# 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

| Task | en | hi | es | de | fr | zh |
|---|---|---|---|---|---|---|
| Natural Language Inference (NLI) | 392K | | 392K | 392K | 392K | |
| Question Answering (QA) | 88.0K | 82.4K | 81.8K | 80.0K | | |
| Part Of Speech (POS) | 21.2K | 13.3K | 28.4K | 166K | | 7.9K |
| Named Entity Recognition (NER) | 20K | 5K | 20K | 20K | 20K | 20K |
| Paraphrase Identification (PA) | 49.4K | | 49.4K | 49.4K | 49.4K | 49.4K |

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

$$P_{\mathcal{D}}(i) = q_i^{1/\tau} / \left( \sum_{k=1}^{n} q_k^{1/\tau} \right)$$

  - Parameterized sampling: P(D) is learnable

# Parameterized sampling

The probability is parameterized by $P_{\mathcal{D}}(i) = e^{\psi_i} / \sum_j e^{\psi_j}$

Alternate update:

$$\psi^* = \underset{\psi}{\operatorname{argmin}}\ J(\theta^*(\psi), \mathcal{D}_{dev}) \quad (3)$$

- $$\theta^*(\psi) = \underset{\theta}{\operatorname{argmin}}\ E_{x,y\sim P(T;\psi)}[l(x,y;\theta)] \quad (4)$$

Where $J(\theta, \mathcal{D}_{dev})$ is the objective on development data

Reward function

$$R(x,y;\theta_t) \approx \underbrace{\nabla J(\theta_t, \mathcal{D}_{dev})^T}_{g_{dev}} \cdot \underbrace{\nabla_\theta l(x,y;\theta_{t-1})}_{g_{train}} \quad (5)$$

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

Update parameter:

$$\psi_{t+1} \leftarrow \psi_t + R(x,y;\theta_t) \cdot \nabla_\psi log(P(x,y;\psi)) \quad (7)$$

# Algorithm

**Algorithm 1** Our Meta-learning Approach

**Input:** $\mathcal{D}_{train}$ set of TLPs for meta training (Also $\mathcal{D}_{dev}$ for parametrised sampling)
Sampling Strategy (Temperature / MultiDDS)

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

**Initialize** $P_D(i)$ depending on the sampling strategy
**while** not converged **do**
  ▷ *Perform Reptile Updates*
  Sample $m$ TLPs $T_1, T_2, \ldots, T_m$ from $\mathcal{M}$
  **for** i = 1,2,...,m **do**
    $\theta_i^{(k)} \leftarrow U_i^k(\theta)$, denoting $k$ gradient updates from $\theta$ on batches of TLP $T_i$
  **end for**
  $\theta \leftarrow \theta + \frac{\beta}{m} \sum_{i=1}^{m} (\theta_i^{(k)} - \theta)$
  **if** Sampling Strategy ← MultiDDS **then**
    **for** $\mathcal{D}_{train}^i \in \mathcal{D}_{train}$ **do**
      $R(i;\theta) \leftarrow cos(g_{dev}, g_{train})$, $g_{dev}$ is gradient on $\{\mathcal{D}_{dev}\}$ and $g_{train}$ is gradient on $\mathcal{D}_{train}^i$
    **end for**
    ▷ *Update Sampling Probabilities*
    $d_\psi \leftarrow \sum_{i=1}^{n} R(i;\theta) \cdot \nabla_\psi log(P_\mathcal{D}(i;\psi))$
    $\psi \leftarrow$ GradientUpdate$(\psi, d_\psi)$
  **end if**
**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

| Model | SS | QA (F1) | | | | NLI (Acc.) | | | | PA (Acc.) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | en | hi | es | de | en | es | de | fr | en | es | de | fr | zh |
| Baselines | | 79.94 | 59.94 | 65.83 | 63.17 | 81.39 | 78.37 | 76.82 | 77.30 | 92.35 | 89.75 | 87.45 | 89.61 | 83.32 |
| Lang-Limited MTL | | 69.80 | 53.24 | 62.29 | 58.91 | 80.49 | 76.10 | 75.18 | 74.94 | 93.75 | 87.75 | 85.35 | 88.55 | 80.49 |
| Task-Limited MTL | | 74.04 | 57.77 | 64.28 | 61.47 | 80.95 | 78.15 | 75.90 | 77.14 | 93.65 | 86.65 | 86.25 | 86.82 | 81.24 |
| All TLPs MTL | | 63.22 | 42.94 | 54.05 | 51.61 | 80.05 | 76.48 | 74.86 | 76.18 | 93.50 | 90.30 | 88.45 | 89.71 | 82.66 |
| Lang-Limited | Temp | -0.04 | -0.24 | -0.27 | +0.07 | +0.06 | +0.39 | +0.03 | -0.70 | +0.45 | +0.05 | +0.35 | +0.40 | -0.06 |
| | mDDS | +0.07 | -0.12 | +0.06 | +0.14 | +0.02 | -0.61 | -0.80 | -0.60 | -0.25 | -0.05 | 0.00 | -0.30 | -1.41 |
| Task-Limited | Temp | +0.55 | +0.43 | +0.50 | +0.40 | +1.65 | +1.12 | +1.25 | +0.79 | +0.20 | -0.15 | -0.55 | +0.85 | -0.15 |
| | mDDS | +0.21 | +0.62 | -0.67 | +1.06 | +1.32 | +1.10 | +1.39 | +0.48 | +0.50 | -0.65 | -0.35 | +1.45 | +1.06 |
| All TLPs | Temp | +0.53 | +0.47 | +0.32 | +0.47 | +1.90 | +1.22 | +1.45 | +0.95 | +0.35 | +0.45 | +1.20 | +1.05 | +0.85 |
| | mDDS-Lang | +0.08 | +0.50 | -1.57 | +0.08 | +0.76 | +0.26 | -0.10 | +0.32 | +0.25 | +0.85 | +0.75 | +0.75 | +1.11 |
| | mDDS-Task | +0.18 | +0.60 | +0.11 | +0.54 | +1.50 | +0.90 | +0.72 | +0.72 | +0.10 | +0.80 | +1.27 | +1.10 | +1.16 |

| Model | SS | NER (Acc.) | | | | | | POS (Acc.) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | en | hi | es | de | fr | zh | en | hi | es | de | zh |
| Baselines | | 93.23 | 95.72 | 95.84 | 97.32 | 95.48 | 94.34 | 96.15 | 93.57 | 96.02 | 97.37 | 92.60 |
| Lang-Limited MTL | | 92.54 | 92.67 | 95.14 | 96.40 | 94.38 | 92.97 | 95.08 | 92.43 | 95.19 | 97.19 | 89.71 |
| Task-Limited MTL | | 93.51 | 93.94 | 95.77 | 97.09 | 95.27 | 93.72 | 95.70 | 93.34 | 95.73 | 97.35 | 92.52 |
| All TLPs MTL | | 92.28 | 91.95 | 94.90 | 96.18 | 94.38 | 92.53 | 94.70 | 91.89 | 95.10 | 97.03 | 89.92 |
| Lang-Limited | Temp | +0.60 | +0.06 | +0.09 | +0.24 | -0.09 | -0.47 | -0.06 | -0.01 | +0.10 | +0.04 | -0.17 |
| | mDDS | -0.21 | -0.85 | -0.20 | -0.10 | -0.57 | -0.55 | -0.27 | -0.02 | -0.19 | -0.06 | -0.37 |
| Task-Limited | Temp | +0.79 | -0.46 | 0.00 | -0.07 | -0.18 | -0.51 | -0.22 | -0.05 | -0.21 | +0.02 | -0.09 |
| | mDDS | -0.10 | -1.61 | 0.00 | -0.16 | -0.33 | -0.69 | -0.38 | -0.02 | -0.22 | +0.05 | -0.12 |
| All TLPs | Temp | -0.15 | -0.70 | +0.13 | 0.00 | -0.16 | -0.39 | -0.22 | -0.09 | -0.21 | +0.03 | -0.16 |
| | mDDS-Lang | -0.16 | -0.09 | +0.11 | -0.08 | -0.14 | -0.65 | -0.21 | -0.10 | -0.11 | +0.03 | -0.17 |
| | mDDS-Task | -0.27 | -0.42 | +0.08 | -0.14 | -0.07 | -0.58 | -0.22 | -0.14 | -0.19 | +0.02 | -0.09 |