Few-shot Text Classification with Distributional Signatures

Yujia Bao, Menghua Wu, Shiyu Chang, Regina Barzilay

ICLR 2020
Motivation

1. In meta-learning, learning solely from words is not enough
   a. Matching information
   b. Interaction
   c. Underlying distribution

2. Model word’s distributional signatures across classes

3. Using this distributional signature as attention weight
Settings

support set
(K × N examples)

query set
(L × N examples)

source pool

a) traditional episode

b) our extension
Framework

b) **Ridge regressor**: Constructing representations

\[ f_{ebd}(\cdot) \rightarrow \phi(x) \]

- **Source pool**
  - media
  - impact
  - tech

- **Support set**
  - religion
  - beauty
  - games

- **This gorgeous grandma proves beauty has no expiration date**

\[ x \]

\[ \begin{array}{c|c}
\text{w} & s(w) \\
\hline
\text{a after and april year yen} & 0.002 \ 0.137 \ 0.002 \ 0.099 \ 0.030 \ 0.422 \\
\end{array} \]

\[ \begin{array}{c|c}
\text{w} & t(w) \\
\hline
\text{a after and april year yen} & 0.017 \ 0.019 \ 0.011 \ 0.051 \ 0.004 \ 0.098 \\
\end{array} \]

\[ \text{b) Attention Generator} \]

\[ \text{biLSTM} \rightarrow \alpha \]

\[ \text{c) Ridge regressor: training from the support set} \]

\[ \Phi_S \rightarrow W \rightarrow \hat{Y}_S \]

\[ \text{d) Ridge regressor: inference on the query set} \]

\[ Y_Q \rightarrow \mathcal{L}^{CE} \rightarrow \hat{Y}_Q \]
Attention Generator

Distribution from the pool

\[ s(x_i) := \frac{\varepsilon}{\varepsilon + P(x_i)} \]

Distribution learned from the support set

\[ t(x_i) := \mathcal{H}(P(y | x_i))^{-1} \]

Attention weight learned from the support set

\[ h = \text{biLSTM}([s(x); t(x)]) \]

\[ \alpha_i := \frac{\exp(v^T h_i)}{\sum_j \exp(v^T h_j)} \]
Ridge regressor

Construct sentence representation

Learn from support set

\[ \mathcal{L}^{RR}(W) := \| \Phi_S W - Y_S \|_F^2 + \lambda \| W \|_F^2 \]

\[ W = \Phi_S^T (\Phi_S \Phi_S^T + \lambda I)^{-1} Y_S \]

Predict on query set

\[ \hat{Y}_Q = a \Phi_Q W + b \]
<table>
<thead>
<tr>
<th>Method</th>
<th>Alg.</th>
<th>20 News 1 shot</th>
<th>20 News 5 shot</th>
<th>Amazon 1 shot</th>
<th>Amazon 5 shot</th>
<th>HuffPost 1 shot</th>
<th>HuffPost 5 shot</th>
<th>RCV1 1 shot</th>
<th>RCV1 5 shot</th>
<th>Reuters 1 shot</th>
<th>Reuters 5 shot</th>
<th>FewRel 1 shot</th>
<th>FewRel 5 shot</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVG</td>
<td>NN</td>
<td>33.9</td>
<td>45.8</td>
<td>46.7</td>
<td>60.3</td>
<td>31.4</td>
<td>41.5</td>
<td>43.7</td>
<td>60.8</td>
<td>56.5</td>
<td>80.5</td>
<td>47.5</td>
<td>60.6</td>
<td>43.3</td>
</tr>
<tr>
<td>IDF</td>
<td>NN</td>
<td>38.8</td>
<td>51.9</td>
<td>51.4</td>
<td>67.1</td>
<td>31.5</td>
<td>42.3</td>
<td>41.9</td>
<td>58.2</td>
<td>57.8</td>
<td>82.9</td>
<td>46.8</td>
<td>60.6</td>
<td>44.7</td>
</tr>
<tr>
<td>CNN</td>
<td>FT</td>
<td>33.0</td>
<td>47.1</td>
<td>45.7</td>
<td>63.9</td>
<td>32.4</td>
<td>44.1</td>
<td>40.3</td>
<td>62.3</td>
<td>70.9</td>
<td>91.0</td>
<td>54.0</td>
<td>71.1</td>
<td>46.0</td>
</tr>
<tr>
<td>AVG</td>
<td>PROTO</td>
<td>36.2</td>
<td>45.4</td>
<td>37.2</td>
<td>51.9</td>
<td>35.6</td>
<td>41.6</td>
<td>28.4</td>
<td>31.2</td>
<td>59.5</td>
<td>68.1</td>
<td>44.0</td>
<td>46.5</td>
<td>40.1</td>
</tr>
<tr>
<td>IDF</td>
<td>PROTO</td>
<td>37.8</td>
<td>46.5</td>
<td>41.9</td>
<td>59.2</td>
<td>34.8</td>
<td>50.2</td>
<td>32.1</td>
<td>35.6</td>
<td>61.0</td>
<td>72.1</td>
<td>43.0</td>
<td>61.9</td>
<td>41.8</td>
</tr>
<tr>
<td>CNN</td>
<td>PROTO</td>
<td>29.6</td>
<td>35.0</td>
<td>34.0</td>
<td>44.4</td>
<td>33.4</td>
<td>44.2</td>
<td>28.4</td>
<td>29.3</td>
<td>65.2</td>
<td>74.3</td>
<td>49.7</td>
<td>65.1</td>
<td>40.1</td>
</tr>
<tr>
<td>AVG</td>
<td>MAML</td>
<td>33.7</td>
<td>43.9</td>
<td>39.3</td>
<td>47.2</td>
<td>36.1</td>
<td>49.6</td>
<td>39.9</td>
<td>50.6</td>
<td>54.6</td>
<td>62.5</td>
<td>43.8</td>
<td>57.8</td>
<td>41.2</td>
</tr>
<tr>
<td>IDF</td>
<td>MAML</td>
<td>37.2</td>
<td>48.6</td>
<td>43.6</td>
<td>62.4</td>
<td>38.9</td>
<td>53.7</td>
<td>42.5</td>
<td>54.1</td>
<td>61.5</td>
<td>72.0</td>
<td>48.2</td>
<td>65.8</td>
<td>45.3</td>
</tr>
<tr>
<td>CNN</td>
<td>MAML</td>
<td>28.9</td>
<td>36.7</td>
<td>35.3</td>
<td>43.7</td>
<td>34.1</td>
<td>45.8</td>
<td>39.0</td>
<td>51.1</td>
<td>66.6</td>
<td>85.0</td>
<td>51.7</td>
<td>66.9</td>
<td>42.6</td>
</tr>
<tr>
<td>AVG</td>
<td>RR</td>
<td>37.6</td>
<td>57.2</td>
<td>50.2</td>
<td>72.7</td>
<td>36.3</td>
<td>54.8</td>
<td>48.1</td>
<td>72.6</td>
<td>63.4</td>
<td>90.0</td>
<td>53.2</td>
<td>72.2</td>
<td>48.1</td>
</tr>
<tr>
<td>IDF</td>
<td>RR</td>
<td>44.8</td>
<td>64.3</td>
<td>60.2</td>
<td>79.7</td>
<td>37.6</td>
<td>59.5</td>
<td>48.6</td>
<td>72.8</td>
<td>69.1</td>
<td>93.0</td>
<td>55.6</td>
<td>75.0</td>
<td>52.6</td>
</tr>
<tr>
<td>CNN</td>
<td>RR</td>
<td>32.2</td>
<td>44.3</td>
<td>37.3</td>
<td>53.8</td>
<td>37.3</td>
<td>49.9</td>
<td>41.8</td>
<td>59.4</td>
<td>71.4</td>
<td>87.9</td>
<td>56.8</td>
<td>71.8</td>
<td>46.1</td>
</tr>
<tr>
<td>OUR</td>
<td></td>
<td><strong>52.1</strong></td>
<td><strong>68.3</strong></td>
<td><strong>62.6</strong></td>
<td><strong>81.1</strong></td>
<td><strong>43.0</strong></td>
<td><strong>63.5</strong></td>
<td><strong>54.1</strong></td>
<td><strong>75.3</strong></td>
<td><strong>81.8</strong></td>
<td><strong>96.0</strong></td>
<td><strong>67.1</strong></td>
<td><strong>83.5</strong></td>
<td><strong>60.1</strong></td>
</tr>
<tr>
<td>OUR w/o t(·)</td>
<td></td>
<td>50.1</td>
<td>67.5</td>
<td>61.7</td>
<td>80.5</td>
<td>42.0</td>
<td>60.8</td>
<td>51.5</td>
<td>75.1</td>
<td>76.7</td>
<td>93.7</td>
<td>66.9</td>
<td>83.2</td>
<td>58.1</td>
</tr>
<tr>
<td>OUR w/o s(·)</td>
<td></td>
<td>41.9</td>
<td>60.7</td>
<td>51.1</td>
<td>75.3</td>
<td>40.1</td>
<td>60.2</td>
<td>48.5</td>
<td>72.8</td>
<td>78.1</td>
<td>94.8</td>
<td>65.8</td>
<td>82.6</td>
<td>54.2</td>
</tr>
<tr>
<td>OUR w/o biLSTM</td>
<td></td>
<td>50.3</td>
<td>66.9</td>
<td>61.9</td>
<td>80.9</td>
<td>42.2</td>
<td>63.0</td>
<td>51.8</td>
<td>74.1</td>
<td>77.2</td>
<td>95.4</td>
<td>66.4</td>
<td>82.9</td>
<td>58.3</td>
</tr>
<tr>
<td>OUR w EBD</td>
<td></td>
<td>39.7</td>
<td>57.5</td>
<td>56.5</td>
<td>76.3</td>
<td>40.6</td>
<td>58.6</td>
<td>48.6</td>
<td>71.5</td>
<td>81.7</td>
<td>95.8</td>
<td>61.5</td>
<td>80.9</td>
<td>54.8</td>
</tr>
</tbody>
</table>
Generalization and Overfitting

(a) CNN+PROTO

(b) OUR
PCA visualization

(a) $s(\cdot)$

(b) $t(\cdot)$

(c) OUR