Intermediate-Task Training Robustness to Domain Divergence in Multilingual Cross-Lingual Transfer on XTREME

Assignee Research

June 23, 2026

Abstract

Transfer learning from large language models (LLMs) has emerged as a powerful technique to enable knowledge-based fine-tuning for a number of tasks, adaptation of models for different domains and even languages. However, it remains an open question, if and when transfer learning will work, i.e. leading to positive or negative transfer. In this paper, we analyze the knowledge transfer across three natural language processing (NLP) tasks - text classification, sentimental analysis, and sentence similarity, using three LLMs - BERT, RoBERTa, and XLNet - and analyzing their performance, by fine-tun

1 Introduction

This paper examines: The (In)Effectiveness of Intermediate Task Training For Domain Adaptation and Cross-Lingual Transfer Learning. Research question: Does the effectiveness of intermediate-task training for cross-lingual transfer on XTREME degrade when the intermediate task domain diverges significantly from the target task domain in multilingual settings?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.1/10.

3 Results

13 papers retrieved. 15 claims extracted; 11 independently verified. Quality review score: 7.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Fine-tuning a post-intermediate task training transfer learnt RoBERTa LLM outperformed other models in three out of six	✓	0.28
BERT outperformed other models in three out of six tasks, specifically in text classification (domain adaptation and cro	✓	0.16
XLNet was consistently the worst performing model in all experiments conducted in the study.	✓	0.16
For the Text Classification target task 'SMS Spam', the BERT-F model achieved an accuracy of 0.71, which is higher than	×	0.15
For the Text Classification target task 'French/Spanish Movie Reviews', the BERT-IF model achieved an accuracy of 0.67,	✓	0.15
For the Sentiment Analysis target task 'GoEmotions', the BERT-IF model achieved the highest accuracy of 0.89 among all t	×	0.08
For the Sentiment Analysis target task 'French/German Emotions', the RoBERTa-F model achieved an accuracy of 0.72, outpe	×	0.12
For the Sentence Similarity target task 'Stock Ticker', the RoBERTa-F model achieved an accuracy of 0.72, while the BERT	✓	0.16
For the Sentence Similarity target task 'PAWS-X', the RoBERTa-F model achieved an accuracy of 0.74, which is the highest	×	0.09
In the study methodology, intermediate task training was performed by training each pre-trained LLM for 100 epochs using	✓	0.27
Fine-tuning for target tasks was performed by training for 10 epochs, regardless of whether intermediate task training w	✓	0.19
For all tasks, both intermediate task training and fine-tuning utilized 70% of the dataset for training and the remainin	✓	0.15
During transfer learning, all model weights were updated and no layers were frozen.	✓	0.20
The dataset used for intermediate task training was at least an order of magnitude larger than the dataset used for fine	✓	0.19
Similar trends where RoBERTa and BERT outperform XLNet in transfer learning have been reported for clickbait detection,	✓	0.16

References

- http://arxiv.org/abs/1910.03548v2
- http://arxiv.org/abs/2210.01091v2
- http://arxiv.org/abs/2003.11080v5