Project Proposal for the course Applied Deep Learning:

Project Background

The field of NLP and language modeling is rapidly evolving and remains an active area of research. The emergence of Large Language Models (LLMs), both open and closed, such as GPT, LLaMA, Claude, Mistral, and others, has demonstrated tremendous impact and wide-ranging applications. However, while these models achieve strong performance across a wide variety of tasks, they typically contain many billions of parameters. Additionally, although many LLMs now support multiple languages but they perform underwhelmingly on medium and low resource languages, there remains a strong imbalance between high- and low-resource languages. Expanding language coverage further increases model size and computational demands.

In practice, it can be beneficial to pretrain-distill a large teacher model into a smaller student model. LLMs are generally over-parameterized and require substantial computational resources for fine-tuning or even general deployment. Knowledge distillation addresses these challenges by transferring knowledge from a large teacher model to a more lightweight student model, ideally resulting in a smaller model that maintains competitive performance.

However, knowledge distillation introduces several key challenges. Due to their reduced parameter capacity, student models have limited representational power, which may negatively impact performance. *Efficiently transferring knowledge from teacher to student is an active research question.* Furthermore, for medium- and low-resource languages, it can be difficult to obtain high-quality tokenizers or embeddings, identify strong teacher models, and collect sufficient training data for effective distillation.

In this project, we will explore the development of a framework for pretraining-distilling a large language model into a smaller student model. This work will examine high-resource, medium-resource, and low-resource language scenarios, including cases where the teacher model does not provide support for the target language. A successful approach would enable efficient deployment of compact models for underrepresented languages, and may also serve as a blueprint for distilling large models into specialized, deployable models with competitive task performance, including for tasks that the larger model was not initially trained on.

Data Situation

In the following GitHub repository — https://github.com/AI4Bharat/IndicLLMSuite?tab=readme-ov-file, there is openly accessible data that can be used for large-scale pretraining, fine-tuning/instruction, and evaluation for 22 Indic languages. There is also support for data pipelines, and the authors address data scarcity through synthetic translation, romanization, and cleaning pipelines.

This dataset allows us to use high-, medium-, and low-resource Indic languages and provides original text, English translations, and romanized versions. Romanization is a standard technique in NLP (and an

active research area) , and this dataset enables us to perform the same distillation setup across multiple languages, ensuring fair and consistent comparison.

Project Approach and Outline

Our plan is to perform pre-training distillation, regular distillation during training, and post-training distillation. We will use romanization and logits-based similarity. We also discussed the possibility of incorporating feature-based similarity or embedding alignment as potential extensions, depending on time availability. Our aim is to train a small student model with ‘only’ 100 million or fewer parameters on indic languages and come up with the training recipe for all low resource languages.

We will apply the same procedure to all six languages (1 high resource, 2 medium resource, 3 low resource languages) to ensure comparability. Model performance will be evaluated using perplexity, next-token prediction, BLEU scores and other evaluations criteria (still a discussion in progress). At this time, we have not yet selected a specific knowledge-distillation algorithm, loss function(s) for the various tasks, or a specific teacher model, although a possible candidate would be RoBERTa and other multilingual base models.

We have also identified a large survey (https://arxiv.org/pdf/2503.12067) that discusses the current state of research on distillation. We plan to use this work to guide our decisions on algorithm selection, loss functions, and implementation details.

The main goal of this project can be summarized by the following questions:

“How well do low-resource languages perform compared to high- and medium-resource languages? and can we implement small language student models for these languages ? ”

,
“Can we identify a workable and implementable framework for performing distillation and large-scale compression across low-, medium-, and high-resource languages while still achieving comparable performance?”

,

“Identification of further research directions in this topic”