UCG
Univerzitet Crne Gore
1974

English version below

Naslov: Tehnologije koje omogućavaju (ogromne) velike jezične modele – hoće li nastaviti rasti?

Govornik: PROF. AVI MENDELSON – Računarske nauke, Tehnion

Sažetak

Veliki jezički modeli (VJM), kao što su chatGPT, Bert i Gemini, dominiraju u mnogim aplikacijama koje koristimo svakodnevno, kao što su poštanske usluge, bankarski sistemi, investicijske odluke, odgovaranje na školske zadatke i druge. Da bi podržali takvu kompleksnu raznolikost aplikacija i modela korištenja, VJM moraju upravljati eksponencijalnim rastom broja parametara i sličnim rastom kompleksnosti. Kao rezultat, održavanje takvih modela postaje vrlo skupo, ti sistemi postaju osjetljiviji na sigurnosne napade, a kvalitet rezultata se s vremenom smanjuje.

Moje predavanje će biti podijeljeno u tri dijela; tokom prvog dijela ću ukratko objasniti što je VJM i koje tehnologije se koriste za njegovu podršku. Drugi dio predavanja bit će posvećen ranjivostima koje se nalaze u modelima VJM i tačnosti informacija koje sistem pruža. Posljednji dio će biti posvećen novim izazovima i potencijalnim smjerovima istraživanja.

Kratka biografija

Avi Mendelson je IEEE član i profesor na Odsjeku za računarske nauke na Tehnionu, Izrael. Ima mješavinu industrijskog i akademskog iskustva u nekoliko različitih područja, kao što su arhitektura računara, sigurnost hardvera, hardverski akceleratori i arhitektura za mašinsko učenje, sa naglaskom na arhitekturama koje podržavaju obuku ogromnih modela učenja.

Diplomirao je na odsjeku za računarske nauke na Tehnionu, BSC (1979) i MSC (1982) te je doktorirao (1990) na Univerzitetu Massachusetts u Amherstu (UMASS).

Među njegovim industrijskim poslovima, radio je 11 godina kao viši istraživač i glavni inženjer u Intelu. Među njegovim dostignućima je bio glavni arhitekta za CMP (višejezgreni na čipu) karakteristiku prvih dvojezgrenih procesora koje je Intel razvio. Posljednjih nekoliko godina fokusiran je na aktivnosti vezane za inovacije; npr., upravljao je akademskim aktivnostima za Microsoft R&D Izrael.

Avi Mendelson je bio član Savjeta ACM-Europe (2009-2014) i Odbora guvernera IEEE računarskog društva (2017-2019 kao član i drugi potpredsjednik). U suradnji sa sljedećom ISCA konferencijom, suorganizovao je radionicu o "novom pristupu za obuku velikih jezičnih modela" (https://llm-gnn.org/)

Tit*le:* *Technologies that enable (Huge)Large Language Models – will they continue to scale?*

Speaker**: **Prof. Avi Mendelson – Computer Science Technion

Abstract

Large Language Models (LLMs), such as chatGPT, Bert, and Gemini, dominate many of our daily used applications, such as mail services, banking systems, investment decisions, answering class assignments, and more. To be able to support such a complex variety of applications and usage models, LLMs need to deal with an exponentially growing number of parameters and similar growth in complexity. As a result, maintaining such models is becoming very costly, and such systems are becoming more vulnerable to security attacks, and the quality of the results reduces over time.

My talk will be divided into three parts; during the first part, I will briefly explain what LLM is and what technologies are being used to support it. The second part of the talk will be devoted to the vulnerabilities found in the LLM models and the accuracy of the information the system provides. The last section will be dedicated to new challenges and potential research directions.

Short Bio

Avi mendelson is an IEEE Fellow and a Professor at the Computer Science department at the Technion, Israel. He has a blend of industrial and academic experience in several different areas, such as Computer architecture, Hardware Security, Hardware accelerators, and Architecture for machine learning, emphasizing architectures that support the training of Huge Learning Models.

He graduated from the CS department, Technion, BSC (1979) and MSC (1982) and got his Ph.D. (1990) from the University of Massachusetts at Amherst (UMASS).

Among his industrial jobs, he worked 11 years as a senior researcher and principal engineer at Intel. Among his achievements was being the chief architect of the CMP (multicore-on-chip) feature of the first dual-core processors Intel developed. For the last few years, he has been focusing on innovation-related activities; e.g., he used to manage the academic-related activities for Microsoft R&D Israel.

Avi Mendelson was a member of the ACM-Europe Council (2009-2014) and the IEEE Computer Society Board of Governors (2017-2019 as a member and second VP). In conjunction with the next ISCA conference, he co-organized a workshop on a "new approach for training Large Language Models" (https://llm-gnn.org/)

Naslov: DataFlow SuperComputing za BigData DeepAnalytics

Govornik: VELJKO MILUTINOVIĆ - Profesor Univerziteta u Beogradu i gostujući profesor Univerziteta Crne Gore, vanredni profesor Tehničkog univerziteta u Grazu, Austrija, vanredni profesor Univerziteta Indiana u Bloomingtonu, SAD

Sažetak:

Ovaj mini-kurs na licu mjesta ili online, ili potpuni kurs o DataFlow programiranju, analizira suštinu DataFlow SuperComputinga, definira njegove prednosti i osvjetljava povezani programski model. Naglasak je na pitanjima od interesa za primjene matematike, računarstva, fizike, geonauka i građevinarstva. DataFlow paradigma (koja sama predstavlja primjenu teorije grafova), u poređenju sa ControlFlow paradigmom, nudi: (a) Ubrzanja od najmanje 10x do 100x i ponekad mnogo više (ovisi o algoritamskim karakteristikama najvažnijih petlji i prostorno/temporalnim karakteristikama Big Data toka, itd.), (b) Potencijale za bolju preciznost (ovisi o karakteristikama optimizirajućeg kompajlera i operativnog sistema, itd.), (c) Smanjenje potrošnje energije od najmanje 10x (ovisi o brzini sata i unutrašnjoj arhitekturi, itd.), i (d) Smanjenje veličine za više od 10x (ovisi o implementaciji čipa i tehnologiji pakovanja/hlađenja, itd.). Što su podaci veći, i što je veća mogućnost ponovne upotrebe pojedinačnih podataka (što je tipično za ML), veće su prednosti dataflow paradigme nad control flow paradigmom. Međutim, programerska paradigma je drugačija i mora se savladati. Tešnja istraživanja govornika su snažno utjecala četiri različita dobitnika Nobelove nagrade: (a) od Richarda Feynmana je naučeno da će buduće računarske paradigme biti uspješne samo ako se količina komunikacije podataka minimizira;

(b) od Ilye Prigogine je naučeno da se entropija računarskog sistema može minimizirati ako se prostorni i vremenski podaci dekupliraju; (c) od Daniela Kahnemana je naučeno da sistemski softver treba nuditi opcije vezane za aproksimativno računanje; i (d) od Tima Hunta je naučeno da sistemski softver treba biti sposoban trgovati latenciju za preciznost. Posebna pažnja je posvećena dva dataflow sistema koarhitektirana od strane govornika, jedan prilagođen fizici silicija (Honeycomb Flow za visoko iterativni kod) i drugi prilagođen fizici GaAs (Sistolički niz za Gram Schmidtovu ortogonalizaciju), oba za primjene u složenoj matematici i matematici za računarsku fiziku i geo/atmo/astro/svemirsku fiziku.

Kratka biografija:

Veljko Milutinović je doktorirao na Univerzitetu u Beogradu u Srbiji, bio je na fakultetskim pozicijama u SAD-u (nedavno na Univerzitetu Purdue i Univerzitetu Indiana), u Evropi (nedavno na Tehničkom univerzitetu Beč i Tehničkom univerzitetu u Grazu), i priznat je za DARPA-in prvi GaAs (galijum arsenid) mikroprocesor na 200MHz (oko decenije prije glavnog toka) i DARPA-in prvi GaAs sistolički niz sa 4096 CPU-ova, te za razne inovacije u domenu dataflow paradigme (aritmetičke operacije i mapiranja grafova). Njegovi trenutni akademski istraživački i industrijski razvojni interesi su u ubrzanju dataflow-a kompleksnih matematičkih algoritama koji zahtijevaju nisku potrošnju energije i visoku brzinu, za aplikacije intenzivne matematike u računarskoj fizici i geonaukama. Ima preko 6000 citata na Google Scholar-u i h-indeks 40. Ova prezentacija je ranije održana kao jednosatno predavanje ili kao kurs koji traje cijeli semestar, na relativno velikom broju vodećih univerziteta u SAD-u (MIT, NEU, UMass, Harvard, Michigan, Ohio, Illinois, Wisconsin, Purdue, Indiana, NYU, Columbia, FIU, FAU, itd.).

English version below

Title: DataFlow SuperComputing for BigData DeepAnalytics

Speaker: VELJKO MILUTINOVIĆ - Professor of the University of Belgrade and Visiting Professor of the University of Montenegro, Adjunct Professor of the Technical University of Graz, Austria Adjunct Professor of the University of Indiana in Bloomington, USA

Abstract:

This on-site or on-line mini-course or a full-blown course on DataFlow Programming, analyses the essence of DataFlow SuperComputing, defines its advantages and sheds light on the related

programming model. The stress is on issues of interest for Applications of Mathematics, Computing, Physics, Geo Sciences, and Civil Engineering. The DataFlow paradigm (which itself represents an application of graph theory), compared to the ControlFlow paradigm, offers: (a) Speedups of at least 10x to 100x and sometimes much more (depends on the algorithmic characteristics of the most essential loops and the spatial/temporal characteristics of the Big Data Streem, etc.), (b) Potentials for a better precision (depends on the characteristics of the optimizing compiler and the operating system, etc.), (c) Power reduction of at least 10x (depends on the clock speed and the internal architecture, etc.), and (d) Size reduction of well over 10x (depends on the chip implementation and the packiging/cooling technology, etc.). The bigger the data, and the higher the reusability of individual data items (which is typical of ML), the higher the benefits of the dataflow paradigm over the control flow paradigm. However, the programming paradigm is different, and has to be mastered. The ongoing research of the speaker has been highly influenced by four different Nobel Laureates: (a) from Richard Feynman it has been learned that future computing paradigms will be successful only if the amount of data communications is minimized;
(b) from Ilya Prigogine it has been learned that the entropy of a computing system could be minimized if spatial and temporal data get decoupled; (c) from Daniel Kahneman it has been learned that the system software should offer options related to approximate computing; and (d) from Tim Hunt it has been learned that the system software should be able to trade latency for precision. Special attention is given to two dataflow systems co-architected by the speaker, one tuned to physics of Silicon (Honeycomb Flow for Higlhy Iterative Code) and the other tuned to physics of GaAs (Systolyc Array for Gram Schmidt Orthogonalization), both for applications in complex math, and math for computational physics and geo/atmo/astro/space-physics.

Short bio:

Veljko Milutinovic received PhD from the University of Belgrade in Serbia, has been on faculty positions in the USA (more recently at Purdue University and Indiana University), in Europe (more recently at Technical University of Vienna and Technical University of Graz), and is credited for the DARPAs first GaAs (Gallium Arsendie) microprocessor at 200MHz (about a decade before mainstream) and the DARPAs first GaAs Systolic Array with 4096 CPUs, plus for various innovations in the domain of the dataflow paradigm (arithmetic operations and graph mappings). His current academic research and industrial development interests are in dataflow acceleration of complex mathematical algorithms needing low power and high speed, for math-intensive applications in computational physics and geo sciences. He has over 6000 Google Scholar citations and h=40. This presentation has been delivered before, as a one-hour talk or a full-semester course, at a relatively large number of leading universities of the USA (MIT, NEU, UMass, Harvard, Michigan, Ohio, Illinois, Wisconsin, Purdue, Indiana, NYU, Columbia, FIU, FAU, etc...).