everybody is talking about AI generative Ai and large language models because generative AI is growing massively many people are confused and also curious to know more about generative AI recently there has been a lot of changes when it comes to Opportunities when it comes to career paths when it comes to transitioning to this particular field of data science and AI earlier days companies used to ask for machine learning experience natural language processing or deep learning nowadays companies are also asking for generated Ai and large language models implementations fine tuning and many more but can we directly jump into generative AI there are so many students that have come to me and have asked me sir can we directly jump into generative Ai and um you know create our career in that path because there are a lot of opportunities as a generative AI Specialist or as a generative AI engineer well the answer is unfortunately no you cannot directly jump into generative AI without having prior knowledge on machine learning NLP deep learning and various other Concepts because if you know them then only you can go ahead and do some R&D or some research around generative AI well do you know that generative AI large language models the core concept is coming from the mathematics statistics machine learning NLP and majorly Transformers and of course neural networks but what are Transformers many people are confused about Transformers and to be very honest on the internet right now there are very few videos on Transformers that is giving you a concise understanding and a detailed explanation about Transformers this video is going to solve that problem this video is an end to endend video on Transformers and the self attention technique in Transformers because these are all the basics of generative AI if you understand Transformers next following videos will be around various other topics that will be able to help you in your generative AI Journey just to let you know that this particular video is a part of my data science and AI Masters program that has been recently launched doing really good it's a 140 Plus hours of program that is costed around that is you know priced at around 99 INR but I'm giving away it at 79 INR in case you are interested to know about it details will be in the description please go through it and this video is a part of that Series so if you watch out this video if you know uh if you're able to relate if you're able to see the way of teaching 100% I can guarantee you that you will be able to follow rest of my videos because I try to explain in a very Layman St now the prerequisite for this video is basic understanding on NLP and machine learning but even if you don't have that understanding I will also request you to go through this video understand in case you have any questions let me know in the comment section reach out to me on LinkedIn or Whatsapp if you need my help see you in the video and let's learn [Music] Transformers hi welcome to this video on Transformers so in this video majorly we are going to cover what are Transformers and what happened in the space of machine learning in the space of NLP deep learning and what were the problems that had occurred which led to Transformers nowadays let me tell you all you see like chat GPT or any kind of AI tool or generative AI tool doing Machine level translational activities doing generative AI generating text generating images is generating videos each and every tool is leveraging the concept of Transformers so Transformers are going to be your basic knowledge your base knowledge before you get started with all your knowledge about generative a so what are Transformers now traditionally speaking I assume that you all have your basic understanding on the various deep learning concept such ASN CNN and RNN we we have also covered RNN in details where we talked about lstms grus bidirectional lstms and so on right so we all known is mostly used for traditional tabular data if you have a data set in a tabular format Ann will be able to solve it if you have your data set in in image format or in audio format or in videos format convolutional neural networks will be able to to solve it and when it comes to sequential data we use RNN networks and sequential data includes your textual data especially in the field of NLP right so what are Transformers Transformers are basically neural networks so I'm writing it as NN these are neural networks and their architecture is based on sequence to sequence task so basically they have been created in such a way that they can solve sequence to sequence tasks with ease that is the main concept of Transformers so there were some issues with traditional RNN models which they were not able to solve obviously we will be talking about those problems uh in the coming uh minutes there were certain amount of problems which Transformers is eventually solving them and let me also tell you when transform the official paper of Transformer was launched I think it was uh introduced around 2017 now when it was launched it was launched by Google brain

the paper was launched by Google brain and it was around I I could be minutely wrong uh I'm not sure whether it is 2017 or 16 but yeah I can check it out uh you can also Google it out Google brain transform paper if you search on Google you will get it around 2017 they released a paper and that paper was basically a revolutionary paper which means they even didn't know the Google brain people the researchers even didn't know what Transformers will be doing in 2024 and we all know what's happening in the world right now people are going crazy for chat GPS people are going crazy for various AI tools billions and billions of investment is happening around the space of AI right now even we are in a situation where people are confused and also scared what will happen to their jobs however let me tell you there is no there is no problem uh with respect to AI engineering jobs nobody can replace us at least for the next 5 to 10 years so don't worry about it so coming back to this topic of Transformers as I told you Transformers are are neural network architecture for sequence to sequence tasks which basically means your input is also a sequence and your output is also a sequence now what are the examples of this the examples of this are many you can talk about machine translation you can talk about question and answer uh question and answer systems chat Bots text summarization and many more okay so basically you have sequences in your input and your having sequences in your output sequence of sentences right now on the left hand of the screen you can see a beautiful architecture which looks little bit difficult to understand and grasp but do not worry we'll cover that so this part is basically your encoder and this part is basically your decoder okay so I'm expecting you all know the concepts of encoder and decoder right now apart from that you can see the input embeddings the output embeddings and so on right now to summarize Transformer architecture is nothing but a neural network architecture it is based on the principles of sequence to sequence it is also encoder decoder concept and special thing let me write it down here first of first of all it is a neural network architecture second of all it is a sequence to sequence architecture third is encoder decoder concept sorry encoder decoder concept and the first and the fourth is the most important point which we will have a separate video which is called as self attention now this is based on the principles of attention Okay what is attention we will be explaining you that now let me jump back to the historical part as I told you there was a paper that was launched or introduced in 2017 by Google brains okay you can Google it out you will find it so after this there was a revolution and now we all know how famous Transformers are and Transformer based models are right chat GPT is nothing but a Transformer many of you have been using Char GP might not know what's behind the hoods it is nothing but a Transformer okay so it's a revolution now what is the impact of Transformers let me also show you that the impact of let me use this impact of Transformers now there are multiple impacts of Transformers okay so let's say so let's say and talk about the first impact the first impact is the NLP re revolutionizing that means there were some traditional issues with NLP which are currently being resolved by Transformers so in the space of NLP if you talk about the research first came theistic approaches right all theistic approaches then came all the various machine learning statistical concepts of solving problem then come your RNN then comes your lstms to solve some of the problems that we're having with RNN and finally there are still some problems with lstms where Transformers came into picture okay now the Great example is chat GP right the second impact of Transformers is democratizing ai democratizing ai now what do you mean by that now using pre-train models you can actually build a lot of AI tools so there are already a lot of pre-train models of Transformers available on hugging F and various other platforms you can leverage those and you can build lot of AI products right now if you see any product anything like you pick any use case generation of images generation of text automated resume U resume parsing um some tools for giving you AI uh resume based AI Services um Insurance AI claims AI based claims AI based Bots AI based girlfriends AI based boyfriends whatnot a lot of things happening in the space and that is because of leveraging all the pre-end models by the Transformers the third is multimodel capabilities multi model capabilities okay now what do you mean by that so it means it you it basically works with any kind of data okay so apart from textual data you also can work on voice you can also work on videos you can also work on various other things so it's not just only with texts you can work with a lot of things and that's where generative AI comes into picture and generative AI is not about text people have a confusion generative AI is textual related no textual related is large language models where you only deal with

textual data we have generation of images d e we have generation of videos now we have Sora so it has multimodal capabilities then comes the acceleration of gen AI as I told you because of transformers this concept came into picture generative Ai and lastly we have the unification of deep learning unification of deep learning which means try to understand it's a very important concept before Transformers before all these generative AI Concepts we only had some neural networks right Ann to solve tabular data CNN to solve images data or voice data RNN to solve some uh textual data now we have Transformers to solve any kind of data that is the unification of deep learning so with the power of Transformers you can actually do a lot of things right it's fascinating it's a very important topic there are some complex it is to understand the architecture but I will try to make sure that your understanding is easier so let me get started with one of the previous papers that came in 2014 so there was a paper something like sequence to sequence neural networks and this was by IA I forgot the name wait one second sequence to sequence uh 2014 paper yeah Ilia Su Suk sus K if you just search this thing sequence to sequence neural networks or the search sequence to sequence learning with neural networks paper 2014 you will get some uh arxiv.org platform I can also give you the link of this one in this video you can probably uh type in yourself let me just copy P this so this is the link okay it's not that difficult link so just type it out and you will be redirected to this particular page In Case by the time you are watching this video this paper is not available as I told you just Google it out you will be able to understand okay now sequence to sequence neural networks was introduced in 2014 there was a paper by Ilia basically it was to solve sequence to sequence problems using the encoder decoder technique okay now if you talk about the architecture this is how the architecture looks like so let me draw and then here we have an embedding here we have another embeding and then let's say you your input goes here okay uh obviously input goes from here also okay now let's say my input is very simple my name is s okay now we have lstm here we have lstm here we have lstm here we have LM here and this is basically passing the data and eventually this passes the data to the decoder so we have HT and CT now this is your encoder this is your decoder right inside your decoder also you have some cells which passes the data and you finally have your output here okay so you can have a soft Max layer and then you can have the final output so something like this this was the architecture okay now try to understand this architecture it is very simple to understand the left side is your encoder your right side is your decoder okay now basically data goes from here like this okay now what happens here is so this is a step by step input right that means first goes my the the word my second goes the word name third goes is fourth goes sat and eventually all your data is in your encoder then you are passing the data to the decoder right so basically encoder processes and maintains the hidden State and process the full sentence and creates a context Vector which is CT right so the entire text as the sentence is read then only we are generating CT and then we are passing it to the decoder right that is the thing now decoder again basically goes through step by- step process to eventually do the final output now try to understand this thing this particular architecture works really good for smaller sentences let's say my name is satjit if you want to do a translation yeah it will work because it understands this is one sentence it understand this is an entity how to process the data how to translate it will be able to perform well but for larger data if the Corpus is huge the input is huge it will not be able to solve this problem because it is taking everything and then processing it and then sending it to the decoder so it will not work now the problem of this and and this was was in 2014 so you can remember you can understand that from here till 2017 lot of things happened incrementally and finally in 2017 we got this paper on um paper by Google brains on Transformers which is actually solving all this problem using Transformers you can train like millions and terabytes and gigabytes and petabytes whatever bytes you can call those kind of data and that's how we have chat GPT right so in 2014 this was a very initial idea very good architecture but it had flaws on larger data and then comes another paper I forgot the year that was um in this in which year this paper was live I don't remember uh but you can just Google it out the paper name is neural machine translation by jointly uh um learning to align and translate now this was by uh damitri sorry dri or just one second let me search yeah zimit zimit B okay if you search you will be able to get it so again this paper solved some of these flaws but um it was not the best so anyways let's talk about this paper now this paper was basically uh in which attention concept was used this is the first paper which

basically talked about attention concept which we know right now is one of the most important thing in transformers in Transformers without attention you will not be able to solve any problem attention is the main focus in Transformers which was tossed in this paper now what do you mean by attention so try to understand again this thing so let's say I have an architecture and I have some hidden cells so we have input when we are passing let's say this is my input turn of the lights now this is your H1 this is your H2 this is your H3 this is your H4 and finally you have so this paper basically focused on one concept which is attention what this paper told was do we actually need the entire sentence let's say we are okay I did not talk about the problem statement the the problem statement here is let's say I want to translate English to Hindi or maybe English to any any of your local languages that you are uh you know familiar with I'm just giving an example with respect to Hindi and you can see this paper was all about machine translation right it was using the concepts of sequence to sequence and it is solving some of the sequence to sequence problems with respect to machine translation so basically this paper talked about a concept that while we are solving this problem of converting English to Hindi now if I convert this into Hindi what will be the Hindi turn of the lights uh again I I'm not great with Hindi however Hindi was my major during my school days but I might be grammatically wrong uh e and bu chy I could make a mistake but I I think you are understanding right or or or if in your in your odia let's say I'm I'm basically coming from odisa let's say I'm converting this data into odia so light Bond Coro something similar so or else if you want to convert it into any other language so the core concept here is let's say we are translating this to this do we actually need the entire sentence now for for example turn off basically means Bund right turn off means Bund so to to have Bund if you have two inputs let's say X1 and X2 I think you are solving the problem to solve to find light you just need X4 you don't need X1 X2 X 3 and X4 right so the main concept that was uh you know talked about in this paper was instead of sending the full input context Vector that the way we were sending here we were paring the entire data and then sending the context Vector this paper basically talked about instead of sending the full input context Vector um can we just send some of the puts and still get things sorted out at the decoder level can decoder understand yes decoder can understand right so let's say from here let me try to draw this thing okay I'll I'll okay let me draw this so let's say you have this I'm drawing the decoder okay so basically understand so encoder and decoder are connected right so let's say Here Comes My First second third and fourth and let's say we have 1 2 3 and four so here I'm predicting light here I'm predicting BND here I'm predicting Coro and here there is a end of sentence here it is the beginning of sentence so here is where so basically this line is coming here sorry let me let me draw it properly so what's happening is this line is basically coming here and it is being fed here okay and what's happening is now this is a traditional encoder decoder technique that I'm talking about right so here what's happening is this is your starting and this is your uh light this is your Bund this is your Cur right so the concept was can we use only certain inputs to kind of uh generate the decoder outputs yes we can and that's where the concept of this architecture came into picture so this architecture is something like this so you have your input or maybe you can consider it as your encoder this is your decoder and these are your four inputs right this is your turn of the lights and here we have light Bond C okay and then end so what's happening is so let's say just give me one second so this basically tells you that here you have your inputs right you have the start here you have light here you have Bund and here you have Cur so what it is telling is instead of the traditional approach where we were directly passing this sorry uh maybe we'll do it here instead of the traditional approach which is something like this we have encoder and decoder we were doing something like this right and the very first approach if you remember same right encoder decoder and something like this right so you are passing the HT and you are passing the CT as well right in this particular approach what it is telling is individually you will be passing the context you do not have the context of the entire sentence however you individually you are passing this context so I'm just rejecting this one okay this is just for demonstration this line does does not exist right now C1 C2 C3 C4 maybe C1 depends on uh let's say H1 and H2 let's say C2 is only dependent on H3 and H4 let's say right so you basically do not need the full input sentences to solve this particular problem now here the formula is very simple CI at any any input in your decoder is summarized as I i JH I which basically means this is your attention attention WS okay so attention is something like if you are trying to

predict light that means this H4 is very very important to you because your mind is focused on this word when you're translating this in your head turn off the lights okay lights BND lights when you're are targeting first output lights okay lights is basically coming from my fourth word so think from a human brain point of view when you're trying to translate from the Hindi perspective the output of Hindi perspective each word is actually dependent on a specific input right so that is your attention layer so to predict CI at any any point this is the formula summation of l i j hi okay now in this technique translation quality increased after this paper was released translational quality increased attend but there was still some problems attention based encoder decoder still had issues in sequential learning which basically means when you're are transferring the input through sequential data and when you are dealing with huge data set it is still slower so in large data set we cannot perform forget about fine-tuning fine tuning and pre-tuning is completely a different topic it is very very difficult to process larger data okay now also try to understand this concept let's say I'll randomly take something from the internet just give me one second let's say okay let's say I got some um input from the internet maybe I'll be using that here as a text okay now I just got a text paragraph from our uh captain of the Indian team based on our captain of our Indian team like who led to our 2011 Victory I think every Indian should know about him but anyways that's not the focus the focus here is the entire paragraph try to understand this thing now if I ask you to pause this video and try to create the translation in your head in your local language is this possible I mean from a human point of view it is still difficult right when you are processing Mahendra Singh dhoni commonly known as msoni okay so Mahendra Singh dhoni Jo Ms is an Indian cricketer who is widely regarded okay cricketer Jo one of the greatest so what's happening what's happening in your brain is you are not able to process the entire thing if I ask you you have to process everything at one go and then give a translation not possible even human brains will fail human brains will also fail right so what is the solution the solution here is the solution here is breaking down breaking down the text to chunks okay okay so anyways jumping into our last paper so this was one of the problems jumping into our last paper that is where everything started in 2017 the paper's name was attention is all you need attention is all you need and that was a very catchy title right we all need attention in our day-to-day life also apart from studying apart from neural networks in our family we need attention from our wives from our girlfriend from our kids from our parents everybody needs attention right so it is a very important topic that was tossed by Google brains now try to also understand that whenever we work whenever researchers works on research papers they basically work on incremental strategy that means somebody wrote a paper then we reev revolutionized to another topic toic then we are revolutionizing to another topic right this is the usual flow but this paper did not take any inputs from the previous paper in fact this paper was something basically like a Time time travel like it was written in such a way that somebody from the future 2025 or 2030 basically came to those researchers years told that this is what we need to do in 202 25 20 2030 and they wrote it I mean it is amazing and all the advance advancements that's happening is based on this paper and this is where your architecture comes into picture which is your uh which is your Transformer architecture I will take this into uh on the screen so that it is visible to you now this is where the concept of self attention comes that means your brain when you are processing large data you are basically processing Chun by Chun right Mahindra Singh commonly known as Ms okay mahra sing Ms is an Indian cricketer cricketer who is widely regard so you're taking English sentences Chong by Chong by Chong and you're feeding it to your memory and your memory when you're feeding it that is the self attention you are trying to give to your brains and that is available in this beautiful architecture so let's try to talk about the beautiful architecture and what are the features of this Transformers and in the next video we shall be talking about the self attention mechanism so this has a better architecture sorry so this has a better architecture stable architecture I would say stable architecture with very robust and revolutionizing revolutionizing and these four points are not enough I mean whatever you say how much you want to praise it is less because this is a revolutionizing thing that the world is seeing right so we will be talking about this architecture in details for anyways but this is just an overall idea about transform s in the next video we shall be jumping into the self attention technique hi so let's talk about our next topic which is self attention in the previous video I tossed this topic of self attention right now we will be talking about

this in depth because this is one of the most important topic when it comes to Transformers it is the core concept basically now first let's talk about some NLP applications in NLP applications what are we solving what is the most important requirement what is the most important requirement in an NLP problem can you please pause this video and try to answer this so I know the answer that you might be thinking is see there are different answers that you might be thinking the most important requirement could be data the most important requirement could be uh something else uh the most require uh most important requirement could be the conversions of words to text or the most important could be intent intent identification right but I would say this is sorry words to text my bad words to vectors right words to vectors so in my knowledge this is by far one of the most important requirement right most important problem now why it is the most important problem or the most important requirement because ultimately you will be passing your data to computers right and computers only process numbers right they only know numbers so you need to convert your words to vectors or numbers and then pass it to your computer model right so in the past we have seen how NLP has evolved starting with some of the basic Tech basic techniques let's say one hot en coding let's say we talk about cat rat and Matt let's say 1 0 0 0 1 0 0 0 1 so cat can be represented as 1 0 0 rat can be represented as this and Matt can be represented this right similarly moving ahead we also had some techniques like bag of words right let's say I love sorry my bad I love cats cats love humans so how many words I love cats humans so one 2 2 1 right this is how your bag of words technique works right and similarly we had lot of other techniques but there was a revolutionizing concept called as word embeddings right if you remember word embeddings that was taught in NLP now word embeddings were able to solve some of the problems that the initial issues had right because it was based on the semantic meaning now try to also understand this thing let's say you're training a large Corpus of data right you feed it to your neural networks basically they usually understand the words right understand the words and that is how the advanced word embedding techniques work they understand the words and convert it into a nend dimensional vectors right if you remember we talked about all these Concepts in NLP right so if you if you remember CB or skip GRS or glove or not glove or fast text we know that if you have a large Corpus of data you can convert it into a 300 dimensional Vector if you remember fast text right we were able to convert it into a 300 dimensional vectors so I'm talking about that particular concept so you have a large Corpus of data you feed it to your neural networks understand the words and convert it into a n dimensional data now imagine in this particular scenario I'm talking about a five dimensional data okay so let's say we have King which is let's say 7.19 43 right and we have Queen let's say we have 6.1 9.19 something like that now here imagine your first concept first Vector uh Dimension is basically your let's say how rich they are maybe right and king and queen are usually Rich that's why you see a higher number right now if you try to visualize this into a two dimensional graph because I cannot visualize it on a five dimensional graph so on two Dimension you can see the king and the queen Vector is somewhat similar based on the cosine similarity concept right here comes the concept of cos Theta right and let's say you have another keyword let's say happy or something else let's say um let's say cricket so Cricket could be somewhere here right it is completely different from other words now try to understand the concept of attention so I'm I was just talking about the revolutionized on from different concepts and now we are going to enter the era of Transformers which is basically solving the problem of attention now what do you mean by attention before that let me try to draw a simple diagram so we have let's say an encoder and we have a decoder okay so here let's say we have some Neons and here we have some Neons okay now let's say what goes here is a simple line let's say we will take the same line that we considered in the last session turn of the lights okay now now internally this thing happens and you have this right and let's say here you have light Bund Caro and here you have the end okay now this so let me draw the full box around it so this is the full box around it right now try to understand this is your decoder and this is your encoder right now encoder basically creates a summary of the entire text right so the first step is encoder creates a summary so the entire text summary it creates and then it converts it into let me not use numbers I'll directly um do a arrow Arrow based um architecture or flow flowchart so we have encoder that creates a sentence uh creates a summary then basically it converts it into Vector agree with me yes now as input passes to the decoder let's say you pass the vector

information to the decoder decoder basically works on a stepbystep approach right you can see stepbystep approach and then it basically predicts light whatever it is light Bund C so step by step it is predicting you can see the first output is light then Bund then C now why do we need attention try to understand this thing now when you're talking about huge Corpus let's say in the previous example I was talking about a huge Corpus about Ms D right let me just quickly get exactly the same thing or maybe maybe uh maybe a different text about donon just one second let me use Wikipedia and get those text and we will be using that for our example so let's say I want to take this entire sentence and I will decrease the font so that it fits to our screen and we will be able to understand so try to understand this concept now here what's happening is you are having this big input right now again the same activity try to pause the video and take as much time as you want and try to translate it in your head without without using Google translate or without using any kind of translational tool now if you start doing so it could be a nightmare for you to translate this thing even though you will be able to understand your brain will be able to understand each and everything but translation will be difficult let's try that let's try to pause this video and try it out now when you try doing this uh translation thing right what your brain processes is line by line right Ms dhoni no also known as Mahindra sing donon is one of the India's most iconic and successful cricketers born on July so here we our brain takes a pause okays so Ms MRA sing successful I don't know successful uh prid bid cricketer right so what's happening is you are taking chunk by chunk feeding it to your neural networks your brain your brain is processing it slowly right just think this way if your brain which is a super powerful neural networks is facing this problem why can't a simple encoder decoder technique will not have a problem so this was the problem problem statement with any kind of encoder decoder technique when it comes to Big sentences when you are trying to summarize it and pass it to the decoder there are some techniques like Focus words you cannot have Focus words while you are transferring the data from encoder to decoder because you're transferring the entire Corpus or the entire paragraph right so if human have challenges how machine will work now the concept of the previous paper that we talked about that paper on the machine translation right for prediction of this do we actually need the entire input the question is valid right because we logically need only this for prediction of light we only need lights if we are able to create a model something like this which can understand for each and every prediction which input from encoder it is using then our life is sorted right and that is the problem we are solving using Transformers for Bund what do we need we need turn off so we need basically H1 and H2 for light for light we need H4 we don't need the entire Corpus so the solution was very simple human human being which were not able to transfer the translate the whole paragraph we usually do it chunk by chunk right we also have an attention mechanism on important words so in this case in decoder tal to 1 so this is your decoder tal to 1 right here what's happening is we only need the encoder tal to 4 and so on right so let me redraw the structure uh I think we can use the same structure so that is where the concept of Transformer came okay I will redraw this structure uh okay I I we can reuse it actually we can we can reuse it so let me just uh copy paste this uh it's okay okay we will create new okay uh because we need to create fresh so that it looks clear because there are so many arrows and all it will make you confused so we have this encoder we have this we have this decoder we have all this let me draw and then I will explain so this is a connection and here comes your input and uh let's say here you have H1 and we have outputs here okay so let me draw it in a different uh color so we have light Bund curve and this is end right and here as an input comes starting and this is light this is Bund this is Cur and similarly in your input turn of the lights now try to understand this thing that these are nothing but your H1 H2 H3 and H4 and these are nothing but your S1 S2 S3 and S4 okay now instead of the traditional approach we are introducing C1 which is your attention unit C1 C2 C3 C4 okay so these are your attention units now at time step two in your output layer I'm talking about time step two in the output layer what all inputs do you need so let's say you need time step two is this okay what inputs do you need you need S1 this thing you need yes what else do you need you need so this is your y let's say this is your y0 this is your y1 this is your Y2 and this is your Y3 so this you need right y1 you need and S1 you need right now this is the traditional thing right in your decoder layer you need two inputs which is this and this now I'm just talking about the Trad traditional encoder decoder technique right you are processing the entire input which is turn off the

lights from your encoder you are passing it to your decoder right and in your decoder at time step two you only need S1 that is coming from the previous uh unit which is from this block and you need this y1 right which is light so for now to predict you need this and you need this to predict what to predict Bund and there is a problem because you are taking the entire text or entire input sentence from the encoder to decoder now the question that was coming is do we actually need the entire text the entire Corpus we talked about the problem right the problem is this neural networks will not be able to process large amount of data so we need attention so here the problem changed from this and now using Transformers at T is equal to 2 if you want to expect an output called as Bund you basically need y1 S1 and on top of that you also need let me draw the bracket you also need something from the encoder what do you need you need basically let's say H4 or maybe combination of combination of H okay let's say for light you need combination of H right you need H4 for Bund you need combination of H1 and H2 I'm just giving an example so H1 and H2 that is what you need right and that's where this concept of C1 C2 C3 C4 comes in C1 C2 C3 3 C4 are nothing but your attention units or you can call call it as attention weights now in your vanilla encoder decoder technique at time is equals to I what happens is you basically take y of i-1 and S of I -1 and in this now in this architecture at time is equals to I you are basically taking I of i - one sorry y of i- One S of i- one and C of I and this is nothing but your attention input okay now what is CI just think about it what is CI CI is a vector scalar or a matrix so think about it what is CI because CI is nothing but coming from multiple hidden states of your encoder right what is the purpose of CI the purpose of CI is to identify at decoder tals to I which CI is important what is the dimension so try to understand the dimension of CI is exactly same like hi if hi is three dimensions CI is also three dimensions if hi is four dimension CI is four dimensions okay so I I hope this is clear right now I will move on to little bit more complexity so the formula at CI is nothing but denoted as alpha 1 H1 plus Alpha 2 H2 plus Alpha 3 H3 plus Alpha 4 H4 now what is alpha alpha is nothing but the inputs from these are alpha 1 Alpha 2 Alpha 3 Alpha 4 similarly we also have multiple Alphas here now if you want to draw this just take it down and try to redraw this because it is already complex so try to understand Alpha is the connection from your encoder to the decoder layer so let me draw using black lines okay so let's say black line is even confusing okay similarly for second one you also have this right similarly for third one you have this for fourth one you have this that means all of these are connected right so to reduce the confusion instead of this complex thing which which will eventually have how many Alphas Alphas will be 4 cross 4 right because four in the input or in the decoder and four in the encoder so it will become 16 Alphas right so let me undo it so that we will reduce the number of alpha lines and to reduce the confusion so how many Alphas do we have four in the encoder and four in the decoder so total 16 Alphas right I'm just talking about this one alpha 1 Alpha 2 Alpha 3 Alpha 4 and that is what it is C1 is nothing but alpha 1 of H1 Alpha 2 of H2 Alpha 3 of H3 and Alpha 4 of H4 simple right similarly for C2 let's say if you want to generalize it you can call it as alpha 1 one alpha one2 alpha 1 3 Alpha 1 14 so C2 becomes Alpha 21 H1 Alpha 22 H2 Alpha 23 H3 Alpha 24 H4 I hope you're able to get right there are 16 Alphas alpha 1 1 to Alpha 14 for my first encoder to all the decoders alpha 21 to Alpha 24 from my second encoder to all the decoders alpha 312 Alpha 34 Alpha 412 Alpha 44 right that is what it it is so the generalized formula is my C of I is nothing but summation of alpha i j h of J now consider encoder has J units and uh decoder has I units Alpha how many Alpha we will have I cross J right in this case luckily I and J is same that's why it is 4 cross 4 right this is the generalized formula okay now let's say you are considering Alpha 21 what is Alpha 21 Alpha 21 is basically here with respect to this one right so for BND so Alpha 2 one wait wait wait yeah yeah basically yeah I I also got confused so Alpha 21's output is BND right for Alpha so for for this particular cell in your decoder what is your input in the decoder the input is this right Alpha 21 is basically what is coming from this keyword called as turn right that is what we need to understand what is the important what is the role right and that is all about your attention let's say you talk about Alpha 21 Alpha 21 is nothing but combination of H1 + S1 right what is S1 S1 is this thing and what is H1 H1 is this part right so try to understand this can also be written as as this is a combination of H1 and H one it can also be written as function of H1 and S1 now what do you mean by function in neural networks your function could be anything a very simple function is nothing but a Ann architecture or a Ann model right so what is Ann we all know we have inputs we

have some hidden layers and we have output right so What's Happening Here What's Happening Here is we have bunch of neurons so let's say I'm just using a magnifying glass and I'm just trying to expand this what happens is you have a neural network what is going as an input is from the decoder we have the S1 from the encoder we have the the H1 and the output is nothing but Alpha 21 now I think you will be able to understand what is Alpha 21 right Alpha 21 I will repeat is for my second object in decoder which is my this block my this block I'm talking about this part only that is my Alpha 21 right so how does my Alpha 21 gets calculated it is calculated on based on this the input of X S1 which is this right this and the output here this they are combined goes through NN architecture and then Alpha 2 one is calculated and we all know Ann is nothing but a beautiful neural networks which has a concept of back propagation and we can adjust this right that is basically the attention we are talking about mathematically so that's all about Transformers and the self attention technique in Transformers I know there were prerequisites of some neural network techniques or some NLP techniques in order to understand Transformers but uh trust me uh yes we are studying Transformers there are some pre and there are some post videos that will basically help you to get more understanding on Transformers if you are a part of of the data science and AI Masters program definitely you'll be able to relate what I'm talking about in case you're interested in my program let me know or else reach out to me on Whatsapp and we can discuss we can check your resume we can check your background and then only I can guide you whether this is the right program for you or not in case you have any other questions let me know that's all about it see you in the next video [Music]