8/19: Course Overview & Logistics

- Language Models: A Guide for the Perplexed
- Artificial Intelligence — The Revolution Hasn't Happened Yet | by Michael Jordan | Medium
- The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence | First Monday
- ACL Is Not an AI Conference
- ACL is not an AI Conference (?)
- Natural Language Processing RELIES on Linguistics
- The Ultimate Guide to Word Embeddings
- Neural Networks, Manifolds, and Topology -- colah's blog
- RLHF: Reinforcement Learning from Human Feedback
- Language Processing Pipelines · spaCy Usage Documentation

8/21: Machine learning foundations: Logistic regression

- Logistic Regression (Section 5.2) in Speech and Language Processing (3rd ed. draft)

8/26: Tokenizer; Morphology

- Word and Subword Tokenization (Section 2.5) in Speech and Language Processing (3rd ed. draft)
- Neural Machine Translation of Rare Words with Subword Units (the first paper that applied BPE for tokenization)
- `Let's build the GPT Tokenizer' by Andrej Karpathy (practical tour of BPE with a focus on LLMs)
- Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP (an excellent survey on tokenization)

8/28: Neural networks foundations: Feedforward neural networks

- Logistic Regression (Section 5.3) in Speech and Language Processing (3rd ed. draft)
- Neural Networks (until Section 7.4) in Speech and Language Processing (3rd ed. draft)
- Neural Networks, Manifolds, and Topology -- colah's blog

9/4: Vector Semantics and Embeddings

- Vector Semantics and Embeddings (Sections 6.1, 6.2., 6.4, 6.8+) in Speech and Language Processing (3rd ed. draft)
- [1310.4546] Distributed Representations of Words and Phrases and their Compositionality (the paper that introduced skip-gram)
- https://fasttext.cc/docs/en/unsupervised-tutorial.html (check out this tutorial for how to train your own word embeddings easily)

- Understanding and Creating Word Embeddings | Programming Historian (see how historians can use word embeddings)

9/9: Vector Semantics and Embeddings (Continued)

- Jurafsky and Martin: Speech and Language Processing Section 7.4

9/11: Language Modeling

- Jurafsky and Martin: Speech and Language Processing Sections 3.1-3.5; Section 7.6 (Feedforward Neural Language Modeling); Chapter 8 (RNNs & LSTMs)

9/16: Machine Translation: Seq2Seq

- Jurafsky and Martin: Speech and Language Processing Section 13.2

9/18: Machine Translation: BLEU, Decoding, Attention

- Attention ⇒ Jurafsky and Martin: Speech and Language Processing Section 9.1
- Decoding ⇒ Jurafsky and Martin: Speech and Language Processing Section 10.2
- MT Evaluation ⇒ Jurafsky and Martin: Speech and Language Processing Section 13.6

9/23: Transformer

- Jurafsky and Martin: Speech and Language Processing Chapter 9
- Attention Is All You Need - The paper where the transformer architecture is introduced
- The Illustrated Transformer – Jay Alammar - Helpful illustration
- [Public, Approved] Intro to Transformers - Slides by Lucas Beyer

9/25: Pretraining & Finetuning

- Pretraining with MLM & finetuning encoder-only MLM-pretrained models ⇒ Jurafsky and Martin: Speech and Language Processing Chapter 11
- LM pretraining & finetuning ⇒ Jurafsky and Martin: Speech and Language Processing Section 10.3

9/25 (cont.): Pretraining Data

- Jurafsky and Martin: Speech and Language Processing Section 10.3.2

10/14: Prompting

- Jurafsky and Martin: Speech and Language Processing Chapter 12
- The Prompt Report: A Systematic Survey of Prompting Techniques - Useful survey/overview

10/16: RLHF

- The original work where RLHF is proposed: Deep reinforcement learning from human preferences
- Using RLHF to train a summarization model using human feedback: Learning to summarize from human feedback
- This work is commonly referred to as InstructGPT; the first work to bring RLHF for further adjusting large language models and what lead to ChatGPT: Training language models to follow instructions with human feedback

Additional blogs that can help to understand:

- Illustrating Reinforcement Learning from Human Feedback (RLHF)
- RLHF: Reinforcement Learning from Human Feedback

10/21: Transformer types & Practical considerations

- Decoder-only transformer ⇒ Jurafsky and Martin: Speech and Language Processing Figure 9.15 in Section 9.5
- Classification head / classifier head ⇒ Jurafsky and Martin: Speech and Language Processing Figure 11.9 in Section 11.4.1 (illustrated with encoder-only transformer)
- Encoder-only transformer ⇒ Jurafsky and Martin: Speech and Language Processing Section 11.1
  - "Remember that we said the models of Chapter 9 are sometimes called decoderonly, because they correspond to the decoder part of the encoder-decoder model we will introduce in Chapter 13. By contrast, the masked language models of this chapter are sometimes called encoder-only, because they produce an encoding for each input token but generally aren't used to produce running text by decoding/sampling. That's an important point: masked language models are not used for generation. They are generally instead used for interpretative tasks."
- Encoder-decoder transformer ⇒ Jurafsky and Martin: Speech and Language Processing Figure 13.5 and 13.6 in Section 13.3
- Hyung Won Chung's lecture (recording, slides)

10/23: Weight and Key-value Cache Quantization

- KV cache original paper: Efficiently Scaling Transformer Inference
- State-of-the-art KV cache quantization: KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache

Additional blogs that can help to understand:

- A Gentle Introduction to 8-bit Matrix Multiplication for transformers at scale using Hugging Face Transformers, Accelerate and bitsandbytes
- Unlocking Longer Generation with Key-Value Cache Quantization
- Understanding Quantization for LLMs | by LM Po | Medium

- Transformers KV Caching Explained | by João Lages | Medium

10/28: (Q)LoRA

- Jurafsky and Martin: Speech and Language Processing Section 10.5.3 (PEFT & LoRA)
- LORA original paper: LoRA: Low-Rank Adaptation of Large Language Models
- QLoRA original paper: QLoRA: Efficient Finetuning of Quantized LLMs
- PEFT survey: Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey

Additional blogs that can help to understand:

- Making LLMs even more accessible with bitsandbytes, 4-bit quantization and QLoRA
- QLoRA: Fine-Tuning Large Language Models (LLM's)

10/28 (cont.): QA landscape

- QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension

10/30: Dense retrieval; Answer extraction

- Jurafsky and Martin: Speech and Language Processing Sections 14.1–14.2
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Section 4.2 for answer extraction

Additional blogs that can help to understand:

- How to Build an Open-Domain Question Answering System? | Lil'Log

11/4: Retrieval augmented generation (RAG); Summarization

- ACL 2023 Tutorial: Retrieval-based Language Models and Applications a comprehensive tutorial
-