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 early exit mechanism as an adaptive inference technique to further reduce inference cost.

To align closely with the original XLM-R pretraining setup, I developed a Hugging Face dataset loader[1] to filter out a multilingual distillation dataset (about 27 GB) from CC100, ensuring proportional representation across 100 languages. The distillation process was conducted over five epochs, with the dataset split into five language-aligned partitions to allow for memory-efficient training. Multiple layer mapping configurations and intermediate checkpoints were explored to evaluate how layer depth and distillation data size affect downstream task performance.

The distilled models were evaluated on two multilingual benchmarks:

- XNLI (sentence-level inference) and
- WikiANN-NER (token-level named entity recognition)

The evaluation covered six typologically and resource-diverse languages (English, Russian, Hindi, Turkish, Swahili, and Urdu). Both accuracy/F1 and efficiency metrics (FLOPs, inference time, and memory usage) were measured to quantify trade-offs.

[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 (5.4 GB data)	78.9	65.7	10.88	25.6	1.17
Distilled XLM-R (27 GB data)	78.5	64.4	10.88	25.6	1.17
XLM-R Base	80.2	67.8	10.88	25.6	1.17
XLM-R Large	81.6	70.5	38.69	82.0	2.18

Table 1: Model performance in Swahili evaluated using WikiANN-NER benchmark under zero-shot setting.

Findings, partially shown in Table 1, indicate that modest distillation (e.g., 5.4 GB distillation data used) can yield strong results, likely due to the prior multilingual knowledge in the pre-trained student. Additionally, early exit applied to the distilled model offers further reductions in FLOPs, with minimal loss in accuracy/F1. However, the performance loss in low- and medium-resource languages is relatively greater.

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.