

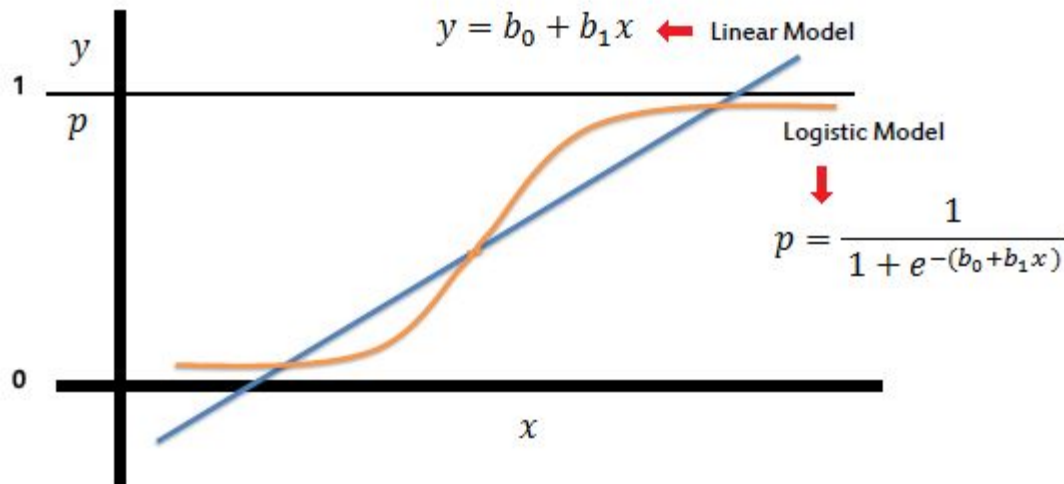
Text Generation

Data Science Club
Generative Models Workshop

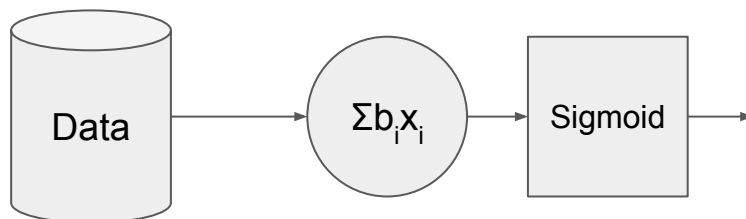
Deep Learning Review

First steps: Logistic Regression

Remember how logistic models work: They have a set of parameters (b_0 , b_1) and output a probability corresponding to a certain data input (x).



Neural networks take this idea and extend it.

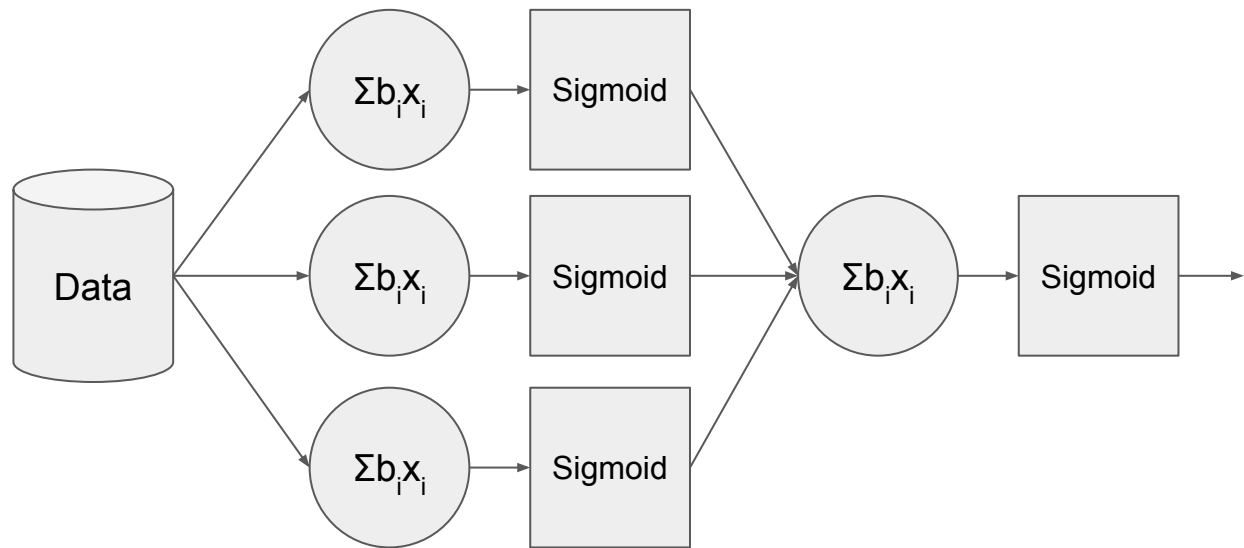


Next step: A Neural Network

We can instead have multiple logistic regressions in parallel that each have different values for b .

Now each one of these outputs can be used as the data input for a next logistic regression.

During the training process the model learns to diversify each set of b values in a way that gives more useful information.



1 layer with 3 neurons

What is a Text Generation model?

A model that inputs any type of data and returns text data.

Many models of these models are text to text, but we can also do speech to text, image to text, etc.

Text to Text models

Some examples of Text to Text models are:

Previous words to Next word

Language 1 to Language 2

Question to Answer

...

Requirements to Code

Data Representation

Input and/or outputs can be represented as either letters or words. GPT series uses words. But they split it into “subwords” based on part of speech, etc.

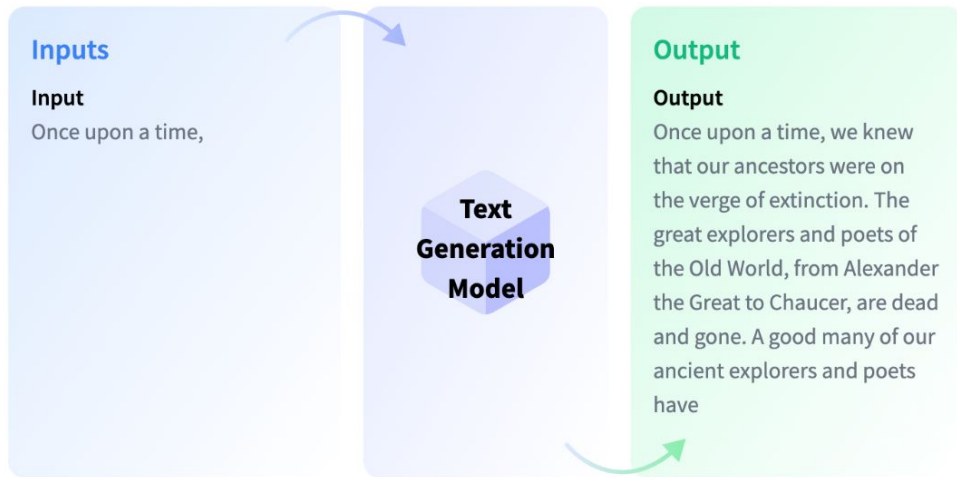
However, this data needs to be vectorized. Usually this means a one-hot-encoding with the dictionary of all words.

Inputs can also use word2vec to turn words into numbers representing semantics. However, this cannot effectively be used for outputs.

Text Generation Demo

[What is Text Generation? - Hugging Face](#)

Generating text is the task of producing new text. These models can, for example, fill in incomplete text or paraphrase.



Compatible libraries

🤖 Transformers

⚡ Text Generation demo

using `gpt2`

📄 Text Generation

Example 5 ▾

My name is Lewis and I like to think of myself as a person who's very, very funny, very strong. In this interview with the New York Times, he talks of being a real jerk and how the show he was on made him do |

Compute

⌘+Enter

0.8

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.777 s

</> JSON Output


🖥 Maximize

Question and Answering

Interestingly, many tasks besides predicting the rest of the paragraph can be accomplished just by reformatting the input to change the most likely prediction.

 **Text Generation demo**

using `gpt2`

 Text Generation Example 5 ▾

Q: What is my name?

N: My name is David. I have lived in Mexico for a very, very long time. I am 22 years old. And I am a professional wrestler by profession. My name is David, also


Compute

⌘+Enter

0.9

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.793 s

</> JSON Output

 Maximize

Methods Used in Training

Attention

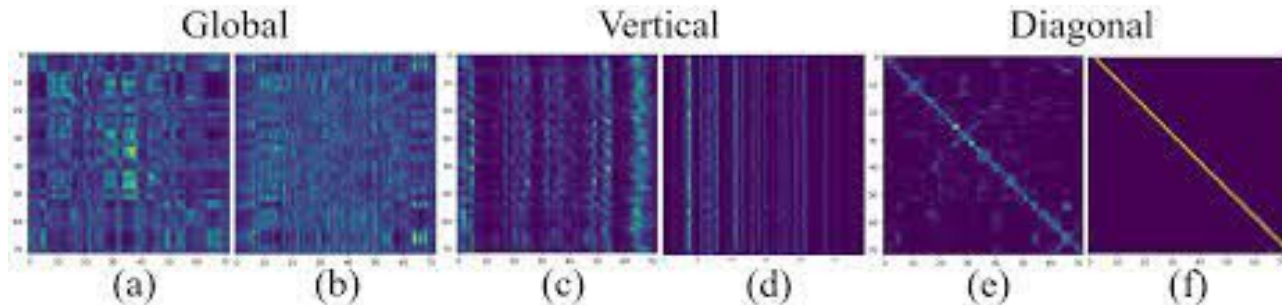
Recurrent neural networks

Data selection/Prompt Tuning

From GPT-3 paper:

We use the same model and architecture as GPT-2 [RWC+19], including the modified initialization, pre-normalization, and reversible tokenization described therein, with the exception that we use alternating dense and locally banded sparse attention patterns in the layers of the transformer, similar to the Sparse Transformer [CGRS19].

Attention



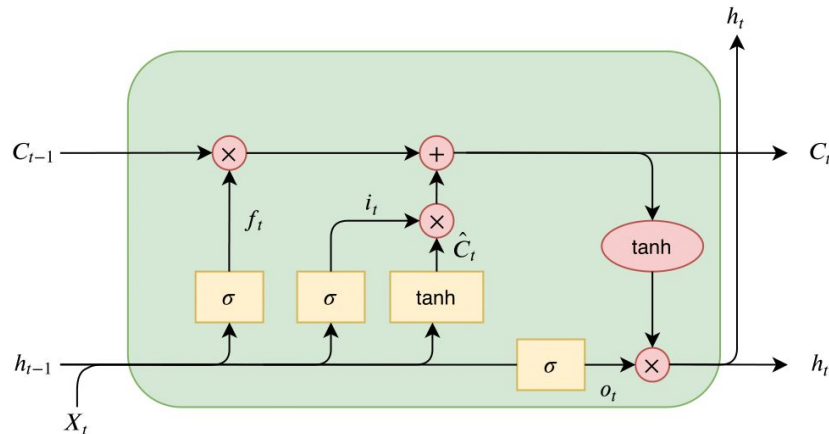
This is a way for models to focus on certain parts of the input data while lowering the importance of other parts. Determining the importance of words beforehand makes the inference process much simpler and therefore more accurate.

Each word in the input sequence is associated with a query vector and a key vector (trained). Determining the importance of a word to other words is done by taking a query vector attributed to a specific word in the sequence performing a dot product on each key of the other words. Then we softmax these scores to get attn weights.

RNNs

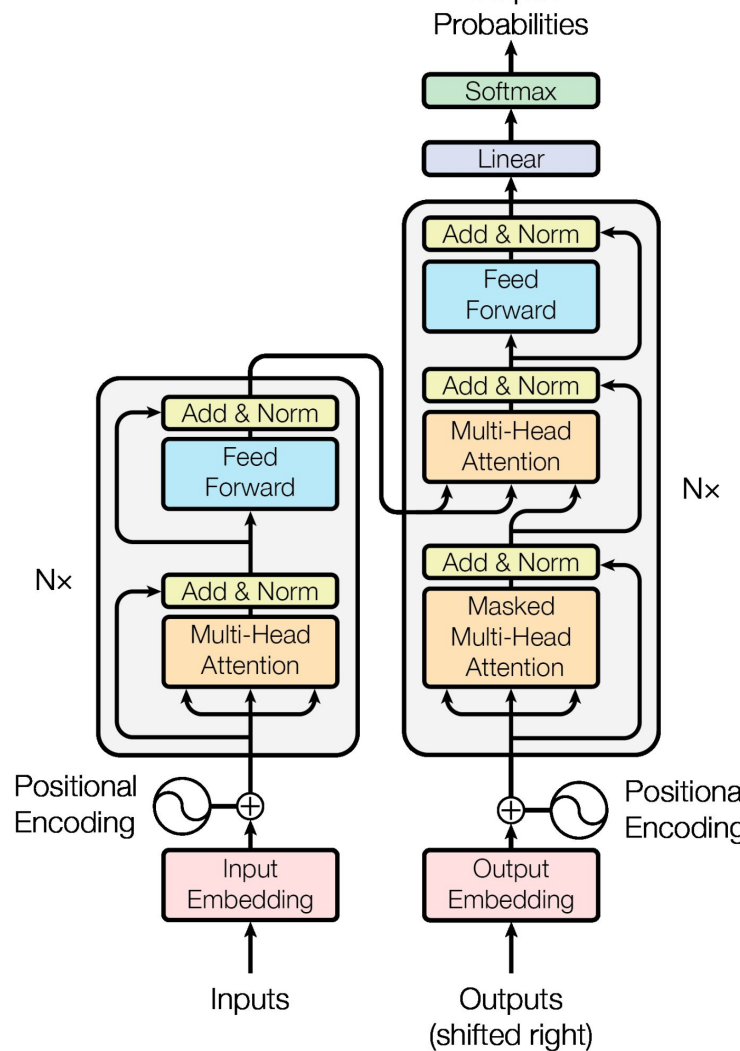
Traditional RNNs and LSTMs aim to solve a common problem of sequences of inputs corresponding to sequences of outputs, without a one-to-one mapping of between them.

To achieve this, some sort of memory mechanism is needed. Previous inputs should affect new inputs in a sequence. There are a number of ways to do this.



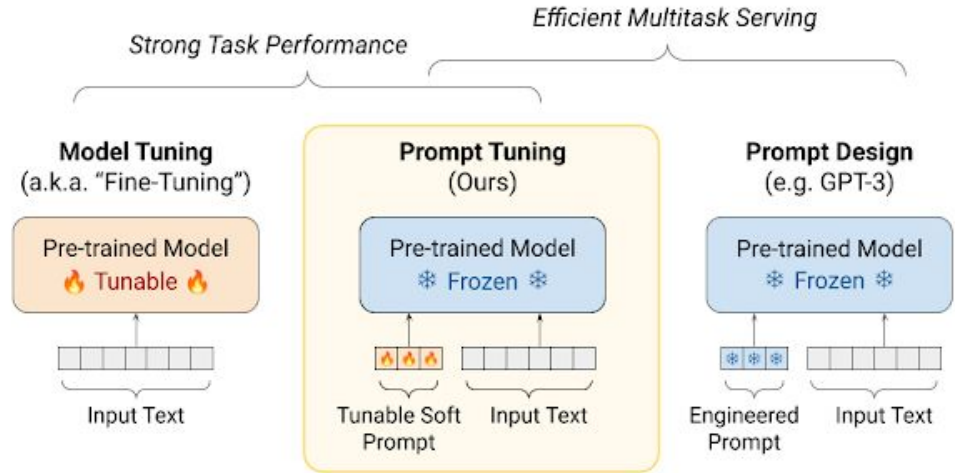
Transformers

Process: First, encode each new word using the encoding of the previous words as extra input. Then, the model uses self-attention to weigh the significance of each part of the input data differently and provide context for any position in the input sequence. This allows the model to generate a probability distribution over the next word in the sequence. The model selects the most likely next word and generates a new vector representation based on the new input. This process is repeated until the desired length of text has been developed.



What is a Generative Pre-trained Transformer (GPT)

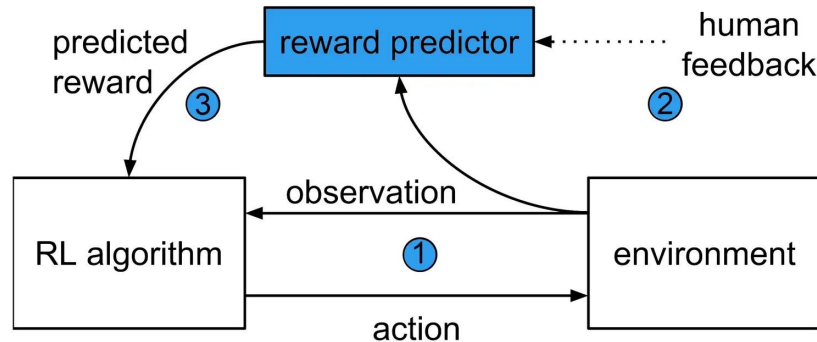
GPT models are based on a transformer architecture that has been pre-trained on vast amounts of text data using unsupervised learning. The pre-training process involves training the model to predict missing words or next words in a sentence, and then fine-tuning the model on a specific downstream task such as language translation, text classification, or question answering.



Reinforcement Learning in ChatGPT

ChatGPT is based on the original GPT-3 model, but has been further trained by using human feedback to guide the learning process with the goal of fixing the model's misalignment issues and improving its request responses.

The technique used is called Reinforcement Learning from Human Feedback (RLHF). We newly train a reward model based on human demonstrations. Then fine-tune the language model with reinforcement learning that uses the reward function.



Some numbers

300 billion words, ~570 GB of carefully selected training data.

GPT 3: ~200 billion parameters

GPT 3.5: ~10 trillion parameters

GPT 4: ~100 trillion parameters

GPT 5: ???

Discussion

How well can LLM do on inputs it hasn't seen?

How can we remove social biases?

How can we prevent adversarial inputs?



Horace He

@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

1/4

g's Race	implementation, math				
nd Chocolate	implementation, math				
triangle!	brute force, geometry, math				
	greedy, implementation, math				
Numbers	brute force				
ine Line	implementation				
or Stairs?	implementation				
Loves 3 I	math				
s	implementation, math				
	greedy, implementation, sortings				
	greedy, implementation				
Cat?	implementation, strings				
Actions	data structures, greedy, implementation, math				
Interview Problem	brute force, implementation, strings				
vers	brute force, implementation, strings				
nd Suffix Array	strings				
ther Promotion	greedy, math				
Forces	greedy, sortings				
d and Append	implementation, two pointers				
ng Directions	geometry, implementation				