FOUNDATION OF DIGITAL BUSINESS

Surojit Mookherjee

Vinod Gupta School of Management

Indian Institute of Technology Kharagpur

Week 06

Lecture 27

Lecture 27 : Introduction to Gen AI and LLMs

Good morning. Now I move to my module 8, which will be talking about Generative AI, the next productive frontier. So, this will be an overview of what language models are. I will introduce LLMs and how they work, as well as the economic potential of generative AI, but it will not be a technical session. So, we will not get into too much technicality because that is not the purpose of this course. So, the picture I took from Economic Times, dated 21st June 2024, and the topic of the article was developing an app to write a story in 5 to 10 minutes. So, this is a generative AI tool whichever it is I do not know what was used that it says that you

maybe a plot, and it can generate a story in 5 to 10 minutes. So, I can say that a boy is growing up in a village town and then goes to finishes education goes to a join a company in the city out for movies, etcetera, etcetera. Whatever such plots I can give as prompts—a few pointers, so to say. And then ask the tool to write a story, maybe in 500 words, 1000 words, 2000 words, or 200 words—whatever length I can fix for the story.

And then the tool does that it precisely comes up with the story I can change the plot—maybe somebody dies, or they go abroad, or change jobs, or something happens—an accident happens. So, I can keep on changing the plot, so to say. And that tool will keep on changing the story—all within minutes. So, you can imagine how this is going to disrupt—or has the potential to disrupt—the content writing or creative writing business.

So, to say, all of us today can become published authors of stories, novels, poems, and poetries. And now, all of you must be knowing. So, we can also now create videos. So, we can become filmmakers. So, you can write film script and then do the film all with the

help of this tool and sitting at your home and obviously, at very minimal cost and the best part is in minimal time. So, in a days time few hours time you can come up with some finished good finished product.

So, this is going to play a havoc in our today's economy. The current potential for the next productivity frontier which everybody is talking about and hence I am covering it in the course because it is such a discussed topic today and it is going to impact every business, there is no doubt about that. So, this will become an essential tool for any digital transformation any industry will be undertaking. The first thing they will be thinking is Gen AI tools.

Its impact on productivity could add trillions of dollars in value to the global economy. About 75 percent of value that GenAI use cases could deliver falls across primarily four areas, customer operations, marketing and sales, software engineering and R&D. So, in this lecture we will not talk about movie making or writing stories etcetera just to initiate the topic I gave those examples, but we will stick to business applications. Jenna you will have a significant impact across all industry sectors, it has the potential to change the anatomy of work augmenting the capabilities of individual workers by automating some of their individual activities.

So, we will keep go on explaining this each of these in the subsequent slides and lectures. The pace of workforce transformation is likely to accelerate given increases in the potential for technical automation. JNI can substantially increase labor productivity across the economy and the era of JNI is just beginning because it is so easy to use, it is easily available to everybody, it is available on your laptop, it is available on your phone. So, it is very difficult to resist not to use JNI.

So, it can create lot of problems. In the education system for example, it is creating many problems because any assignment we give to the students they can use use general GenAI to generate whatever answers, assignments, or projects, etcetera. Now, how do you stop that? Anyway, so we will keep discussing the problems, but let us talk about the positive things, the good things.

So, what is generative AI to begin with? Next-generation AI technology is derived from deep learning neural networks known as foundation models and uses powerful transformers to produce high-quality content—text, images, or code—from data. GenAI is based on training data that can include text, images, audio, and employs large language

models. So, this is known as multimodal input. So, you have various modes of input like text, images, etc.

And this is, you can say, roughly and simplistically, the basic definition of what GenAI is. It uses deep learning neural networks known as foundation models and powerful transformers to produce high-quality content—be it text, images, or code—from data. And how do these models add value? They make it possible to use AI in a business environment. These models continuously learn from a broad set of unlabeled data.

So, you are not giving labeled input to help summarize, extract, generate, classify, and answer questions. So, all of these are big activities: summarize—if I give a 1000-page document to a GenAI tool, it can summarize it into 2 pages. You can just imagine the savings in productivity is doing just one example 1000 page document I do not have to read and summary it can do it for me in minutes and bring it down to 2 pages of summary. With foundation models AI becomes scalable and can theoretically extend into any domain. These models can minimize a number of steps in data collection by reducing label data requirement giving your team's ability.

to fine tune where needed you can create customized AI models to fit enterprise needs. So, once you have a LLM available with you say chat GPT for example, then you can fine tune that to use it for various applications specific applications because this is a general model foundation model LLM it is there big one it is pre-trained model for you. Now, you use it to fine tune to put it to specific uses. Language understanding starts from the simple yet sophisticated ability to predict the next word in a given sentence.

What effective architecture for this is known as it here we use the transformer. The goal is to create an AI model to receive the input and produce relevant and sensible outputs. I will give an input you give me an output sensible output that is your job. So, whatever input I give corresponding output you have to give. These models have shown abilities that resemble human thought and expression when operated in large scale.

It is almost behaving like a human being. I ask somebody to give me something and he gives. So, what I ask a question, what is the weather today, going to be today and then somebody gives me an answer. The same thing the answer is given by this tool. So, ChatGP if you take it is the full name is Generative Pre-trained Transformer.

So, Generative Pre-trained Transformer. It is an AI chatbot. Basically, ultimately it is a chatbot developed by OpenAI company AI designed to generate human like text

responses to prompts or questions if the keyword is human like. GPD part refers to the type of large language model it uses which is straight and vast amounts of text data to produce convincing language outputs. Almost all models are built on top of few fundamentals.

Some of these are computers understand numbers, hence data text data has to be converted to a number. One of the method used is called tokenization commonly, but we will again will not talk about this because this will become more technical. So, all that you need to do this all text or whatever everything finally has to be converted to a number, otherwise the computers cannot use that. Numbers have hidden statistical or logistical patterns within them that can be figured out mathematically. So, it is not just random gibberish.

and finding these patterns manually is nearly impossible. So, we try to create a neural network architecture to figure these things out this is where that neural network comes in the deep neural network or the artificial neural network. Once that is done you have your AI model to generate the new responses based on everything learned from the training data. So, just if you take a look at this picture. So, these are the various input data multimodal various data types text image speech etcetera which you are using to train your foundation model.

So, all whatever data is available with you more the better obviously. And once you have done this the model is ready then you can use it for following tasks just small list of tasks can centralize the information from all the data from various modalities that is what it has done the pre trained model. And now you are using it this one model can then be adapted to a wide range of downstream task. So, you can fine tune this overall standard model to do work for you question answering, sentiment analysis, information extraction,

image captioning, object recognition, instruction following etcetera. So, this is the core principle you use input to develop one general model then fine tune that model for specific tasks. How to create for model for say generating text? Collect the right kind of data for training usually it will be text, turn all the data into numbers which we have talked about a neural network is basically a assume it is a black box fancy calculator that uses math to produce a result based on the data you give.

and then you train the network. So, which means basically you are adjusting the variables in that math which you are using with various methods. So, it turns learns to give you the correct result for the data it receives and all these various methods etcetera what I am

talking about is the domain of machine learning experts the technical guys. Results are in numbers which we then change back into the original form like words.

I have talked about tokens which really are the basic units of data processed by a model typically representing words, sub words or characters in natural language processing models like transfer models. For example, there is one called Word which is developed by Google and GPT by OpenAI and they are created by breaking down text into smaller pieces using a tokenizer. So, first you break them now into smaller elements and then you convert them using a library to numerical values. tokens are the input units which feed the NLP models.

For example, a sentence I love AI might be tokenized into 3 parts I love and AI or smaller sub words like I, then LO, then hash as V, then A, then hash as I depending on what sort of tokenizer is being used. Each token each of this token is mapped to a numerical ID from a predefined vocabulary or a library which the model is using for processing. Now, it is becoming more technical. The tokens are crucial for handling text data determining context and managing the input size the more the token the more is the input load.

So, maximum token limits in models like GPT are typically 512 or 20,000, 2048 these are just I mean for your reference you can learn what you are interested you can why I giving you this is if you are interested further you can more do more studies on this from the net or from any open source courses. In large language models, the number of token effects the computational cost and memory usage as models process sequences of tokens to predict the next token or the general output. So, if you increase the input load obviously, the computation time load will be more. There is another term called parameters factor.

So, what is the parameters? Parameters are the numerical values of the weights and biases that define a machine to learning models behavior. They are learned during training to minimize the models error. So, to give you a classic example that straight line model of y is equal to mx plus c. So, the bias is the c, the intercept value when your x is 0, y value when x is 0 and m is the slope.

So, that is the weight. So, this is the simplest example of parameters definition to understand what is the weight and what is the bias. In models like linear regression, parameters are coefficients of weights for each feature and the intersect is the bias. In neural networks, parameters include weights of connections between neurons and biases

in layers. You have multiple layers and you have neurons in each of these layers and the neurons are connected to the next layers or other neurons through thing called synapse.

This is just a replica of the how the human brain works. So, these parameters will include the weights of the connections, because when some one calculation is done in any node, there will be a weight factor and that same calculation will transfer to the next node. So, that will introduce the bias. So, there will be a bias and a weight factor for even each of these computation output is happening.

And these parameters can go up to millions or billions for example, GPT-3 has 175 million parameters. more the number of parameters better accuracy will get from the model. So, that they are spending huge amount of money when they are developing their versions by increasing the number of parameters. So, the accuracy of the model becomes better and better and better. Parameters encode the models learned knowledge data etc.

And the parameters are optimized using techniques like then gradient descent and etcetera to reduce the loss function mean squared error which I had explained when I was talking about linear regression in my previous lecture. Then and also cross entropy etcetera for classification. So, if you are interested you can learn about this and get more knowledge on these areas of regression analysis. The key differences are the tokens are data inputs, while parameters are internal model components learned numbers basically their numbers and tokens are text units.

Tokens are processed by the model, parameters define how the model processes them. In practices to say in an NLPA model like GPT takes a sequence of tokens as inputs, processes them through layers defined by parameters and outputs predictions. So, outputs are predictions. The number of parameters impacts the model complexity and capacity while the number of tokens affects the input size and computational cost. So, just to give you an overall idea of how the LLM model is developed because it is being used so much I thought you should know a brief little bit about the how the LLM technology is the base

of the technology and how it is developed and what are used for developing the model. But if you want to have more detailed knowledge feel free you can get into these areas then you will get more and more technical. But of course, if you are interested you can always get into that because this is one tool which is going to be there for some more time and to be extensively used by most of us in various functions. So, wherever we work or wherever we study education, industry, business, government every area this tool is going to be used very much. The other thing is called attention in the context of AI.

It is all about paying attention to what is important. The task is to write a short summary that captures the most of the story. So, how can you make a computer know this? Suppose you want a computer to write a short summary of a big piece of content. So, we need to find a way to give more importance to certain parts of the input.

That is exactly what attention does. This is the work of the transformer model. In technical terms, the attention mechanism calculates the weights determining how much focus to put on each part of the input data. The process enables the model to prioritize the information. So, attention is nothing but introducing a mathematical step before the computational output, and there is something called self-attention.

To give an example to understand, imagine you are in a room full of people talking about different topics. Everybody is talking about something. Now, you are trying to follow a particular conversation about a favorite movie. Some people are talking about a movie that you like, and others are talking about some other movie or some other thing, like politics, etc. So, to do this, you focus more on the people talking about that movie, even though other conversations are happening at the same time.

So, what it means is you are giving priority or giving more weightage to this particular group who is discussing your favorite topic, and you are giving lesser weightage to groups talking about other topics. This is how the transformer model functions. So, this way you can build a great understanding of which conversation is the most important to you in the context of your favorite movie. So, this is similar to what the self-attention does in AI models.

So, anyways I think this gives you a good brief idea of how a LLM model is constructed of and how it was developed and what are the principal things which are involved in a LLM model and some commonly used terms like token, parameters, transformer etcetera attention model. So, this is how and it shows you the large language models, the work they do text generation, classification, knowledge answering, translation, dialogue generation etcetera. And these are some of the models which is being commercially available Blender, Bord and there is a huge list with openAI, Gemini, Bloom, Sphere, LMD etcetera. And then you have things like data center tooling.

By a company called human first you are hosting by hugging face playgrounds and prompt engineering So, these are some specific applications of GenAI models which you can use and customers can use it for particular purpose. So, it is a huge world you can have from internet you can search there are the large number of LLMs, the LLMs to

small language models and so many things are this is one of the most happening area today in probably computer science I can say. Now in my concluding session for this lecture session I will show you tell you something how that bias thing comes into play because this bias

is very important feature or factor of behind this development of these models. So, here using this training data I have shown you the picture about the inputs training Then you have the modular diversity. So, many people the person who is training his or her influence can also influence the bias part and architectures and objectives. What is my objective?

All of these can introduce a part of the training a bias sources. So, when you say bias, the bias has to come from somewhere. So, what are the potential sources? One is of course, the data itself you know the bias data, data has bias. So, you start with a garbage in you get a garbage out etcetera same philosophy.

Then I as an individual might have certain targets when I am developing this model, I want to influence that. So, that is a bias definitely it is a bias like if you take any Chinese GPT tool. And you ask a question about say India or say Arunachal Pradesh, it will give you a different answer altogether. So, it will probably give different names what is the capital of Arunachal Pradesh states instead of Tawang or whatever they will say some other name some Chinese name. Because in the Chinese map they are showing that as may be as a part of extended part of China and then giving names etcetera.

So, that is the bias from the both modular diversity and the objectives. what is your objective. So, you want to put so that any question against China. So, all Chinese GPT will have to be handled in a different way. So, they will give a different answer right or wrong and they are using that to train the tool because the tool can be used by anybody once you open it up anybody can open it use it.

So, if I ask did you know something start from China the virus. It will obviously not say yes or no or whatever; it will give something hallucinated or provide a different answer. So, that is how it has been trained. So, for any questions related to China, you have to be careful, and this is the type of answer that should be expected. So, these form the intrinsic bias of the foundation model.

Now, there could be some adoption bias. Sources per model data mechanism again modulus, because you know that is a general model which you now want to adapt for a

particular task. So, you fine-tune that general model. So, even the fine-tuning process can also insert bias. So, you have data mechanism, data mechanism, and modulus also again similar here, like this can additionally add some bias.

So, what is finally the bias going to do? It is going to affect the user experience; it will have some extrinsic harm. When you are talking about bias, we normally talk about the negative thing. So, it is causing some harm to the user. So, there can be representational bias, performance disparities, abuse, or stereotypes.

So, representational bias means selecting or preferring not preferring, say, like the example I talked about earlier about Amazon hiring not selecting black women candidates. So, there is a bias against black women candidates, and their applications were getting rejected. Similarly, in performance evaluations, you can introduce bias to highlight negative aspects about certain groups or cohorts—whatever you choose to emphasize. Yes, we know stereotyping, and we also understand it. So, these are the types of harms that can result from all these bias factors. This is my reference slide.

So, for all my sessions you will have this common reference slide which gives the from books which I have referred to and certain things use for further information and knowledge. So, with that, I end this session. Thank you very much.