1. AI as we know it today is actually quite dumb yes this includes chat GPT stable diffusion Sora and all the other state-of-the-art models that we have right now they're still very incapable and inefficient and the future generation of AI will look very different from what we have now here I'm going to explain why the current generation is so limited and what the future generation of AI will look like first we need to understand the mean mechanics of AI as we know today all AI is based on the neuron
2. Network which is designed based on the human brain this is basically a network of nodes in which information flows through from one end to the other now this is going to be a very simplified explanation of how a neural network works I'm explaining this for people without a technical background in AI so if you do have experience in AI feel free to skip this section each do in a neural network is called a node or neuron and each line of nodes is called a layer you might have heard of the term deep learning or deep neuron networks
3. this is basically a neuron network with many layers hence it is very deep each node determines how much information flows through to the next layer now again this is an oversimplification there are a lot of settings like weights and biases and activation functions but basically just think of this neuron Network as a series of dials and knobs which determine how much information flows through to the next layer here's a simple example let's say we have this neuron Network which is designed to determine whether an image is a cat or a
4. dog for its input we would feed it an image of a cat or a dog and this image would be broken down into Data also known as tokens which are then fed through this neuro Network eventually after the data flows through all these layers it reaches the end layer which would conclude whether the image is a cat or a dog now what about training a model how does that work well a neural network needs to undergo usually millions of rounds of training to learn how to do something here's an example of how one round of training would look
5. like let's say you input an image of a dog and then this image would be broken down into data which flows through this neuron Network and it spits out the answer this is a dog well in that case since it got the answer correct it's likely that these dials and knobs which we can also refer to as weights are set correctly if it gets the answer right well we don't really need to tweak these weights further however what if it gets it wrong what if it says that this is a cat well in that case it would incur a penalty and this penalty would cause the
6. weights in this neuron Network to be updated so that this penalty would be minimized in the future specifically the weights would be updated from the last layer to the next layer back to the next layer back in a process which is called back propagation all the way until it reaches the first layer of nodes and usually one round of training isn't good enough so the network would undergo millions of rounds of training where the weights would be slightly tweaked to minimize the penalty incurred from any errors and this goes on and on until
7. finally we reach the configuration of dials and knobs so that this neuron Network can very accurately determine whether any image is a cat or a dog and this is how AI models that we know today are trained as well so for example GPT is basically a neuron network but these dials and knobs are optimized for understanding natural language stable diffusion is another neuro Network where the dials and kns are optimized for image generation now again this is very much an oversimplification and the architecture or basically the design of the neuron
8. network is also very important for example how many layers should we have how many nodes in each layer should we have there are also many different architectures such as the Transformer model for large language models or lstm for time series data or convolutional neuron networks for object detection and image classification but in a nutshell the backbone behind all these AI models is just a neural network which has a preconfigured set of dials and knobs to do the job accurately so now that you understand how the
9. current generation of AI Works let's look at the biggest limitations of this first of all once the model is finished training the weights or basically these dials and knobs are fixed in value when the user asks chat GPT something or when the user uses stable Fusion to generate an image these dials and knobs do not change in value in other words all the AI models that we have today are fixed think of this as a brain that cannot learn or get any smarter for example GPT 4 cannot continue learning and become smarter and smarter with time if we want
10. a smarter model well we need to train a new generation of GPT such as GPT 40 or GPT 5 or whatever you want to call it same with stable diffusion for example stable diffusion 2 cannot get smarter and generate better images as we use it more and more in order for it to improve we currently need to train a new generation also known as stable diffusion 3 and once stable diffusion 3 is finished training well that's as smart as it gets and if you don't think it's good enough well you need to train a new model so basically all the AI
11. models that we have today are fixed in their intelligence and their capab abilities again think of this as a brain that has stopped growing and cannot learn or get smarter but this is not how the human brain works there's a term called neuroplasticity which refers to how the brain can reorganize or reconfigure Itself by forming new neural connections over time in order to adapt to new environments or learn new things and that's exactly what the next generation of AI can do which we'll talk about in a second but there's another
12. huge limitation of current AI models they are extremely inefficient and computes intensive as you may know AI is designed based on the architecture of the human brain so let's compare it to the efficiency of the human brain right now gpt3 has 175 billion parameters this was trained using thousands of gpus over several weeks or several months the total power required for training gpt3 was estimated to be around 1,287 megawatt hours of electricity this is roughly equivalent to the monthly electricity consumption of 1,500 homes
13. in the USA now keep in mind gpt3 was completed in 2020 that's 4 years ago the latest version GPT 4 is closed source so we don't actually know its architecture or how long it took to train but we do know that it has around around 1.76 trillion parameters 10 times more than GPT 3 keep in mind that the amount of computations required scales exponentially as the parameter size gets larger so from a rough calculation GPT 4 could have taken around 41,000 megawatt hourss of energy to train that's enough energy to power
14. around 47,000 homes in the US for a month the compute used to create create these state-of-the-art models that we know today such as GPT 4 or clae 3 or Gemini 1.5 Pro requires massive data centers and a lot of energy that's why Tech Giants are scrambling to invest and build even bigger data centers because they know that compute is the main limitation here and that's exactly why Microsoft and open AI are planning a $ 100 billion Stargate project to build the biggest data center in the world all of this is for more compute now contrast
15. this to the human brain some might say the human brain is still more intelligent than GPT 4 at least in some regards the human brain only uses 175 kilowatt hours in an entire year and it gets this energy in the form of calories from the food we eat so training GPT 4 is estimated to require approximately 234,000 times more energy than what the human brain uses in an entire year in other words the energy required to train GPT 4 Once could power the human brain for over 234,000 years now I gave this comparison to show you that there's something
16. fundamentally wrong with AI models today they are very energy inefficient and they take up a lot of compute it's not even close to the efficiency of the human brain so the next generation of AI has to solve this efficiency problem as well otherwise it will not be sustainable so to summarize the major limitations of current AI models is number one they are fixed and unable to improve or learn further after being trained and number two they're also very energy intensive and inefficient these are the two biggest problems of the
17. current generation of AI now let's enter the Next Generation we aren't there yet but there are a few possible architectures that are being discussed and developed as we speak the first architecture is called liquid neural networks Now liquid neural networks are designed to mimic the flexibility or the plasticity of the human brain the human brain is very flexible and can reorganize or reconfigure itself over time and this ability allows the brain to adapt to new situations or learn new skills or compensate for injury and
18. disease for example when you learn something new your brain changes structurally and functionally to accommodate the new information learning a new language can lead to changes in the brain structure and function such as increased density of gray matter in the left hemisphere the brain can also reconfigure itself to recover from injury for example after a traumatic brain injury physical therapy and cognitive exercises can help rewire the brain to regain lost functions and for people who've lost a sense like sight or
19. hearing the brain will reorganize itself to compensate for the loss and make other senses become more acute so this flexibility this plasticity is exactly what liquid neuron networks are designed to have liquid neuron networks can adapt in real time to new data this means that the configuration of the neuron Network can change as it receives new inputs and that's why it's called liquid these Connections in the network and these dials and knobs are fluid so they can change dynamically over time liquid neuron networks also retain what they
20. have learned while incorporating new information this is similar to how our brains can remember old information while learning new things so here's how liquid neuro networks work they have three main components much like a traditional neuron Network it has an input layer which receives the input data but then in the middle we have this liquid layer otherwise known as a res res this is the core component of a liquid neuron Network and it's basically a large recurrent neuron Network think of this as a big bowl of water in which
21. each Splash creates a ripple these ripples are basically the neurons in this network reacting to inputs the reservoir acts as a dynamic system that transforms the input data into a high dimensional representation called Reservoir States and this reservoirs Rich Dynamics and Transformations capture the complex temporal patterns in the input data and then finally we have the output layer this layer receives the reservoir States and Maps them to the desired output using what is called a readout function in layman terms this is
22. a layer that looks at the ripples in the reservoir and tries to understand what it all means it takes the dynamic patterns from the reservoir and makes predictions or decisions from it the key aspect of liquid neural networks is this Reservoir layer which remains untrained during the entire learning process only the output layer is trained to map the reservoir states to the Target outputs in other words to understand what these ripples mean and because this Reservoir remains fluid and flexible throughout time it's not fixed in value that allows
23. this liquid Neer Network to basically adapt to new data and learn new things here's how you would train a liquid neural network the connections between neurons in their reservoirs are set up randomly at the start these connections typically stay the same and don't change during training next you would feed the input layer some data and when this data is broken down into tokens and it reaches the reservoir layer it causes the neurons in the reservoir to react and create complex patterns much like ripples in water so as this input data
24. creates ripples you basically observe and analyze the patterns created in the reservoir over time and that's exactly what the readout layer does it learns to recognize these patterns it's like learning ahuh this is what caused this type of Ripple and that is what caused this other type of Ripple and eventually after lots and lots of rounds of training the readout layer can make accurate predictions based on observed patterns again note that only the readout layer is trained which is simpler and faster because you're not
25. adjusting anything in the reservoir layer this is much quicker and needs less compute compared to traditional neuron networks that's because in neuron networks that we know today all the weights including those in the hidden layers are trainable this means more parameters to optimize leading to longer training times and higher computational requirements but in liquid neuron networks you don't adjust the weights of the reservoir during training only the readout layer is trained and this significantly reduces the computational
26. burden during trainings since fewer parameters need to be optimized plus it's a lot faster to train thanks to our sponsor bright data bright data is an all-in-one platform designed to help businesses collect high quality web data at scale this is especially useful for AI companies which require huge amounts of diverse and high quality training data to build robust and unbiased AI models collecting this training data manually can be timec consuming and prone to errors and that's where bright data comes in with bright data you can
27. access high volume high quality web data effortlessly from parsed validated data sets to custom scraping Solutions they've got you covered get parsed and clean data sets ready to use on demand customize any data set to fit your specific needs and benefit from reliable delivery and full compliance in fact every 15 minutes their customers scrape an enough data to train chat GPT from scratch that's a lot of data to say the least they have many tools like the web scraper API the proxy manager and unblocking Technologies to help automate
28. your data scraping at scale allowing you to build reliable data sets to train any AI or llm visit the link in the description below to learn more it's a lot faster for these liquid neuron networks to converge at an Optimum and because of this Reservoir where the weights and configurations can change dynamically depending on the data that you feed it liquid neuron networks can potentially be much smaller than traditional neuron networks which have fixed weights and connections and this offers a lot more efficient learning and
29. inference so for example researchers at MIT were able to Pilot a drone using a liquid neuron network with only 20,000 parameters which is very tiny compared to state-of-the-art AI models such as GPT 4 which often have over a trillion parameters just think about that 20,000 parameters versus over a trillion parameters so these smaller sizes generally translate to faster inference and lower computational requirements liquid neuron networks are also way less memory intensive again since you don't train the reservoir weights memory usage
30. is much lower during training compared to traditional neuron networks where the gradients and the parameters for all layers must be stored in memory liquid neur networks are particularly good at processing temporal data due to their Dynamic Reservoir so they excel in tasks that involve complex time series data now you might be wondering well how can these liquid neuron networks actually be applied in the real world so here are some use cases as we race to build fully autonomous AI robots these robots will be deployed in the real world and often
31. times they might encounter situations that they 've never seen before during training for example there could be unpredictable environments in search and rescue missions but with liquid neuron networks these robots can adapt to changing conditions and learn new tasks on the Fly and eventually we're going to have these autonomous robots in our houses helping us do chores and other tasks but maybe you have a certain way of folding clothes or doing the laundry or cooking that the robot was never trained on so with a traditional neuron
32. Network these robots aren't able to learn new skills after being deployed but with liquid neuron networks built into a humanoid robot it can learn new tasks that you teach it and this robot will become a lot more personalized for you and then we have autonomous driving there's no doubt that self-driving cars will eventually become the future but current Technologies still do not perform well especially in challenging environments or new conditions again this is because traditional neuron networks can only do well on data that
33. they were trained on they're not able to adapt to new environments but with liquid neuron networks autonomous vehicles can navigate complex and dynamic environments by continuously learning and training from sensor data and adjusting their behavior accordingly it's constantly training and improving over time now as I've mentioned before liquid neuron networks often incorporate recurrent connections making them suitable for processing time series data so it's great for things like weather prediction and of course stock trading
34. the stock market is filled with Ever Changing Trends and Cycles so it's close to impossible for one fixed algorithm or formula to beat the market however because liquid neural networks can adapt to everchanging data it can optimize trading strategies in real time to maximize profits in other words you could be constantly streaming the latest Market data to this liquid neuron Network which would change its configuration to adapt to this data in real time to help you maximize profits another use case would be Healthcare
35. liquid neuron networks can be used in wearable devices to monitor patients in real time adapting to changes in the patients's conditions and predicting potential health issues before they become critical in cyber security liquid neuron networks can continuously learn from Network traffic and user Behavior to adapt Access Control policies and detect anomalies or unauthorized access access attempts yet another use case would be streaming services such as Netflix they can use Liquid neuron networks to adapt to each user's viewing
36. habits and preferences providing more personalized content recommendations another use case would be smart City management for example liquid neuron networks can optimize traffic flow in real time by learning from traffic patterns and changing traffic lights accordingly to reduce congestion and improve efficiency energy management is also very relevant smart grids can use Liquid neuron networks to Balance power supply and demand in real time improving efficiency and reducing costs by adapting to consumption patterns however
37. although liquid neuron networks seem promising it does have its limitations this is still a relatively New Concept in the field of neuron networks and research on them is still in its early stages compared to more traditional architectures while liquid neuron networks show promising theoretical benefits such as the ability to process continuous data streams and adapt on the Fly there is still a lack of real world results demonstrating their superiority on a large scale many researchers are likely waiting for more compelling
38. Benchmark results before investing significant effort into liquid neuron networks also as I mentioned previously they're particularly suited for temporal or sequence data so for for tasks that do not involve time such as identifying images of cats or dogs traditional neuron networks might actually be more effective and straightforward to implement also the Dynamics within this Reservoir layer can be very complex and difficult to interpret and this makes it challenging to understand how the reservoir processes these inputs it
39. would be quite hard to fine-tune it for Optimal Performance finally there is a lack of standardized support and fewer established Frameworks for four liquid neuron networks compared to traditional neural networks and this can make implementation and experimentation more challenging so all in all liquid neuron networks are still a very early concept and an area of active research unlike traditional neuron networks that are fixed and need to be retrained with a large data set to learn new information liquid neuron networks can update their
40. knowledge incrementally with each new piece of data this offers a flexible and adaptive model which could potentially become infinitely smarter over time now liquid neuron networks aren't the only possibility that could become the next generation of AI we have another type of neuron Network which is designed to mimic the human brain even more than traditional neural networks and this brings us to spiking neuron networks these are closely inspired by the way neurons in our brains communicate using discrete spikes or action potentials you
41. see in the human brain which is basically a network of neurons each neuron doesn't immediately fire to the next set of neurons when it receives input instead the input has to build up to a certain threshold and once it passes this threshold then it fires to the next set of neurons and after it fires it goes back to its resting state well spiking neuron networks are designed to mimic this Behavior so here's how it works the architecture is quite similar to traditional neuron networks however for each neuron it
42. waits to receive signals or spikes from other neurons think of these spikes as like little electric pulses the input data such as an image or a sound is turned into these spikes that move through this neural network for example if it's a loud sound it might generate more spikes while a quiet sound might generate fewer spikes now each neuron in the network collects incoming spikes imagine a bucket collecting drops of water as more spikes come in the bucket fills up and when the neuron gets enough spikes in other words when it reaches a
43. certain threshold it fires a spike to the next set of neurons and after firing it resets and starts collecting again from zero so instead of using continuous signals like traditional neuron networks spiking neuron networks uses spikes which are basically bursts of activity at discrete time points to process information in other words spiking neuron networks incorporate time into their processing with neurons firing only when their potential exceeds a certain threshold now there are different methods and algorithms to train a spiking neural
44. network and there currently isn't a standard way that's set in stone so this is still an active field of research one common method is called Spike timing dependent plasticity or stdp this method is inspired by how the brain strengthens or weakens connections between neurons so if one neuron spikes just before another then the connection between them gets stronger if it spikes just after then the connection gets weaker it's like learning which connections are important based on the timing of the spikes and speaking of timing it's the
45. exact timing of spikes that matters it's not just about how many spikes there are but when they happen now stdp is only one method to tr TR the spiking neuron networks there are a few other ones which are beyond the scope of this conversation but like traditional neuron networks spiking neuron networks have to undergo millions of rounds of training with a lot of data and eventually the configuration of the network and all its parameters will reach an Optimum State now again I'd like to remind you that this is a very simplified explanation of
46. spiking neuron networks and I've left out a lot of mathematical details but in a nutshell that's how it works so you might be wondering well what are the benefits of spiking neural networks first of all it's designed to mimic the human brain even more by implementing this spiking mechanism so in theory maybe we could reach a superior level of intelligence compared to the current generation of AI if we Mimi the human brain even more but the biggest benefit of spiking neuron networks is their efficiency if you remember at the
47. In the beginning we compared the energy consumption of the human brain versus a current state-of-the-art model like GPT 4 which requires huge data centers and huge amounts of compute that's because traditional neuron networks are always active each input of data activates the entire neural network so you have to do an insane amount of Matrix multiplications across the entire network just to do one round of training or inference however for spiking neural networks they only use energy where spikes occur while the rest of the
48. neuron Network remains inactive this makes it a lot more energy efficient plus spiking neuron networks are particularly suitable for neuromorphic chips which are designed to mimic the human brain now neuromorphic chips are a huge topic and deserves its own full so let me know in the comments if you'd like me to make a video on this as well so how can these spiking neuron Networks actually be applied to the real world well because these neuron networks can encode and process information in the timing of spikes this is great for
49. processing temporal data this makes them great for adaptive and autonomous systems plus this Spike timing dependent plasticity which I mentioned before where the timing of the spikes influences the strength of the connections in the network this can lead to more robust and adaptative learning capability so this Dynamic learning can make spiking neuron networks suitable for autonomous systems such as self-driving where the AI has to learn and adapt to changing environments or it can be used in Realtime processing like
50. predicting the stock market or patient monitoring and personalized medicine and of course autonomous robots now although spiking neuron networks offer some huge benefits especially regarding Energy Efficiency they do have some limitations setting up and programming spiking neuron networks is more complicated compared to traditional neuron networks this spiking behavior of course adds a layer of complexity making them harder to design and understand training spiking neur networks is also quite difficult current neuron networks use
51. methods like back propagation to adjust their parameters but this process doesn't work well with these discrete time-based spikes researchers are still trying to find an effective training algorithm for spiking neuron networks also given this additional dimension of time spiking neuron networks might actually require more computational resources to simulate this is because they need to track and process spikes over time which can be computationally expensive yet another limitation is that running spiking neuron networks
52. efficiently often requires specialized Hardware such as neuromorphic chips which are not widely available or standardized compared to Conventional Computing Hardware neuromorphic chips are optimized for this Spike based processing and are still being developed and that's why for example Sam Alman is investing millions of dollars into a neuromorphic chip company called rain finally while spiking neuron networks show promising results especially for time-based data they often lag behind current neuron networks for non-time
53. based data they often underperform compared with current AI models particularly for complex tasks this is partly due to the challenges in training spiking neuron networks effectively and as with liquid neuron networks spiking neuron networks are also relatively new so there are fewer tools and Frameworks available for developing spiking neuron networks compared to current AI models this makes experimentation and development slower and more difficult but anyways that sums up what could potentially be the next generation of AI
54. to bring it all back the current generation of AI is very energy inefficient requiring huge amounts of compute plus it can't learn new things after being trained if we want to achieve AGI or ASI we need to essentially create something as efficient and as fluid as the human brain which can constantly learn new things and adapt to changing environments these are the two essential things that new types of neuron networks such as liquid neuron networks and spiking neuron networks can solve at least in theory however these are still
55. relatively new and they are still being developed but the potential could be massive imagine an AI that can keep learning and get infinitely smarter let me know what you think about these neuron networks in the comments below things are happening so fast in the world of AI it's quite hard to keep up with all the technological innovations that are happening right now so if I've missed any other groundbreaking architectures that are worth mentioning please let me know in the comments below and I'll try to do a video on that as