Meta-Learning for Effective Multi-task and Multilingual Modelling

Ishan Tarunesh, Sushil Khyalia, Vishwajeet Kumar, Ganesh Ramakrishnan, Preethi Jyothi
EACL 2021
**Motivation**

Address multitask - multilingual modeling

Address sampling strategy for meta learning in multitask scenario

<table>
<thead>
<tr>
<th>Task</th>
<th>en</th>
<th>hi</th>
<th>es</th>
<th>de</th>
<th>fr</th>
<th>zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Language Inference (NLI)</td>
<td>392K</td>
<td></td>
<td>392K</td>
<td></td>
<td>392K</td>
<td></td>
</tr>
<tr>
<td>Question Answering (QA)</td>
<td>88.0K</td>
<td>82.4K</td>
<td>81.8K</td>
<td>80.0K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part Of Speech (POS)</td>
<td>21.2K</td>
<td>13.3K</td>
<td>28.4K</td>
<td>166K</td>
<td></td>
<td>7.9K</td>
</tr>
<tr>
<td>Named Entity Recognition (NER)</td>
<td>20K</td>
<td>5K</td>
<td>20K</td>
<td>20K</td>
<td>20K</td>
<td>20K</td>
</tr>
<tr>
<td>Paraphrase Identification (PA)</td>
<td>49.4K</td>
<td></td>
<td>49.4K</td>
<td></td>
<td>49.4K</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Dataset matrix showing datasets that are available (green) from the XTREME Benchmark. The number of training instances are also mentioned for each available dataset.
Task selection and Data sampling

- **Task selection**
  - **Task limited**: select all languages for a given task
  - **Lang-limited**: select all tasks for a given language

- **Data sampling**
  - Temperature-based heuristic
    \[ P_D(i) = \frac{q_i^{1/\tau}}{\left( \sum_{k=1}^{n} q_k^{1/\tau} \right)} \]
  - Parameterized sampling: P(D) is learnable
Parameterized sampling

The probability is parameterized by  \[ P_D(i) = e^{\psi_i} / \sum_j e^{\psi_j} \]

Alternate update:

\[ \psi^* = \arg \min_{\psi} J(\theta^*(\psi), D_{dev}) \quad (3) \]

\[ \theta^*(\psi) = \arg \min_{\theta} E_{x,y \sim P(T; \psi)} [l(x, y; \theta)] \quad (4) \]

Where \( J(\theta, D_{dev}) \) is the objective on development data

Reward function

\[ R(x,y; \theta_t) \approx \left( \nabla J(\theta_t, D_{dev}) \right)^T g_{dev} \cdot \nabla l(x,y; \theta_{t-1}) g_{train} \quad (5) \]

\[ \approx \cos(g_{dev}, g_{train}) \quad (6) \]

Update parameter:

\[ \psi_{t+1} \leftarrow \psi_t + R(x,y; \theta_t) \cdot \nabla_{\psi} \log(P(x,y; \psi)) \quad (7) \]
Algorithm 1 Our Meta-learning Approach

**Input:** $\mathcal{D}_{\text{train}}$ set of TLPs for meta training
(Also $\mathcal{D}_{\text{dev}}$ for parametrised sampling)

Sampling Strategy (Temperature / MultiDDS)

**Output:** The converged multi-task multilingual model parameters $\theta^*$

1: **Initialize** $P_D(i)$ depending on the sampling strategy
2: **while** not converged do
3: ▷ **Perform Reptile Updates**
4:  Sample $m$ TLPs $T_1, T_2, \ldots , T_m$ from $\mathcal{M}$
5:  for $i = 1, 2, \ldots , m$ do
6:  $\theta_i^{(k)} \leftarrow U_i^k(\theta)$, denoting $k$ gradient updates from $\theta$ on batches of TLP $T_i$
7:  end for
8: $\theta \leftarrow \theta + \frac{\beta}{m} \sum_{i=1}^{m} (\theta_i^{(k)} - \theta)$

9:  if Sampling Strategy $\leftarrow$ MultiDDS then
10:  for $\mathcal{D}_{\text{train}}^i \in \mathcal{D}_{\text{train}}$ do
11:  $R(i; \theta) \leftarrow \cos(g_{\text{dev}}, g_{\text{train}})$, $g_{\text{dev}}$ is gradient on $\{\mathcal{D}_{\text{dev}}\}$ and $g_{\text{train}}$ is gradient on $\mathcal{D}_{\text{train}}^i$
12:  end for
13: ▷ **Update Sampling Probabilities**
14:  $d_\psi \leftarrow \sum_{i=1}^{n} R(i; \theta) \cdot \nabla_\psi \log(P_D(i; \psi))$
15:  $\psi \leftarrow \text{GradientUpdate}(\psi, d_\psi)$
16:  end if
17: end while
Baselines

Baseline: train supervised on the target task-language pair
Task-limited MTL: multitask model on the same task
Lang-limited MTL: multitask model on the same language
All TLPs MTL: multitask model on all tasks and languages
<table>
<thead>
<tr>
<th>Model</th>
<th>SS</th>
<th>QA (F1)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>hi</td>
<td>es</td>
<td>de</td>
<td>en</td>
<td>es</td>
<td>de</td>
<td>fr</td>
<td>en</td>
<td>es</td>
<td>de</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td>79.94</td>
<td>59.94</td>
<td>65.83</td>
<td>63.17</td>
<td>81.39</td>
<td>78.37</td>
<td>76.82</td>
<td>77.30</td>
<td>92.35</td>
<td>89.75</td>
<td>87.45</td>
</tr>
<tr>
<td>Lang-Limited MTL</td>
<td></td>
<td>69.80</td>
<td>53.24</td>
<td>62.29</td>
<td>58.91</td>
<td>80.49</td>
<td>76.10</td>
<td>75.18</td>
<td>74.94</td>
<td>93.75</td>
<td>87.75</td>
<td>85.35</td>
</tr>
<tr>
<td>Task-Limited MTL</td>
<td></td>
<td>74.04</td>
<td>57.77</td>
<td>64.28</td>
<td>61.47</td>
<td>80.95</td>
<td>78.15</td>
<td>75.90</td>
<td>77.14</td>
<td>93.65</td>
<td>86.65</td>
<td>86.25</td>
</tr>
<tr>
<td>All TLPs MTL</td>
<td></td>
<td>63.22</td>
<td>42.94</td>
<td>54.05</td>
<td>51.61</td>
<td>80.05</td>
<td>76.48</td>
<td>74.86</td>
<td>76.18</td>
<td>93.50</td>
<td>90.30</td>
<td>88.45</td>
</tr>
<tr>
<td>Lang-Limited</td>
<td>Temp</td>
<td>-0.04</td>
<td>-0.24</td>
<td>-0.27</td>
<td>+0.07</td>
<td>+0.06</td>
<td>+0.39</td>
<td>+0.03</td>
<td>-0.70</td>
<td>+0.45</td>
<td>+0.05</td>
<td>+0.35</td>
</tr>
<tr>
<td></td>
<td>mDDS</td>
<td>+0.07</td>
<td>-0.12</td>
<td>+0.06</td>
<td>+0.14</td>
<td>+0.02</td>
<td>+0.61</td>
<td>-0.80</td>
<td>-0.60</td>
<td>-0.25</td>
<td>+0.05</td>
<td>+0.35</td>
</tr>
<tr>
<td>Task-Limited</td>
<td>Temp</td>
<td>+0.55</td>
<td>+0.43</td>
<td>+0.50</td>
<td>+0.40</td>
<td>+1.65</td>
<td>+1.12</td>
<td>+1.25</td>
<td>+0.79</td>
<td>-0.20</td>
<td>-0.15</td>
<td>-0.55</td>
</tr>
<tr>
<td></td>
<td>mDDS</td>
<td>+0.21</td>
<td>+0.62</td>
<td>-0.67</td>
<td>+1.06</td>
<td>+1.32</td>
<td>+1.10</td>
<td>+1.39</td>
<td>+0.48</td>
<td>+0.50</td>
<td>-0.65</td>
<td>-0.35</td>
</tr>
<tr>
<td>All TLPs</td>
<td>Temp</td>
<td>+0.53</td>
<td>+0.47</td>
<td>+0.32</td>
<td>+0.47</td>
<td>+1.90</td>
<td>+1.22</td>
<td>+1.45</td>
<td>+0.95</td>
<td>+0.35</td>
<td>+0.45</td>
<td>+1.20</td>
</tr>
<tr>
<td></td>
<td>mDDS-Lang</td>
<td>+0.08</td>
<td>+0.50</td>
<td>-1.57</td>
<td>+0.08</td>
<td>+0.76</td>
<td>+0.26</td>
<td>-0.10</td>
<td>+0.32</td>
<td>+0.25</td>
<td>+0.85</td>
<td>+0.75</td>
</tr>
<tr>
<td></td>
<td>mDDS-Task</td>
<td>+0.18</td>
<td>+0.60</td>
<td>+0.11</td>
<td>+0.54</td>
<td>+1.50</td>
<td>+0.90</td>
<td>+0.72</td>
<td>+0.72</td>
<td>+0.10</td>
<td>+0.80</td>
<td>+1.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>SS</th>
<th>NER (Acc.)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>en</td>
<td>hi</td>
<td>es</td>
<td>de</td>
<td>fr</td>
<td>zh</td>
<td>en</td>
<td>hi</td>
<td>es</td>
<td>de</td>
<td>zh</td>
<td>en</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td>93.23</td>
<td>95.72</td>
<td>95.84</td>
<td>97.32</td>
<td>95.48</td>
<td>94.34</td>
<td>96.15</td>
<td>93.57</td>
<td>96.02</td>
<td>97.37</td>
<td>92.60</td>
<td>95.08</td>
</tr>
<tr>
<td>Lang-Limited MTL</td>
<td></td>
<td>92.54</td>
<td>92.67</td>
<td>95.14</td>
<td>96.40</td>
<td>94.38</td>
<td>92.97</td>
<td>95.08</td>
<td>92.43</td>
<td>95.19</td>
<td>97.19</td>
<td>92.52</td>
<td>94.70</td>
</tr>
<tr>
<td>Task-Limited MTL</td>
<td></td>
<td>93.51</td>
<td>93.94</td>
<td>95.77</td>
<td>97.09</td>
<td>95.27</td>
<td>93.72</td>
<td>96.15</td>
<td>93.57</td>
<td>96.02</td>
<td>97.37</td>
<td>92.60</td>
<td>95.08</td>
</tr>
<tr>
<td>All TLPs MTL</td>
<td></td>
<td>92.28</td>
<td>91.95</td>
<td>94.90</td>
<td>96.18</td>
<td>94.38</td>
<td>92.53</td>
<td>96.15</td>
<td>93.57</td>
<td>96.02</td>
<td>97.37</td>
<td>92.60</td>
<td>95.08</td>
</tr>
<tr>
<td>Lang-Limited</td>
<td>Temp</td>
<td>+0.60</td>
<td>+0.06</td>
<td>+0.09</td>
<td>+0.24</td>
<td>-0.09</td>
<td>-0.47</td>
<td>-0.06</td>
<td>-0.01</td>
<td>+0.10</td>
<td>+0.04</td>
<td>+0.17</td>
<td>-0.27</td>
</tr>
<tr>
<td></td>
<td>mDDS</td>
<td>-0.21</td>
<td>-0.85</td>
<td>-0.20</td>
<td>-0.10</td>
<td>-0.57</td>
<td>-0.55</td>
<td>-0.27</td>
<td>-0.02</td>
<td>-0.19</td>
<td>+0.04</td>
<td>+0.17</td>
<td>-0.27</td>
</tr>
<tr>
<td>Task-Limited</td>
<td>Temp</td>
<td>+0.79</td>
<td>-0.46</td>
<td>0.00</td>
<td>-0.07</td>
<td>-0.18</td>
<td>-0.51</td>
<td>-0.22</td>
<td>-0.05</td>
<td>-0.21</td>
<td>+0.02</td>
<td>-0.09</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td>mDDS</td>
<td>-0.10</td>
<td>-1.61</td>
<td>0.00</td>
<td>-0.16</td>
<td>-0.33</td>
<td>-0.69</td>
<td>-0.22</td>
<td>-0.14</td>
<td>-0.19</td>
<td>+0.02</td>
<td>-0.09</td>
<td>-0.22</td>
</tr>
<tr>
<td>All TLPs</td>
<td>Temp</td>
<td>-0.15</td>
<td>0.03</td>
<td>+0.13</td>
<td>0.00</td>
<td>-0.16</td>
<td>-0.39</td>
<td>-0.22</td>
<td>-0.09</td>
<td>-0.21</td>
<td>+0.03</td>
<td>+0.16</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td>mDDS-Lang</td>
<td>-0.16</td>
<td>-0.09</td>
<td>+0.11</td>
<td>-0.08</td>
<td>-0.14</td>
<td>-0.65</td>
<td>-0.22</td>
<td>-0.10</td>
<td>-0.11</td>
<td>+0.03</td>
<td>+0.17</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td>mDDS-Task</td>
<td>-0.27</td>
<td>-0.42</td>
<td>+0.08</td>
<td>-0.14</td>
<td>-0.07</td>
<td>-0.58</td>
<td>-0.22</td>
<td>-0.14</td>
<td>-0.19</td>
<td>+0.02</td>
<td>-0.09</td>
<td>-0.22</td>
</tr>
</tbody>
</table>