Order of Intermediate Tasks in Sequential Multi-Task Learning for Zero-Shot Cross-Lingual Transfer on XTREME-R

Assignee Research

July 8, 2026

Abstract

Multilingual models jointly pretrained on multiple languages have achieved remarkable performance on various multilingual downstream tasks. Moreover, models finetuned on a single monolingual downstream task have shown to generalize to unseen languages. In this paper, we first show that it is crucial for those tasks to align gradients between them in order to maximize knowledge transfer while minimizing negative transfer. Despite its importance, the existing methods for gradient alignment either have a completely different purpose, ignore inter-task alignment, or aim to solve continual learning

1 Introduction

This paper examines: Sequential Reptile: Inter-Task Gradient Alignment for Multilingual Learning. Research question: Does the order of intermediate tasks (sequential multi-task learning) affect zero-shot cross-lingual transfer performance on XTREME-R compared to single-task transfer?.

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. 14 claims extracted; 10 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
STL (Single Task Learning) is a model trained on each single language individually.	✓	0.18
The base MTL model's objective is defined as Eq. 1 without the regularizer Ω(φ).	×	0.14
GradNorm prevents training from being dominated by a single task by adaptively weighting each task gradient.	✓	0.22
PCGrad projects a task gradient onto other task gradients if the inner product between them is negative.	✓	0.21
GradVac alters task gradients to match the empirical moving average of cosine similarity between task gradients.	✓	0.24
RecAdam prevents catastrophic forgetting by penalizing the 2 distance from the pretrained model.	✓	0.17
Reptile performs inner-optimization individually for each task.	✓	0.18
Sequential Reptile aligns gradients across tasks by composing the inner-learning trajectory with all tasks.	✓	0.19
In synthetic experiments, three local optima were defined at coordinates x1=(0, 10), x2=(0, 0), and x3=(10, 0).	✓	0.18
In synthetic experiments, the MTL objective was optimized from the initialization point (20, 5).	×	0.11
In synthetic experiments, all baselines except Reptile and Sequential Reptile converged to one of the MTL local minima.	×	0.12
In synthetic experiments, the solution found by Reptile exhibited very low cosine similarity.	×	0.09
Sequential Reptile finds a trade-off between minimizing MTL loss and maximizing cosine similarity due to implicit enforc	✓	0.23
Table 1 reports F1 and EM scores on the TYDI-QA dataset for Question Answering tasks.	✓	0.16

References

- https://arxiv.org/abs/2110.02600
- http://arxiv.org/abs/1909.09587v1
- https://arxiv.org/abs/2202.13083