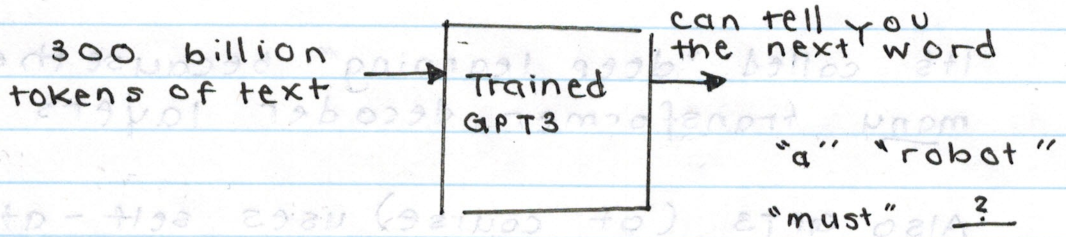
From Jay Alammar's blog.

## HOW GPT3 WORKS (WITH PICS!)

GPT3 is a massive language model with this general purpose: pass it text as input → model outputs things it learned during its training period (this process has been completed, @ $4.6 million).

E.g. objective:



### HOW TO MAKE TRAINING EXAMPLES

Use the "window technique" which "zooms out" or adds a word to string of text

the window is also callet the "context window". GPT3's window is 2048 tokens wide.

| Eg / | Next word / Correct output |
|---|---|
| ① second law | of |
| ② second law of | robotics |
| ③ second law of robotics | : |
| ④ second law of robotics : | a |
| ⑤ second law of robotics: a | robot |

If GPT3 gets the output wrong during this training, that's OK. We calculate the error & update the model. ∴ each new version of the model is more accurate than the last.

Note: when we say "prediction" we mean matrix manipulation.

i.e. "each sucessive set of training leads to better predictions".

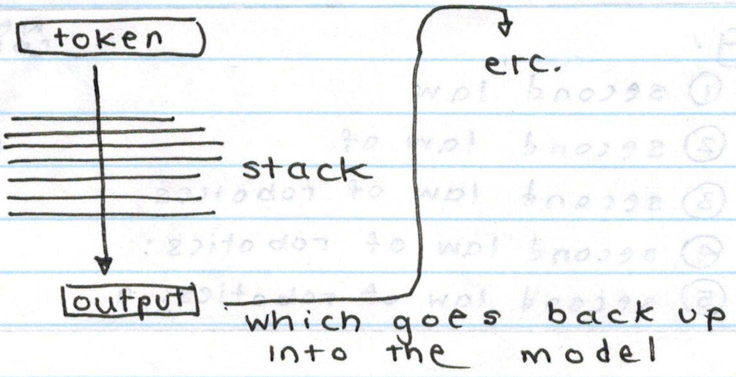
## IN-DEPTH LOOK AT PREDICTION AFTER GPT3 is TRAINED

word → vector
vector → compute* prediction as vector
use computed prediction $\xrightarrow{\text{turn back into}}$ word output

It's called "deep learning" because there are many transformer decoder layers (deep!)

Also GPT3 (of course) uses self-attention, particularly alternating dense- & sparse-attention layers.

Every token goes through all layers in stack.


token
stack
output
which goes back up into the model
etc.