Dynamic Memory Induction Networks for Few-Shot Text Classification

Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, Xiaodan Zhu
Motivation

Observation: good prototypes yield better performance

Literature:

Average prototype (Prototypical Network)

Query-based weighted sum (Hybrid Attention Prototypical network)

Induction network (Dynamic Routing)

This paper try to advance the prototypes using dynamic memory with knowledge from supervised task
Overview

Supervised Learning
- Base Set
  - Batch

Pretrained Encoder
- Support Set
  - Class 1
  - Class 2
  - Class 3
- Query Set
  - Query

Dynamic Memory Module
- Sample Vector
- Class Vector

Query-enhanced Induction Module
- Query Vector

Classifier
- Classification Score

Meta Learning
Three-stage matching network

1. Dynamic Memory Instance Encoder

\[ q' = DMR(M, q) \]

\( q' \) is a memory matrix \( M \) (here \( W_{\text{base}} \))

\( q \) is a sample vector \( q \in \mathbb{R}^d \).

2. Dynamic Memory Query-based Prototype

\[ e_c = DMR(\{e'_{c,s}\}_{s=1,...,K}, e_q). \]

3. Classifier

\[ s_{q,c} = \tau \cdot \cos(e_q, e_c) = \tau \cdot \overline{e_q}^T \overline{e_c}. \]
Dynamic Memory Routing

\[ q' = DMR(M, q) \]

1. Linear transformation and squash

\[ \hat{m}_{ij} = \text{squash}(W_jm_i + b_j), \]
\[ \hat{q}_j = \text{squash}(W_jq + b_j), \]

2. Pearson Correlation Coefficient

\[ p_{ij} = \tanh(PCCs(\hat{m}_{ij}, \hat{q}_j)), \]
\[ PCCs = \frac{Cov(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}}. \]

3. Routing iteration

\[ d_i = \text{softmax}(\alpha_i), \]
\[ \alpha_{ij} = \alpha_{ij} + p_{ij}\hat{m}_iv_j. \]

4. Vector representation

\[ \hat{v}_j = \sum_{i=1}^{n}(d_{ij} + p_{ij})m_{ij}, \]
\[ v_j = \text{squash}(\hat{v}_j). \]
**Algorithm 1** Dynamic Memory Routing Process

Require: $r$, $q$ and memory $M = \{m_1, m_2, ..., m_n\}$

Ensure: $v = v_1, v_2, ..., v_l, q'$

1: for all $m_i, v_j$ do
2:     $\hat{m}_{ij} = \text{squash}(W_j m_i + b_j)$
3:     $\hat{q}_j = \text{squash}(W_j q + b_j)$
4:     $\alpha_{ij} = 0$
5:     $p_{ij} = \tanh(\text{PCCs}(\hat{m}_{ij}, \hat{q}_j))$
6: end for
7: for $r$ iterations do
8:     $d_i = \text{softmax}(\alpha_i)$
9:     $\hat{v}_j = \sum_{i=1}^{n} (d_{ij} + p_{ij}) \hat{m}_{ij}$
10:    $v_j = \text{squash}(\hat{v}_j)$
11:   for all $i, j$: $\alpha_{ij} = \alpha_{i,j} + p_{ij} \hat{m}_{ij} v_j$
12:   for all $j$: $\hat{q}_j = \frac{\hat{q}_j + v_j}{2}$
13:   for all $i, j$: $p_{ij} = \tanh(\text{PCCs}(\hat{m}_{ij}, \hat{q}_j))$
14: end for
15: $q' = \text{concat}[v]$
16: Return $q'$
<table>
<thead>
<tr>
<th>Model</th>
<th>5-way Acc.</th>
<th>10-way Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-shot</td>
<td>5-shot</td>
</tr>
<tr>
<td>BERT</td>
<td>30.79±0.68</td>
<td>63.31±0.73</td>
</tr>
<tr>
<td>ATAML</td>
<td>54.05±0.14</td>
<td>72.79±0.27</td>
</tr>
<tr>
<td>Rel. Net</td>
<td>59.19±0.12</td>
<td>78.35±0.27</td>
</tr>
<tr>
<td>Ind. Net</td>
<td>60.97±0.16</td>
<td>80.91±0.19</td>
</tr>
<tr>
<td>HATT</td>
<td>60.40±0.17</td>
<td>79.46±0.32</td>
</tr>
<tr>
<td>LwoF</td>
<td>63.35±0.26</td>
<td>78.83±0.38</td>
</tr>
<tr>
<td>DMIN</td>
<td>65.72±0.28</td>
<td>82.39±0.24</td>
</tr>
</tbody>
</table>
# Ablation study

<table>
<thead>
<tr>
<th>Model</th>
<th>Iteration</th>
<th>1 Shot</th>
<th>5 Shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o DMM</td>
<td>3</td>
<td>81.79</td>
<td>90.19</td>
</tr>
<tr>
<td>w/o QIM</td>
<td>3</td>
<td>82.37</td>
<td>90.57</td>
</tr>
<tr>
<td>DMIN</td>
<td>1</td>
<td>82.70</td>
<td>90.92</td>
</tr>
<tr>
<td>DMIN</td>
<td>2</td>
<td>82.95</td>
<td>91.18</td>
</tr>
<tr>
<td>DMIN</td>
<td>3</td>
<td><strong>83.46</strong></td>
<td><strong>91.75</strong></td>
</tr>
</tbody>
</table>