Pontificia Universidad Católica de Chile
Escuela de Ingeniería
IIC3697 Aprendizaje Profundo

Proyecto de Curso

Fecha Máxima de Entrega: 7 de Julio

Estimad@s, llegó la hora de comenzar el proyecto de curso, en el cual podrán poner en práctica los conocimientos que han adquirido este semestre sobre Aprendizaje Profundo (*Deep Learning*). Dado que el curso es nivel 3000, es decir, nivel postgrado, como parte de su metodología deben también ejercitar su capacidad para realizar estudio independiente e investigación en esta área, lo cual es el principal objetivo del proyecto de curso.

En esta oportunidad podrán desarrollar su proyecto en grupos de hasta 4 integrantes. Los grupos pueden formarse libremente. Si necesitan ayuda, envíen un correo utilizando la plataforma Canvas o escriban directamente a la ayudante jefe *acota@uc.cl*.

1 Introducción

Los Grandes Modelos de Lenguaje o LLMs (Large Language Models), han revolucionado el área de la IA gracias a sus avanzadas capacidades cognitivas. Al entrenarse con grandes y variados corpus de texto, aprenden patrones relevantes y operaciones sobre ellos, lo que les permite generar texto coherente, responder preguntas, entre otras tareas.

En el mundo se hablan más de 7000 idiomas, pero los recursos digitales disponibles son muy desiguales. Más del 90% se consideran idiomas de bajos recursos en términos de la cantidad de datos disponibles (low-resource languages), mientras que menos del 10% están en la categoría de altos recursos (high-resource languages). Para reducir esta brecha, la comunidad de NLP (*natural lenguaje processing*) ha impulsado iniciativas específicas, por ejemplo, este año se realizó LoResLM, el primer workshop sobre LLMs para idiomas de bajos recursos, que promueve la inclusión lingüística en NLP. En este contexto, gracias a su capacidad de generalizar con pocos ejemplos (*few-shot*), los LLMs se perfilan como una herramienta prometedora para desarrollar recursos y aplicaciones lingüísticas en idiomas con baja disponibilidad de datos.

2 Descripción

El proyecto de curso se centrará en fortalecer la traducción automática de los LLMs en idiomas con pocos recursos. El objetivo es diseñar y aplicar una metodología que incremente la calidad de traducción de un modelo para un idioma con escasa disponibilidad de datos, reutilizando las representaciones aprendidas al entrenar un traductor con idiomas de alta disponibilidad. La idea es aprovechar representaciones intermedias que capturen paralelismos semánticos entre idiomas. Dada la alta demanda de cómputo de estos modelos, se recomienda emplear arquitecturas que mantengan la carga computacional en niveles manejables. La estrategia propuesta consiste en: (1) elegir algún idioma con abundantes datos para entrenar (o partir desde un modelo ya entrenado) y (2) adaptar ese modelo a un idioma con escasos datos utilizando un número mínimo de ejemplos. Para ello pueden emplearse datasets como Tatoeba (más de 400 idiomas) o Flores-200 (más de 200 idiomas), ambos con corpus paralelos.

En lo referente a modelos, estudios recientes [Khade O. et al., 2025; Liang X. et al., 2025] demuestran que técnicas como LoRA (Low-Rank Adaptation) o sus derivados son una herramienta atractiva para aplicaciones en el contexto de idiomas de bajos recursos, por ejemplo: maratí, hindi o malayo. LoRA es una técnica para adaptar los modelos grandes preentrenados a tareas específicas o datasets, la cual congela los pesos del modelo preentrenado e inyecta matrices de bajo rango en capas específicas del modelo original, de tal forma que al realizar el fine-tuning al modelo, solo los parámetros que son parte de estas matrices inyectadas son actualizados [Hu et al., 2021].

Algunos temas de investigación que podrían explorar en sus proyectos:

- Estudiar métricas adecuadas para medir la calidad de la traducción
- Analizar el efecto de utilizar idiomas de la misma familia lingüística que el idioma de bajos recursos
- Estudiar el impacto de la técnica de tokenización utilizado (BPE multilingüe, SentencePiece, etc.)
- Observar cómo varía el desempeño en función de la cantidad de datos disponibles para el lenguaje con pocos recursos.

3 Lecturas recomendadas

Como punto de partida es importante que realicen una revisión bibliográfica que les permita familiarizarse con la temática del proyecto de curso y así definir en qué enfocar su trabajo. Acá una lista de lecturas que pueden revisar y considerar como parte de su informe de avance: :

- Hu E. et al., LoRA: Low-Rank Adaptation of Large Language Models, 2021, `https://arxiv.org/pdf/2106.09685`
- NLLB Team and R. et al. No Language Left Behind: Scaling Human-Centered Machine Translation, 2022, `https://arxiv.org/pdf/2207.04672`
- Khade O. et al., Challenges in adapting multilingual LLMs to Low-Resource Languages using LoRA PEFT Tuning, 2025, `https://aclanthology.org/2025.chipsal-1.22.pdf`
- Liang X. et al., Toward Low-Resource Languages Machine Translation: A Language-Specific Fine-Tuning With LoRA for Specialized Large Language Models, 2025, `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10918960`
- Popovic M., ChrF: character n-gram F-score for automatic MT evaluation, 2015, `https://aclanthology.org/W15-3049.pdf`
- Zhao M., et al., A multilingual BPE Embedding Space for Universal Sentiment Lexicon Induction, 2019, `https://aclanthology.org/P19-1341.pdf`
- Moskvoretskii V. et al., Low-Resource Machine Translation through the Lens of Personalized Federated Learning, 2024, `https://aclanthology.org/2024.findings-emnlp.514.pdf`
- Chronopoulou A. et al., Language-Family Adapters for Low-Resource Multilingual Neural Machine Trans- lation, 2023, `https://aclanthology.org/2023.loresmt-1.5.pdf`

4 Nivel de Complejidad Esperado

El proyecto constituye un componente fundamental del curso, tanto desde el punto de vista metodológico como en su impacto en la evaluación final. IIC3697 es un curso de 10 créditos semanales, por lo que se espera que una parte sustancial de ese tiempo se dedique al desarrollo riguroso de este trabajo. Es fundamental que el proyecto esté orientado a profundizar su formación en los contenidos del curso, mediante un análisis cuidadoso y una implementación sólida.

Como criterio general, los resultados cuantitativos no serán suficientes por sí solos: un proyecto que presente buenos resultados pero que consista esencialmente en una copia de una implementación pública, sin un esfuerzo en comprensión o una adaptación significativa, será evaluado negativamente.

5 Fechas Importantes

Informe de Avance: Fecha entrega 23/06
Este informe de avance no tiene nota pero será tomado en cuenta al momento de evaluar el informe final. El objetivo de este informe es demostrar que el desarrollo del proyecto se ha realizado en forma sistemática durante el período de tiempo asignado. El formato de entrega es un breve documento en pdf indicando los avances a la fecha. Como mínimo, este informe debe incluir: i) Una descripción del trabajo a realizar, ii) Una revisión bibliográfica, y iii) Una descripción de sus avances en la implementación, en lo posible un archivo con los avances del código implementado a la fecha, no importa que aún no sea funcional.

Entrega Final: Fecha entrega 07/07
El informa final debe ser documentado en formato de artículo académico. En la página web del curso se publicarán templates en latex y Microsoft Word que pueden ser usados para este fin. El informe debe entregarse en formato PDF. El documento final debe incluir las siguientes secciones:

- Breve introducción describiendo las aristas principales de su propuesta.
- Revisión bibliográfica que incluya al menos 2 trabajos relevantes distintos a los indicados en la sección 2.
- Descripción de la metodología propuesta, incluyendo un diagrama de bloques y una breve descripción de cada bloque.
- Resultados experimentales.
- Descripción de las principales conclusiones de su trabajo.
- Detalle de las referencias en formato paper.

Adicionalmente, deben entregar un jupyter notebook con su código y el material relevante a su trabajo.