INTRODUCTION TO LARGE LANGUAGE MODELS (LLMS)

PROF. TANMOY CHAKRABORTY
Department of Electrical Engineering
IIT Delhi

PROF. SOUMEN CHAKRABORTI
Department of Computer Science & Engineering
IIT Bombay

PRE-REQUISITES : Mandatory: Machine Learning, Python Programming Optional: Deep Learning

INTENDED AUDIENCE : UG and PG students in CSE, EE, ECE, IT, Maths, etc.

INDUSTRY SUPPORT : All those industries whose work involves machine learning, such as Google, Microsoft, Adobe, IBM, Accenture, JP Morgan, Wipro, Flipkart, Amazon, etc.

COURSE OUTLINE :

This course introduces the fundamental concepts underlying Large Language Models (LLMs). It starts with an introduction to the various problems in NLP, and discusses how to approach the problem of language modeling using deep learning. It describes the architectural intricacies of Transformers and the pre-training objectives of the different Transformer-based models. It also discusses the recent advances in LLM research, including LLM alignment, prompting, parameter-efficient adaptation, hallucination, bias and ethical considerations. This course prepares a student to comprehend, critique and approach various research problems on LLMs.

ABOUT INSTRUCTOR :

Professor Tanmoy Chakraborty holds the Rajiv Khemani Young Faculty Chair in AI and an Associate Professor at IIT Delhi. He leads the Laboratory for Computational Social Systems (LCS2), a research group that primarily focuses on building economical, interpretable and faithful language models and applying them specifically to mental health and cyber-informatics. He served as visiting professor at MPI Saarbrucken, TU Munich and TU Darmstadt. Tanmoy did his PhD as a Google PhD Scholar at IIT Kharagpur and postdoc at University of Maryland. Tanmoy has received numerous faculty fellowships, including the Ramanujan, DAAD, Humboldt, ELISE, PECFAR, and several faculty awards from industries like Google, LinkedIn, JP Morgan, IBM, and Adobe. He has authored two textbooks -- "Social Network Analysis" and "Introduction to Large Language Models", on which he has been offering NPTEL courses. He is an ACM Distinguished Speaker. More details may be found at tanmoychak.com.

Prof.Soumen Chakrabarti is a Professor of Computer Science at IIT Bombay. He works on linking unstructured text to knowledge bases and exploiting these links for better search and ranking. Other interests include link formation and influence propagation in social networks, and personalized proximity search in graphs. He has published extensively in WWW, ACL, EMNLP, NeurIPS, ICML, AAAI, IJCAI, SIGKDD, VLDB, SIGIR, ICDE and other conferences. He won the best paper award at WWW 1999. He was coauthor on the best student paper at ECML 2008. His work on keyword search in databases got the 10-year influential paper award at ICDE 2012. He got his PhD from University of California, Berkeley and worked on Clever Web search and Focused Crawling at IBM Almaden Research Center. He has also worked at Carnegie-Mellon University and Google. He received the Bhatnagar Prize in 2014 and the Jagadis Bose Fellowship in 2019.

COURSE PLAN :

Week 1

1. Course Introduction
2. Introduction to NLP (NLP Pipeline, Applications of NLP)

Week 2

1. Introduction to Statistical Language Models
2. Statistical Language Models: Advanced Smoothing and Evaluation

Week 3

1. Introduction to Deep Learning (Perceptron, ANN, Backpropagation, CNN)
2. Introduction to PyTorch

Week 4

1. Word Representation
   a. Word2Vec, fastText
   b. GloVe
2. Tokenization Strategies

Week 5

1. Neural Language Models
   a. CNN, RNN
   b. LSTM, GRU
2. Sequence-to-Sequence Models, Greedy Decoding, Beam search

3. Other Decoding Strategies: Nucleus Sampling, Temperature Sampling, Top-k Sampling
4. Attention in Sequence-to-Sequence Models

Week 6

1. Introduction to Transformers
   a. Self and Multi-Head Attention
   b. Positional Encoding and Layer Normalization
2. Implementation of Transformers using PyTorch

Week 7

1. Pre-Training Strategies: ELMo, BERT (Encoder-only Model)
2. Pre-Training Strategies: Encoder-decoder and Decoder-only Models
3. Introduction to HuggingFace

Week 8

1. Instruction Tuning
2. Prompt-based Learning
3. Advanced Prompting Techniques and Prompt Sensitivity
4. Alignment of Language Models with Human Feedback (RLHF)

Week 9

1. Open-book question answering: The case for retrieving from structured and unstructured sources;retrieval-augmented inference and generation
2. Retrieval augmentation techniques
   a. Key-value memory networks in QA for simple paths in KGs
   b. Early HotPotQA solvers, pointer networks, reading comprehension
   c. REALM, RAG, FiD, Unlimiformer
   d. KGQA (e.g., EmbedKGQA, GrailQA)

Week 10

1. Knowledge graphs (KGs)
   a. Representation, completion
   b. Tasks: Alignment and isomorphism
   c. Distinction between graph neural networks and neural KG inference

Week 11

1. Parameter-efficient Adaptation (Prompt Tuning, Prefix Tuning, LoRA)
2. An Alternate Formulation of Transformers: Residual Stream Perspective
3. Interpretability Techniques

Week 12

1. Overview of recently popular models such as GPT4, Llama 3, Claude 3,Mistral, and Gemini
2. Ethical NLP – Bias and Toxicity
3. Conclusion