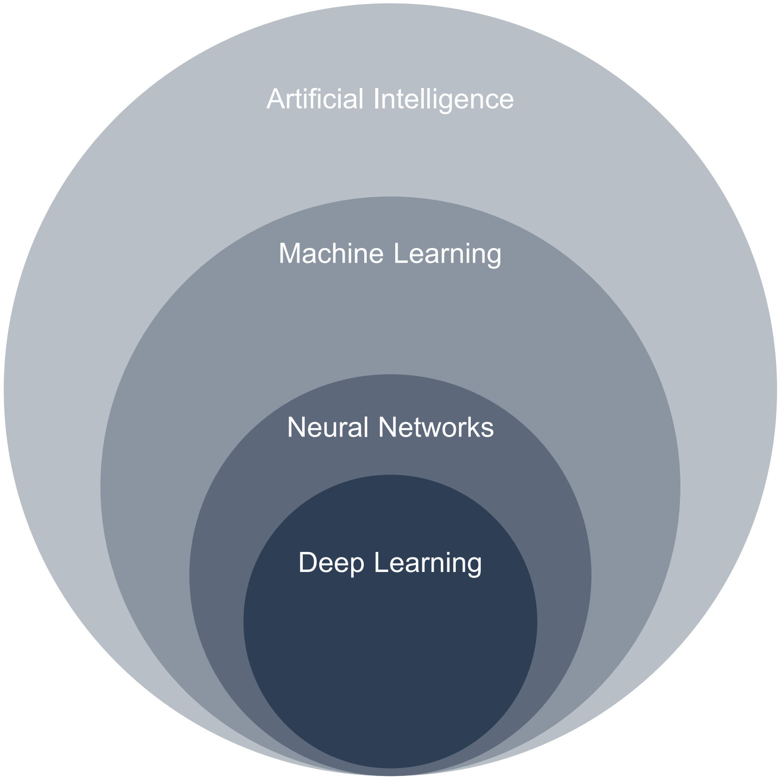
University of Illinois at Urbana-Champaign
Dept. of Electrical and Computer Engineering

ECE 101: Exploring Digital Information Technologies for Non-Engineers

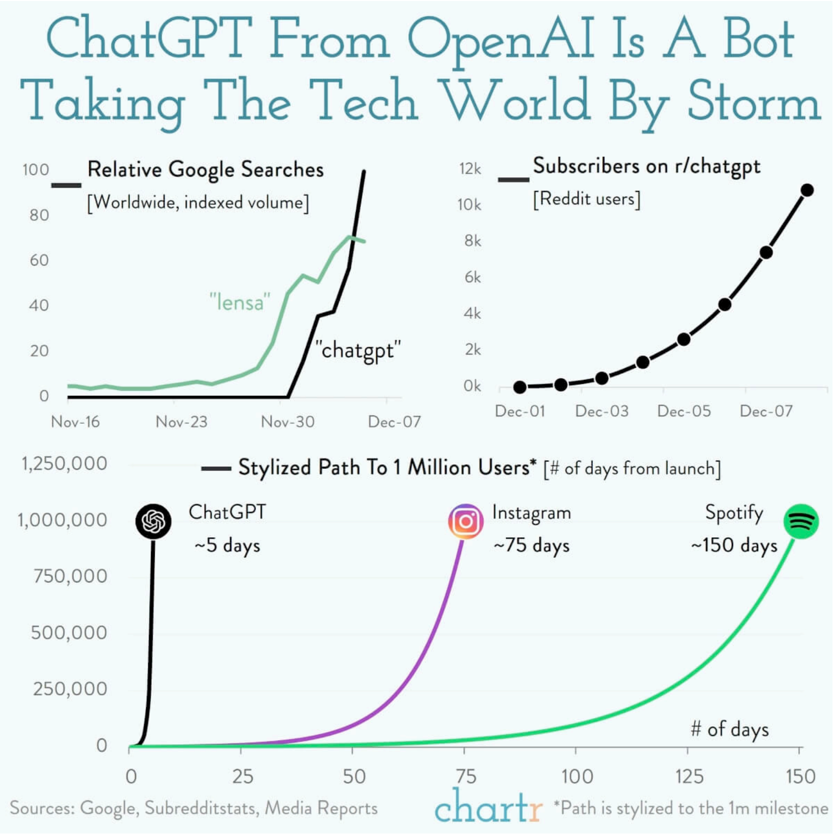
Neural Networks and NLP

Neural Nets for AI



Neural Nets in Recent Years

- **2017**—A team of Google researchers proposes a new simple network architecture, the Transformer. Transformers enabled advancements in generative models compared to older long short-term memory models
- **2018**—OpenAI releases GPT (Generative Pre-trained Transformer), a language model that achieves state-of-the-art performance on various natural language processing tasks
- **2021**—DALL.E, a transformer-based neural network-based system developed by OpenAI, generates images from textual descriptions
- **2022**—ChatGPT releases GPT-3.5, an AI tool that reached one million users within five days. The tool can access data from the web from up to 2021.



ChatGPT From OpenAI Is A Bot Taking The Tech World By Storm

Relative Google Searches [Worldwide, indexed volume]

Subscribers on r/chatgpt [Reddit users]

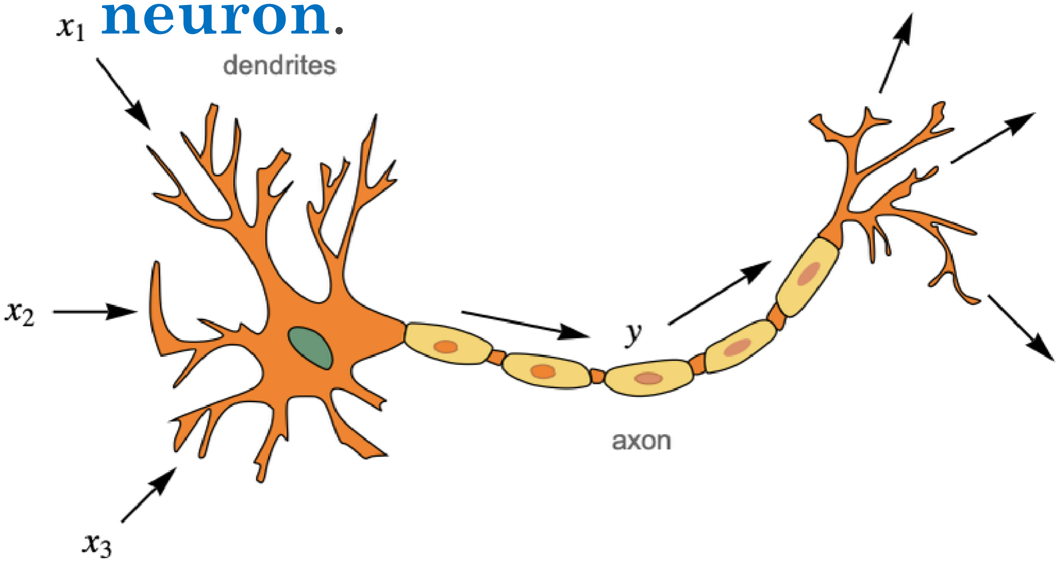
Stylized Path To 1 Million Users* [# of days from launch]

Sources: Google, Subredditstats, Media Reports

chartr *Path is stylized to the 1m milestone

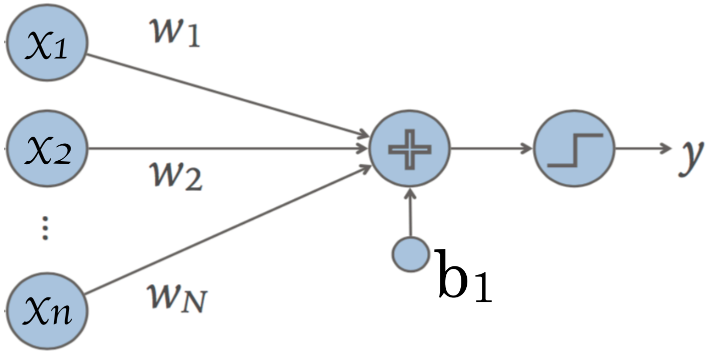
Perceptron

A **perceptron,** the simplest form of an artificial neural network **mimics a human neuron**.

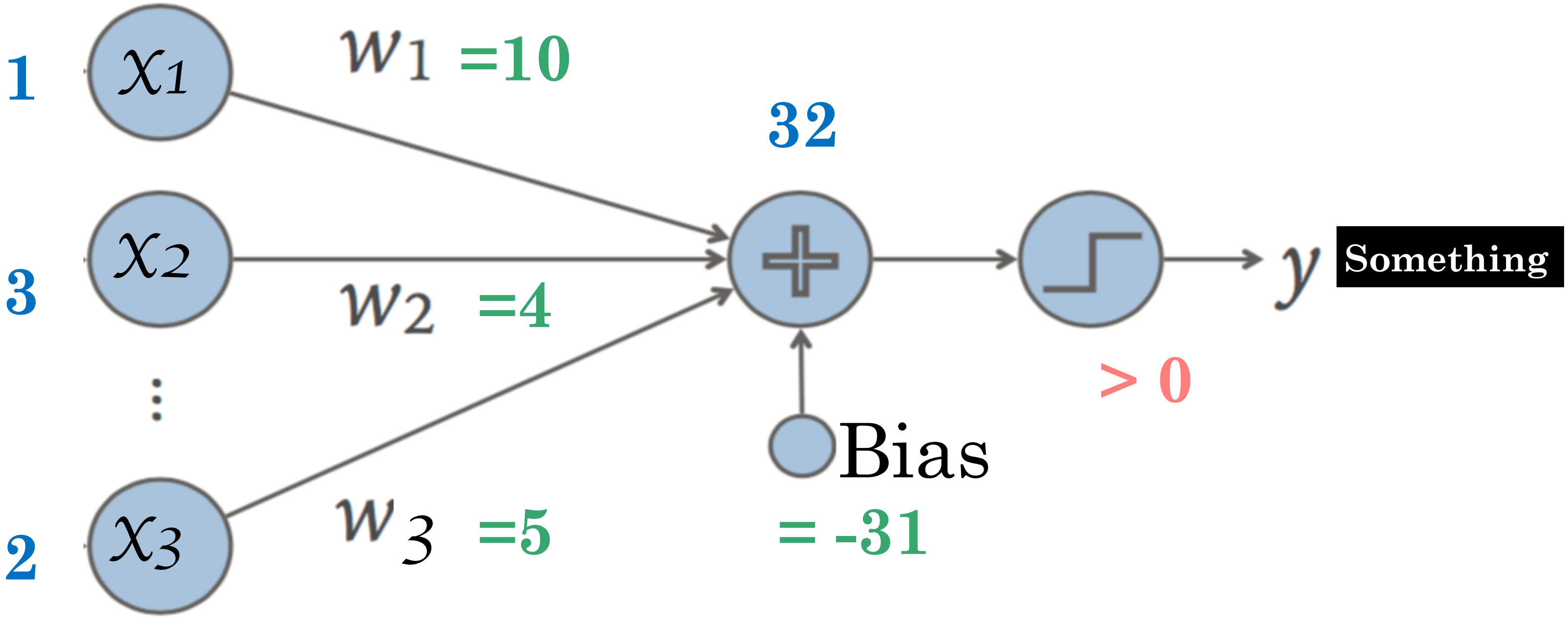


Inputs (on the left) are multiplied by weights, then summed together with a bias.

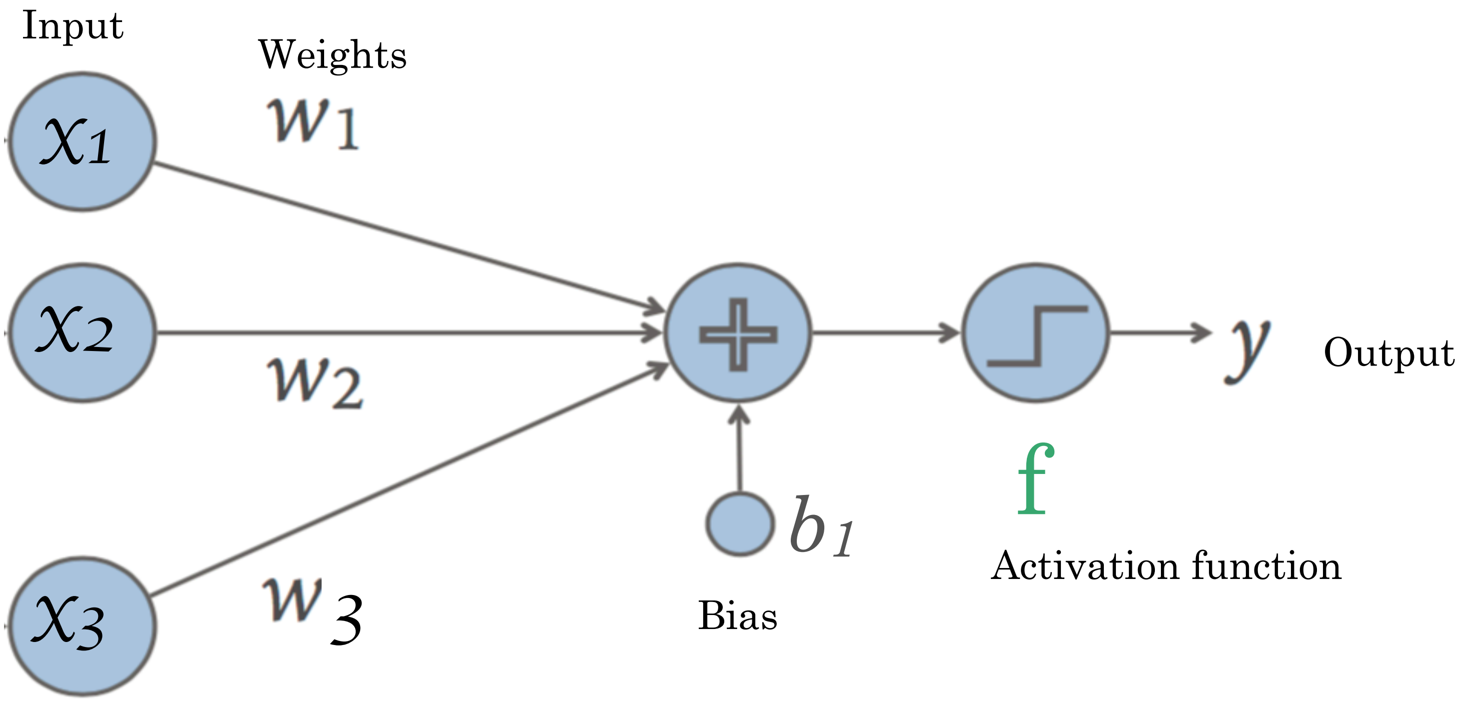
The sum is then converted to +1 or -1.



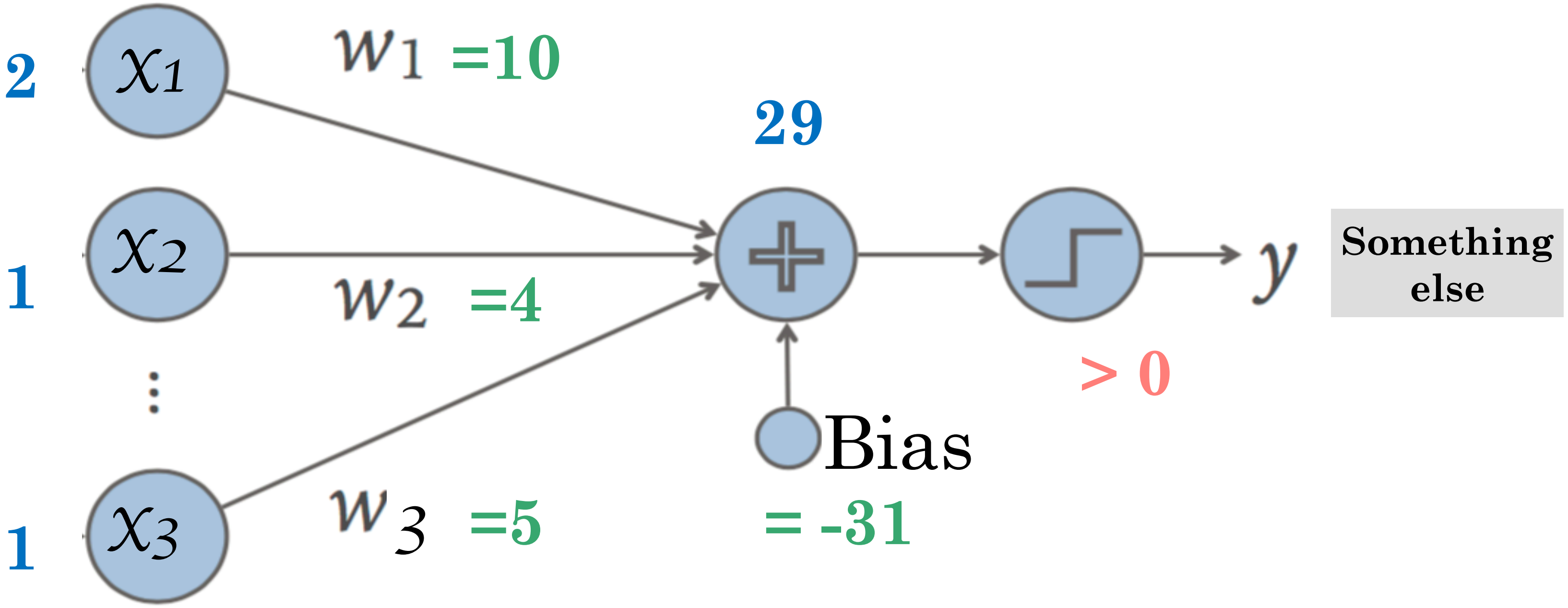
Computation in a Perceptron



Computation in a Perceptron



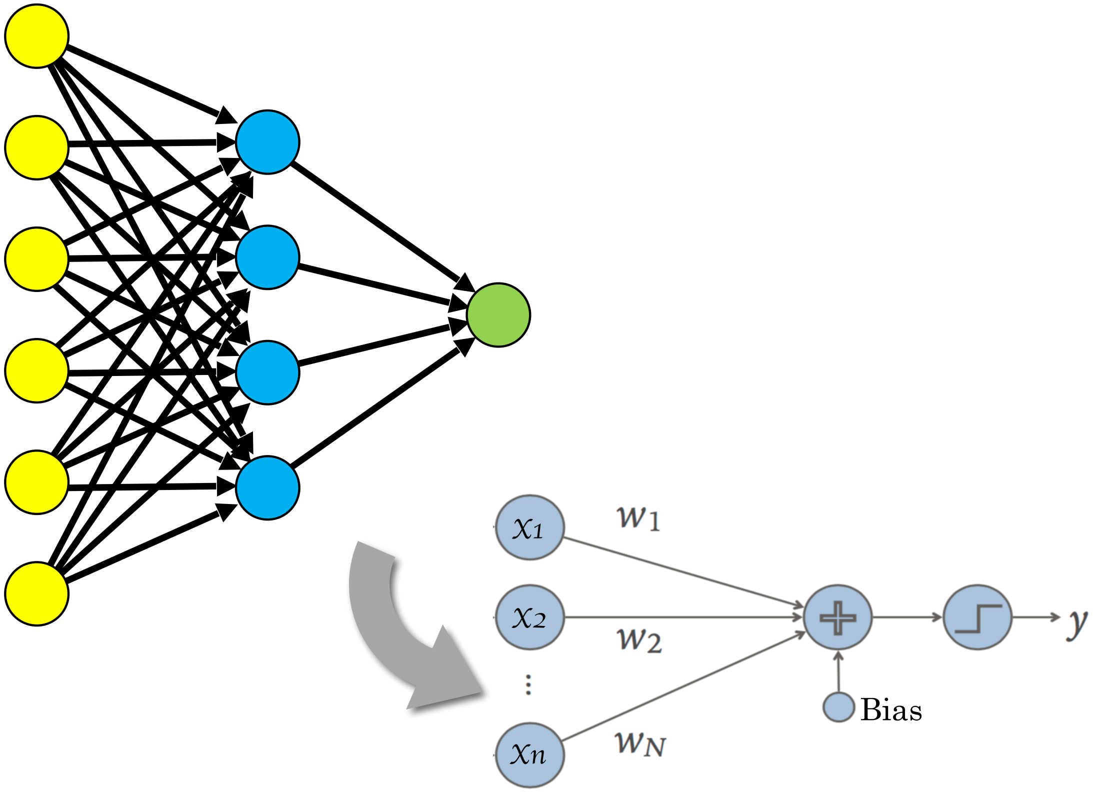
Computation in a Perceptron



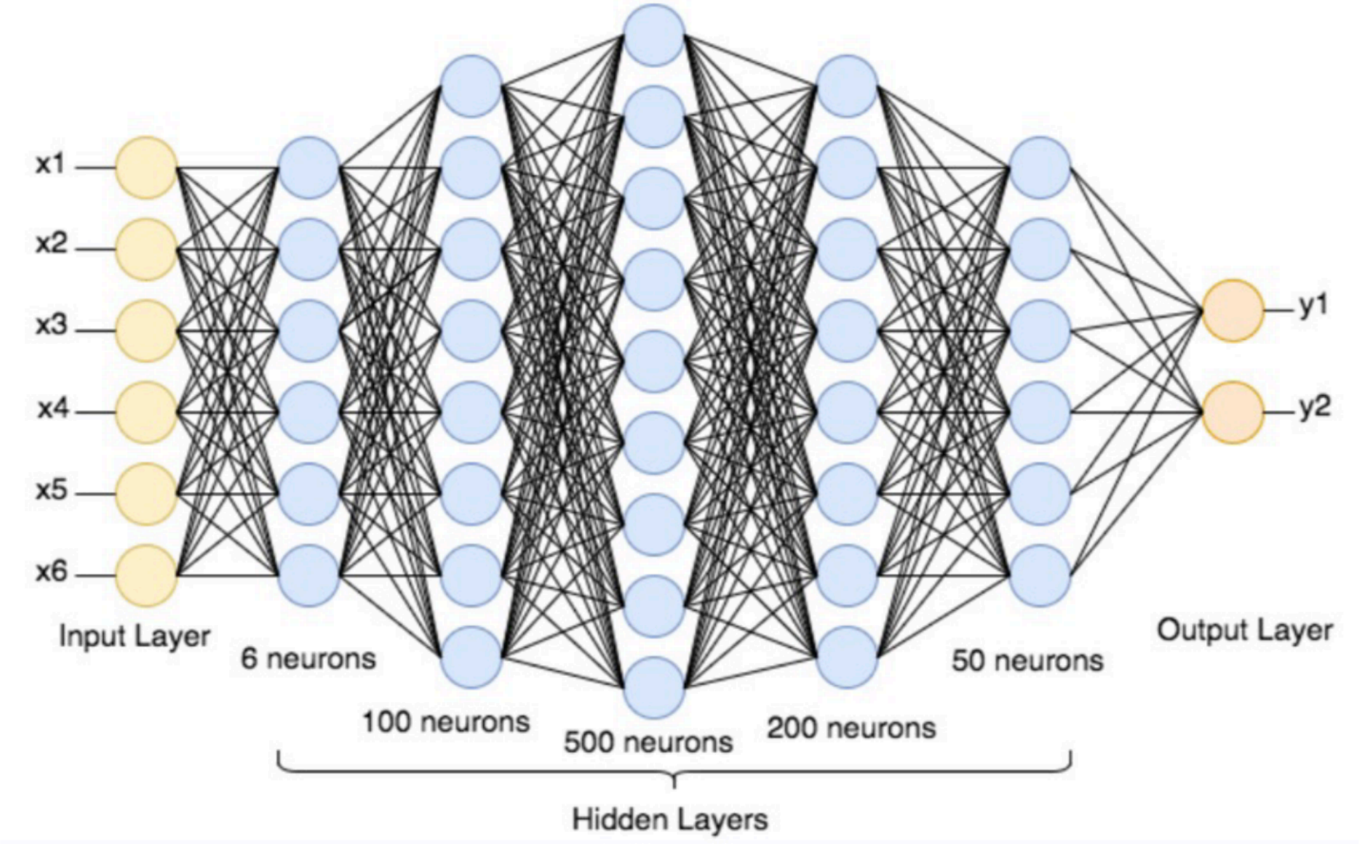
Neural Networks Consist of Many Artificial Neurons

To **perform more complex** (non-linear) **tasks**,

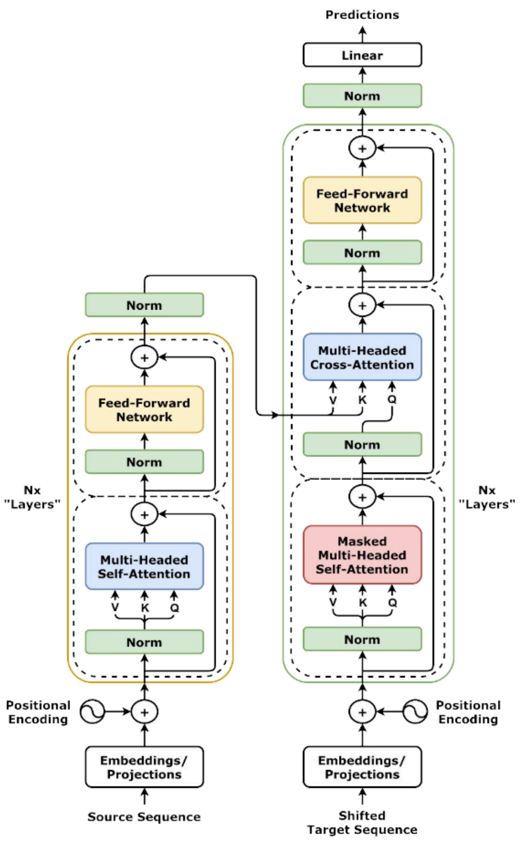
- perceptrons can be **connected in a network**
- by using the output of one perceptron
- as the input to a second,
- then a third, and so forth.



Neural Networks Consist of Many Artificial Neurons

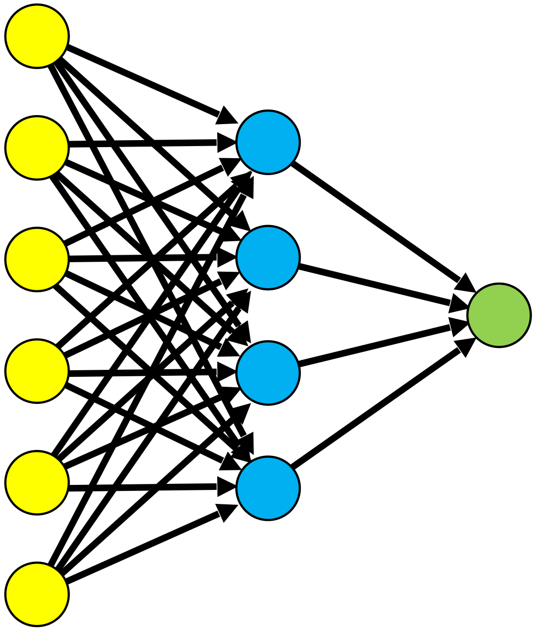


Neural Networks Consist of Many Artificial Neurons



Input, Output and Hidden Layers

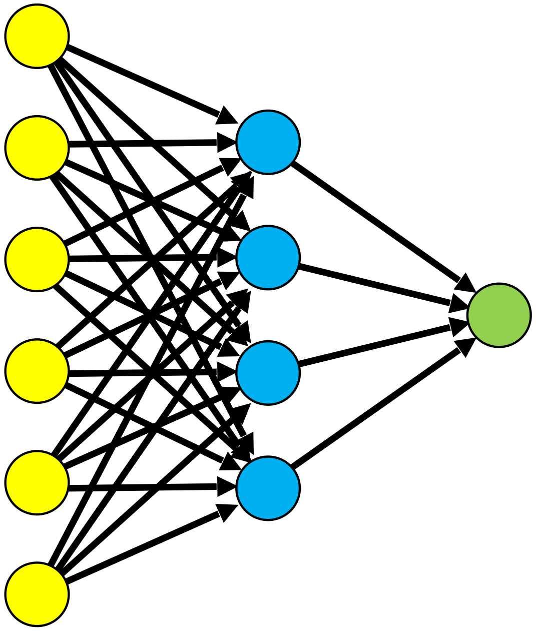
- **Input layer** (in yellow)
- Every input is connected to every node in the **hidden layer** (in blue)
- **Output layer** (in green) consists of a single output perceptron
- In a fully connected network, every node is connected every node in the next layer



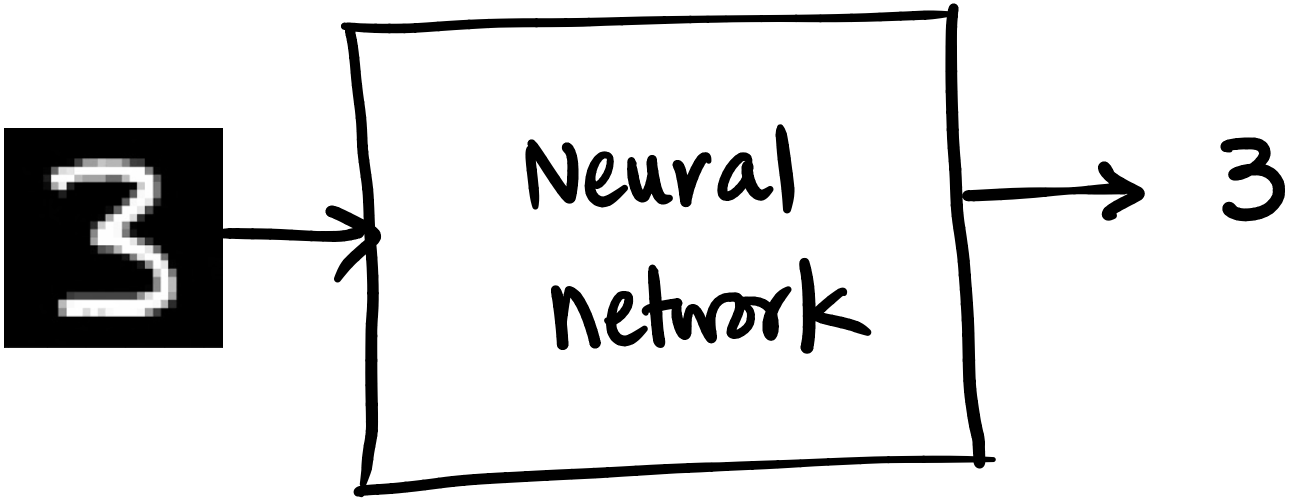
The Network is a Complex Function made up of Simple Parts

Think of it like a brain
made of tiny decision-makers.

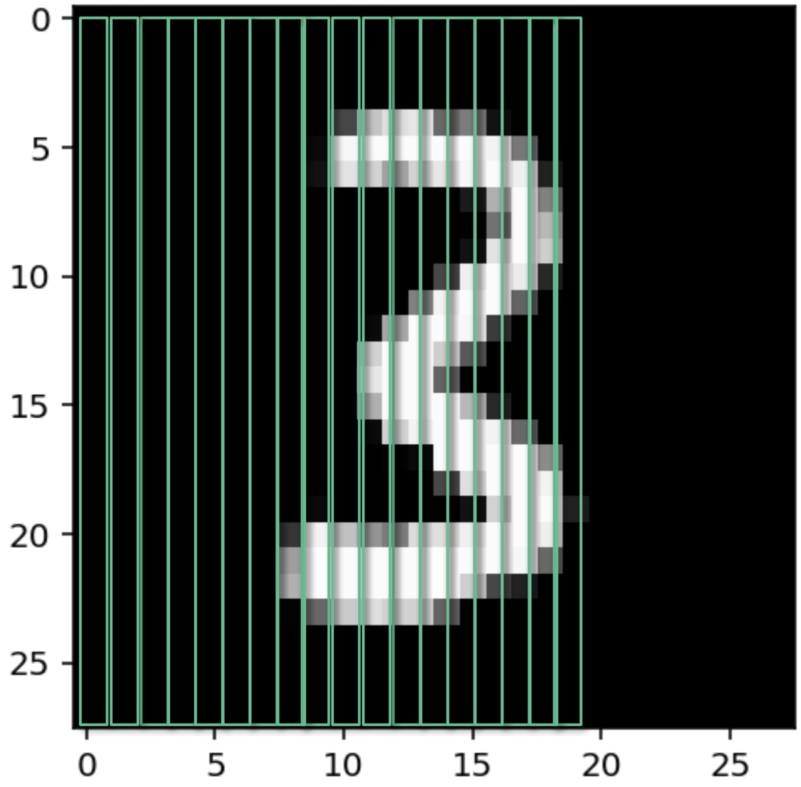
Each one looks at a piece of information
and passes it along.



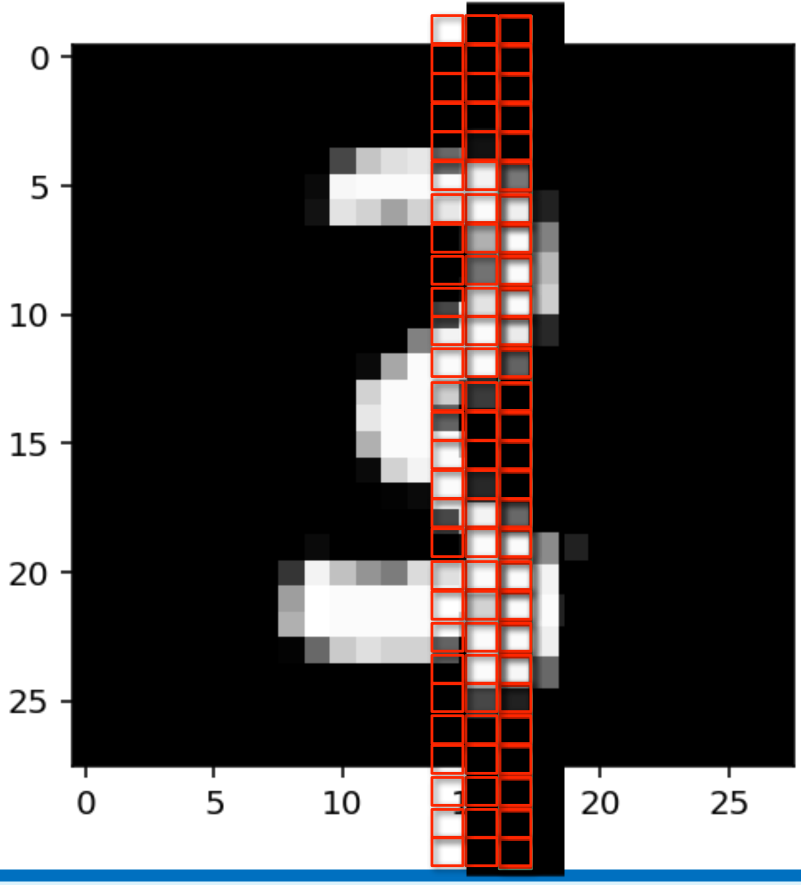
How does a Neural net recognize images?



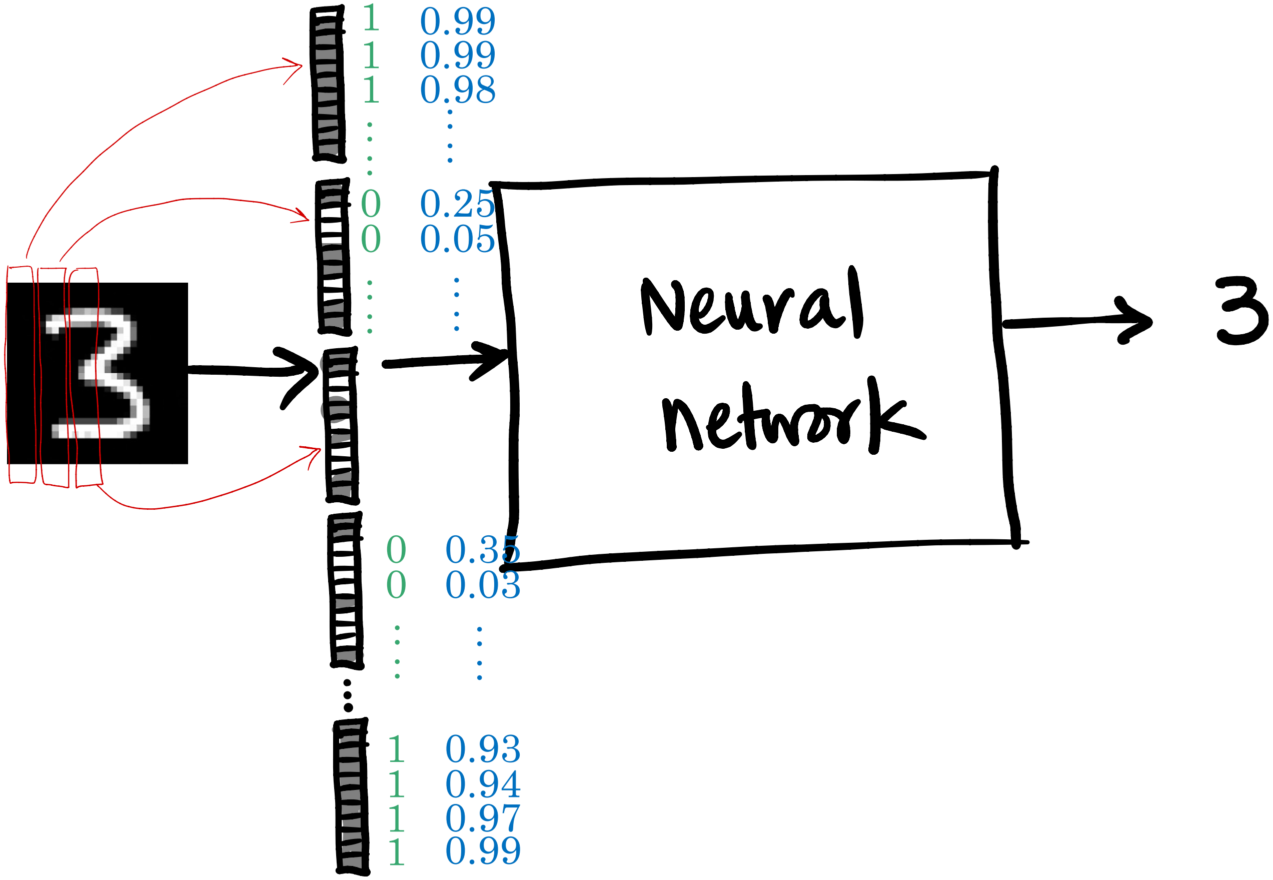
How does a Neural net recognize images?



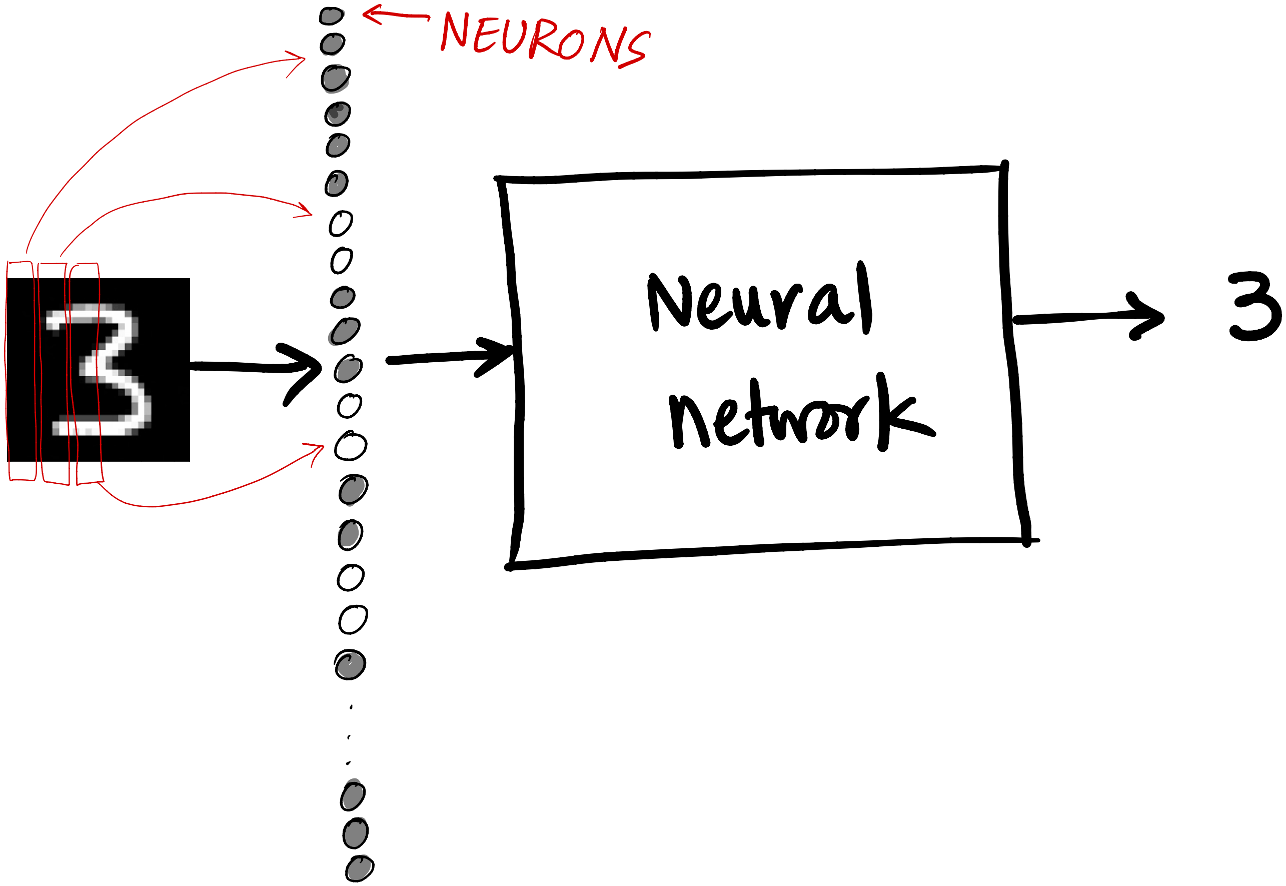
How does a Neural net recognize images?



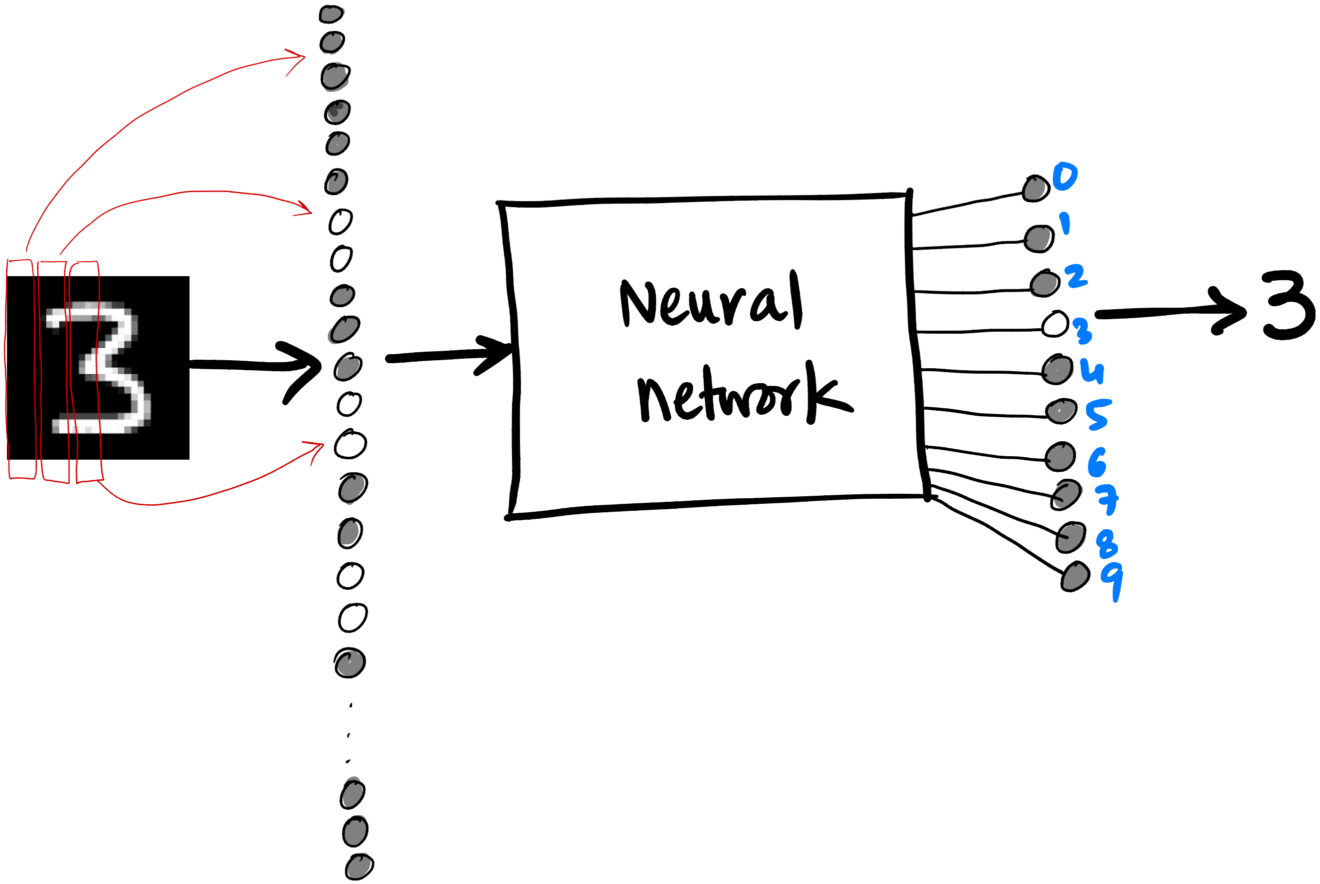
What is the Network Doing?



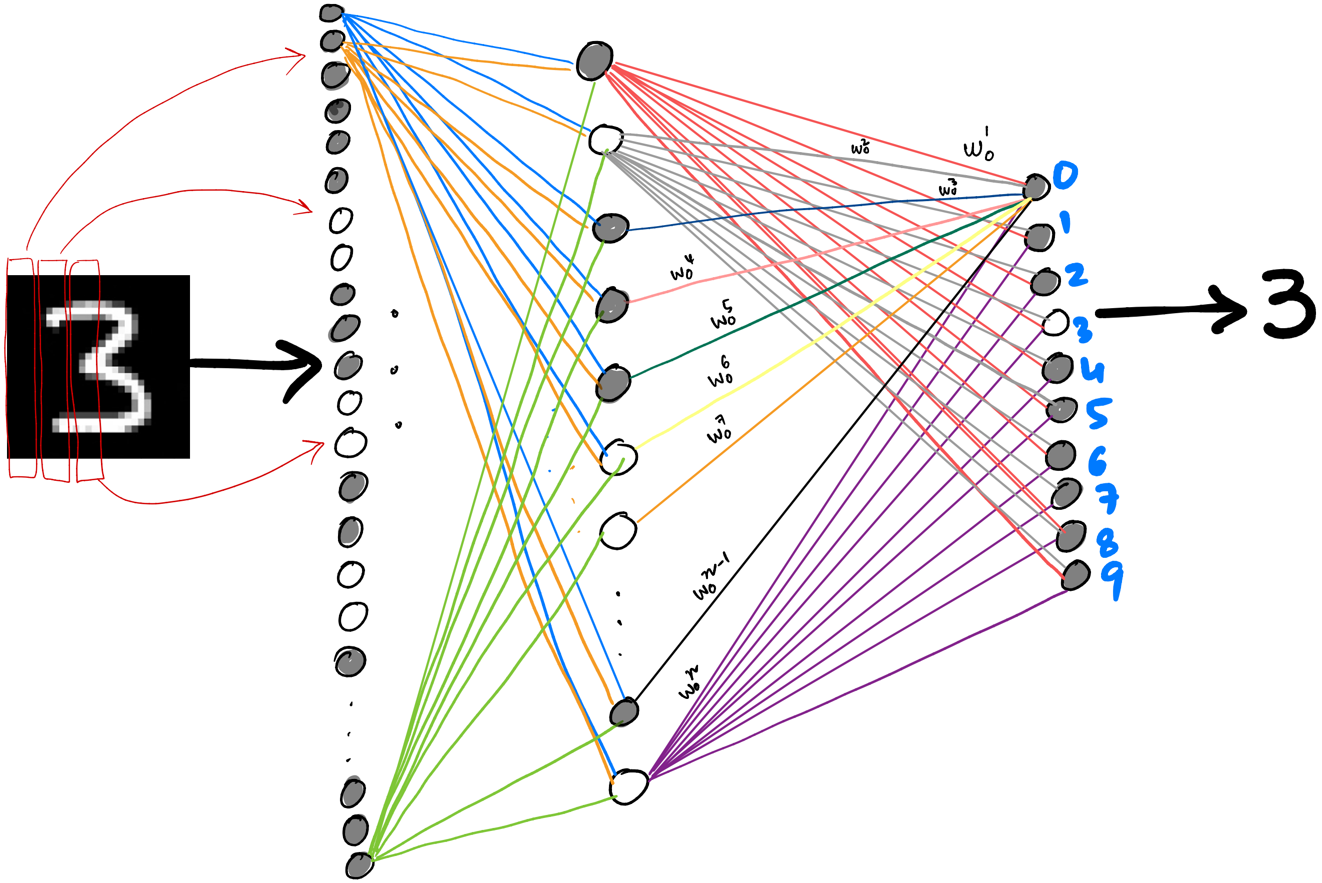
What is the Network Doing?



What is the Network Doing?

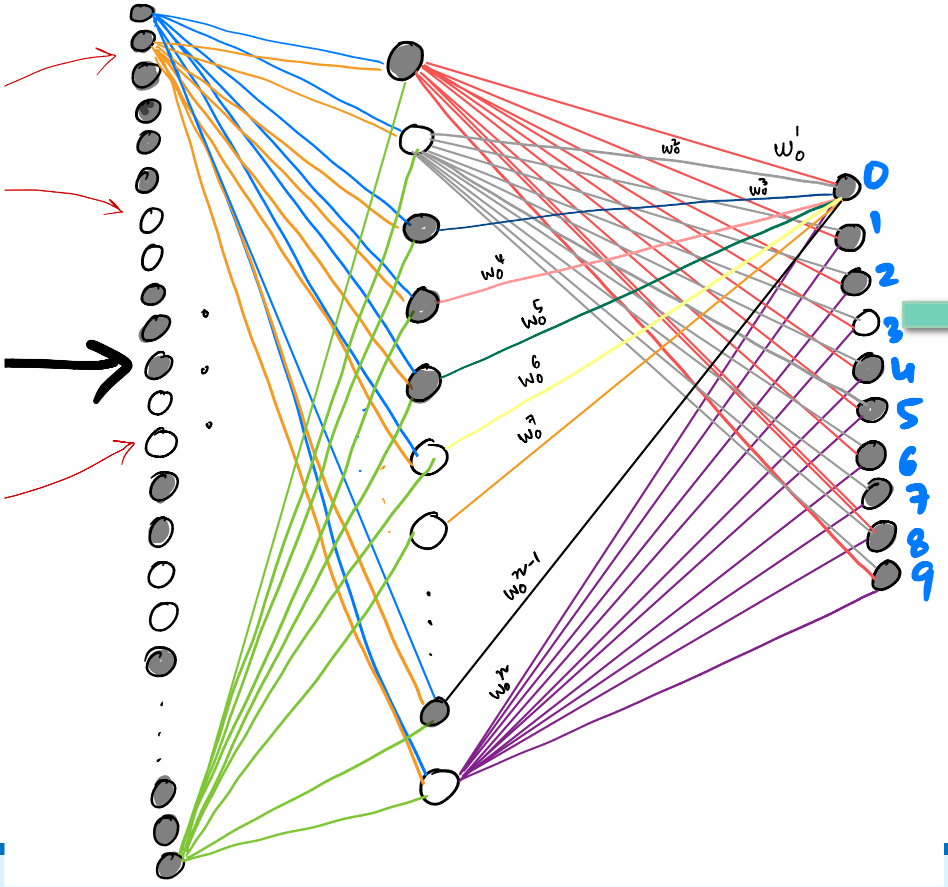


What is the Network Doing?



What is the Network Doing?

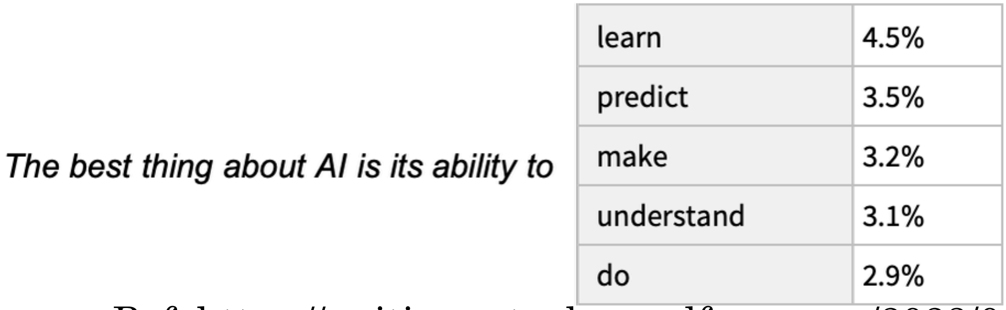
The cat sat on the



Box
Hat
Bat
Mat
Rug
Dog
Grass
Couch
Throne
Flying carpet

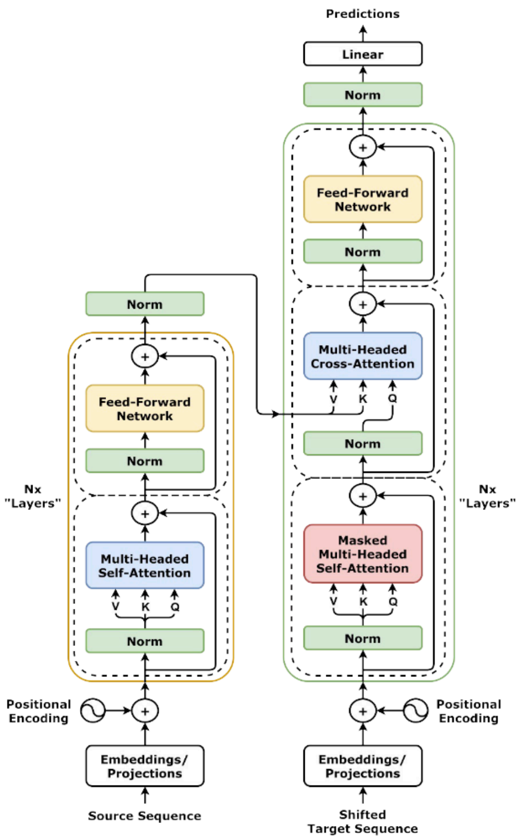
LMM: Large Language Model

- Deep neural networks—fundamentally trying to do is to produce a “reasonable continuation” of whatever text it’s got so far
- By “reasonable” we mean “what one might expec someone to write after seeing what people have written on billions of webpages, etc.”



The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

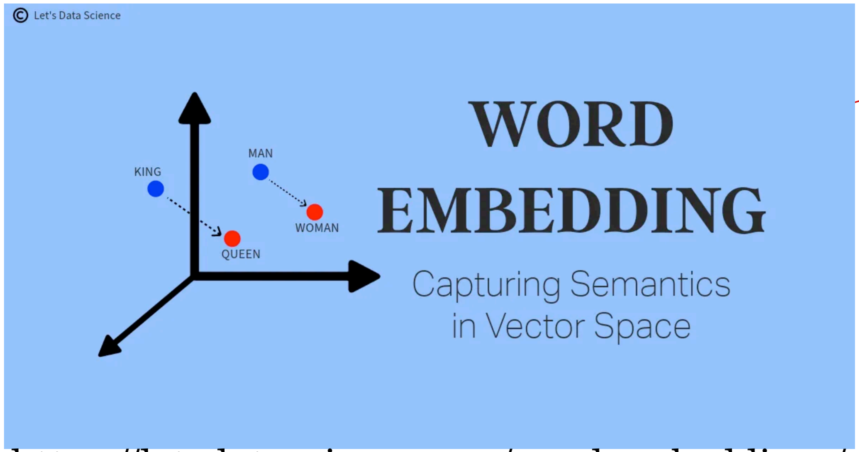


Ref: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

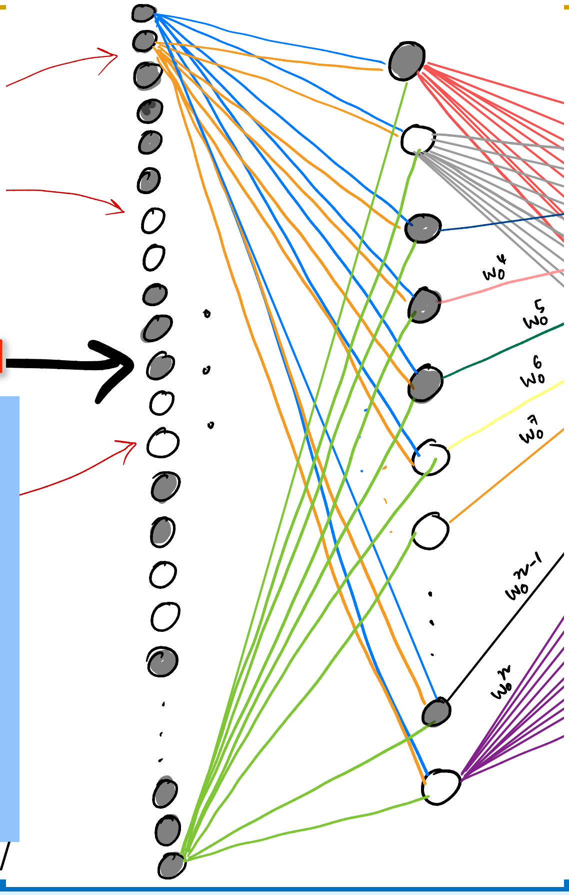
How do we get from words to numbers?

Tokens

The cat sat on the



https://letsdatascience.com/word-embeddings/



Guessing Words

The cat sat on the mat

The cat was chasing the mouse

The cat lapped up the milk

The cat was …

Guessing Words is Easier with Words on Both Sides

Understanding how words fit together is an important element. For example, you hear,

"I took my … [more words]."

What's the next word?

In 1953, journalists* realized that **the words AFTER the missing would help** in guessing.

Everyone else already knew.

*W. L. Taylor, "Cloze procedure: A new tool for measuring readability," *Journalism Bulletin*, 30(4):415–433, 1953.

Google Applied Bidirectional Idea to Create BERT

"I took my … for a walk."

What's the next word?

Dog, perhaps?

Could be lots of words, but dog may be a good MLE choice.

"Cat," "snake," "Ferrari," maybe not so good.

In 2019, Google* realized the same thing, and natural language processing changed forever.

*Everyone except engineers, I meant: J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019. To their credit, the authors DID mention the journalists.

A Model to Understand Text

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)

- revolutionized how machines understand language
- **read text in both directions at once**, which helps it understand the context better (before BERT, many models read text in one direction i.e. left to right)
- first trained on a large amount of text (**pre-training**) to learn general language patterns
- **fine-tuned** to the specific task, like answering questions or sentiment analysis (versatility)

Large Language Models

GPT (Generative Pre-trained Transformer), developed by OpenAI.

1. **Generative Model**: While BERT is great at understanding and processing text, GPT is designed to generate text. This means it can write essays, create dialogue, and even compose poetry.
2. **Transformer Architecture**: Both BERT and GPT use a type of neural network called a Transformer, a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence. GPT uses this architecture to predict the next word in a sentence, making it very good at generating coherent and contextually relevant text.
3. **Pre-training on a Massive Scale**: GPT models are pre-trained on a vast amount of text from the internet. This extensive training helps them generate more natural and diverse text.

Whats the big deal about transformers?

1. **They focus on important words:** Transformers use a mechanism called *attention* to figure out which words in a sentence matter most for understanding meaning.

2. **They process text in parallel:** Unlike older models that read words one by one, transformers look at all words at once, making them faster and better at handling long sentences.

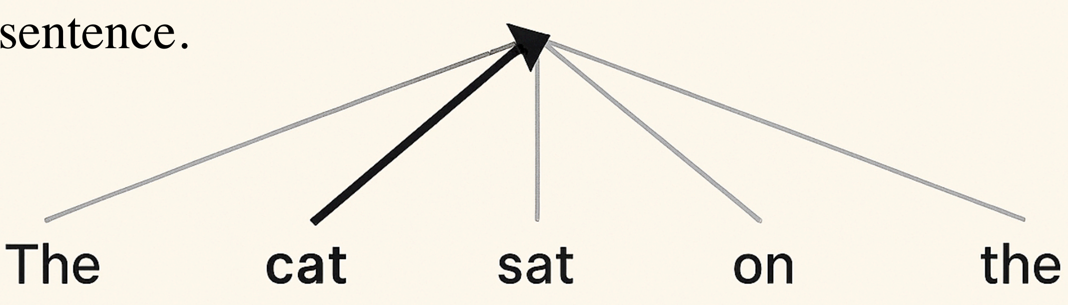
3. **They learn context deeply:** By stacking layers, transformers understand relationships between words and phrases, which helps them generate accurate translations, summaries, and answers.

Attention is all you need …

Seminal paper from Google: https://arxiv.org/abs/1706.03762

Attention in transformers works like this:

- **Spot important words:** For every word in a sentence, the model looks at all other words and decides which ones are most relevant for understanding its meaning.
- **Assign weights:** It gives higher “attention scores” to important words and lower scores to less relevant ones. For example, in “The cat sat on the mat,” the word “cat” is more important to “sat” than “mat.”
- **Combine context:** These weighted relationships are used to create a richer representation of each word, so the model understands not just the word itself but how it relates to others in the sentence.



Examples of NLP Applications

Text Classification: Categorizing text into predefined categories.

- Did the reviewer like a movie?
- What sentiment did they express?
- Did they mention or suggest any movie genres/categories?

Examples of NLP Applications

Interpretation:

- Is anything in a patient's electronic medical records relevant to the patient's current symptoms?
- **Named Entity Recognition (NER):** Identifying and classifying entities like names, dates, and locations in text.

Question Answering: Finding answers to questions within a text

Examples of NLP Applications

- **Text Generation**: Creating new text based on a given prompt.
  - **Dialogue Systems**: Powering chatbots and virtual assistants.
  - **Creative Writing**: Generating stories, poems, and other creative content.
- **How do you use it?**

Close the Loop: Communicate/Actuate

Let's close the loop by returning to the cycle:
sense, **compute**, **communicate**, **actuate**.

Once a smart home unit has understood a human and performed any necessary actions, it **needs to respond verbally**.

The process is similar and uses similar models:

1. **Convert** the response **into an intelligible sequence of words** in the speaker's language.
2. **Convert** the words **into** an audio output, **a synthetic voice**.

Voice Synthesis Allows a “Human” Response

The last step

- is called **voice synthesis**, or text to speech,
- **generation of human voice from text**.

The “voice” can be parametrized

- and thus **tuned to the listener’s preferences**
- or to match their verbal style and accent.

Synthesis is also **useful**

- **for entertainment and accessibility**,
- such as reading aloud for the vision-impaired
  or while humans are busy with other tasks.

Terminology You Should Know from These Slides

- NLP: Natural Language Processing
- LMMs: Large Language Models
- Transformers
- BERT: **Bidirectional Encoder Representations from Transformers**
- GPT: Generative Pre-trained Transformer
- NLP Applications:
- Voice Synthesis