* The tech industry's obsession with AI is beginning to hit its first physical limitation power consumption both training and using AI models is proving to be one of the most energy intensive collection of computational processes that is consumed by the public at large in fact it's estimated that a single gp4 textual request consumes around 30 00 wat hours of energy or about the amount required to charge 60 iPhones this is roughly 1,000 times more energy than what is needed to perform a
* traditional non- AI based request by Google search one study at the Amsterdam School of Business and economics predicts that by 2027 at its current trajectory Global AI processing will consume as much energy as Sweden at around 131 gwatt hours per year when the energy consumption of the human brain is examined it becomes clear that the current approach of AI is unsustainable and grossly inefficient during intense mental activity a brain consumes just one qu of a Food calorie per minute this equates to about 17 hours of intense
* thought for about the same energy consumption of a basic GPT for request this stark contrast between the Energy Efficiency of biological neuros systems and Kent AI models has created a new race to develop the next generation of AI one that more closely mimics our biology the intense power consumption of AI is a direct result of the underlying models of artificial neural networks that the vast majority of AI systems are composed of artificial neural networks Loosely emulate biological systems in their approach to problem solving using
* a complex statistical model to best approximate a solution the basic structure of an artificial neural network consists of a network of interconnected nodes or artificial neurons structured into organized layers data is fed into the network through an input layer each neuron in the input layer represents an input feature or variable the complexity of this input layer depends on the dimensionality of the input data and its complexity and Fidelity the functionality of an artificial neural network comes from the
* inter interaction of information within the core of the network known as the hidden layers hidden layers are situated between the input and output layers and can be structured into a wide variety of configurations depending on the complexity of the task the network is designed for two common architectures for example are convolutional neural networks which are designed for processing grid-like data such as images and recurrent neural networks which are structured for processing time-based sequential data the choice of neural architecture and
* hyperparameters such as the number of layers number of neurons per layer and activation functions depend on the complexity of the task the available training data and the desired performance within each hidden layer are multiple neurons that process and transform the data received from the previous layer this occurs through the application of an activation function to the weighted sum of the inputs while many types of activation functions exist the three most commonly used are sigmoid tan and the highly favored rectified
* linear unit function the hidden layers interfac to an output layer which produces the final predictions or outcomes of the network the number of neurons in the output layer depends on the task the network is designed for and the desired output format neurons and adjacent layers of the network are connected with an Associated weight that determines the strength and importance of the signal passing through it during training the weights are adjusted to minimize the difference between the predicted outputs and the actual targets
* additionally each neuron in the hidden and output layers also has a bias term associated with it the bias term acts as an additional input to the neuron and helps to shift the activation function providing flexibility in the Network's learning process in effect weights and biases store information within the network and create its functionality the flow of information in an artificial neural network typically flows forward from the input layer through the hidden layers to the output layer in order for the weights and biases to take on values
* that perform the intended task a neural network must be trained for the majority of artificial neural network applications a labeled data set is used to train a network in this process a known piece of training data is fed into the network and its output compared to the training data a cost function is used to measure any difference indicating how accurately the network model performs from this cost function a technique known as back propagation is used to propagate this error backwards from the output to the input layer this
* error is used to calculate the gradient of the cost function with respect to each weight and bias it helps determine how much they need to be changed to produce a more accurate output the weights and bies are then adjusted using a process called gradient descent gradient descent is an optimized ization algorithm that is used to find the weights and biases that minimize the cost function pulling the network closer to more accurate outputs based on the training data at its core an artificial neural network is a complex mathematical
* function that Maps input features to Output predictions the weights and biases of the network represent the parameters of this function and the goal of training is to find the optimal values of these parameters that minimize the difference between between the predicted output and the actual targets the training and use of artificial neural networks involve a large amount of mathematical computation primarily in the form of matrix multiplication for propagating parameter effect within the network and calculus
* for back propagation passes where updates to weights and biases are made through gradient descent the number of Matrix multiplications scales with the number of layers and neurons in the network while the number of gradient computations scales with the number of weights and biases let's look at a tiny simple feedforward neural network with an input layer of 1,000 neurons one hidden layer of 500 neurons and an output layer of just 10 neurons during a Ford pass this network would involve 500,000 multiplications and additions
* for the first layer and 5,000 multiplications and additions for the second layer as the network gets larger and utilizes more layers the computing power required Skyrock ETS the V16 architecture for example a convolutional neural network renowned for its Simplicity and Effectiveness in image classification and object detection has just 16 layers yet involves over 138 million parameters the number of floating Point operations required for a single Ford pass through the network is in the order of billions as artificial
* neural networks grow to the levels required for industry changing large language model AIS the computing power required becomes staggering to illustrate the magnitude of computation involved in large neural networks let's consider the already outdated gpt3 model developed by open AI in 2020 gpt3 launched as one of the largest language models at the time bringing public awareness to the incredible power of artificial neural networks at these scales gpt3 at its core is based on a type of neural network architecture
* known as a Transformer that consists of 96 layers each with 12,288 neurons Transformer networks use self attention mechanisms to capture dependencies between words in a sequence weighing the importance of different input parts for predictions when all the supporting mechanisms are considered gpt3 has a staggering 175 billion parameters which include the weights and biases of the network to put this into perspective if each parameter was stored as a 32-bit floating Point number the model would require approximately 700 GB
* of memory training gpt3 required an immense amount of computational power it was trained on a cluster of 1224 Nvidia a100 GPUscollectively creating up to 320 POF flops of mixed Precision performance the training process involved feeding the model with a vast Corpus of text Data comprising approximately 45 tabt of compressed plain text during training the model processed hundreds of billions of individual words or subwords know as tokens the training process took several weeks to complete consuming a significant amount of energy and
* computational resources some estimates placed the energy consumption of training at around 220 megawatt hours or enough to power about 20 average US homes for a year even after training a single Ford pass through the gpt3 model involves a massive number of Matrix operations with 175 billion parameters and 96 layers the number of floating Point operations required for a single Ford pass through gpt3 is estimated to be in the order of trillions for example to process a sequence of 1,000 tokens gpt3 would require approximately 400
* teraflops in 2024 it's estimated that all leading models operate on well past a parameters and this is expected to continue growing until a point of diminished returns is reached with energy consumption being a primary element in this limit currently most AI is derived from the second generation of artificial neural network development characterized by its primary focus on deep learning but to push past the energy requirements associated with it current research on third generation networks is focused on a concept that
* mimics biological systems much closer with spiking neural networks while inspired by biological neurons current artificial neural networks are simplified models that do not fully capture the complexity of biological systems spiking neural networks aim to bridge this Gap by communicating through discrete spikes or pulses with timing of these spikes carrying information when compared to the continuous use of activation functions to compute an output based on the weighted sum of inputs the spiking method of information
* transmission more closely resembles biological neurons operating on discrete events that occur at a certain point of time spiking neural networks receive a series of spikes or a spike train as input and produce a spike train as the output one of the largest advantages to this approach is in Energy Efficiency current artificial neural networks require a constant recalculation of the entire network whenever changes occur making its reaction to new information incredibly energy intensive in contrast spiking neural networks much like
* biological systems only generate spikes when necessary leading to sparse activity and drastically reduced Energy overhead spiking neural networks behave drastically different from traditional activation function-based models in that their memory and functionality operates analogous to the membrane potential mechanism of biological neurons while several neuron models for spiking neural networks exist for determining the relationship between neural membrane potential at the input stage and membrane potential at the output stage
* the most commonly used model is the Leaky integrate and fire threshold model in this model the membrane potential equivalent in a spiking neural network can be increased by excitatory spikes and decreased by inhibitory spikes it also exhibits Decay over time simulating the leakage of electrical charge in biological neurons if a neuron's membrane potential exceeds a threshold the neuron will send a single impulse to each connected Downstream neuron after generating a spike the neuron's membrane potential is reset to a resting value
* after firing a spike the neuron enters a refractory period during which it cannot generate another Spike the refractory period simulates the biological constraints of neurons requiring time to recover before firing again because of of their event-driven nature spiking neural networks produce a continuous asynchronously driven output that reacts dynamically with inputs and the Network's internal structure this is dramatically different from the large parameter function model of traditional artificial neural networks that produce
* a real number output instead of a calculated gradient descent into a solution spiking neural networks approach a goal by dynamically reaching an equilibrium over time within its Network spiking neural network networks communicate with far more information than traditional artificial neural networks due to the timing element of the spiking process known as temporal coding these Spike trains can represent information in a broad range of encodings from simple pulse rates to elaborate timing patterns and even multi-layered coordinated patterns with
* other neuron groups it's theorized that these temporal interactions among groups of neurons create emergent signal processing patterns that can potentially replace place the equivalent functionality of hundreds of neurons in traditional artificial neural networks because time is an encoded property of spiking neural networks information flow it's well suited for processing continual real world sensory information such as spatial temporal data and motion control it can also accomplish this with less Network complexity and Incredibly
* low processing latencies all while eliminating the need for the recurrent structures that introduce temporal awareness intr traditional artificial neural networks while incredibly powerful and versatile spiking neural networks exhibit an inherent incompatibility with current artificial neural network technology reliably encoding and decoding traditional data through a spiking neural network in pulse trains is proving to be difficult though various experimental methods exist for coding real numbers as Spike trains such as rate codes or frequency
* of spikes time to First Spike and the interval between spikes even within the realm of neurobiology research is still ongoing as to how exactly sensory information is encoded processed and reacted to all within 10 milliseconds a response time that supersedes what's possible with basic coding methods spiking neural networks also suffer from a fundamental incompatibility with current artificial neural network training techniques because of the asynchronous nature of spiking neural networks and the difficulty in
* mathematically defining change in spiking information propagation with within the network spiking neural networks are unsuitable for traditional artificial neural network gradient descent-based training methods that perform error back propagation when combined with the challenges of information coding spiking neural networks are proving to be challenging to train in a supervised manner where labeled data is used to provide a specific functionality from the network in fact to date there is no effective supervised Training Method that is
* suitable for spiking neural networks that has Prov provided better performance than second generation networks however they have been demonstrated to be a viable option for unsupervised biologically inspired training methods that work best with generalized prediction clustering and Association of information because traditional artificial neural networks are effectively massive math problems they work well with classic Computing architecture in which a system encompasses a clocked interconnection of CPU memory storage and IO that exchanges
* data and instructions back and forth in a sequential manner to perform computation this heavy Reliance on matrices allows artificial neural networks to scale well with parallel Computing for large scale artificial neural networks this is accomplished using thousands to millions of processor cores using GPUsor dedicated GPU like parallel Computing processors spiking neural networks however do not perform as well on traditional Computing architecture and cannot scale easily on them they are asynchronous behavior and
* Reliance on localized timing Independence makes emulating their behavior in software computationally expensive due to the overhead created by the fundamental incompatibility and how information flows through them when compared to traditional Computing architecture while currently this hinders their use at larger scales that can rival current artificial neural network capabilities it has led to the research and development into an entirely New Field of Hardware Computing architecture based on biking neural networks known as neuromorphic Computing
* neuromorphic devices are based around a processing architecture that physically recreates the properties of a biological neuron this paradigm shift in Computing replaces the synchronous monolithic movement of data and instructions between separate processors and memory with a large array of interconnected artificial neuron elements each with their own localized memory and Signal processing neuromorphic devices can be based on a broad range of mediums such as chemical and fluid systems but semiconductor-based mixed mode analog
* digital ic's are the focus of current research while traditional full digital Computing can be applied to the concept researchers are looking towards analog Computing based on historis or the dependence of the state of a system on its history to create neuronlike functionality within these devices analog processing eliminates the complexity and latency of digital architecture by deriving AR icial neuron functionality from the physical properties of a semiconductor component directly this creates an extremely fast
* reacting Computing element with orders of magnitude less power consumption while analog Computing has always been too noisy and inconsistent for traditional Computing much like in biological systems timecode spiking signals are far more resilient in a noisy and irregular signal environment currently a few key analog semiconductor technologies that store and prod process information in a way that resembles the behavior of biological synapses are at the Forefront of semiconductor-based artificial neuron research memers are
* two terminal devices that change their resistance based on the amount of current that has flown through them phase change memory consists of calogen material sandwiched between two electrodes when a voltage is applied the material heats up and changes its phase from Amorphis to crystalline changing its electrical resistance ferroelectric Field Effect transistors are three terminal devices that use a ferroelectric material as the gate dialectric when a voltages appli to the gate the ferroelectric material polarizes changing the conductivity of
* the channel between the source and drain electrodes spintronic devices use the spin of electrons to store and process information they typically consist of a magnetic material sandwiched between two non-magnetic electrodes when a current is passed through the device the spin of the electrons aligned with the magnetic field changing the devices resistance by 2014 the first neuromorphic chip would be introduced called True North True North consists of 4,096 cores each containing 256 programmable simulated neurons totaling
* just over a million neurons each neuron has 256 programmable synapses that transmit signals between them resulting in over 268 million programmable synap true North's design allows for efficient memory computation and communication handling within each neurosynaptic core bypassing traditional Computing architecture bottlenecks this results in a low 70 M power consumption and a power density 1,000th that of a conventional microprocessor in 2017 Intel would introduce loow e a neuromorphic chip fabricated using Intel's 14 nanometer
* process that features 128 clusters of 1,2 4 artificial neurons each totaling 131,072 simulated neurons and around 130 million synapses although less powerful than IBM's True North it offered far more flexibility becoming a powerful tool for energy efficient real world spiking neural network-based problem solving research by September 2021 lye 2 would be released featuring over a million simulated neurons with faster speeds higher bandwidth intership Communications increas capacity a more compact size and improved
* programmability compared to its predecessor the lowy HE2 chip would become the basis for the h point in 2024 becoming the world's largest neomorphic system consisting of 1,152 lowy HE2 processors the system supports up to 1.15 billion neurons and 128 billion synapses across 14,545 neuromorphic processing course while consuming just 2600 watts of power it also includes over 2,300 embedded x86 processors for ancillary computations it's estimated that halap point has the neuron capacity of roughly equivalent to
* that of an alow brain or the cortex of a capucin monkey L he 2 based systems have demonstrated the ability to perform inference and optimization using 100 times less energy at speeds up to 50 times faster than existing GPU based architecture as of 2024 despite ongoing research there are no commercially available analog-based AI chips though some research-based and smallscale ic's have been developed while neuromorphic development is pushing forward on mature digital architecture the industry is optimistic about a breakthrough into a
* hybrid analog future as research progresses we can expect neuromorphic systems with greater neuron capacities faster processing speeds and improved Energy Efficiency to revolutionize the AI field with self-contained AI drastically advancing in fields such as Robotics and autonomous systems within the coming years as deeper models and more sophisticated Hardware evolve in the pursuit of a third generation of neural networks the missing link between prediction algorithms and true intelligence May soon begin to emerge a
* great way to bridge the gap in understanding of this Revolution and appreciate the ground breaking advancements propelling the field forward is brilliant.org brilliant is where you discover the thrill of learning with thousands of captivating interactive lessons in math data analysis programming and AI designed to Unleash Your Potential and transform you into a confident Problem Solver brilliant is an Innovative learning platform that stands out for its use of a first principles approach that enables you to build a solid foundation of
* understanding each lesson is brimming with interactive problem solving exercises allowing you to actively engage with with Concepts this technique has been shown to be six times more effective than simply viewing lecture videos moreover all Brilliance content is developed by a distinguished team of award-winning Educators researchers and Industry experts from prestigious institutions such as MIT Caltech Duke and renowned companies like Microsoft and Google brilliant immerses you in active problem solving because truly
* grasping a concept demands more than just mere observation and memorization you need to experience it by engaging in Hands-On learning you will not only build Real World Knowledge on specific topics but also develop critical thinking skills that make you a better thinker overall investing in Daily learning is Paramount for personal and professional development and Brilliant makes this convenient and enjoyable with captivating digestible lessons that seamlessly integrated into your daily routine you can build genuine knowledge
* in just a few minutes each day say goodbye to aimless scrolling and embrace a more rewarding way to spend your free time a great introduction to the evolving technology behind lm's is Brilliance how l L M's work course in this series of lessons you'll take a peak under the hood of today's most popular large language models to understand how they work and the challenges of creating them all while building a solid comprehension of their capabilities to try everything brilliant has to offer for free for a full 30 days
* visit brilliant.org newmind or click on the link in the description below you'll also get 20% off an annual premium subscription for

Made with ❤️ by[Glasp](https://glasp.co/?ref=youtube-summary)