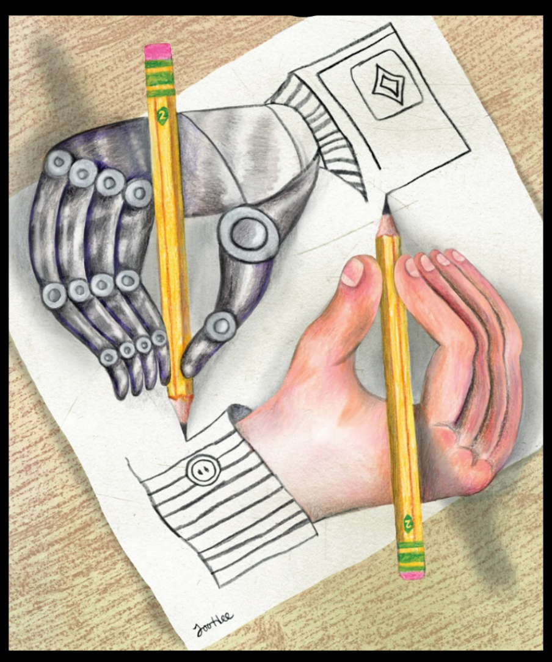
Self-Attention Mechanism of ChatGPT

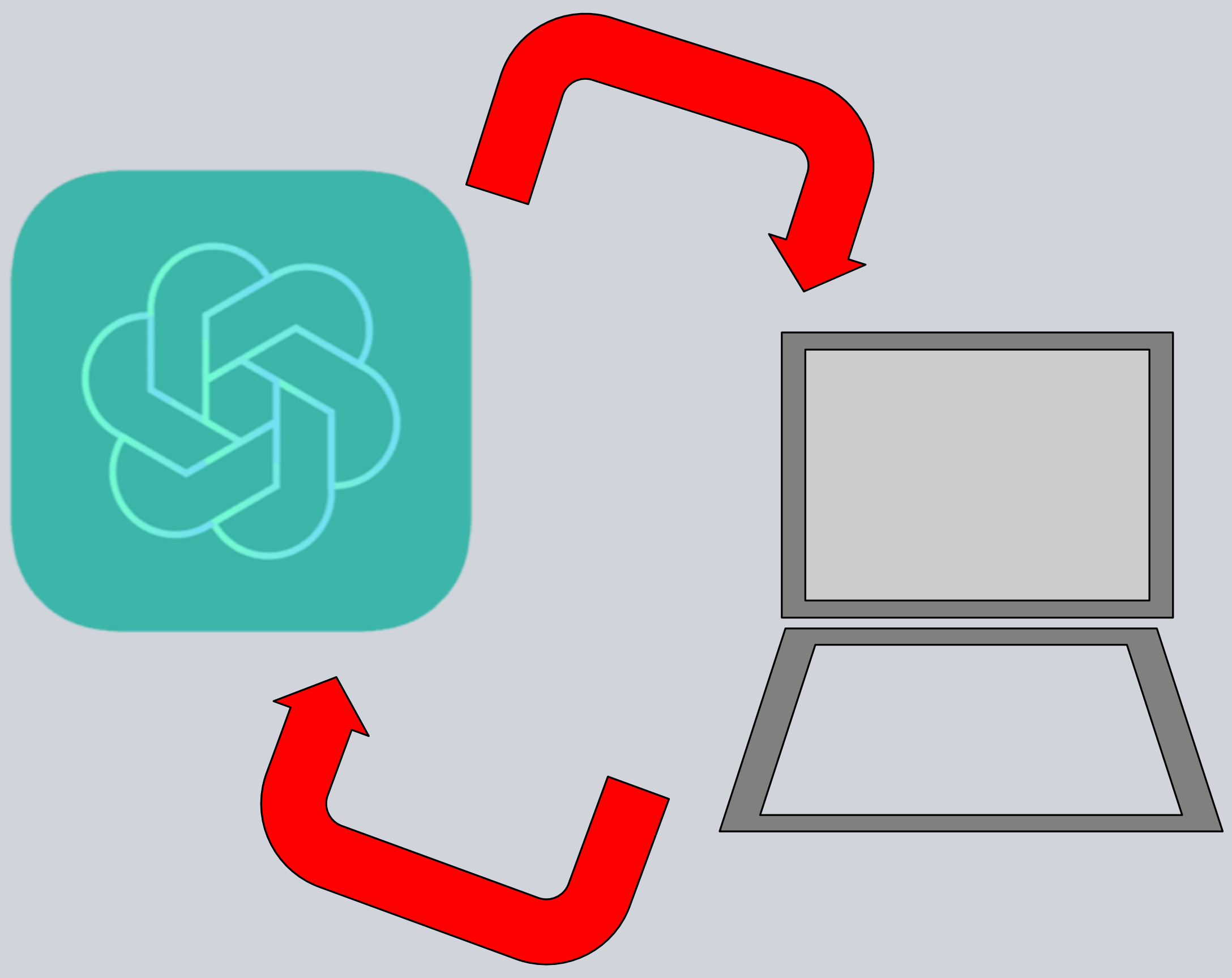
Team Members: Fardeen Abir, Nathan Campos, Anthony Edeza, Jose De Flores Santiago, Edward (Jong Yeon Kim) Fjura, Kenneth Lieu, Kevin Mateo, Michael Nguyen, Roberto Reyes, Maggie Yang
Faculty Advisor: Dr. Yuqing Zhu
Computer Science Liaison: Dr. Russ Abbott
Department(s) of Computer Science
College of Engineering, Computer Science, and Technology
California State University, Los Angeles



Background

ChatGPT is a language model created by OpenAI, based on the GPT-3.5 architecture. It is a type of artificial intelligence that can understand and generate natural language, which means it can read and write like a human would.

Based on Andrej Karpathy's "Let's build GPT: from scratch, in code, spelled out." The team has created a report that contains the simplified understanding of the core mechanism that makes ChatGPT work.



Self-Attention

The overall goal of this code is to construct for each token a picture (in terms of features) of the token that will follow it. This picture will be used (in code not included here) to find tokens that best match that feature portrait.

Overview

Tokens

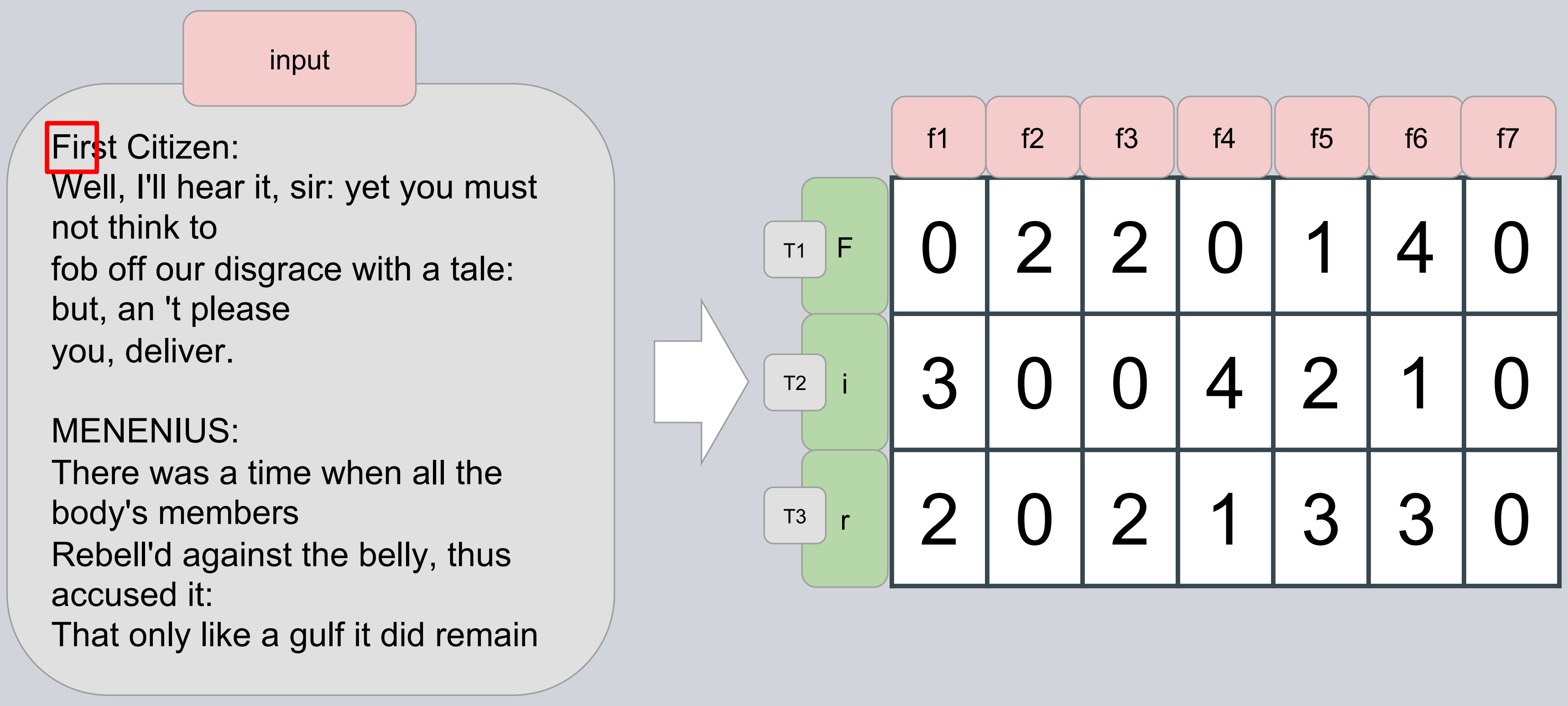
Weights

Output

Tokens

Each token t gives the tokens that precede it weights based on the extent to which those preceding tokens help token t predict the token that follows it. In other words:

- Each token t paints a picture of what preceding tokens should look like (in terms of their features) so that those preceding tokens will be useful to t for predicting the token that will follow t. That picture is expressed as a vector called q (for query).
- Each token paints a picture of itself, i.e., a self-portrait, that characterizes that token in terms of its features. These self-portraits are expressed as vectors called k (for key).



input

First Citizen:
Well, I'll hear it, sir: yet you must not think to
fob off our disgrace with a tale: but, an 't please
you, deliver.

MENENIUS:
There was a time when all the body's members
Rebell'd against the belly, thus accused it:
That only like a gulf it did remain

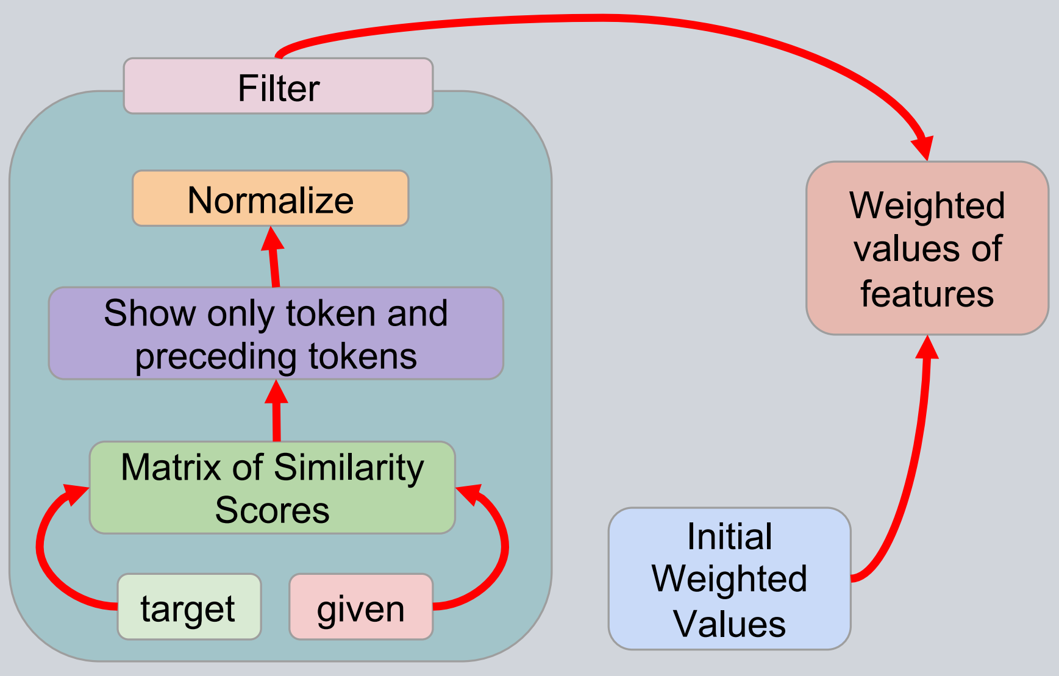
		f1	f2	f3	f4	f5	f6	f7
T1	F	0	2	2	0	1	4	0
T2	i	3	0	0	4	2	1	0
T3	r	2	0	2	1	3	3	0

Weights

Independently, each token paints a picture of the features that characterize tokens that it expects will appear sometime in the future. This portrait is expressed as the value (v) vector. In other words, a token's v vector is its way of setting the context for what comes next.

weights =

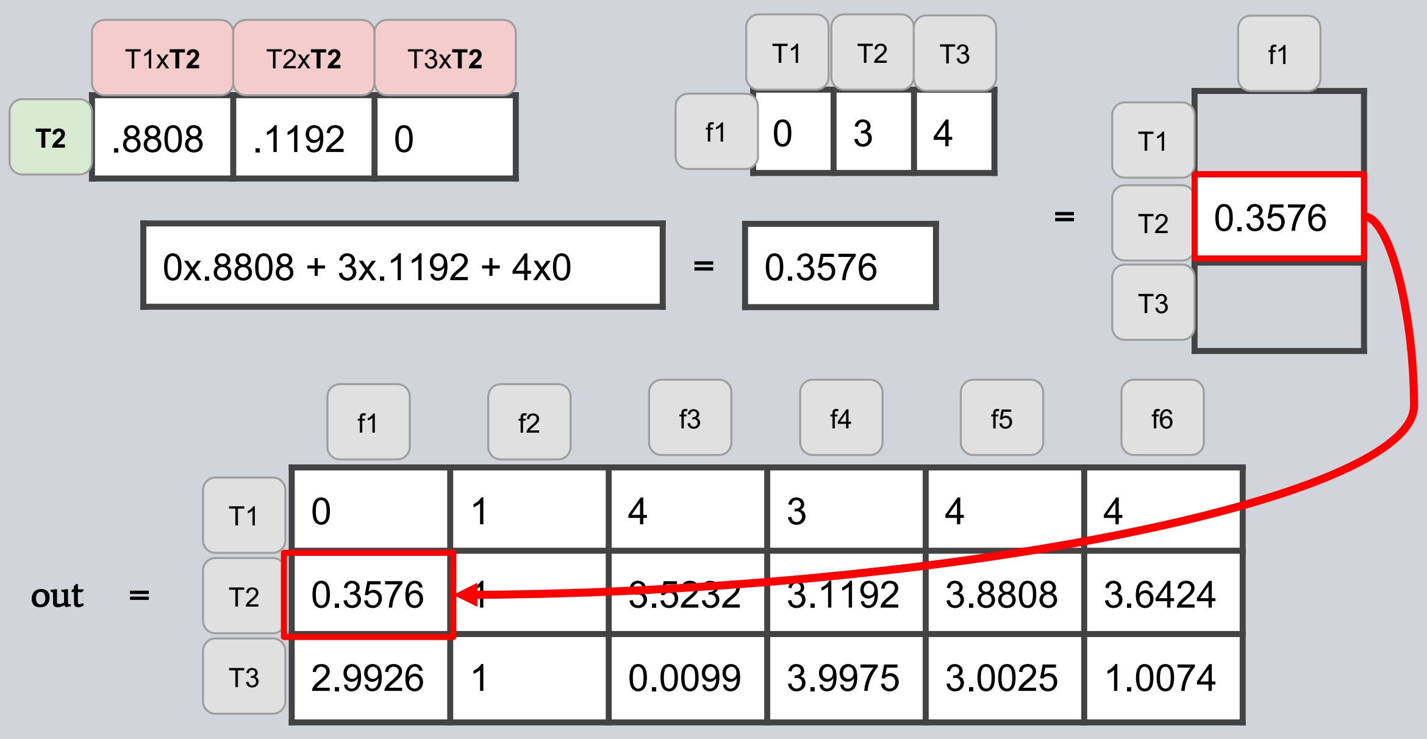
	T1	T2	T3
T1	47	57	45
T2	53	51	37
T3	41	47	31



Filter
Normalize
Show only token and preceding tokens
Matrix of Similarity Scores
target
given
Weighted values of features
Initial Weighted Values

Output

We combine these two collections of weights to achieve our goal: enable token t to characterize as well as possible the features of the token that will follow it. Dot-product of: **(a)** the weights token t gives the preceding tokens and **(b)** the weights feature f is given by those preceding tokens.



	T1x**T2**	T2x**T2**	T3x**T2**
T2	.8808	.1192	0

	T1	T2	T3
f1	0	3	4

0x.8808 + 3x.1192 + 4x0 = 0.3576

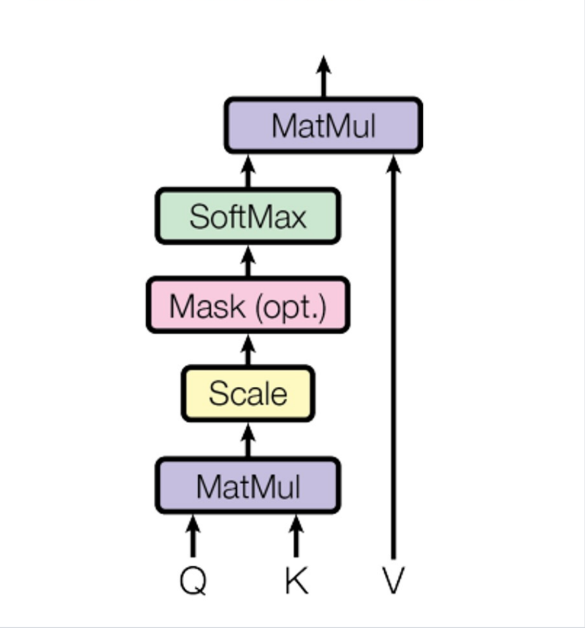
=

	f1
T1	
T2	0.3576
T3	

out =

	f1	f2	f3	f4	f5	f6
T1	0	1	4	3	4	4
T2	0.3576	1	3.5232	3.1192	3.8808	3.6424
T3	2.9926	1	0.0099	3.9975	3.0025	1.0074

Scaled Dot-Product Attention



MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V