via contrastive learning.

When the different types of positive pairs come to contrastive learning, they should be treated under the different standards.
Motivation:
Which twins are more similar?

Identical Twins

Fraternal Twins

Everyone can easily distinguish between these two pairs of twins.
But, can a language model do the same?
Objective:
Keep the margins between the two pairs of twins to help model distinguish the subtle differences.
Challenges:
Contrastive learning with data augmentation

- Data augmentation for gaining positive pairs with less **semantic distortions**

- Adaptive contrastive learning for different types of positive samples to **address sub-optimal issues**
Three main components in this IFCL framework:

- A fusion data augmentation technique
- A training loss function named Twin Loss
- A hippocampus queue mechanism
IFCL Framework:

A fusion data augmentation technique minimizing semantic distortions and increasing diversity of expressions.
IFCL Framework:

- A training loss function named Twin Loss capturing fine-grained semantics and alleviating the sub-optimal issues according to their margins.
IFCL Framework:

- A hippocampus queue mechanism
  - storing the previous mini-batches into a short-term memory and reusing the negative effectively
Method:
How to generate the Identical and Fraternal Twins?

- **Identical twins**: the most similar positive pair
- **Fraternal twins**: the diverse pair with less semantic distortions
Method:

InfoNCE Loss with positive and negative instances

For the set of identical twins \( \{h_i, h_i^+\}_{i=1}^N \) or fraternal twins \( \{h_i, h_i^-\}_{i=1}^N \), we define the function by using negative instances \( \{H_m\}_{m=1}^{k*N} \) stored in the hippocampus queue.

\[
\ell^I_i = -\log \frac{e^{\text{sim}(h_i, h_i^+)/\tau}}{\sum_{j=1}^N e^{\text{sim}(h_i, h_j^+)/\tau} + \varphi},
\]

\[
\varphi = \sum_{m=1}^{k*N} p_m e^{\text{sim}(h_i, H_m)/\tau},
\]
Method:
Twins Loss for fine-grained semantic understanding

- This loss function aims to keep the margins between two types of positive pairs.
- M represents the innate margins between identical and fraternal twins.
- Each M depends on the previous step to prevent sub-optimal optimization problems.

\[
\ell_i^T = \left| e^{\text{sim}(h_i, h_i^+)} - e^{\text{sim}(h_i, h_i^-)} - M_i \right|
\]

\[
M_i = e^{\text{sim}(\text{emb}_i, \text{emb}_i^+)} - e^{\text{sim}(\text{emb}_i, \text{emb}_i^-)}
\]
Method:
Hippocampus Queue Mechanism for reusing instances

- The queue storing the negative is **continuously updated**
- The sample is gradient-free to **save GPU memory**
- The forgetting coefficient focus more on the **latest instance**
Results:

Experiments on semantic textual similarity tasks

- Evaluate using Spearman's correlation metric
- Perform well in both Chinese and English tasks.

### Results of Chinese tasks

<table>
<thead>
<tr>
<th>Method</th>
<th>Chinese STS-B</th>
<th>SimCLUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>55.52</td>
<td>29.89</td>
</tr>
<tr>
<td>BERT-whitening</td>
<td>68.27</td>
<td>-</td>
</tr>
<tr>
<td>SimCSE-BERT</td>
<td>68.91</td>
<td>40.74</td>
</tr>
<tr>
<td>SimCSE-BERT</td>
<td>60.41</td>
<td>40.54</td>
</tr>
<tr>
<td>IFCL-BERT</td>
<td><strong>71.41</strong></td>
<td><strong>44.42</strong></td>
</tr>
</tbody>
</table>

### Results of English tasks

<table>
<thead>
<tr>
<th>Method</th>
<th>STS12</th>
<th>STS13</th>
<th>STS14</th>
<th>STS15</th>
<th>STS16</th>
<th>STS-B</th>
<th>SICK-R</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT_base</td>
<td>39.70</td>
<td>59.38</td>
<td>49.67</td>
<td>66.03</td>
<td>66.19</td>
<td>53.87</td>
<td>62.06</td>
<td>56.70</td>
</tr>
<tr>
<td>BERT-flow_base</td>
<td>58.40</td>
<td>67.10</td>
<td>60.85</td>
<td>75.16</td>
<td>71.22</td>
<td>68.66</td>
<td>64.47</td>
<td>66.55</td>
</tr>
<tr>
<td>BERT-whitening_base</td>
<td>57.83</td>
<td>66.90</td>
<td>60.90</td>
<td>75.08</td>
<td>71.31</td>
<td>68.24</td>
<td>63.73</td>
<td>66.28</td>
</tr>
<tr>
<td>ConSERef_base</td>
<td>64.64</td>
<td>78.49</td>
<td>69.07</td>
<td>79.72</td>
<td>75.95</td>
<td>73.97</td>
<td>67.31</td>
<td>72.74</td>
</tr>
<tr>
<td>SimCSE-BERT_base</td>
<td>68.40</td>
<td>82.41</td>
<td>74.38</td>
<td>80.91</td>
<td>78.56</td>
<td>76.85</td>
<td>72.23</td>
<td>76.25</td>
</tr>
<tr>
<td>VsACL-BERT_base</td>
<td>69.08</td>
<td>81.95</td>
<td>74.64</td>
<td>82.64</td>
<td>80.57</td>
<td>80.23</td>
<td>71.23</td>
<td>77.19</td>
</tr>
<tr>
<td>DCLR-BERT_base</td>
<td>70.81</td>
<td>83.73</td>
<td>75.11</td>
<td>82.56</td>
<td>78.44</td>
<td>78.31</td>
<td>71.99</td>
<td>77.22</td>
</tr>
<tr>
<td>MoCoSE-BERT_base</td>
<td>71.48</td>
<td>81.40</td>
<td>74.47</td>
<td><strong>83.45</strong></td>
<td>78.99</td>
<td>78.68</td>
<td><strong>72.44</strong></td>
<td>77.27</td>
</tr>
<tr>
<td>PT-BERT_base</td>
<td>71.20</td>
<td><strong>83.76</strong></td>
<td><strong>76.34</strong></td>
<td>82.63</td>
<td>78.90</td>
<td>79.42</td>
<td>71.94</td>
<td>77.74</td>
</tr>
<tr>
<td>IFCL-BERT_base</td>
<td><strong>71.57</strong></td>
<td>82.35</td>
<td>75.08</td>
<td>83.03</td>
<td><strong>80.17</strong></td>
<td><strong>80.27</strong></td>
<td>72.16</td>
<td><strong>77.80</strong></td>
</tr>
<tr>
<td>BERT_large</td>
<td>57.73</td>
<td>61.17</td>
<td>61.18</td>
<td>68.07</td>
<td>70.25</td>
<td>59.59</td>
<td>60.34</td>
<td>62.62</td>
</tr>
<tr>
<td>ConSERef_large</td>
<td>70.69</td>
<td>82.96</td>
<td>74.13</td>
<td>82.78</td>
<td>76.66</td>
<td>77.53</td>
<td>70.37</td>
<td>76.45</td>
</tr>
<tr>
<td>SimCSE_large</td>
<td>70.88</td>
<td>84.16</td>
<td>76.43</td>
<td>84.50</td>
<td>79.76</td>
<td>79.26</td>
<td>73.88</td>
<td>78.41</td>
</tr>
<tr>
<td>DCLR-BERT_large</td>
<td>71.87</td>
<td><strong>84.83</strong></td>
<td><strong>77.37</strong></td>
<td><strong>84.70</strong></td>
<td><strong>79.81</strong></td>
<td>79.55</td>
<td>74.19</td>
<td>78.90</td>
</tr>
<tr>
<td>MoCoSE-BERT_large</td>
<td><strong>74.50</strong></td>
<td>84.54</td>
<td>77.32</td>
<td>84.11</td>
<td>79.67</td>
<td>80.53</td>
<td>73.26</td>
<td>79.13</td>
</tr>
<tr>
<td>IFCL-BERT_large</td>
<td>73.88</td>
<td>84.31</td>
<td>76.64</td>
<td>84.01</td>
<td>79.56</td>
<td><strong>81.37</strong></td>
<td><strong>76.30</strong></td>
<td><strong>79.44</strong></td>
</tr>
</tbody>
</table>
Analyse:
What makes the Twins Loss effective?

Reducing the mutual information between positive pairs while preserving task-relevant information is optimal for the task.

- More diverse semantics are preserved
  \[ M_{\text{II}}(h, h^{-}) \text{ is higher than } M_{\text{II}}(h, h^{+}) \]

- Mutual information of positive pairs contains more task-relevant information.
  \[ M_{\text{II}}(h, h^{+}) \approx M_{\text{II}}(h, h^{-}) \approx M_{\text{II,task}} \]

Table 4. Mutual information and task-relevant information. The IFCL-BERT w/o TL means training IFCL-BERT without Twins Loss. The experiments are conducted with EnData and STS-B datasets on Bert-base.

<table>
<thead>
<tr>
<th>Method</th>
<th>( M_{\text{II}}(h, h^{+}) )</th>
<th>( M_{\text{II}}(h, h^{-}) )</th>
<th>( M_{\text{II,task}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFCL-BERT</td>
<td>4.15</td>
<td>4.17</td>
<td>4.31</td>
</tr>
<tr>
<td>IFCL-BERT w/o TL</td>
<td>4.23</td>
<td>4.20</td>
<td>4.58</td>
</tr>
<tr>
<td>SimCSE</td>
<td>4.24</td>
<td>-</td>
<td>4.52</td>
</tr>
</tbody>
</table>
Thank YOU

Q & A