Architecture neuronale pour données textuelles

Janvier 22, 2025

Familles de modèles classiques en apprentissage supervisé

- Modèles de régression pénalisée
- Arbres de décision
- Forêts aléatoires et gradient boosting
- Réseaux de neurones profonds (ou pas)
- Support vector machines (modèles à marge de séparation maximale)

Nous avons besoin d'un vecteur de variables explicatives $X \in \bigtimes_{j=1}^{d} \mathcal{X}_j$.

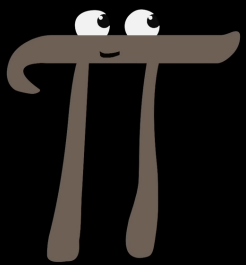
Révolution du mécanisme d'attention en 2018

Équivalente à la révolution des couches de convolution en classification d'images (AlexNet , concours du 30 septembre 2012 au Canada sur la base de données ImageNet).

Les indispensables

- Tokenizer : l'idée est de garder entier les mots très utilisés, mais de découper les mots transformés par la grammaire
- Embedding ou plongement : transformer nos jetons (numérotation) en vecteur dans $\mathbb{R}^d$
- Architecture neuronale dite **transformer**

Generative Pre-trained Transformer



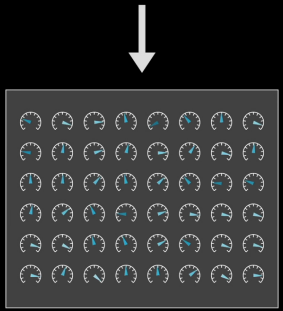
Generative Pre-trained Transformer

The most effective way to learn computer science is to actively engage with the material, practice regularly, and seek help when needed. Here are some specific steps you

Generative Pre-trained Transformer

The most effective way to learn computer science is to actively engage with the material, practice regularly, and seek help when needed. Here are some specific steps you can take to improve your computer science skills: 1. Take online courses: There are many online courses

it for a fact, years ago, in the quiet corner in Soho, that this precious brother had spent her money and left her! He was saying the affectionate word, however, with a far more grudging condescension and patronage than he could have shown if their relative merits and positions had



Weights

What defines the model

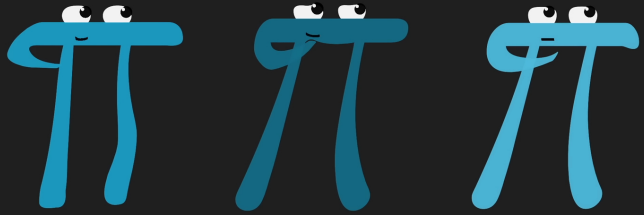
$$\begin{bmatrix} +1.0 & +4.3 & +2.0 & +0.9 & -1.5 & +2.9 & -1.2 & +7.8 \\ +9.2 & -2.3 & +5.8 & +0.6 & +1.3 & +8.4 & -8.5 & -8.2 \\ -9.5 & +6.6 & +5.5 & +7.3 & +9.5 & +5.9 & -0.8 & +5.6 \\ -7.6 & +2.8 & -7.1 & +8.8 & +0.4 & -1.7 & -4.7 & +5.4 \\ -0.9 & +1.4 & -9.5 & +2.3 & +2.2 & +2.3 & +8.8 & +3.6 \\ -2.8 & -1.2 & +3.9 & -8.7 & +3.3 & +3.4 & -5.7 & -7.3 \end{bmatrix} \quad \begin{bmatrix} -3.7 & -2.7 & +1.4 & -1.2 & +9.7 & -7.9 & -5.8 & -6.7 \\ +3.0 & -4.9 & -0.7 & -5.1 & -6.8 & -7.7 & +3.1 & -7.2 \\ -6.0 & -2.6 & +6.4 & -8.0 & +6.7 & -8.0 & +9.4 & -0.6 \\ +9.4 & +2.1 & +4.7 & -9.1 & -4.3 & -7.5 & -4.0 & -7.5 \\ -3.6 & -1.7 & -8.6 & +3.8 & +1.3 & -4.6 & +0.5 & -8.0 \\ +1.5 & +8.5 & -3.6 & +3.3 & -7.3 & +4.3 & -4.2 & -6.3 \end{bmatrix}$$

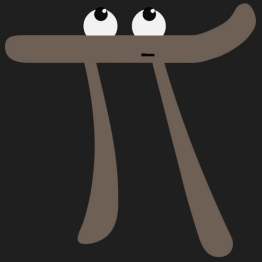
$$\begin{bmatrix} +1.7 & -9.5 & +6.5 & -9.8 & +3.5 & -4.6 & +4.7 & +9.2 \\ -5.0 & +1.5 & +1.8 & +1.4 & -5.5 & +9.0 & -1.0 & +6.9 \\ +3.9 & -4.0 & +6.2 & -2.0 & +7.5 & +1.6 & +7.6 & +3.8 \\ +4.5 & +0.0 & +9.0 & +2.9 & -1.5 & +2.1 & -9.5 & -3.9 \\ +3.2 & -4.2 & +2.3 & -1.4 & -7.2 & -4.0 & +1.4 & +1.8 \\ +1.5 & +3.0 & +3.0 & -1.4 & +7.9 & -2.6 & -1.3 & +7.8 \end{bmatrix} \quad \begin{bmatrix} +6.1 & +4.0 & -7.9 & +8.3 & +4.2 & +9.9 & -6.9 & +7.3 \\ -6.7 & +2.3 & -7.4 & +6.9 & +6.1 & +1.4 & -1.8 & -8.5 \\ +3.9 & -0.9 & +4.4 & +7.3 & +9.4 & +7.0 & -9.7 & -2.8 \\ +4.6 & -6.5 & +0.4 & -8.8 & -5.9 & -9.5 & +5.8 & -5.5 \\ -3.1 & +8.5 & +4.0 & -9.3 & -6.6 & +2.4 & +1.5 & -5.2 \\ +8.6 & +2.3 & +0.7 & +1.8 & +4.6 & -3.7 & -2.0 & -5.7 \end{bmatrix}$$

Data

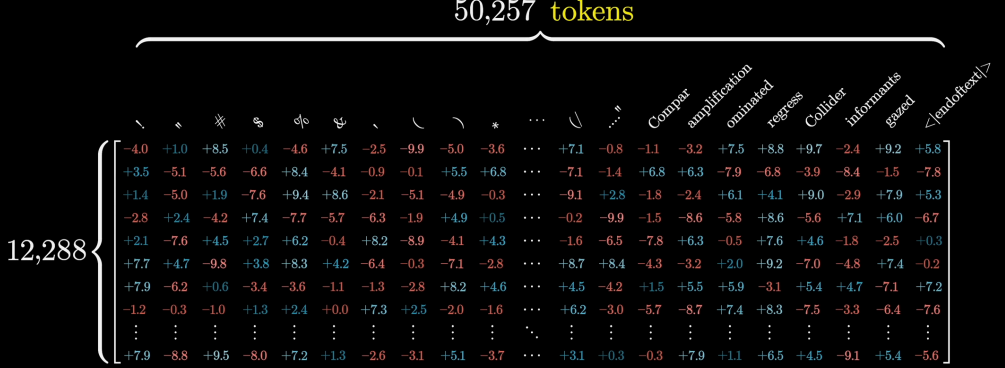
What the model processes

$$\begin{bmatrix} -6.2 \\ +8.8 \\ +4.7 \\ -0.2 \\ -5.4 \\ -4.9 \\ -8.8 \\ -1.3 \end{bmatrix} \quad \begin{bmatrix} -3.7 \\ +3.9 \\ -2.4 \\ -6.3 \\ -9.4 \\ -8.6 \\ +3.6 \\ -0.9 \end{bmatrix} \quad \begin{bmatrix} +0.7 \\ +7.9 \\ +9.7 \\ -5.6 \\ +3.2 \\ -4.7 \\ -9.5 \\ +5.1 \end{bmatrix} \quad \begin{bmatrix} -3.6 \\ -2.3 \\ +1.7 \\ +6.6 \\ +2.6 \\ +7.4 \\ -4.5 \\ +5.9 \end{bmatrix} \quad \begin{bmatrix} -6.2 \\ +9.0 \\ -9.4 \\ +3.7 \\ +8.9 \\ +4.6 \\ -4.9 \\ -5.7 \end{bmatrix} \quad \begin{bmatrix} +0.4 \\ -9.4 \\ -5.8 \\ -1.5 \\ -2.5 \\ -0.7 \\ -4.4 \\ +1.7 \end{bmatrix} \quad \begin{bmatrix} +7.2 \\ -7.6 \\ +0.3 \\ -7.3 \\ +4.3 \\ -2.1 \\ +1.3 \\ -6.3 \end{bmatrix} \quad \begin{bmatrix} -7.0 \\ -0.2 \\ -2.9 \\ +8.7 \\ +5.3 \\ +4.9 \\ +8.0 \\ -8.2 \end{bmatrix}$$



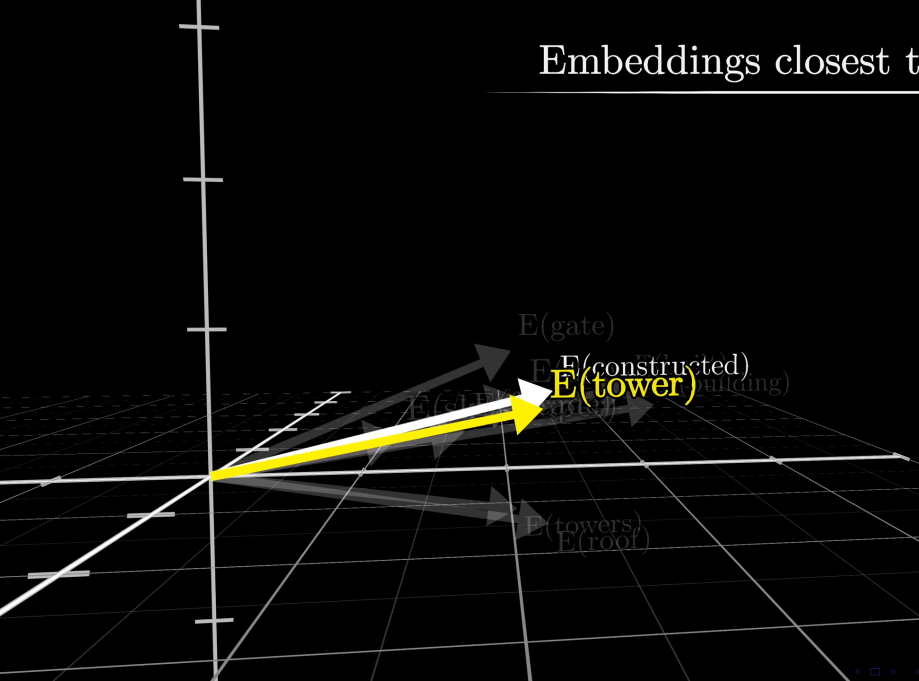


Total parameters $= 12{,}288 \times 50{,}257 = \boxed{617{,}558{,}016}$



$W_E =$ Embedding matrix

Embeddings closest to E(tower)



E(towers)

E(gate)

E(building)

E(skyscraper)

E(roof)

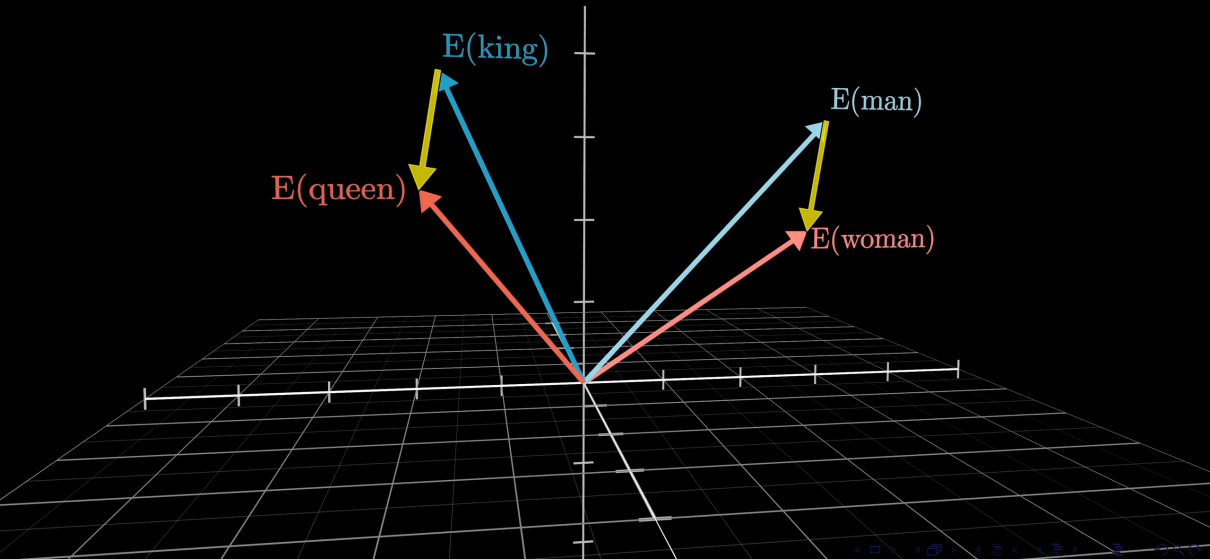
E(built)

E(dome)

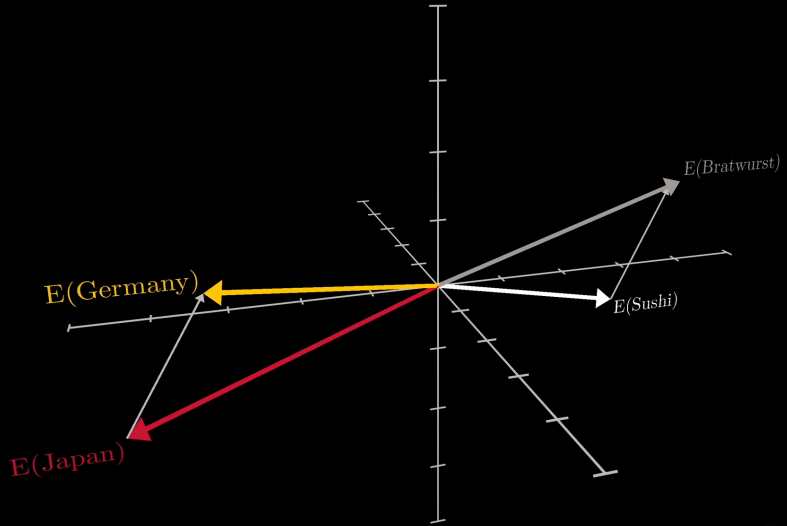
E(facade)

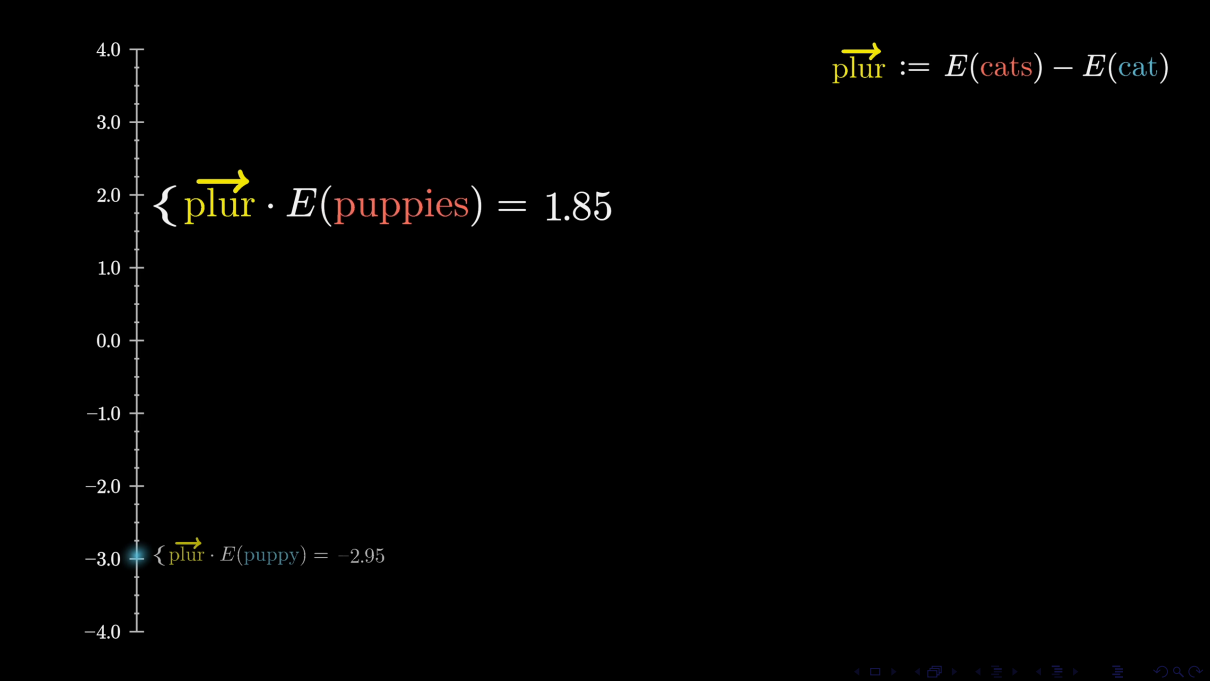
E(constructed)

E(queen) - E(king) ≈ E(woman) - E(man)



E(Sushi) + E(Germany) - E(Japan) ≈ E(Bratwurst)




$\overrightarrow{\text{plur}} := E(\text{cats}) - E(\text{cat})$
4.0
3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−4.0
$\{\overrightarrow{\text{plur}} \cdot E(\text{puppies}) = 1.85$
$\{\overrightarrow{\text{plur}} \cdot E(\text{puppy}) = -2.95$

$$E(\text{queen}) \approx E(\text{king}) + E(\text{woman}) - E(\text{man})$$

quill

Harry Potter was a highly unusual boy in many ways. For one thing, he hated the summer holidays more than any other time of year. For another, he really wanted to do his homework but was forced to do it in secret, in the dead of night. And he also happened to be a wizard.

It was nearly midnight, and he was lying on his stomach in bed, the blankets drawn right over his head like a tent, a flashlight in one hand and a large leather-bound book (A History of Magic by Bathilda Bagshot) propped open against the pillow. Harry moved the tip of his eagle-feather quill down the page, frowning as he looked for something that would help him write his essay, "Witch Burning in the Fourteenth Century Was Completely Pointless discuss."

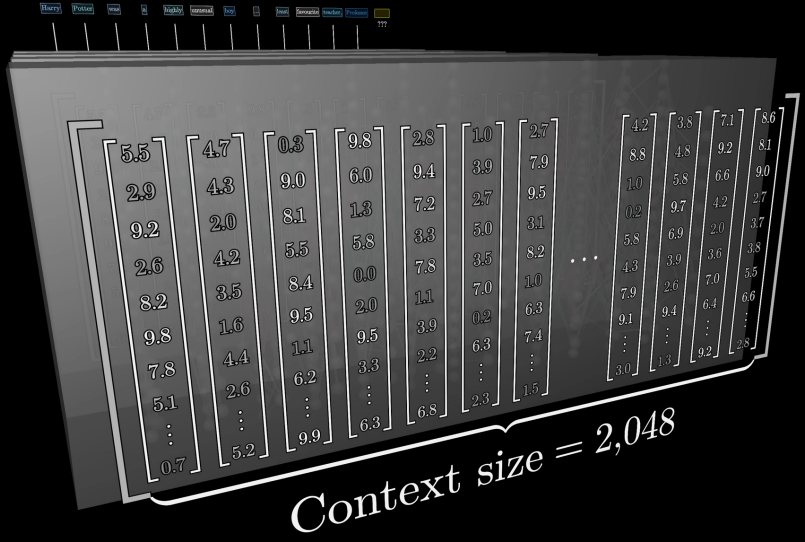
The quill paused at the top of a likely-looking paragraph. Harry Pushed his round glasses up the bridge of his nose, moved his flashlight closer to the book, and read:

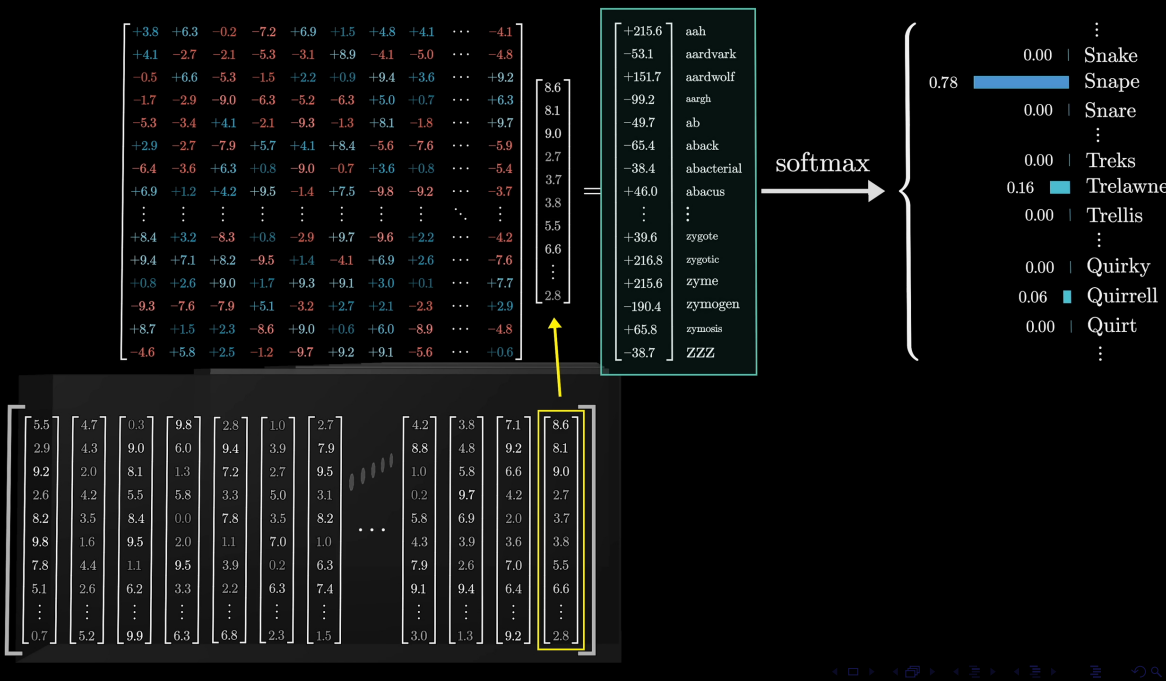
Non-magic people (more commonly known as Muggles) were particularly afraid of magic in medieval times, but not very good at recognizing it. On the rare occasion that they did catch a real witch or wizard, burning had no effect whatsoever. The witch or wizard would perform a basic Flame Freezing Charm and then pretend to shriek with pain while enjoying a gentle, tickling sensation. Indeed, Wendelin the Weird enjoyed being burned so much that she allowed herself to be caught no less than forty-seven times in various disguises.

Harry put his quill between his teeth and reached underneath his pillow for his ink bottle and a roll of parchment. Slowly and very carefully he unscrewed the ink bottle, dipped his quill into it, and began to write, pausing every now and then to listen, because if any of the Dursleys heard the scratching of his quill on their way to the bathroom, he'd probably find himself locked in the cupboard under the stairs for the rest of the summer.

The Dursley family of number four, Privet Drive, was the reason that Harry never enjoyed his summer holidays. Uncle Vernon, Aunt Petunia, and their son, Dudley, were Harry's only living relatives. They were Muggles, and they had a very medieval attitude toward magic. Harry's dead parents, who had been a witch and wizard themselves, were never mentioned under the Dursleys' roof For years, Aunt Petunia and Uncle Vernon had hoped that if they kept Harry as downtrodden as possible, they would be able to squash the magic out of him. To their fury, they had been unsuccessful. These days they lived in terror of anyone finding out that Harry had spent most of the last two years at Hogwarts School of Witchcraft and Wizardry. The most they could do, however, was to lock away Harry's spellbooks, wand, cauldron, and broomstick at the start of the summer break, and forbid him to talk to the neighbors.

This separation from his spellbooks had been a real problem for Harry, because his teachers at Hogwarts had given him a lot of holiday work. One of the essays, a particularly nasty one about shrinking potions, was for Harry's least favorite teacher, Professor


Harry
Potter
was
a
highly
rational
boy
.
It
was
Quirrell
teacher
Professor
???
5.5 2.9 9.2 2.6 8.2 9.8 7.8 5.1 ⋮ 0.7
4.7 4.3 2.0 4.2 3.5 1.6 4.4 2.6 ⋮ 5.2
0.3 9.0 8.1 5.5 8.4 9.5 1.1 6.2 ⋮ 9.9
9.8 6.0 1.3 5.8 0.0 2.0 9.5 3.3 ⋮ 6.3
2.8 9.4 7.2 3.3 7.8 1.1 3.9 2.2 ⋮ 6.8
1.0 3.9 2.7 5.0 3.5 7.0 0.2 6.3 ⋮ 2.3
2.7 7.9 9.5 3.1 8.2 1.0 6.3 7.4 ⋮ 1.5
...
4.2 8.8 1.0 0.2 5.8 4.3 7.9 9.1 ⋮ 3.0
3.8 4.8 5.8 9.7 6.9 3.9 2.6 9.4 ⋮ 1.3
7.1 9.2 6.6 4.2 2.0 3.6 7.0 6.4 ⋮ 9.2
8.6 8.1 9.0 2.7 3.7 3.8 5.5 6.6 ⋮ 2.8
Context size = 2,048


softmax
aah
aardvark
aardwolf
aargh
ab
aback
abacterial
abacus
zygote
zygotic
zyme
zymogen
zymosis
ZZZ
0.00 Snake
0.78 Snape
0.00 Snare
0.00 Treks
0.16 Trelawney
0.00 Trellis
0.00 Quirky
0.06 Quirrell
0.00 Quirt

Total weights: 175,181,291,520
Organized into 27,938 matrices

GPT-3

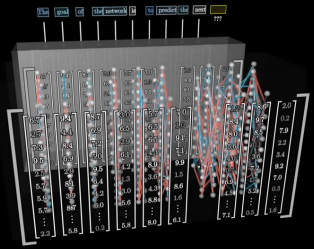
Embedding	$d\_embed^{12,288} * n\_vocab^{50,257} = 617,558,016$
Key	
Query	
Value	
Output	
Up-projection	
Down-projection	
Unembedding	$n\_vocab^{50,257} * d\_embed^{12,288} = 617,558,016$

1,235,116,032

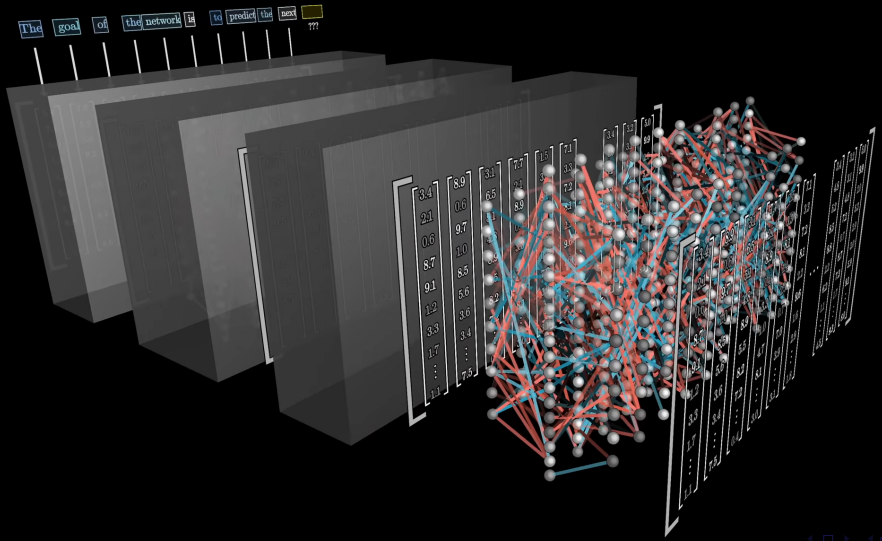
Generative Pre-trained Transformer

computer science is to practice coding regularly. Start by working on simple exercises and gradually move on to more complex projects. 4. Participate in coding challenges and competitions: Coding challenges and competitions provide a great opportunity to put your skills to the test and learn from others

drew back from the window, and the Doctor looked for explanation in his friend's ashy face. "They are," Mr. Lorry whispered the words, glancing fearfully round at the locked room, "murdering the prisoners. If you are sure of what you say; if you really have the power you think you



Transformer



Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to

A Convenient Lie

Let's | pretend | that | tokens | are | always | simply | words

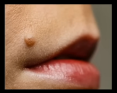
The Truth

This | process | (| known | fan | cifully | as | token | ization |) | frequently | subdiv | ides | words

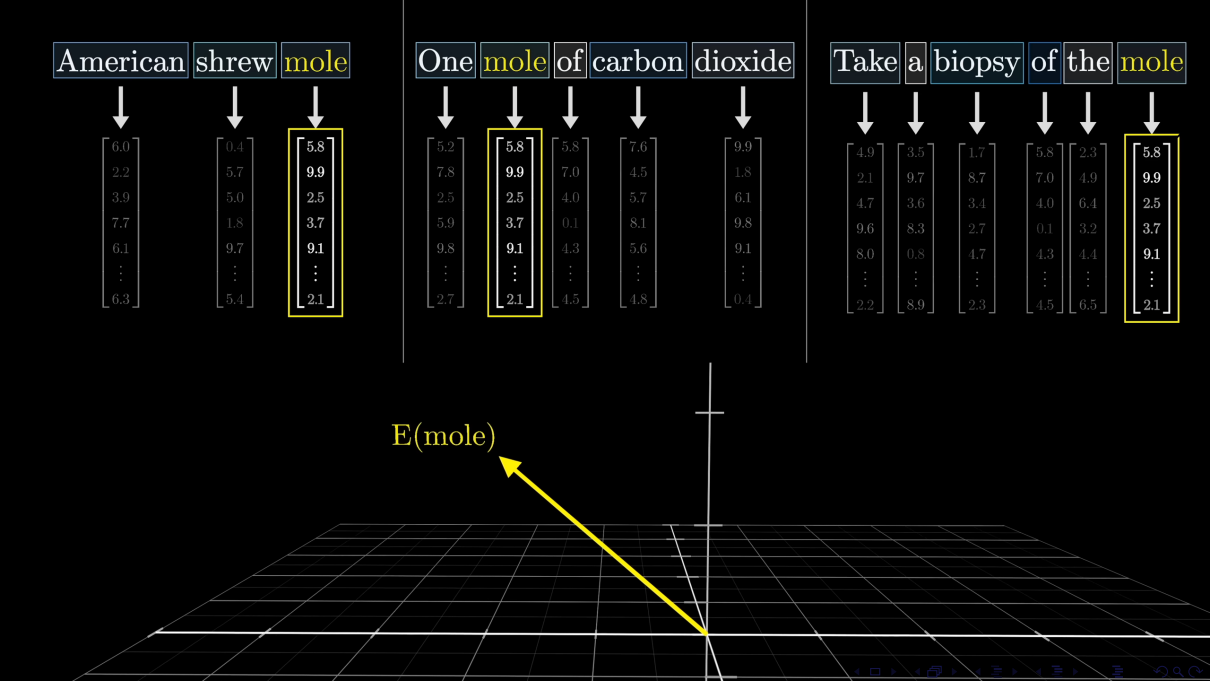
American shrew mole

6.02×10^{23}

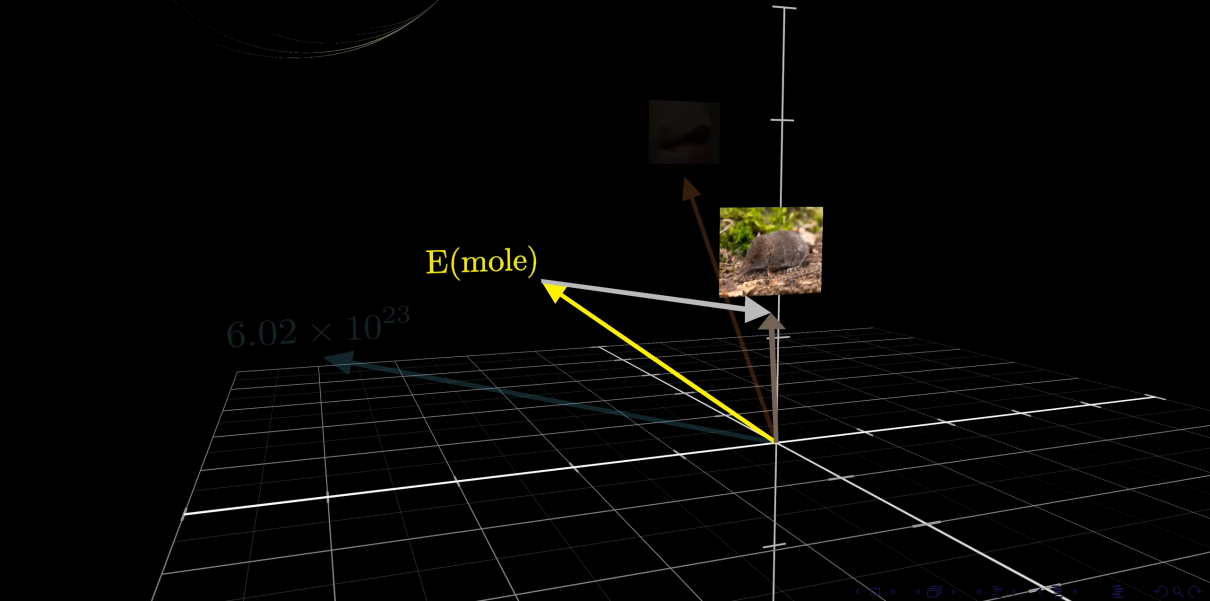
One mole of carbon dioxide



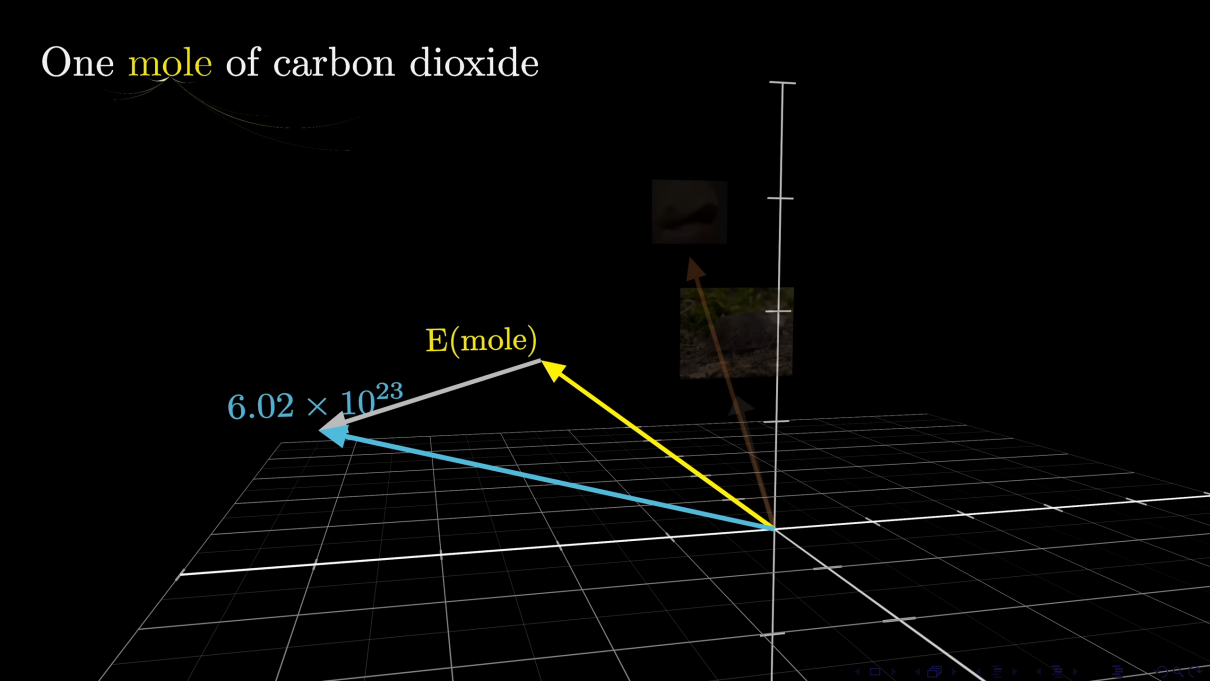
Take a biopsy of the mole


American shrew mole
One mole of carbon dioxide
Take a biopsy of the mole
E(mole)

American shrew mole



One mole of carbon dioxide



Short distance

One mole of carbon dioxide

Long distance

Harry Potter was a highly unusual boy in many ways. For one thing, he hated the summer holidays more than any other time of year. For another, he really wanted to do his homework but was forced to do it in secret, in the dead of night. And he also happened to be a wizard.

It was nearly midnight, and he was lying on his stomach in bed, the blankets drawn right over his head like a tent, a flashlight in one hand and a large leather-bound book (A History of Magic by Bathilda Bagshot) propped open against the pillow. Harry moved the tip of his eagle-feather quill down the page, frowning as he looked for something that would help him write his essay, "Witch Burning in the Fourteenth Century Was Completely Pointless discuss."

The quill paused at the top of a likely looking paragraph. Harry Pushed his round glasses up the bridge of his nose, moved his flashlight closer to the book, and read:

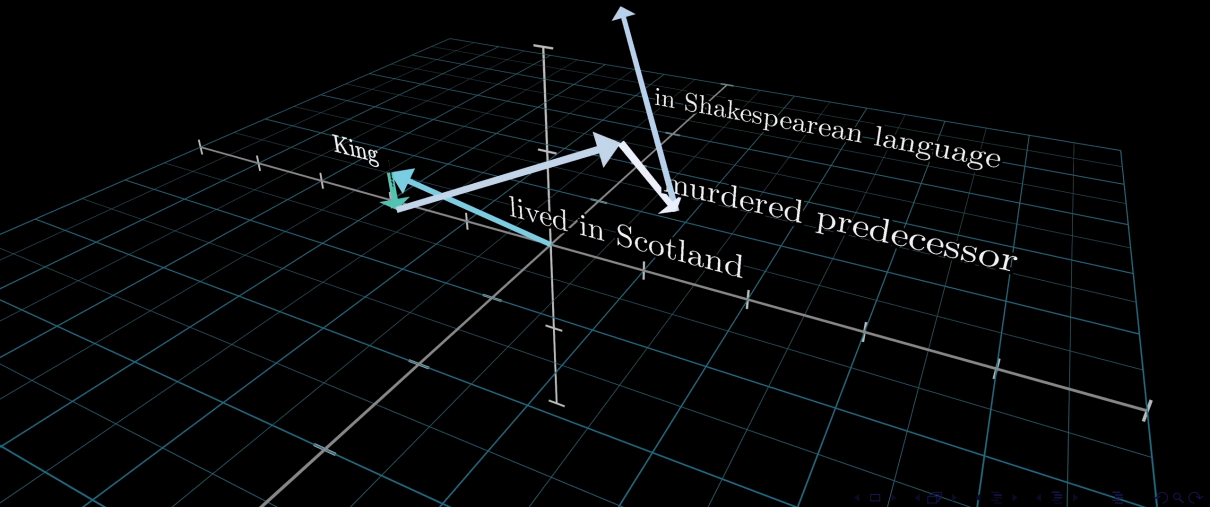
Non-magic people (more commonly known as Muggles) were particularly afraid of magic in medieval times, but not very good at recognizing it. On the rare occasion that they did catch a real witch or wizard, burning had no effect whatsoever. The witch or wizard would perform a basic Flame Freezing Charm and then pretend to shriek with pain while enjoying a gentle, tickling sensation. Indeed, Wendelin the Weird enjoyed being burned so much that she allowed herself to be caught no less than forty-seven times in various disguises.

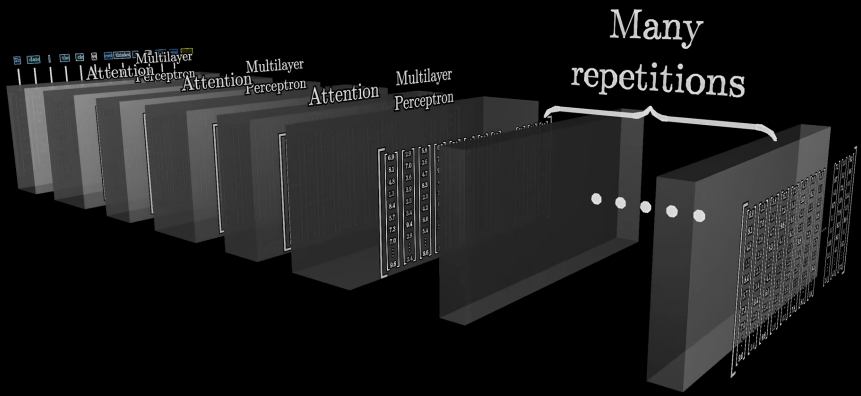
Harry put his quill between his teeth and reached underneath his pillow for his ink bottle and a roll of parchment. Slowly and very carefully he unscrewed the ink bottle, dipped his quill into it, and began to write, pausing every now and then to listen, because if any of the Dursleys heard the scratching of his quill on their way to the bathroom, he'd probably find himself locked in the cupboard under the stairs for the rest of the summer.

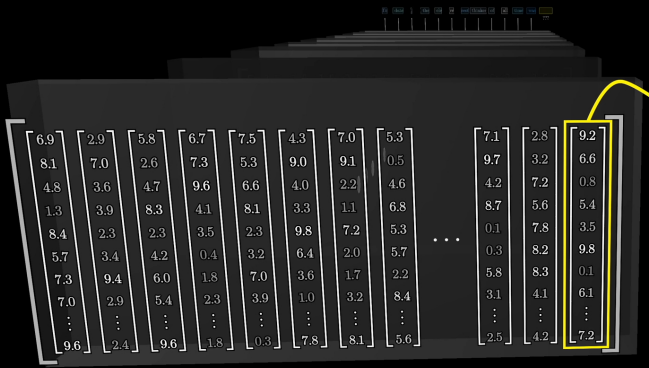
The Dursley family of number four, Privet Drive, was the reason that Harry never enjoyed his summer holidays. Uncle Vernon, Aunt Petunia, and their son, Dudley, were Harry's only living relatives. They were Muggles, and they had a very medieval attitude toward magic. Harry's dead parents, who had been a witch and wizard themselves, were never mentioned under the Dursleys' roof. For years, Aunt Petunia and Uncle Vernon had hoped that if they kept Harry as downtrodden as possible, they would be able to squash the magic out of him. To their fury, they had been unsuccessful. These days they lived in terror of anyone finding out that Harry had spent most of the last two years at Hogwarts School of Witchcraft and Wizardry. The most they could do, however, was to lock away Harry's spellbooks, wand, cauldron, and broomstick at the start of the summer break, and forbid him to talk to the neighbors.

This separation from his spellbooks had been a real problem for Harry, because his teachers at Hogwarts had given him a lot of holiday work. One of the essays, a particularly nasty one about shrinking potions, was for Harry's least favorite teacher, Professor

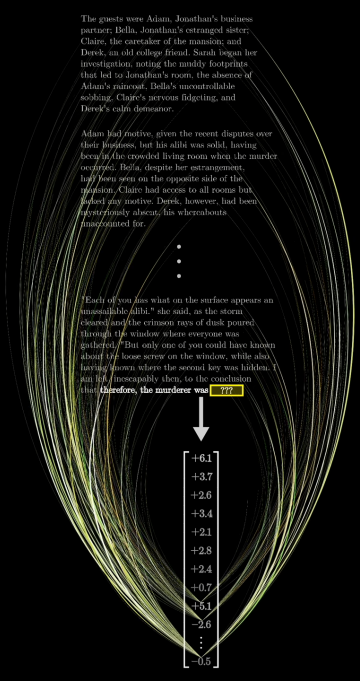
The King doth wake tonight and takes his rouse ...



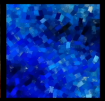

Attention
Multilayer Perceptron
Attention
Multilayer Perceptron
Attention
Multilayer Perceptron
Many repetitions



the	8.82%
probably	4.37%
John	4.04%
Sir	3.66%
Albert	3.63%
Ber	3.31%
a	2.90%
Isaac	2.01%
undoubtedly	1.58%
arguably	1.33%
Im	1.16%
Einstein	1.13%
Ludwig	1.04%


The guests were Adam, Jonathan's business partner; Bella, Jonathan's estranged sister; Claire, the caretaker of the mansion; and Derek, an old college friend. Sarah began her investigation, noting the muddy footprints that led to Jonathan's room, the absence of Adam's raincoat, Bella's uncontrollable sobbing, Claire's nervous fidgeting, and Derek's calm demeanor.
Adam had motive, given the recent disputes over their business, but his alibi was solid, having been in the crowded living room when the murder occurred. Bella, despite her estrangement, had been seen on the opposite side of the mansion. Claire had access to all rooms but lacked any motive. Derek, however, had been mysteriously absent, his whereabouts unaccounted for.
"Each of you has what on the surface appears an unassailable alibi," she said, as the storm cleared and the crimson rays of dusk poured through the window where everyone was gathered. "But only one of you could have known about the loose screw on the window, while also having known where the second key was hidden. I am left inescapably then, to the conclusion that therefore, the murderer was ???."
+6.1
+3.7
+2.6
+3.4
+2.1
+2.8
+2.4
+0.7
+5.1
−2.6
−0.5

a fluffy blue creature roamed the verdant forest


a	fluffy	blue	creature	roamed	the	verdant	forest
1	2	3	4	5	6	7	8

Position

12,288

1	2	3	4	5	6	7	8
6.0	5.6	8.8	1.6	4.5	5.2	0.3	7.2
0.2	5.8	8.0	6.1	7.1	0.5	1.6	3.1
3.0	5.7	7.0	1.2	8.6	2.0	6.2	3.9
6.5	6.5	1.0	8.4	9.7	0.2	5.7	2.1
2.9	6.5	9.1	8.0	8.5	7.9	2.4	1.8
6.1	4.3	7.1	5.6	0.1	2.2	9.2	9.3
4.2	8.9	9.9	4.0	3.6	3.4	6.1	7.3
1.3	3.6	1.5	0.7	7.2	9.2	5.3	4.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.0	4.3	8.6	6.9	1.7	7.0	5.8	2.3

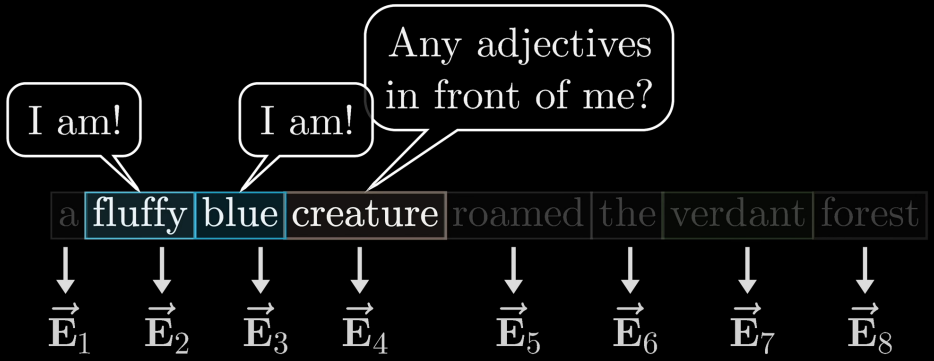


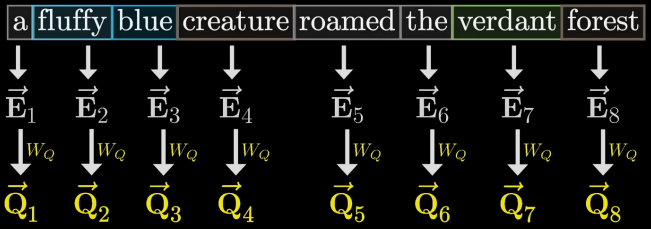


a | fluffy | blue | creature | roamed | the | verdant | forest

$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$

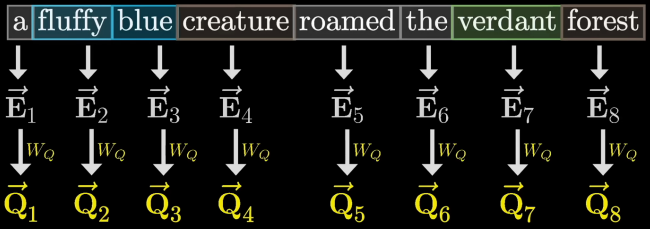
$\vec{E}_1'$ $\vec{E}_2'$ $\vec{E}_3'$ $\vec{E}_4'$ $\vec{E}_5'$ $\vec{E}_6'$ $\vec{E}_7'$ $\vec{E}_8'$



I am!
I am!
Any adjectives in front of me?
a
fluffy
blue
creature
roamed
the
verdant
forest
$\vec{\mathbf{E}}_1$
$\vec{\mathbf{E}}_2$
$\vec{\mathbf{E}}_3$
$\vec{\mathbf{E}}_4$
$\vec{\mathbf{E}}_5$
$\vec{\mathbf{E}}_6$
$\vec{\mathbf{E}}_7$
$\vec{\mathbf{E}}_8$



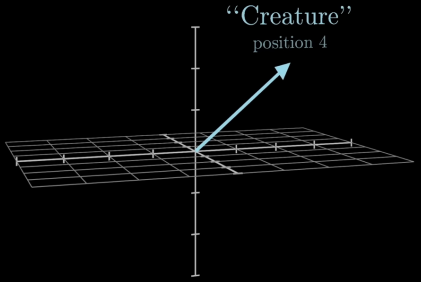
Any adjectives
in front of me?

$$W_Q \begin{bmatrix} +7.5 & -3.2 & +9.1 & -5.3 & +8.9 & +8.7 & +5.9 & +2.6 & +7.4 & -4.1 & \cdots & +2.3 \\ -9.6 & -3.0 & -7.0 & +9.5 & -0.4 & -0.1 & +2.8 & -2.6 & -7.2 & +6.4 & \cdots & +0.2 \\ -5.5 & -8.0 & +7.2 & +9.4 & +9.1 & +8.0 & +5.4 & -3.3 & -8.3 & -1.8 & \cdots & -7.3 \\ -8.8 & +4.5 & -9.7 & +5.4 & -7.0 & -8.3 & -8.1 & +3.4 & -5.0 & -1.6 & \cdots & +7.1 \\ +4.5 & -4.5 & -7.3 & -8.8 & -3.9 & -4.7 & -0.9 & +3.6 & +3.9 & -4.3 & \cdots & -6.3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -9.0 & +5.9 & -8.4 & +0.4 & -3.8 & +1.5 & +9.1 & +2.9 & -9.2 & -1.4 & \cdots & +0.7 \end{bmatrix} \vec{E}_i \begin{bmatrix} 2.9 \\ 2.4 \\ 1.0 \\ 0.2 \\ 9.2 \\ 6.6 \\ 7.8 \\ 2.8 \\ 5.8 \\ 0.6 \\ \vdots \\ 9.7 \end{bmatrix} = \vec{Q}_i \begin{bmatrix} +310.6 \\ -95.2 \\ -2.1 \\ -152.0 \\ -123.2 \\ \vdots \\ -12.7 \end{bmatrix}$$



Embedding space

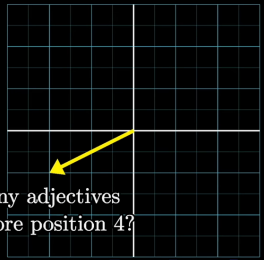
12,288-dimensional

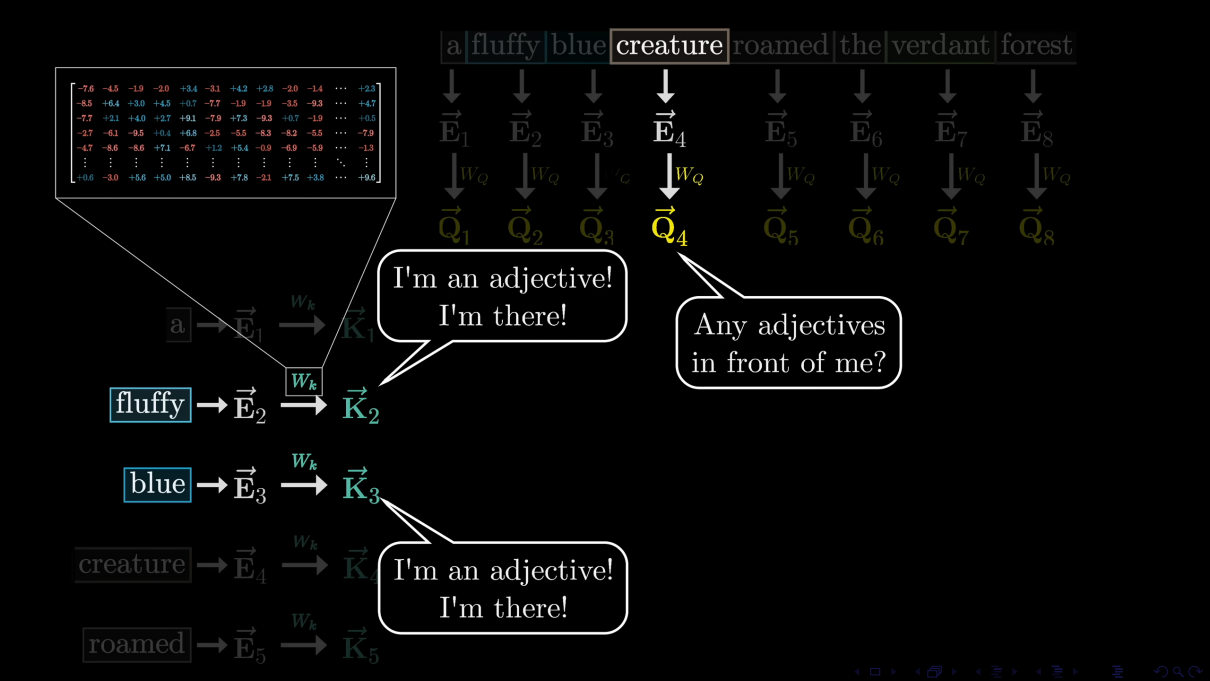


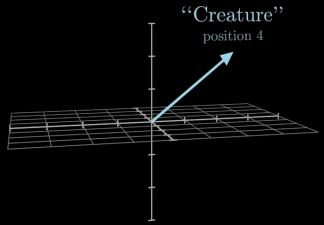
W_Q

Query/Key space

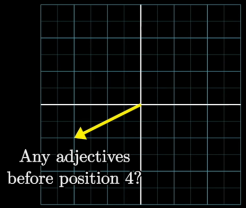
128-dimensional




a fluffy blue creature roamed the verdant forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
W_Q
$\vec{Q}_1$ $\vec{Q}_2$ $\vec{Q}_3$ $\vec{Q}_4$ $\vec{Q}_5$ $\vec{Q}_6$ $\vec{Q}_7$ $\vec{Q}_8$
I'm an adjective!
I'm there!
Any adjectives
in front of me?
W_k
fluffy → $\vec{E}_2$ → $\vec{K}_2$
blue → $\vec{E}_3$ → $\vec{K}_3$
I'm an adjective!
I'm there!



W_Q



a → $\vec{E}_1$ → W_k → $\vec{K}_1$

fluffy → $\vec{E}_2$ → W_k → $\vec{K}_2$

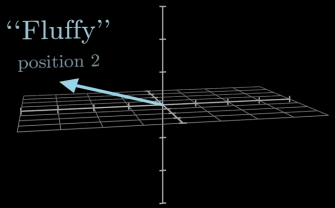
blue → $\vec{E}_3$ → W_k → $\vec{K}_3$

creature → $\vec{E}_4$ → W_k → $\vec{K}_4$

roamed → $\vec{E}_5$ → W_k → $\vec{K}_5$

Embedding space

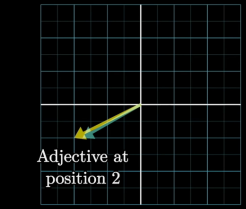
12,288-dimensional



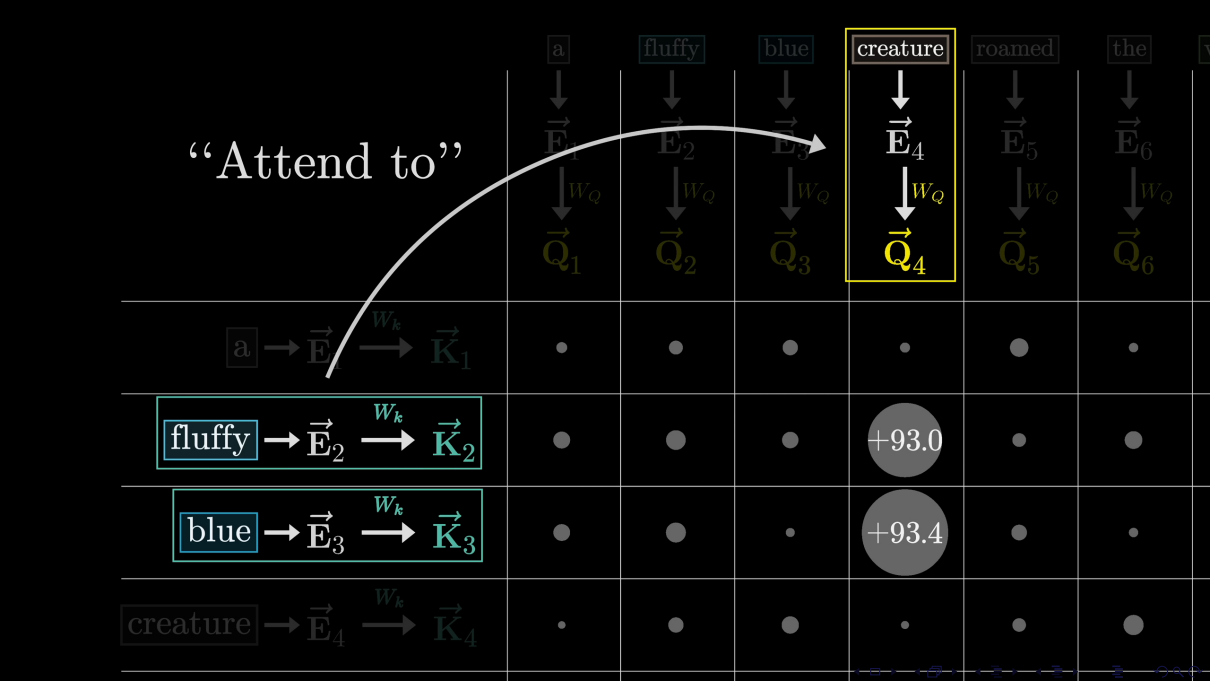
W_K

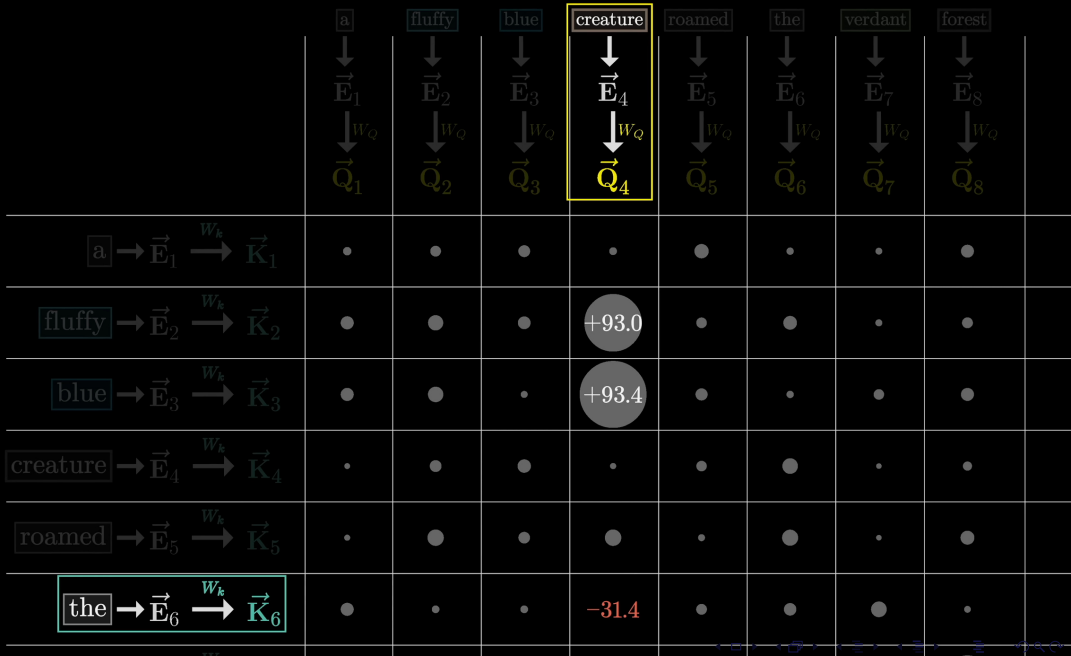
Query/Key space

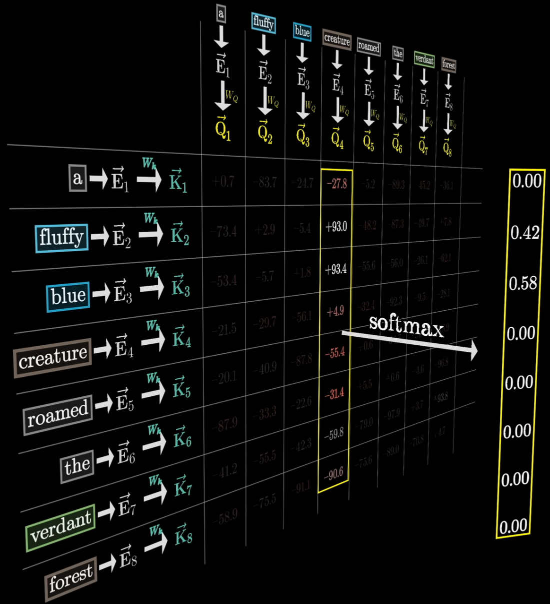
128-dimensional




a
fluffy
blue
creature
roamed
the
verdant
forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
W_Q
$\vec{Q}_1$ $\vec{Q}_2$ $\vec{Q}_3$ $\vec{Q}_4$ $\vec{Q}_5$ $\vec{Q}_6$ $\vec{Q}_7$ $\vec{Q}_8$
a → $\vec{E}_1$ → $\vec{K}_1$
fluffy → $\vec{E}_2$ → $\vec{K}_2$
blue → $\vec{E}_3$ → $\vec{K}_3$
creature → $\vec{E}_4$ → $\vec{K}_4$
roamed → $\vec{E}_5$ → $\vec{K}_5$
the → $\vec{E}_6$ → $\vec{K}_6$
verdant → $\vec{E}_7$ → $\vec{K}_7$
forest → $\vec{E}_8$ → $\vec{K}_8$
W_K
$\vec{K}_1 \cdot \vec{Q}_1$ $\vec{K}_1 \cdot \vec{Q}_2$ $\vec{K}_1 \cdot \vec{Q}_3$ $\vec{K}_1 \cdot \vec{Q}_4$ $\vec{K}_1 \cdot \vec{Q}_5$ $\vec{K}_1 \cdot \vec{Q}_6$ $\vec{K}_1 \cdot \vec{Q}_7$ $\vec{K}_1 \cdot \vec{Q}_8$
$\vec{K}_2 \cdot \vec{Q}_1$ $\vec{K}_2 \cdot \vec{Q}_2$ $\vec{K}_2 \cdot \vec{Q}_3$ $\vec{K}_2 \cdot \vec{Q}_4$ $\vec{K}_2 \cdot \vec{Q}_5$ $\vec{K}_2 \cdot \vec{Q}_6$ $\vec{K}_2 \cdot \vec{Q}_7$ $\vec{K}_2 \cdot \vec{Q}_8$
$\vec{K}_3 \cdot \vec{Q}_1$ $\vec{K}_3 \cdot \vec{Q}_2$ $\vec{K}_3 \cdot \vec{Q}_3$ $\vec{K}_3 \cdot \vec{Q}_4$ $\vec{K}_3 \cdot \vec{Q}_5$ $\vec{K}_3 \cdot \vec{Q}_6$ $\vec{K}_3 \cdot \vec{Q}_7$ $\vec{K}_3 \cdot \vec{Q}_8$
$\vec{K}_4 \cdot \vec{Q}_1$ $\vec{K}_4 \cdot \vec{Q}_2$ $\vec{K}_4 \cdot \vec{Q}_3$ $\vec{K}_4 \cdot \vec{Q}_4$ $\vec{K}_4 \cdot \vec{Q}_5$ $\vec{K}_4 \cdot \vec{Q}_6$ $\vec{K}_4 \cdot \vec{Q}_7$ $\vec{K}_4 \cdot \vec{Q}_8$
$\vec{K}_5 \cdot \vec{Q}_1$ $\vec{K}_5 \cdot \vec{Q}_2$ $\vec{K}_5 \cdot \vec{Q}_3$ $\vec{K}_5 \cdot \vec{Q}_4$ $\vec{K}_5 \cdot \vec{Q}_5$ $\vec{K}_5 \cdot \vec{Q}_6$ $\vec{K}_5 \cdot \vec{Q}_7$ $\vec{K}_5 \cdot \vec{Q}_8$
$\vec{K}_6 \cdot \vec{Q}_1$ $\vec{K}_6 \cdot \vec{Q}_2$ $\vec{K}_6 \cdot \vec{Q}_3$ $\vec{K}_6 \cdot \vec{Q}_4$ $\vec{K}_6 \cdot \vec{Q}_5$ $\vec{K}_6 \cdot \vec{Q}_6$ $\vec{K}_6 \cdot \vec{Q}_7$ $\vec{K}_6 \cdot \vec{Q}_8$
$\vec{K}_7 \cdot \vec{Q}_1$ $\vec{K}_7 \cdot \vec{Q}_2$ $\vec{K}_7 \cdot \vec{Q}_3$ $\vec{K}_7 \cdot \vec{Q}_4$ $\vec{K}_7 \cdot \vec{Q}_5$ $\vec{K}_7 \cdot \vec{Q}_6$ $\vec{K}_7 \cdot \vec{Q}_7$ $\vec{K}_7 \cdot \vec{Q}_8$
$\vec{K}_8 \cdot \vec{Q}_1$ $\vec{K}_8 \cdot \vec{Q}_2$ $\vec{K}_8 \cdot \vec{Q}_3$ $\vec{K}_8 \cdot \vec{Q}_4$ $\vec{K}_8 \cdot \vec{Q}_5$ $\vec{K}_8 \cdot \vec{Q}_6$ $\vec{K}_8 \cdot \vec{Q}_7$ $\vec{K}_8 \cdot \vec{Q}_8$


a
fluffy
blue
creature
roamed
the
"Attend to"
$\vec{E}_1$
$\vec{E}_2$
$\vec{E}_3$
$\vec{E}_4$
$\vec{E}_5$
$\vec{E}_6$
W_Q
$\vec{Q}_1$
$\vec{Q}_2$
$\vec{Q}_3$
$\vec{Q}_4$
$\vec{Q}_5$
$\vec{Q}_6$
W_k
$\vec{K}_1$
$\vec{K}_2$
$\vec{K}_3$
$\vec{K}_4$
+93.0
+93.4

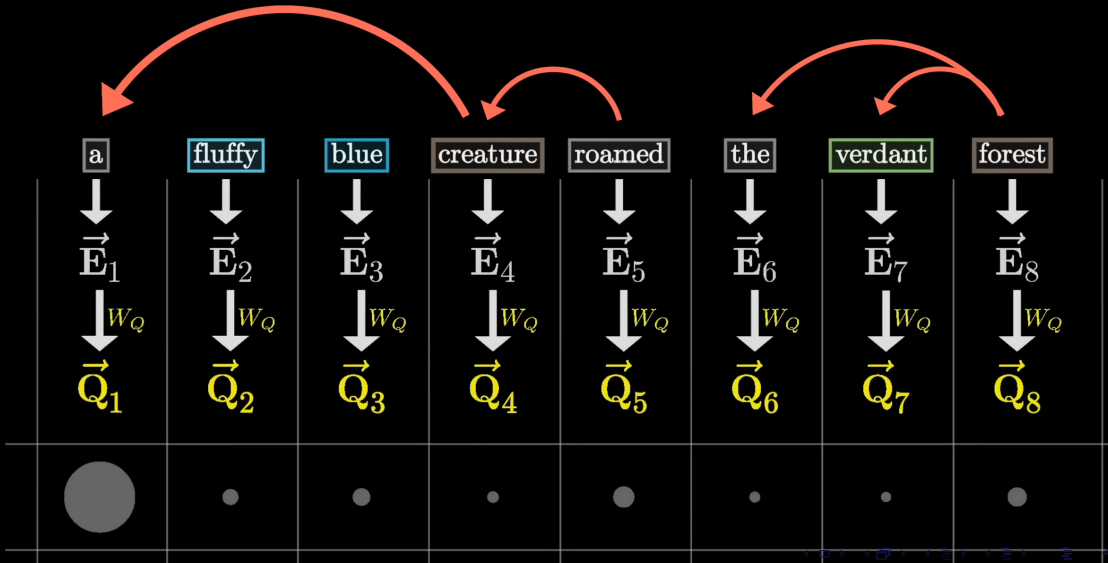

a
fluffy
blue
creature
roamed
the
verdant
forest
$\vec{E}_1$
$\vec{E}_2$
$\vec{E}_3$
$\vec{E}_4$
$\vec{E}_5$
$\vec{E}_6$
$\vec{E}_7$
$\vec{E}_8$
W_Q
$\vec{Q}_1$
$\vec{Q}_2$
$\vec{Q}_3$
$\vec{Q}_4$
$\vec{Q}_5$
$\vec{Q}_6$
$\vec{Q}_7$
$\vec{Q}_8$
a → $\vec{E}_1$ W_k → $\vec{K}_1$
fluffy → $\vec{E}_2$ W_k → $\vec{K}_2$
blue → $\vec{E}_3$ W_k → $\vec{K}_3$
creature → $\vec{E}_4$ W_k → $\vec{K}_4$
roamed → $\vec{E}_5$ W_k → $\vec{K}_5$
the → $\vec{E}_6$ W_k → $\vec{K}_6$
+93.0
+93.4
−31.4



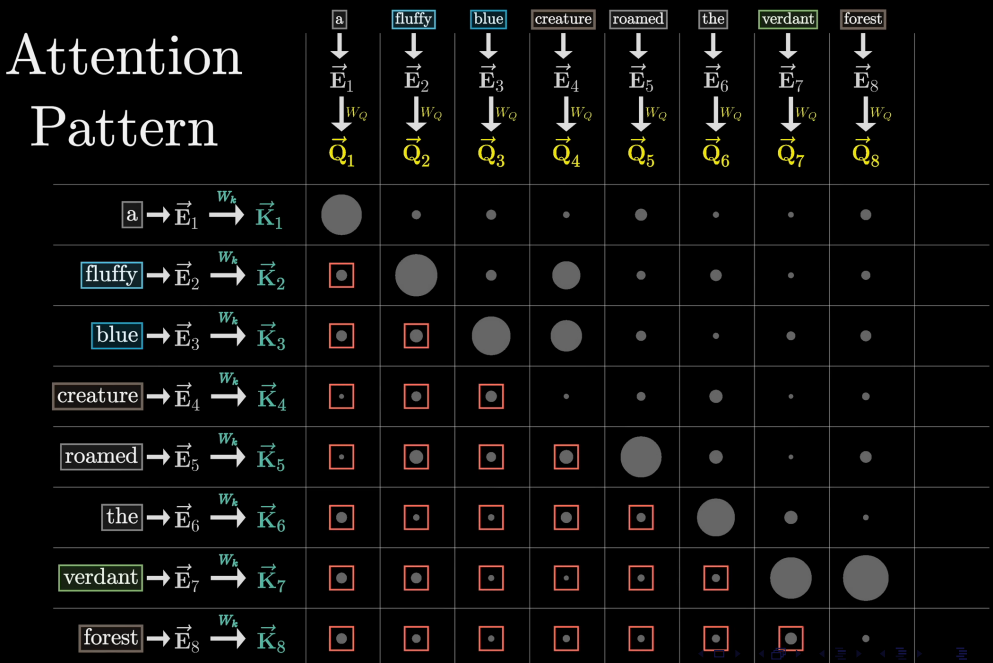
a
fluffy
blue
creature
roamed
the
verdant
forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
W_Q
$\vec{Q}_1$ $\vec{Q}_2$ $\vec{Q}_3$ $\vec{Q}_4$ $\vec{Q}_5$ $\vec{Q}_6$ $\vec{Q}_7$ $\vec{Q}_8$
W_k
$\vec{K}_1$ $\vec{K}_2$ $\vec{K}_3$ $\vec{K}_4$ $\vec{K}_5$ $\vec{K}_6$ $\vec{K}_7$ $\vec{K}_8$
-27.8
+93.0
+93.4
+4.9
-55.4
-31.4
-59.8
-90.6
softmax
0.00
0.42
0.58
0.00
0.00
0.00
0.00
0.00

$$\text{Attention}(Q, K, V) = \text{softmax}(\boxed{\frac{QK^T}{\sqrt{d_k}}})V$$

	Q_1	Q_2	Q_3	Q_4	Q_5	$\cdots$	Q_n	
K_1	$\frac{Q_1 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_1}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_1}{\sqrt{d_k}}$	
K_2	$\frac{Q_1 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_2}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_2}{\sqrt{d_k}}$	
K_3	$\frac{Q_1 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_3}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_3}{\sqrt{d_k}}$	
K_4	$\frac{Q_1 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_4}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_4}{\sqrt{d_k}}$	
K_5	$\frac{Q_1 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_5}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_5}{\sqrt{d_k}}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	


a
fluffy
blue
creature
roamed
the
verdant
forest
$\vec{E}_1$
$\vec{E}_2$
$\vec{E}_3$
$\vec{E}_4$
$\vec{E}_5$
$\vec{E}_6$
$\vec{E}_7$
$\vec{E}_8$
W_Q
$\vec{Q}_1$
$\vec{Q}_2$
$\vec{Q}_3$
$\vec{Q}_4$
$\vec{Q}_5$
$\vec{Q}_6$
$\vec{Q}_7$
$\vec{Q}_8$

Attention Pattern



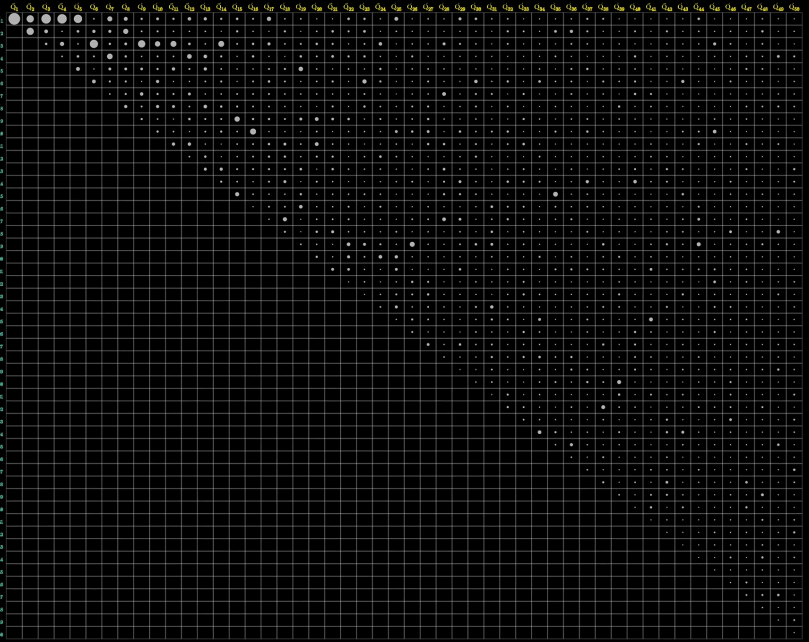
Unnormalized Attention Pattern

+3.53	+0.80	+1.96	+4.48	+3.74	−1.95
$-\infty$	−0.30	−0.21	+0.82	+0.29	+2.91
$-\infty$	$-\infty$	+0.89	+0.67	+2.99	−0.41
$-\infty$	$-\infty$	$-\infty$	+1.31	+1.73	−1.48
$-\infty$	$-\infty$	$-\infty$	$-\infty$	+3.07	+2.94
$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	+0.31

softmax →

Normalized Attention Pattern

1.00	0.75	0.69	0.92	0.46	0.00
0.00	0.25	0.08	0.02	0.01	0.46
0.00	0.00	0.24	0.02	0.22	0.02
0.00	0.00	0.00	0.04	0.06	0.01
0.00	0.00	0.00	0.00	0.24	0.48
0.00	0.00	0.00	0.00	0.00	0.03



Sparse Attention Mechanisms

Blockwise Attention

Linformer

Reformer

Ring attention

Longformer

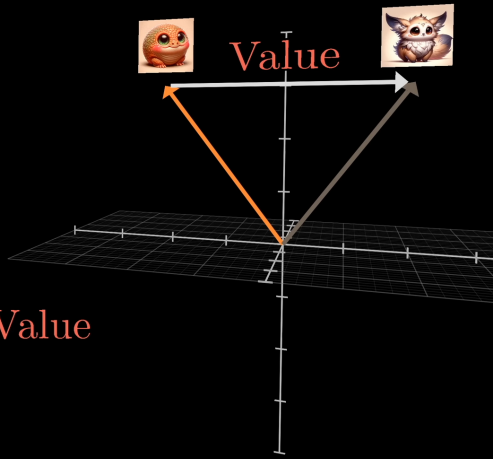
Adaptive Attention Span

$\vdots$

fluffy creature

$$\begin{bmatrix} +9.2 \\ -2.3 \\ +5.8 \\ +0.6 \\ +1.3 \\ +8.4 \\ \vdots \\ -8.2 \end{bmatrix} \quad \begin{bmatrix} -7.6 \\ +2.8 \\ -7.1 \\ +8.8 \\ +0.4 \\ -1.7 \\ \vdots \\ -4.7 \end{bmatrix}$$

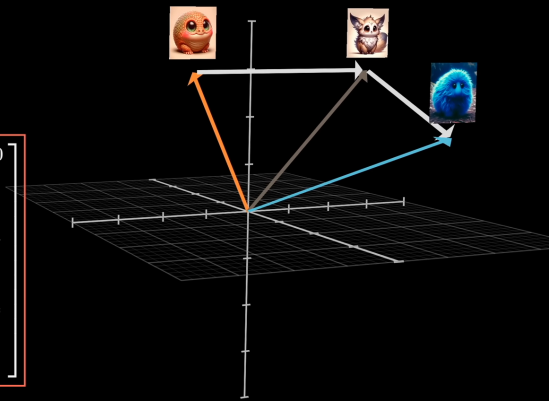
$$W_V \begin{bmatrix} -3.6 & -1.7 & -8.6 & +3.8 & +1.3 & -4.6 & \cdots & -8.0 \\ +1.5 & +8.5 & -3.6 & +3.3 & -7.3 & +4.3 & \cdots & -6.3 \\ +1.7 & -9.5 & +6.5 & -9.8 & +3.5 & -4.6 & \cdots & +9.2 \\ -5.0 & +1.5 & +1.8 & +1.4 & -5.5 & +9.0 & \cdots & +6.9 \\ +3.9 & -4.0 & +6.2 & -2.0 & +7.5 & +1.6 & \cdots & +3.8 \\ +4.5 & +0.0 & +9.0 & +2.9 & -1.5 & +2.1 & \cdots & -3.9 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +1.5 & +3.0 & +3.0 & -1.4 & +7.9 & -2.6 & \cdots & +7.8 \end{bmatrix} \begin{bmatrix} +9.2 \\ -2.3 \\ +5.8 \\ +0.6 \\ +1.3 \\ +8.4 \\ \vdots \\ -8.2 \end{bmatrix} = \begin{bmatrix} -52.4 \\ +89.3 \\ -80.2 \\ -17.8 \\ +7.3 \\ +223.8 \\ \vdots \\ -41.0 \end{bmatrix} \text{Value}$$

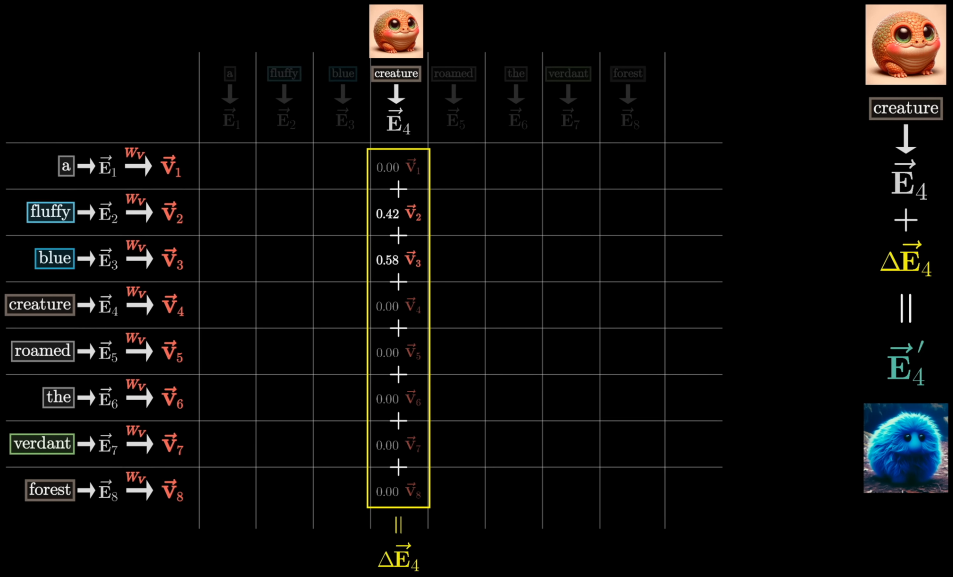


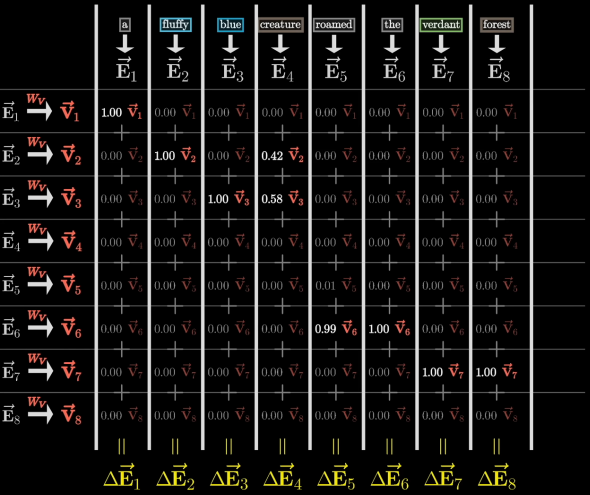
blue fluffy creature

$$\begin{bmatrix} +1.0 \\ +4.3 \\ +2.0 \\ +0.9 \\ -1.5 \\ +2.9 \\ \vdots \\ +7.8 \end{bmatrix} \quad \begin{bmatrix} +9.2 \\ -2.3 \\ +5.8 \\ +0.6 \\ +1.3 \\ +8.4 \\ \vdots \\ -8.2 \end{bmatrix} \quad \begin{bmatrix} -7.6 \\ +2.8 \\ -7.1 \\ +8.8 \\ +0.4 \\ -1.7 \\ \vdots \\ -4.7 \end{bmatrix}$$

$$W_V \begin{bmatrix} -3.6 & -1.7 & -8.6 & +3.8 & +1.3 & -4.6 & \cdots & -8.0 \\ +1.5 & +8.5 & -3.6 & +3.3 & -7.3 & +4.3 & \cdots & -6.3 \\ +1.7 & -9.5 & +6.5 & -9.8 & +3.5 & -4.6 & \cdots & +9.2 \\ -5.0 & +1.5 & +1.8 & +1.4 & -5.5 & +9.0 & \cdots & +6.9 \\ +3.9 & -4.0 & +6.2 & -2.0 & +7.5 & +1.6 & \cdots & +3.8 \\ +4.5 & +0.0 & +9.0 & +2.9 & -1.5 & +2.1 & \cdots & -3.9 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +1.5 & +3.0 & +3.0 & -1.4 & +7.9 & -2.6 & \cdots & +7.8 \end{bmatrix} \begin{bmatrix} +1.0 \\ +4.3 \\ +2.0 \\ +0.9 \\ -1.5 \\ +2.9 \\ \vdots \\ +7.8 \end{bmatrix} = \begin{bmatrix} -103.0 \\ +13.1 \\ +12.6 \\ +95.7 \\ +11.2 \\ +14.4 \\ \vdots \\ +62.1 \end{bmatrix}$$

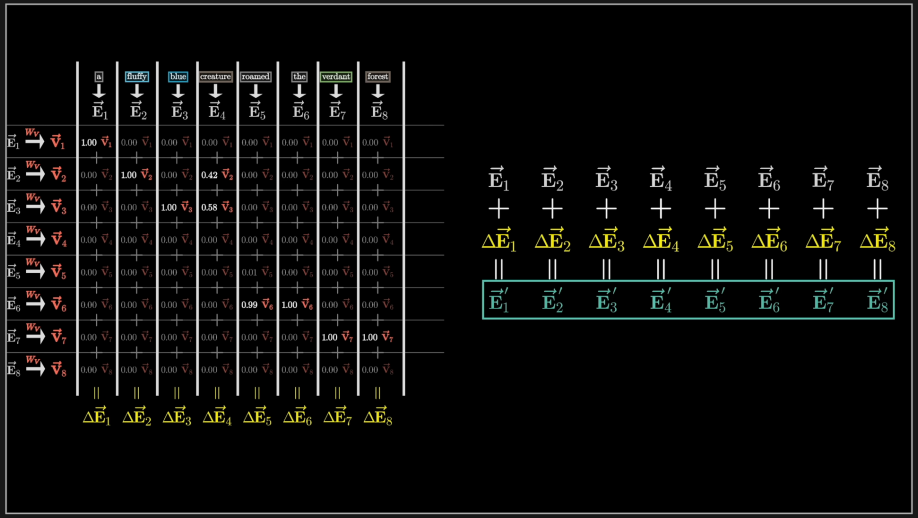



a
fluffy
blue
creature
roamed
the
verdant
forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
$\vec{E}_4$
a → $\vec{E}_1$ W_V → $\vec{V}_1$
fluffy → $\vec{E}_2$ W_V → $\vec{V}_2$
blue → $\vec{E}_3$ W_V → $\vec{V}_3$
creature → $\vec{E}_4$ W_V → $\vec{V}_4$
roamed → $\vec{E}_5$ W_V → $\vec{V}_5$
the → $\vec{E}_6$ W_V → $\vec{V}_6$
verdant → $\vec{E}_7$ W_V → $\vec{V}_7$
forest → $\vec{E}_8$ W_V → $\vec{V}_8$
0.00 $\vec{V}_1$
+
0.42 $\vec{V}_2$
+
0.58 $\vec{V}_3$
+
0.00 $\vec{V}_4$
+
0.00 $\vec{V}_5$
+
0.00 $\vec{V}_6$
+
0.00 $\vec{V}_7$
+
0.00 $\vec{V}_8$
||
$\Delta\vec{E}_4$
creature
$\vec{E}_4$
+
$\Delta\vec{E}_4$
||
$\vec{E}_4'$



$$\begin{array}{cccccccc} \vec{E}_1 & \vec{E}_2 & \vec{E}_3 & \vec{E}_4 & \vec{E}_5 & \vec{E}_6 & \vec{E}_7 & \vec{E}_8 \\ + & + & + & + & + & + & + & + \\ \Delta\vec{E}_1 & \Delta\vec{E}_2 & \Delta\vec{E}_3 & \Delta\vec{E}_4 & \Delta\vec{E}_5 & \Delta\vec{E}_6 & \Delta\vec{E}_7 & \Delta\vec{E}_8 \\ \| & \| & \| & \| & \| & \| & \| & \| \\ \vec{E}_1' & \vec{E}_2' & \vec{E}_3' & \vec{E}_4' & \vec{E}_5' & \vec{E}_6' & \vec{E}_7' & \vec{E}_8' \end{array}$$

One head of attention



Query

12,288

128

$$\begin{bmatrix} -3.7 & +3.9 & -2.4 & -6.3 & -9.4 & -8.6 & +3.6 & -0.9 & \cdots & +0.7 \\ +7.9 & +9.7 & -5.6 & +3.2 & -4.7 & -9.5 & +5.1 & -3.6 & \cdots & -2.3 \\ +1.7 & +6.6 & +2.6 & +7.4 & -4.5 & +5.9 & -6.2 & +9.0 & \cdots & +3.7 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -5.6 & +8.9 & +4.6 & -4.9 & -5.7 & +0.4 & -9.4 & -5.8 & \cdots & -1.5 \end{bmatrix}$$

Key

12,288

128

$$\begin{bmatrix} -2.5 & -0.7 & -4.4 & +1.7 & +7.2 & -7.6 & +0.3 & -7.3 & \cdots & +4.3 \\ -2.1 & +1.3 & -6.3 & -7.0 & -0.2 & -2.9 & +8.7 & +5.3 & \cdots & +4.9 \\ +8.0 & -8.2 & +1.0 & +1.7 & +9.1 & -4.1 & -5.1 & -7.9 & \cdots & -9.6 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +8.5 & +3.4 & +5.6 & -4.3 & +1.7 & -8.6 & -0.3 & +9.5 & \cdots & +7.5 \end{bmatrix}$$

Value

Value

$12{,}288 \times 12{,}288 = 150{,}994{,}944$

Query

1,572,864

$$\begin{bmatrix} -3.7 & +3.9 & -2.4 & -6.3 & -9.4 & -8.6 & +3.6 & -0.9 & \cdots & +0.7 \\ +7.9 & +9.7 & -5.6 & +3.2 & -4.7 & -9.5 & +5.1 & -3.6 & \cdots & -2.3 \\ +1.7 & +6.6 & +2.6 & +7.4 & -4.5 & +5.9 & -6.2 & +9.0 & \cdots & +3.7 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -5.6 & +8.9 & +4.6 & -4.9 & -5.7 & +0.4 & -9.4 & -5.8 & \cdots & -1.5 \end{bmatrix}$$

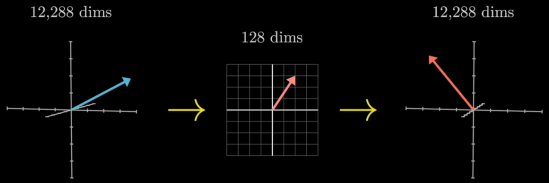
Key

1,572,864

$$\begin{bmatrix} -2.5 & -0.7 & -4.4 & +1.7 & +7.2 & -7.6 & +0.3 & -7.3 & \cdots & +4.3 \\ -2.1 & +1.3 & -6.3 & -7.0 & -0.2 & -2.9 & +8.7 & +5.3 & \cdots & +4.9 \\ +8.0 & -8.2 & +1.0 & +1.7 & +9.1 & -4.1 & -5.1 & -7.9 & \cdots & -9.6 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +8.5 & +3.4 & +5.6 & -4.3 & +1.7 & -8.6 & -0.3 & +9.5 & \cdots & +7.5 \end{bmatrix}$$

12,288

$$12{,}288 \left\{ \overbrace{\begin{bmatrix} -3.2 & +9.1 & -5.3 & +8.9 & +8.7 & +5.9 & +2.6 & +7.4 & \cdots & -4.1 \\ +6.9 & +2.3 & -9.6 & -3.0 & -7.0 & +9.5 & -0.4 & -0.1 & \cdots & +2.8 \\ -2.6 & -7.2 & +6.4 & -6.1 & +0.2 & -5.5 & -8.0 & +7.2 & \cdots & +9.4 \\ +9.1 & +8.0 & +5.4 & -3.3 & -8.3 & -1.8 & -5.3 & -7.3 & \cdots & -8.8 \\ +4.5 & -9.7 & +5.4 & -7.0 & -8.3 & -8.1 & +3.4 & -5.0 & \cdots & -1.6 \\ +1.1 & +7.1 & +4.5 & -4.5 & -7.3 & -8.8 & -3.9 & -4.7 & \cdots & -0.9 \\ +3.6 & +3.9 & -4.3 & -2.4 & -6.3 & +5.7 & -8.8 & +3.9 & \cdots & +5.5 \\ +5.5 & -4.8 & -2.5 & +1.7 & -4.5 & -2.6 & -6.0 & -0.8 & \cdots & -9.0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ +5.9 & -8.4 & +0.4 & -3.8 & +1.5 & +9.1 & +2.9 & -9.2 & \cdots & -1.4 \end{bmatrix}}^{12{,}288} \begin{bmatrix} +0.2 \\ +0.7 \\ +3.6 \\ -4.4 \\ -7.3 \\ -2.1 \\ +9.0 \\ -6.2 \\ \vdots \\ +0.9 \end{bmatrix} = \begin{bmatrix} -198.6 \\ +73.1 \\ -28.2 \\ +119.4 \\ +215.7 \\ +91.8 \\ -29.1 \\ -5.6 \\ \vdots \\ -5.1 \end{bmatrix}$$



“Low rank” transformation

d_input
12,288

d_output
12,288

128

12,288

12,288

$$\begin{bmatrix} +5.0 & -3.3 & \cdots & +7.2 \\ -8.9 & -4.9 & \cdots & -7.8 \\ -3.0 & +4.8 & \cdots & +2.4 \\ +4.2 & -5.8 & \cdots & +3.5 \\ +7.5 & +0.9 & \cdots & -9.3 \\ +4.2 & -9.7 & \cdots & +0.6 \\ +8.4 & -8.1 & \cdots & -9.4 \\ -3.1 & +2.4 & \cdots & -5.7 \\ \vdots & \vdots & \ddots & \vdots \\ +8.9 & -9.8 & \cdots & +2.0 \end{bmatrix}$$

Value↑

$$\begin{bmatrix} -7.8 & +8.9 & -5.3 & +3.8 & -8.7 & +4.6 & +7.6 & -4.5 & \cdots & -2.5 \\ +4.9 & -5.2 & -6.5 & -1.0 & -3.9 & +6.7 & -5.2 & +0.0 & \cdots & +2.7 \\ +7.3 & +8.7 & +5.0 & +4.0 & +9.3 & +9.8 & -1.0 & -8.5 & \cdots & -6.9 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1.8 & +1.0 & -4.5 & -0.9 & -1.9 & -5.0 & +0.1 & -3.8 & \cdots & +0.5 \end{bmatrix}$$

Value↓

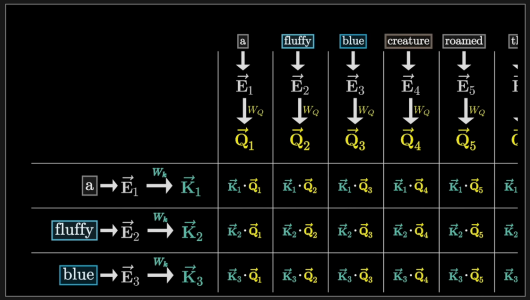
$$\begin{bmatrix} +0.2 \\ +0.7 \\ +3.6 \\ -4.4 \\ -7.3 \\ -2.1 \\ +9.0 \\ -6.2 \\ \vdots \\ +0.9 \end{bmatrix} = \begin{bmatrix} -198.6 \\ +73.1 \\ -28.2 \\ +119.4 \\ +215.7 \\ +91.8 \\ -29.1 \\ -5.6 \\ \vdots \\ -5.1 \end{bmatrix}$$

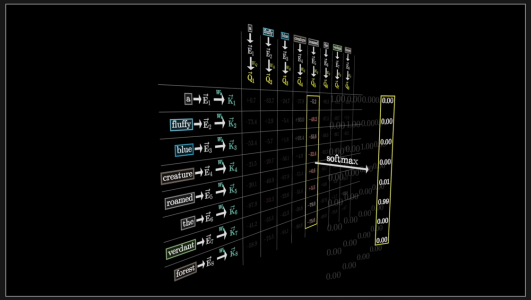
Total weights: 175,181,291,520

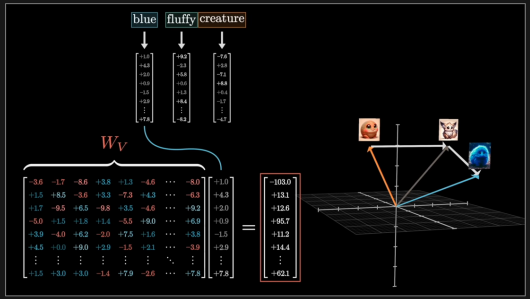
Organized into 27,938 matrices

GPT-3

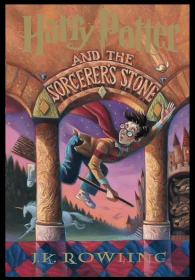
Embedding	$\overset{12,288}{d\_embed} * \overset{50,257}{n\_vocab} = 617,558,016$
Key	$\overset{128}{d\_query} * \overset{12,288}{d\_embed} = 1,572,864$ per head
Query	$\overset{128}{d\_query} * \overset{12,288}{d\_embed} = 1,572,864$ per head
Value$_\downarrow$	$\overset{128}{d\_value} * \overset{12,288}{d\_embed} = 1,572,864$ per head
Value$_\uparrow$	$\overset{12,288}{d\_embed} * \overset{128}{d\_value} = 1,572,864$ per head
Up-projection	$6,291,456$
Down-projection	
Unembedding	$\overset{50,257}{n\_vocab} * \overset{12,288}{d\_embed} = 617,558,016$


a
fluffy
blue
creature
roamed
t
$\vec{E}_1$
$\vec{E}_2$
$\vec{E}_3$
$\vec{E}_4$
$\vec{E}_5$
W_Q
$\vec{Q}_1$
$\vec{Q}_2$
$\vec{Q}_3$
$\vec{Q}_4$
$\vec{Q}_5$
a → $\vec{E}_1$ W_K → $\vec{K}_1$
fluffy → $\vec{E}_2$ W_K → $\vec{K}_2$
blue → $\vec{E}_3$ W_K → $\vec{K}_3$
$\vec{K}_1 \cdot \vec{Q}_1$ $\vec{K}_1 \cdot \vec{Q}_2$ $\vec{K}_1 \cdot \vec{Q}_3$ $\vec{K}_1 \cdot \vec{Q}_4$ $\vec{K}_1 \cdot \vec{Q}_5$ $\vec{K}_1$
$\vec{K}_2 \cdot \vec{Q}_1$ $\vec{K}_2 \cdot \vec{Q}_2$ $\vec{K}_2 \cdot \vec{Q}_3$ $\vec{K}_2 \cdot \vec{Q}_4$ $\vec{K}_2 \cdot \vec{Q}_5$ $\vec{K}_2$
$\vec{K}_3 \cdot \vec{Q}_1$ $\vec{K}_3 \cdot \vec{Q}_2$ $\vec{K}_3 \cdot \vec{Q}_3$ $\vec{K}_3 \cdot \vec{Q}_4$ $\vec{K}_3 \cdot \vec{Q}_5$ $\vec{K}_3$


softmax
a → $\vec{E}_1$ → $\vec{K}_1$
fluffy → $\vec{E}_2$ → $\vec{K}_2$
blue → $\vec{E}_3$ → $\vec{K}_3$
creature → $\vec{E}_4$ → $\vec{K}_4$
roamed → $\vec{E}_5$ → $\vec{K}_5$
the → $\vec{E}_6$ → $\vec{K}_6$
verdant → $\vec{E}_7$ → $\vec{K}_7$
forest → $\vec{E}_8$ → $\vec{K}_8$

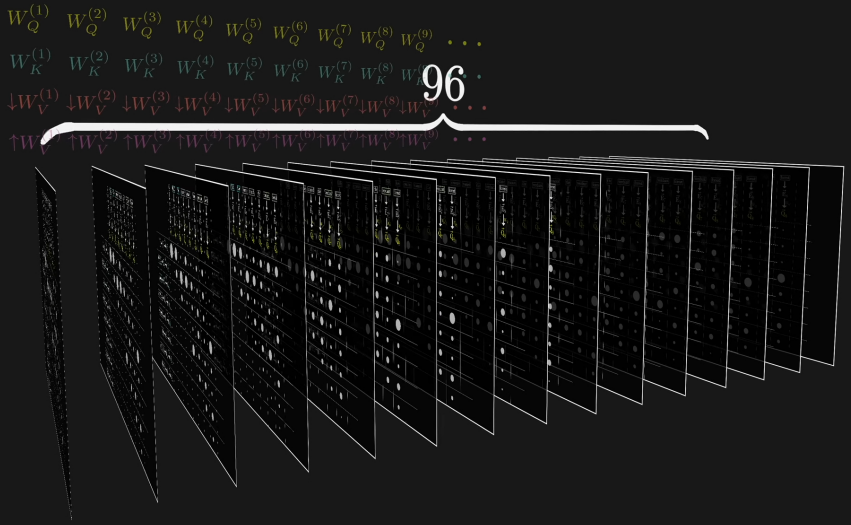

blue
fluffy
creature
W_V
$\vec{V}_1$
$\vec{V}_2$
$\vec{V}_3$
$\vec{V}_4$
$\vec{V}_5$
$\vec{V}_6$
$\vec{V}_7$
$\vec{V}_8$
$\Delta\vec{E}_1$ $\Delta\vec{E}_2$ $\Delta\vec{E}_3$ $\Delta\vec{E}_4$ $\Delta\vec{E}_5$ $\Delta\vec{E}_6$ $\Delta\vec{E}_7$ $\Delta\vec{E}_8$

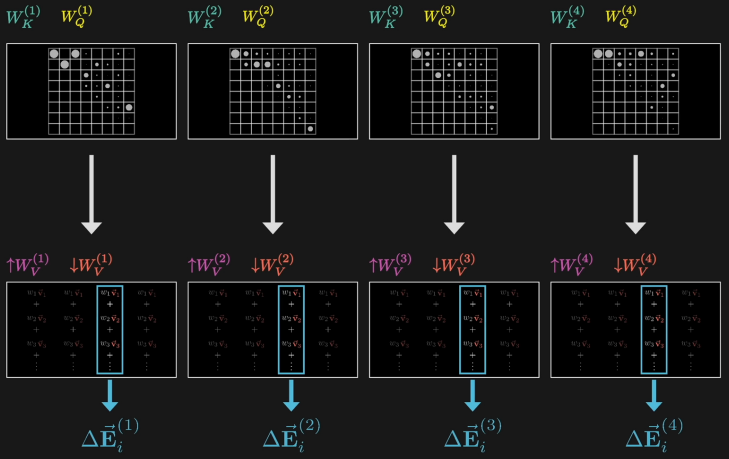
... wizard ... Hogwarts ... Hermione ... Harry



... Queen ... Sussex ... William ... Harry

Multi-headed attention





New embedding

$$\vec{\mathbf{E}}_i + \Delta\vec{\mathbf{E}}_i^{(1)} + \Delta\vec{\mathbf{E}}_i^{(2)} + \Delta\vec{\mathbf{E}}_i^{(3)} + \Delta\vec{\mathbf{E}}_i^{(4)} + \cdots$$

Total weights: 175,181,291,520

Organized into 27,938 matrices

GPT-3

Embedding	$\underset{12{,}288}{d\_embed} * \underset{50{,}257}{n\_vocab} = 617{,}558{,}016$
Key	$\underset{128}{d\_query} * \underset{12{,}288}{d\_embed} * \underset{96}{n\_heads} = 150{,}994{,}944$ per layer
Query	$\underset{128}{d\_query} * \underset{12{,}288}{d\_embed} * \underset{96}{n\_heads} = 150{,}994{,}944$ per layer
Value$_\downarrow$	$\underset{128}{d\_value} * \underset{12{,}288}{d\_embed} * \underset{96}{n\_heads} = 150{,}994{,}944$ per layer
Value$_\uparrow$	$\underset{12{,}288}{d\_embed} * \underset{128}{d\_value} * \underset{96}{n\_heads} = 150{,}994{,}944$ per layer
Up-projection	603,979,776
Down-projection	
Unembedding	$\underset{50{,}257}{n\_vocab} * \underset{12{,}288}{d\_embed} = 617{,}558{,}016$

Total weights: 175,181,291,520

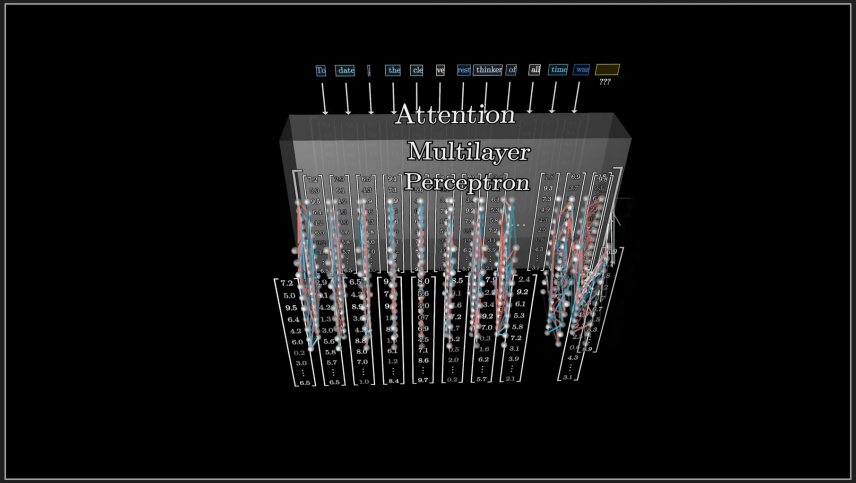
Organized into 27,938 matrices

GPT-3

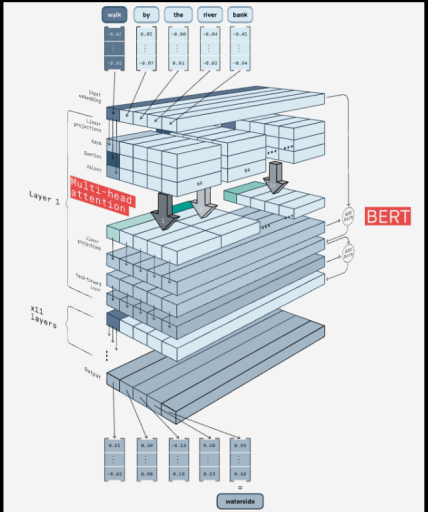
Embedding	$\overset{12{,}288}{d\_embed} * \overset{50{,}257}{n\_vocab} = 617{,}558{,}016$
Key	$\overset{128}{d\_query} * \overset{12{,}288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14{,}495{,}514{,}624$
Query	$\overset{128}{d\_query} * \overset{12{,}288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14{,}495{,}514{,}624$
Value	$\overset{128}{d\_value} * \overset{12{,}288}{d\_embed} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14{,}495{,}514{,}624$
Output	$\overset{12{,}288}{d\_embed} * \overset{128}{d\_value} * \overset{96}{n\_heads} * \overset{96}{n\_layers} = 14{,}495{,}514{,}624$
Up-projection	$57{,}982{,}058{,}496$
Down-projection	
Unembedding	$\overset{50{,}257}{n\_vocab} * \overset{12{,}288}{d\_embed} = 617{,}558{,}016$

About 1/3 of What

Attention is ~~All~~ You Need



En 2020 : on avait que BERT! (version française)



BERT processes a sentence to output contextualized embeddings that are more representative of the true meaning

Un exemple de la corvée du Dr Brunet-Gouet

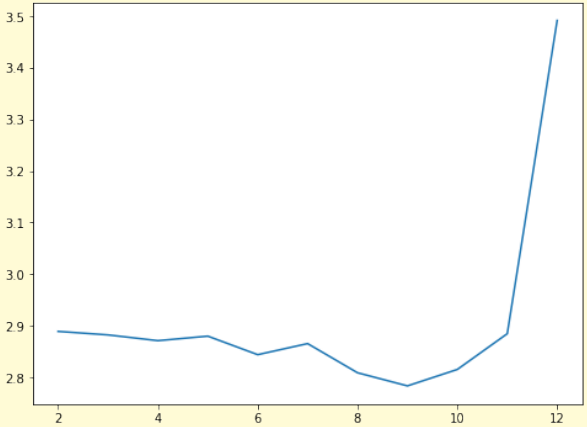
```
 jf+,js+   J'ai 2 frères et deux sœurs.
 &        Je m'entend plutôt bien avec.
 mp-        Ma maman est séparée de mon père.
 jm+        Je m'entend très bien avec ma mère.
si        J'ai deux grandes demi-sœur du côté de mon père.
 &        Elles ont déjà des enfants.
mi,pi    Ma maman à fait cinq enfants, mon père en n'a fait trois.
jm+        Avec ma mère ont aime bien faire les magasins.
```

Bilan : 1648 phrases et 11 labels

- La valence relationnelle : +, - 0. Les relations positives font référence à une bonne entente, à l'expression d'un affect positif et à la coopération. Les relations négatives correspondent à des conflits, des désaccords, l'absence d'une relation normale.
- En l'absence d'information sur la valence, le texte est considéré comme informatif (i) sur les habitudes ou les conditions de vie des personnes.
- Les sujets décrits dans un segment ont été étiquetés comme suit : le répondant je, la mère, le père, la soeur, le frère, la famille et une tierce personne.

Partition apprentissage-test de tailles 1318 et 330.

10 jours de calcul sur NVIDIA A100 Tensor Core 40GB GPU



Résultats numériques

Precision = TP/TP+FP, Recall = TP/TP+FN et F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Label	+	-	0	i	j	f	s	p	m	a	t
Support	72	63	47	146	263	36	57	124	123	72	23
Precision											
Fine-tuning	0.67	**0.85**	0.40	**0.78**	**0.86**	**0.91**	0.97	**0.97**	**0.94**	**0.91**	0.78
Elasticnet lr	0.65(0.69)	0.71(0.75)	0.28(0.33)	0.74(0.61)	0.85(0.81)	0.74(0.79)	0.89(**0.98**)	0.90(0.93)	0.83(0.87)	0.77(0.81)	0.67(0.75)
Gradient Boosting	0.73(0.65)	0.88(0.44)	0.25(**0.45**)	0.75(0.64)	0.84(0.80)	0.75(0.78)	0.88(0.92)	0.84(0.93)	0.79(0.85)	0.83(0.73)	0(**0.83**)
Random Forest	**0.82**(0.70)	0.71(0.73)	0(0.33)	0.71(0.64)	0.82(0.80)	0(0.72)	0.89(0.92)	0.86(0.93)	0.76(0.89)	0.73(0.77)	0(0.67)
SVC	0.68(0.70)	0.79(1)	0(0)	0.76(0.66)	0.85(0.81)	0.68(0.79)	0.91(0.94)	0.90(0.92)	0.83(0.89)	0.82(0.79)	0.50(0.79)
Recall											
Fine-tuning	**0.82**	0.37	**0.38**	**0.81**	0.97	**0.89**	**0.98**	**0.93**	**0.93**	**0.86**	**0.61**
Elasticnet lr	0.56(0.49)	**0.40**(0.10)	0.11(0.02)	0.73(0.75)	0.94(1)	0.39(0.64)	0.68(0.81)	0.76(0.85)	0.82(0.85)	0.67(0.64)	0.26(0.52)
Gradient Boosting	0.49(0.49)	0.22(0.11)	0.02(0.11)	0.75(0.66)	0.97(0.95)	0.17(0.81)	0.49(0.86)	0.70(0.85)	0.69(0.90)	0.47(0.64)	0.00(0.43)
Random Forest	0.38(0.43)	0.08(0.17)	0(0.04)	0.75(0.59)	0.96(0.92)	0.00(0.86)	0.30(0.95)	0.52(0.92)	0.60(0.93)	0.33(0.71)	0.00(0.52)
SVC	0.57(0.46)	0.30(0.05)	0(0)	0.77(0.73)	0.94(**1**)	0.42(0.64)	0.72(0.86)	0.77(0.78)	0.81(0.89)	0.64(0.58)	0.09(0.48)
f1-score											
Fine-tuning	**0.74**	**0.51**	**0.39**	**0.79**	**0.91**	**0.90**	**0.97**	**0.95**	**0.93**	**0.89**	**0.68**
Elasticnet lr	0.60(0.57)	0.51(0.17)	0.15(0.04)	0.74(0.67)	0.89(0.89)	0.51(0.71)	0.77(0.88)	0.82(0.89)	0.83(0.86)	0.72(0.71)	0.38(0.62)
Gradient Boosting	0.58(0.56)	0.35(0.18)	0.04(0.17)	0.75(0.65)	0.90(0.87)	0.27(0.79)	0.63(0.89)	0.77(0.89)	0.74(0.88)	0.60(0.68)	0(0.57)
Random Forest	0.51(0.53)	0.14(0.28)	0(0.08)	0.73(0.61)	0.89(0.86)	0.00(0.78)	0.45(0.93)	0.65(0.92)	0.67(0.91)	0.46(0.74)	0(0.59)
SVC	0.62(0.55)	0.44(0.09)	0(0)	0.77(0.69)	0.90(0.89)	0.52(0.71)	0.80(0.90)	0.83(0.84)	0.82(0.89)	0.72(0.67)	0.15(0.59)

Un exemple de tokenizer : Byte-Pair encoding

```
corpus = [
    "This example is taken from the multitude of tokenizer examples
     explained on the Hugging Face courses.",
    "This is one example of tokenization.",
    "There is several tokenizer algorithms.",
    "Hopefully, you will be able to understand how they are trained and generate tokens.",
]
```

```
{'This': 2, 'Ġexample': 2, 'Ġis': 3, 'Ġtaken': 1, 'Ġfrom': 1, 'Ġthe': 2,
'Ġmultitude': 1, 'Ġof': 2, 'Ġtokenizer': 2, 'Ġexamples': 1, 'Ġexplained': 1,
'Ġon': 1, 'ĠHugging': 1, 'ĠFace': 1, 'Ġcourses': 1, '.': 4, 'Ġone': 1,
'Ġtokenization': 1, 'There': 1, 'Ġseveral': 1, 'Ġalgorithms': 1,
'Hopefully': 1, ',': 1, 'Ġyou': 1, 'Ġwill': 1, 'Ġbe': 1, 'Ġable': 1,
'Ġto': 1, 'Ġunderstand': 1, 'Ġhow': 1, 'Ġthey': 1, 'Ġare': 1,
'Ġtrained': 1, 'Ġand': 1, 'Ġgenerate': 1, 'Ġtokens': 1}
```

Un exemple de tokenizer : Byte-Pair encoding

En partant d'un alphabet

```
[',', '.', 'F', 'H', 'T', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'k',
'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z', 'Ġ']
```

On découpe les mots comme suit

```
{'This': ['T', 'h', 'i', 's'],
'Ġexample': ['Ġ', 'e', 'x', 'a', 'm', 'p', 'l', 'e'],
'Ġis': ['Ġ', 'i', 's'],
'Ġtaken': ['Ġ', 't', 'a', 'k', 'e', 'n'],
'Ġfrom': ['Ġ', 'f', 'r', 'o', 'm'],
'Ġthe': ['Ġ', 't', 'h', 'e'],
'Ġmultitude': ['Ġ', 'm', 'u', 'l', 't', 'i', 't', 'u', 'd', 'e'],
'Ġof': ['Ġ', 'o', 'f'],
'Ġtokenizer': ['Ġ', 't', 'o', 'k', 'e', 'n', 'i', 'z', 'e', 'r'],
'Ġexamples': ['Ġ', 'e', 'x', 'a', 'm', 'p', 'l', 'e', 's'],
'Ġexplained': ['Ġ', 'e', 'x', 'p', 'l', 'a', 'i', 'n', 'e', 'd'], ...}
```

Un exemple de tokenizer : Byte-Pair encoding

On va compter les paires de lettres de l'alphabet

```
('T', 'h'): 3
('h', 'i'): 2
('i', 's'): 5
('Ġ', 'e'): 4
('e', 'x'): 4
('x', 'a'): 3
('a', 'm'): 3
('m', 'p'): 3
('p', 'l'): 4
('l', 'e'): 4
('Ġ', 'i'): 3
('Ġ', 't'): 10
.
.
```

Mise à jour de l'alphabet

```
[',', '.', 'F', 'H', 'T', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i',
'k', 'l', 'm', 'n', 'o', 'p', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y',
'z', 'Ġ', 'Ġt']
```

Un exemple de tokenizer : Byte-Pair encoding

On refait la découpe des mots en fonction des éléments de l'alphabet,

```
{'Ġtrained': ['Ġ', 't', 'r', 'a', 'i', 'n', 'e', 'd'],
 'Ġtokens': ['Ġ', 't', 'o', 'k', 'e', 'n', 's']}
```

deviennent

```
{'Ġtrained': ['Ġt', 'r', 'a', 'i', 'n', 'e', 'd'],
'Ġtokens': ['Ġt', 'o', 'k', 'e', 'n', 's']}
```

Byte-Pair encoding : procédé itératif

Critère d'arrêt : la taille de vocabulaire.

```
['<|endoftext|>', ',', '.', 'F', 'H', 'T', 'a', 'b',
'c', 'd', 'e', 'f', 'g', 'h', 'i', 'k', 'l', 'm',
'n', 'o', 'p', 'r', 's', 't', 'u', 'v', 'w', 'x',
'y', 'z', 'Ġ', 'Ġt', 'en', 'er', 'is', 'ken', 'Ġto',
'Ġe', 'Ġex', 'pl', 'Ġo', 'Ġtoken', 'Ġa', 'Th', 'Ġexa',
'Ġexam', 'Ġexampl', 'Ġexample', 'Ġis', 'Ġth']
```

On peut tokeniser un texte
en

```
['Th','is','Ġis','Ġ','n','o','t','Ġa','Ġtoken','.']
```

Tokenizer : bilan

- Un processus de conversion d'un texte en tokens individuels (mots, sous-mots ou caractères).
- Décomposer le texte en unités exploitables pour le traitement.
- Tokenizers : WordPiece, Byte Pair Encoding, SentencePiece.
- Peut varier considérablement d'une langue à l'autre en raison des différentes structures de jetons.

Que peut-on faire avec tokens ?

Un embedding (plongement) : exemple

```
First Citizen:
Before we proceed any further, hear me speak.

All:
Speak, speak.

First Citizen:
You are all resolved rather to die than to famish?

All:
Resolved. resolved.

First Citizen:
First, you know Caius Marcius is chief enemy to the people.

All:
We know't, we know't.

First Citizen:
Let us kill him, and we'll have corn at our own price.
Is't a verdict?
```

Embedding : exemple

Un tokenizer basique
On obtient l'alphabet :

```
 !$&',-.3:;?ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz
65
```

On s'en sert comme suit : Nous donne :

```
[46, 47, 47, 1, 58, 46, 43, 56, 43]
hii there
```

Un bloc de tokens pour prédire le suivant

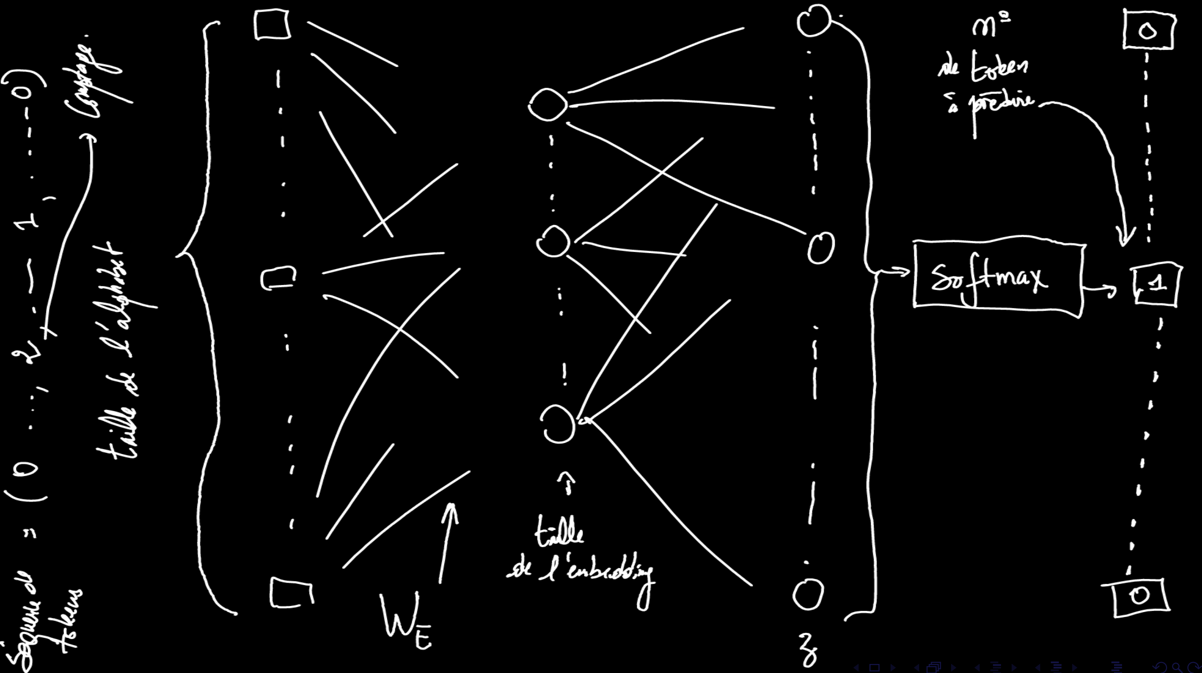
Avec cette séquence de tokens

```
tensor([18, 47, 56, 57, 58,  1, 15, 47, 58])
```

Nous allons construire ce jeu de données d'apprentissage

```
when input is tensor([18]) the target: 47
when input is tensor([18, 47]) the target: 56
when input is tensor([18, 47, 56]) the target: 57
when input is tensor([18, 47, 56, 57]) the target: 58
when input is tensor([18, 47, 56, 57, 58]) the target: 1
when input is tensor([18, 47, 56, 57, 58,  1]) the target: 15
when input is tensor([18, 47, 56, 57, 58,  1, 15]) the target: 47
when input is tensor([18, 47, 56, 57, 58,  1, 15, 47]) the target: 58
```

Il nous reste qu'à encoder les vecteurs et utiliser un réseau de neurones de classification.


Séquence de tokens = (0 ..., 2, ..., 1, ... 0)
Comptage
taille de l'alphabet
W_E
taille de l'embedding
z
Softmax
n° de token à prédire
0
1
0

Embedding : bilan

- On récupère la matrice des poids du réseau de neurones
- Pour l'instant l'ordre des mots n'est pas pris en compte
- Pour prendre en compte l'ordre des mots : on utilise une astuce dite de *Positional Encoding*

Attention : estimation non-paramétrique

Estimateur à noyaux de Nadaraya et Watson en 1964. Pondérer les y_i en fonction de l'emplacement des x_i :

$$f(x) = \sum_{i=1}^{n} \frac{K(x - x_i)}{\sum_{j=1}^{n} K(x - x_j)} y_i,$$

où K est noyau. On peut généraliser cet estimateur à

$$f(x) = \sum_{i=1}^{n} \alpha(x, x_i) y_i,$$

où x est la requête et (x_i, y_i) la paire clé-valeur. La mise en commun de l'attention ici est une moyenne pondérée des valeurs y_i. Le *poids d'attention* $\alpha(x, x_i)$ est attribué à la valeur correspondante y_i sur la base de l'interaction entre la requête x et la clé x_i, modélisée par α.

Attention : estimation non-paramétrique

Pour mieux comprendre la mise en commun de l'attention, il suffit de considérer un *noyau gaussien* défini comme suit

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}).$$

On obtient

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} \alpha(x, x_i) y_i \\ &= \sum_{i=1}^{n} \frac{\exp\left(-\frac{1}{2}(x - x_i)^2\right)}{\sum_{j=1}^{n} \exp\left(-\frac{1}{2}(x - x_j)^2\right)} y_i \\ &= \sum_{i=1}^{n} \text{softmax}\left(-\frac{1}{2}(x - x_i)^2\right) y_i. \end{aligned}$$