THE SCIENCE AND ENGINEERING OF LARGE LANGUAGE MODELS

MUIZENBERG
CAPE TOWN
2025

31 March - 11 April 2025

1. A very brief history of LLMs (from word2vec to transformers), tokenization (including hands-on lab), and a detailed overview of the transformer architecture, including coding attention and a full dense transformer from scratch.

2. Training and Optimization (April 2): Pre-training strategies, improving training efficiency on a single GPU, and building training loops to train a transformer.

3. Advanced Architectures (April 3 - 4): Scaling laws, Mixture of Experts (MoE) architecture, with a potential lab on converting a dense transformer to MoE.

4. Specialized Topics (April 5): A module on Diloco and async training.

5. Adaptation and Efficiency (April 7): Adaptors, LoRA (Low-Rank Adaptation) and sparsity (weight and activation quantization).

6. Scaling (April 8 - 11): Scaling laws, roofline models, profiling (MATmuls and transformers on accelerators), data parallelism, and Fully Sharded Data Parallel (FSDP).

7. Post-Training (April 7-11): Instruction tuning and multiple sessions dedicated to advanced post-training techniques.

8. Hands-on Labs: The workshop emphasizes practical application with numerous lab sessions, allowing participants to solidify their understanding and gain hands-on experience.

Date	Time	Activity	Session details	Who?
Monday 31 March 2025	08:30 - 09:00	registration		
	09:00 - 10:30	lecture	A very brief history of LLMs [from word2vec to transformer) part 1	Amr Khalifa
	10:30 - 11:00	break		
	11:00 - 13:00	lecture	A very brief history of LLMs [from word2vec to transformer) part 2	Amr Khalifa
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture + lab	Tokenization	TA: Makki
	15:30 - 16:00	break		
	16:00 - 18:00	lab	Tokenization lab	TA: Makki
Tuesday 1 April 2025	09:00 - 10:30	lecture	The transformer architecture: overview + Attention,	Amr
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Coding attention from scratch	Amr + TAs
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	The transformer architecture: Dense transformers	Amr Khalifa
	15:30 - 16:00	break		
	16:00 - 18:00	lab	Coding a full dense transformer	Amr + TAs
Wednesday 2 April 2025	09:00 - 10:30	lecture	Pretraining and improving effiency of training at 1 GPU	Amr Khalifa
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Building training loops, training their first transformer	Amr + TAs
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	Mixture of Experts architecture	Andrei Rusu / Kelvin
	15:30 - 16:00	break		
	16:00 - 18:00	lab	Changing the dense transformer into MoE?	Muqeeth
Thursday 3 April 2025	09:00 - 10:30	lecture	Scaling Laws	Kelvin Xu
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Optimizers for LLMs and training dynamics	Razvan Ciuca
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	Diloco	Andrei Rusu
	15:30 - 16:00	break		
	16:00 - 18:00	lab		
Friday 4 April 2025	09:00 - 10:30	lecture	Sparsity and Quantization	Utku Evci
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Sparsity and Quantization lab	Raz and Dereck
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	Adaptors / LoRA (Low-rank adaptation)	Utku Evci
	15:30 - 16:00	break		
	16:00 - 18:00	lab	LoRA from scratch ?	Dereck
Saturday 5 April 2025		weekend activities!		
Sunday 6 April 2025		weekend activities!		
Monday 7 April 2025	09:00 - 10:30	lecture	Post-training: Instruction Tuning	Avi Singh
	10:30 - 11:00	break		
	11:00 - 13:00	lab	SFT lab	Jerry
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	Post training: CoT prompting	Avi Singh
	15:30 - 16:00	break		
	16:00 - 18:00	lab	Post training: Instruction tuning with LoRA tutorial	Jerry
Tuesday 8 April 2025	09:00 - 10:30	lecture	Scaling LLMs part 1	Alban Rrustemi
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Roofline models + Profiling [MATmuls]	Amr + TAs (Makki)
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	Post training part 1: intro to alginment/ RLHF	Gheorghe Comanici
	15:30 - 16:00	break		
	16:00 - 18:00	lab	Post training lab	Eltayeb
Wednesday 9 April 2025	09:00 - 10:30	lecture	Scaling LLMs part 2	Alban Rrustemi
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Sharded matmuls	Amr + TAs
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	Post training part 2: RLHF/ DPO, etc	Gheorghe Comanici
	15:30 - 16:00	break		
	16:00 - 18:00	lab	Post training lab	Eltayeb
Thursday 10 April 2025	09:00 - 10:30	lecture	Scaling LLMs part 3	Alban Rrustemi
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Roofline models + Profiling [Transformers on accelerators]	Amr + TAs
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture	Tentative: Serving LLMs	Alban
	15:30 - 16:00	break		
	16:00 - 18:00	lab	Data parellism lab	
Friday 11 April 2025	09:00 - 10:30	lecture	Scaling LLMs part 4	Alban Rrustemi
	10:30 - 11:00	break		
	11:00 - 13:00	lab	Fully Sharded Data Parallel (FSDP)	Amr + TAs
	13:00 - 14:00	lunch		
	14:00 - 15:30	lecture		
	15:30 - 16:00	break		
	16:00 - 18:00	lab		