Discovering Language-neutral Sub-networks in Multilingual Language Models
Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, Karl Aberer
EPFL

Multilingual Language Models (MultiLMs)

- Pre-trained jointly on raw data from multiple languages
- Fine-tuned for a task using a single high-resource language dataset

Sub-Networks Transfer Well Across Languages

- High performance on cross-lingual transfer across languages (same task)

Language-specific sub-networks have shared multi-lingual components

Investigating Language Neutrality of MultiLMs

Hypothesis:
- Overlap between language-specific sub-networks indicates language neutrality

Extract sub-networks from MultiLMs:
- Using iterative magnitude pruning (Lottery Ticket Hypothesis [Frankle’19])
- Prune for individual language-task pairs

Sparsity Dampens Language Neutrality

Cross-lingual sub-network transfer degrades as the sub-networks get sparser
- Language-specific parameters are retained for the language for which a sub-network is discovered

Language-neutral parameters get pruned at higher sparsity

Sub-Networks Overlap Considerably

- NER sub-networks have the highest overlap
- MLM sub-networks have the lowest overlap
  - Upper layers are specialized to predict language-specific vocabularies
- Absolute overlap and cross-lingual performance are not correlated in a fine-grained manner

<table>
<thead>
<tr>
<th></th>
<th>MLM</th>
<th>XNLI</th>
<th>NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Overlap</td>
<td>68%</td>
<td>85%</td>
<td>94%</td>
</tr>
</tbody>
</table>

The overlap among each task’s sub-networks is high

Evaluate the language neutrality degree of sub-networks:
- Transfer extracted sub-networks across languages
- Fine-tune them on other language-task pairs

Language neutrality → overlap between language-specific sub-networks

Do MultiLMs learn language-neutral parameters?
- They demonstrate promising cross-lingual transfer performance
- Language neutrality is considered a key facilitator of this performance
  - Shared representations that encode similar phenomena across languages

Evaluating transferability of extracted sub-networks on French (Fr) and Urdu (Ur).
- Purple connections are shared in all sub-networks.

✓ Transferred  ✗ Not Transferred

Average of relative cross-lingual transfer performance drop for sub-networks with sparsity levels 50% and 80%. Relative performance change is computed as \( \frac{\text{source performance on language } s \text{ after pruning }}{\text{source performance on language } t \text{ before pruning}} \) where \( x \) and \( t \) are source and target languages and \( J \) is the set of languages for each task.