UPPSALA UNIVERSITET
Department of Linguistics and Philology

Xingming Li
May 2025

Thesis Project Overview

Student

Xingming Li

Academic supervisor

Fredrik Wahlberg

Title

Multilingual NLP Efficiency through Model Compression and Adaptive Inference: Knowledge Distillation and Early Exit on XLM-R

Overview

This thesis investigates how to improve the computational efficiency of a large multilingual transformer model without sacrificing cross-lingual generalization performance much. Specifically, the project focuses on compressing XLM-R Large into XLM-R Base using intermediate layer knowledge distillation and applying the early exit mechanism as an adaptive inference technique to further reduce inference cost.

To closely align with the original XLM-R pretraining configuration, I implemented a custom Hugging Face dataset loader[1] to construct a multilingual distillation dataset (about 27 GB) from CC100, preserving proportional representation across 100 languages. For memory efficiency and cross-lingual coverage, the dataset was partitioned into five language-aligned chunks, with each epoch consuming one chunk while ensuring full dataset coverage over five epochs. The distillation procedure incorporated multiple layer-mapping strategies to explore the effect of varying alignment depth between teacher and student layers. Checkpoints were saved after selected epochs (e.g., after epoch 1, 3 and 5) to analyze the impact of distillation data size on downstream performance. Additionally, I implemented early exit mechanisms on selected distilled models by adding intermediate classifiers and entropy-based stopping criteria. This step allowed further reduction in inference cost by enabling the model to exit early when predictions were sufficiently confident.

[1] `https://huggingface.co/datasets/xmli/filtered_cc100_27gb`

Model	Acc. (%)	F1(%)	GFLOPs	Inference Time (ms)	Memory Usage (GB)
Early Exit Distilled XLM-R	77.8	62.3	7.46	30.6	1.17
Distilled XLM-R	78.9	65.7	10.88	25.6	1.17
XLM-R Large	81.6	70.5	38.69	82.0	2.18

Table 1: Performance of some models evaluated in Swahili using WikiANN-NER benchmark under zero-shot setting.

The distilled models were evaluated on two multilingual benchmarks:

- XNLI (sentence-level natural language inference), and
- WikiANN-NER (token-level named entity recognition).

All models were fine-tuned using only English training data and then evaluated separately on six typologically and resource-diverse languages: English, Russian, Hindi, Turkish, Swahili, and Urdu. This setup allows for measuring both in-language and zero-shot transfer performance. Evaluation includes both task-specific metrics (accuracy for XNLI, accuracy and F1 for WikiANN-NER) and efficiency-oriented metrics (FLOPs, average inference time, and peak memory usage), offering a comprehensive view of performance–efficiency trade-offs across languages and tasks.

Some representative results for Swahili, evaluated using the WikiANN-NER benchmark under a zero-shot setting, are presented in Table 1. These results demonstrate that the distilled model, derived from XLM-R Large, retains a significant portion of the teacher's multilingual performance while operating at a much lower computational cost. Although the distilled model achieves lower F1 and accuracy than the teacher, it requires substantially less inference time, fewer FLOPs and memory resources. Moreover, applying early exit to the distilled model yields further reductions in FLOPs, with only a modest drop in F1 and accuracy. Nonetheless, the measured inference time is slightly higher possibly due to runtime overhead from exit condition checks and implementation bottlenecks. This highlights the need for more optimized implementations of adaptive inference to realize practical latency gains.

In conclusion, this work contributes to the growing need for deployable, efficient multilingual language models, especially in low-resource or edge computing settings, and opens avenues for future research on multilingual generalization under compression.