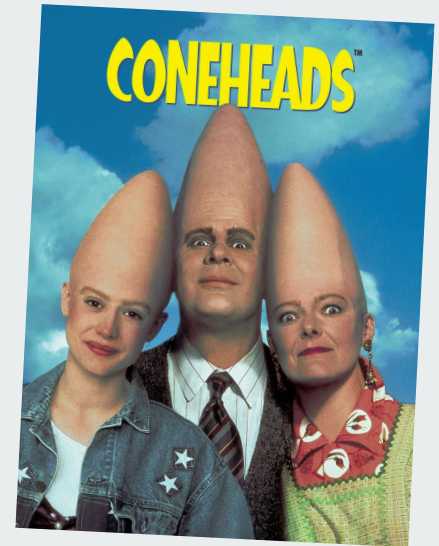

Language models have cone-shaped heads

Nathan Godey - PhD Pizza Seminar





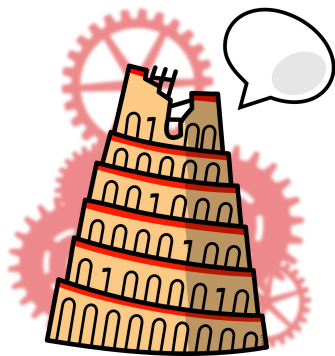
Me



Eric de La Clergerie



Benoît Sagot



ALMAnaCH
Team-project



A cone.



You

Write a story about a cone for a presentation to PhD students.




ChatGPT

Once upon a time in the hallowed halls of Academia, there existed a peculiar cone named Conicus. Conicus wasn't an ordinary cone; it was a mathematical marvel, a three-dimensional embodiment of elegance and precision. This particular cone, however, found itself in an extraordinary situation—it was chosen to be the focal point of a presentation to a room full of inquisitive PhD students.

The esteemed professor, Dr. Octavia Mathesius, had dedicated her life to unraveling the mysteries of mathematics. Her passion for teaching was as boundless as the vast mathematical landscapes she explored. One day, as she prepared to enlighten her

Once
upon
a

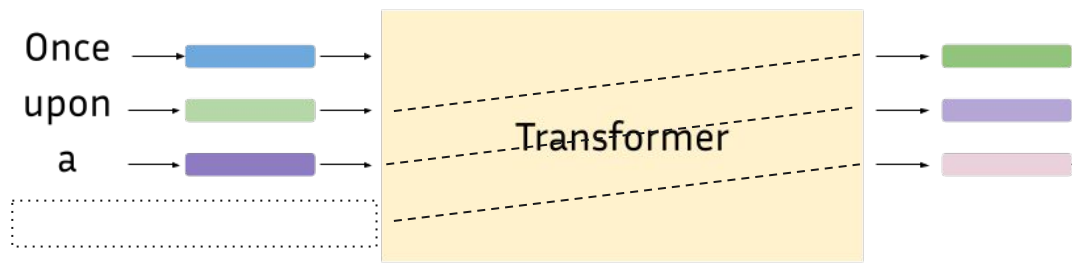
Next word prediction

Once → 
upon → 
a → 

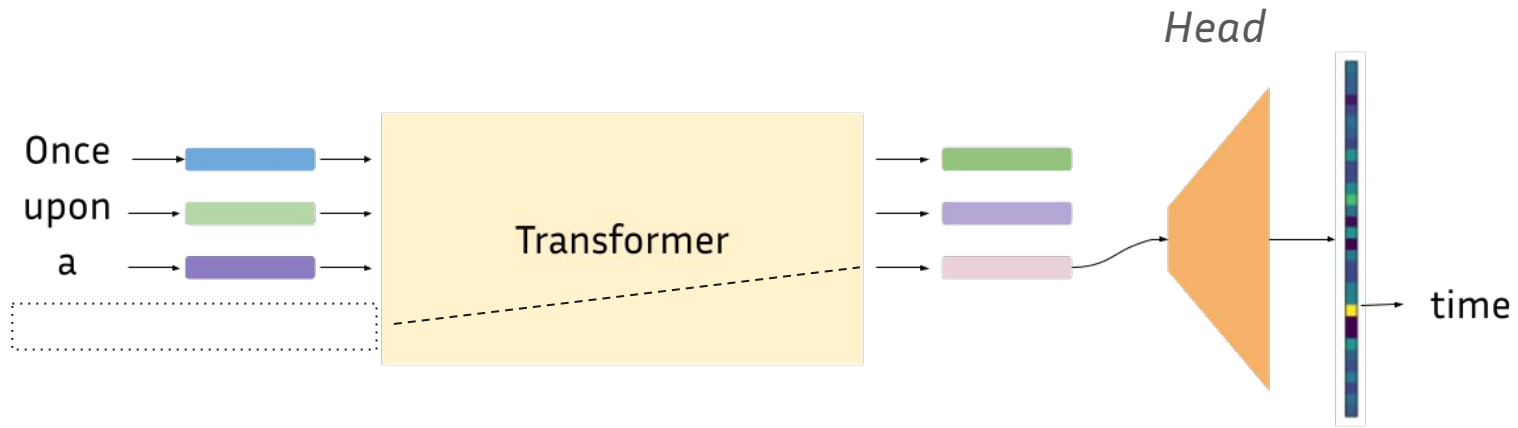
Next word prediction



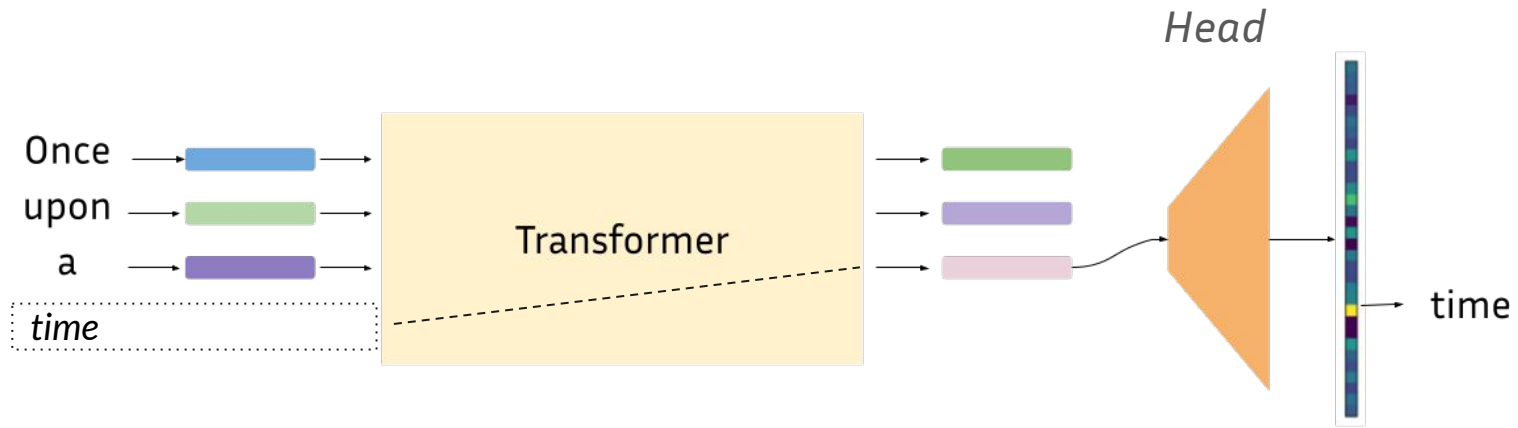
Next word prediction



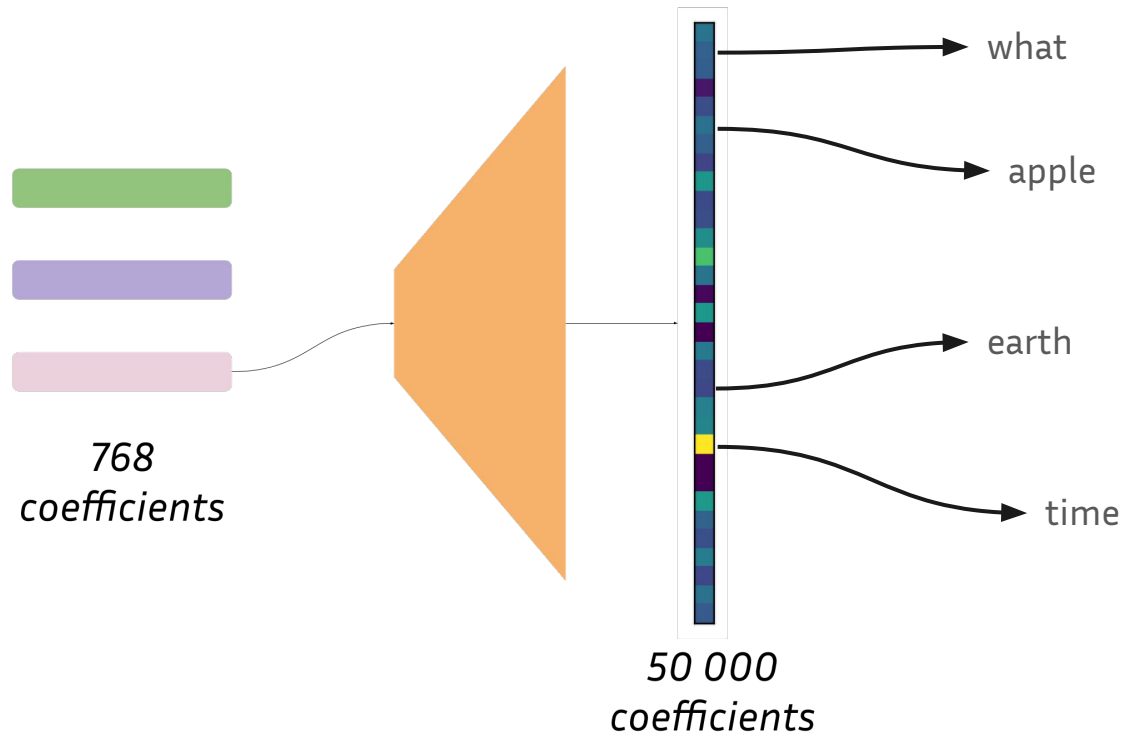
Next word prediction



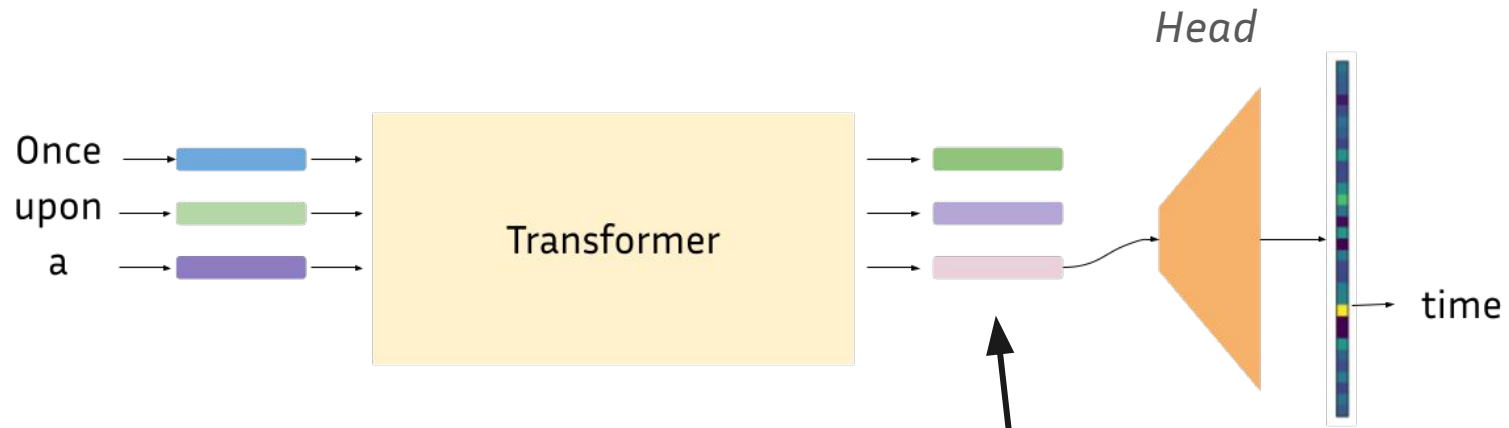
Next word prediction



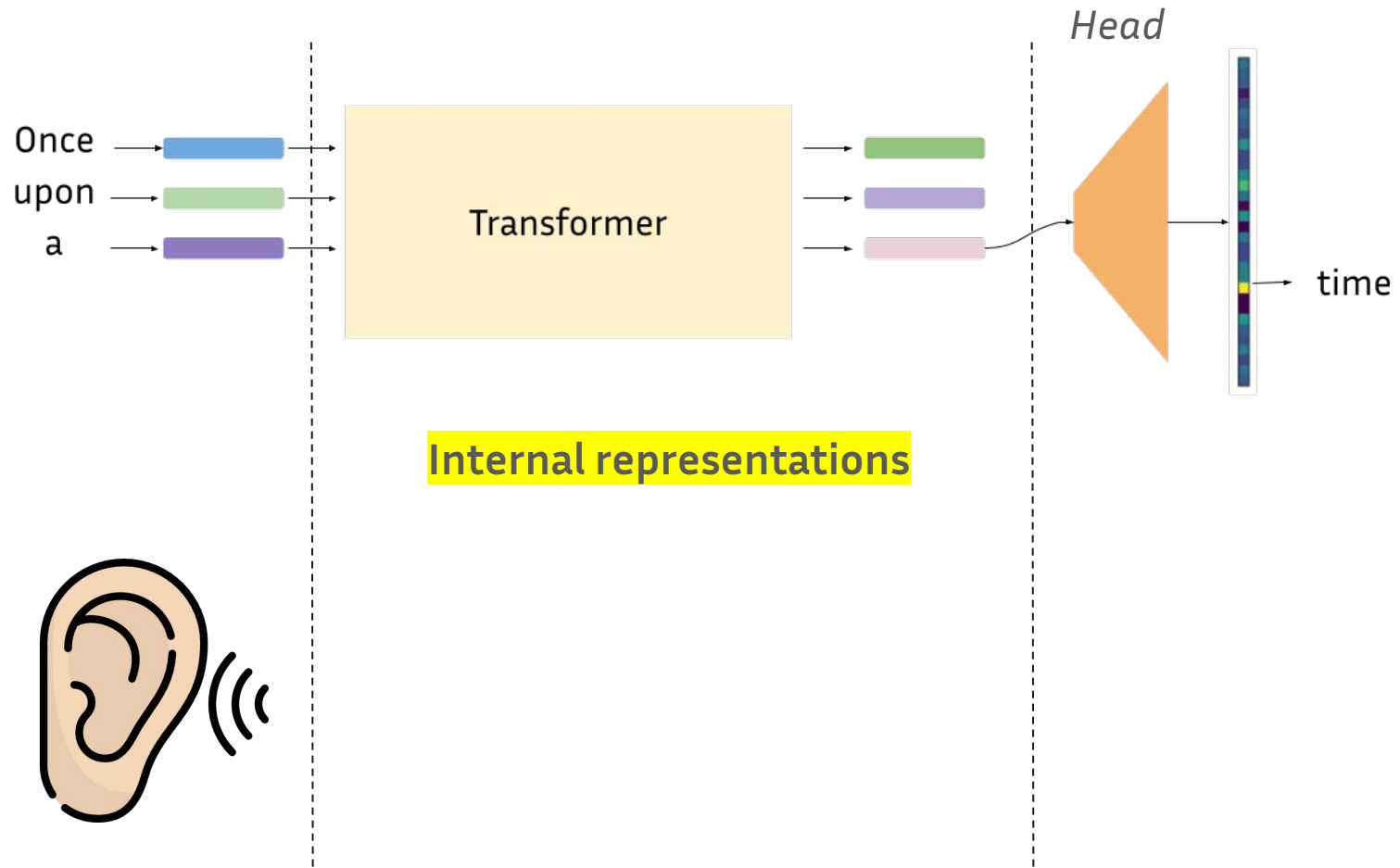
Next word prediction

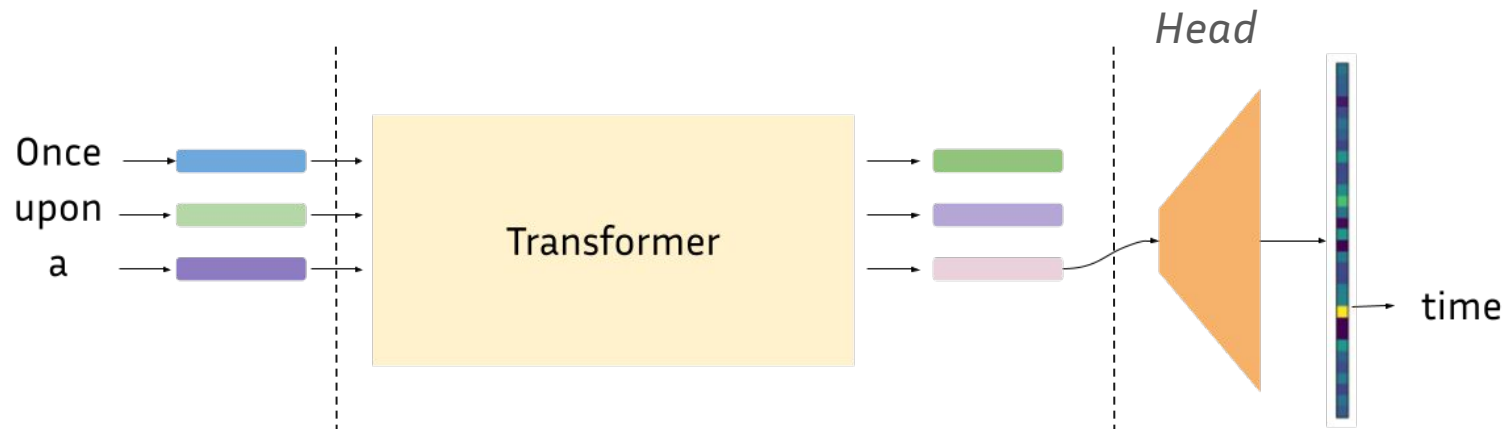


LM head

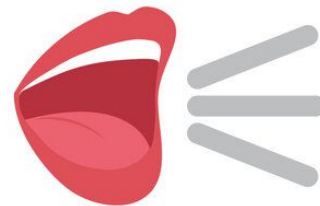


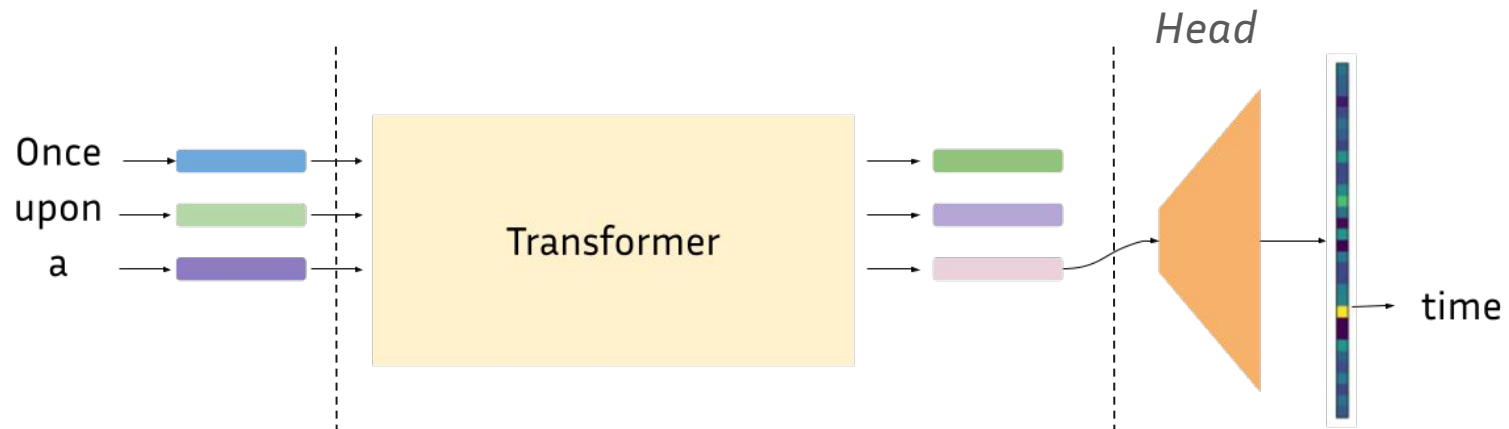
Internal representations





Internal representations

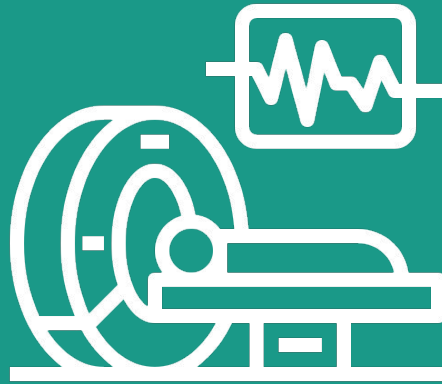




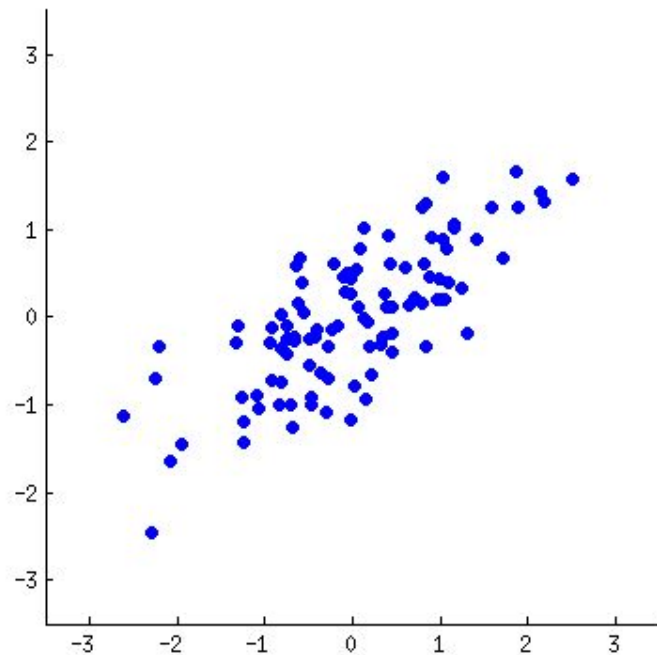
Internal representations



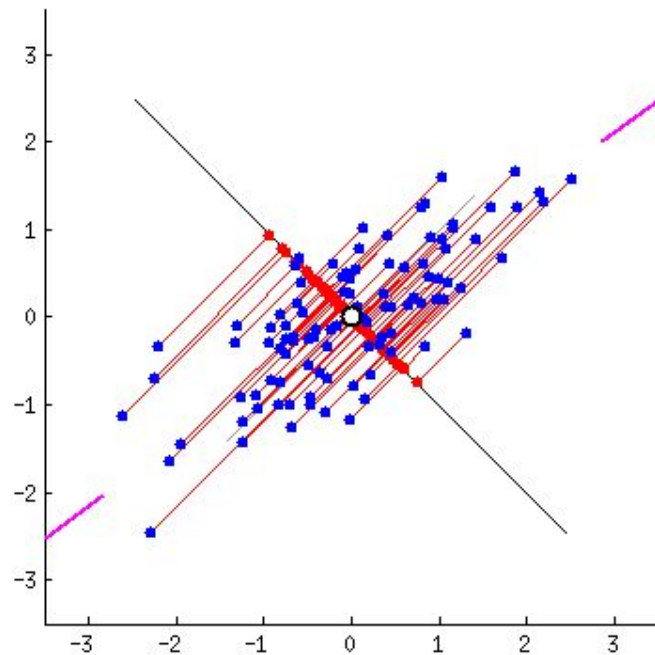
CAT-scanning LMs



high dimensional

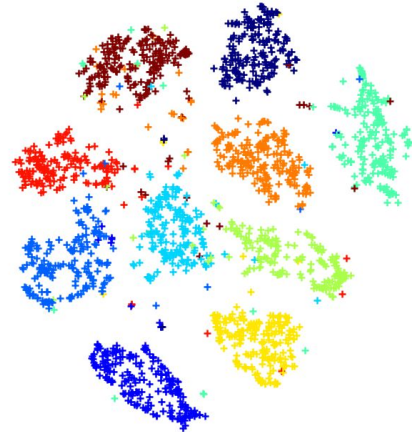


low dimensional



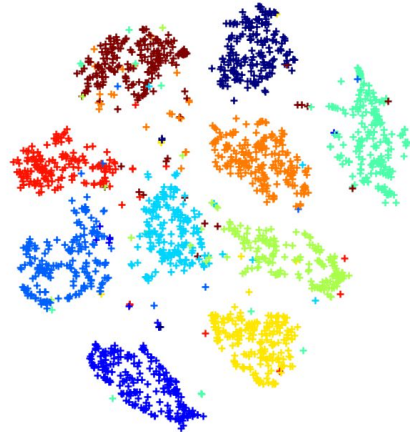
Singular Value Decomposition

EXPECTATION



REALITY

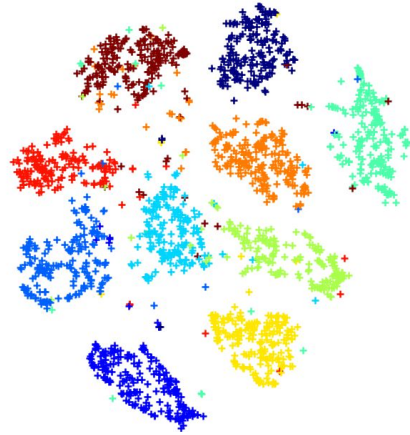
EXPECTATION



REALITY



EXPECTATION



REALITY



Anisotropy, n .

1. Non-uniformity in every direction.

Anisotropy, n .

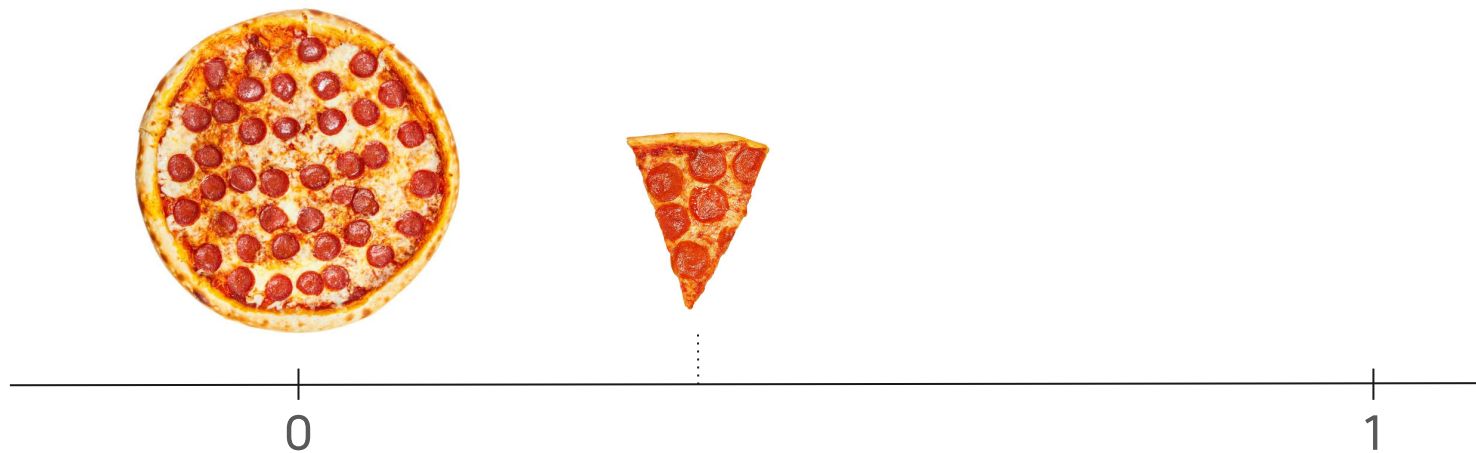
- ~~1. Non-uniformity in every direction.~~
2. How cone are you?



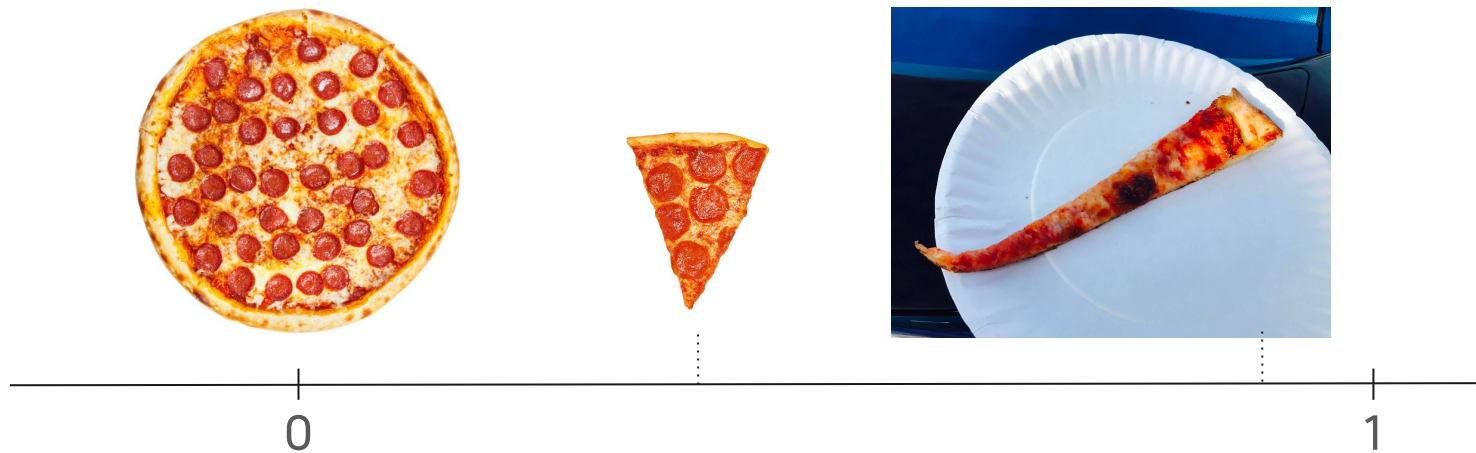
Average cosine-similarity



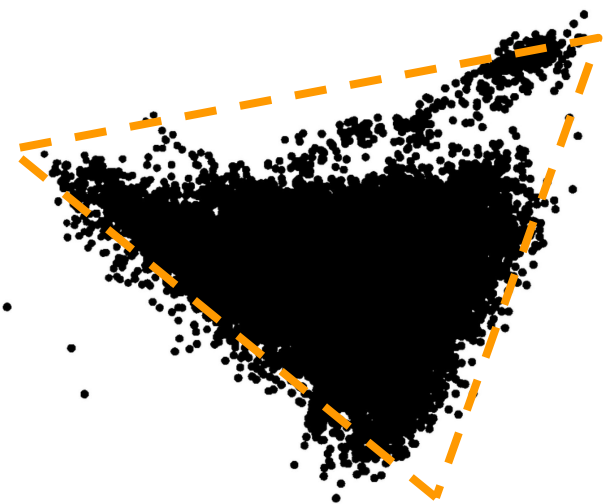
Average cosine-similarity



Average cosine-similarity

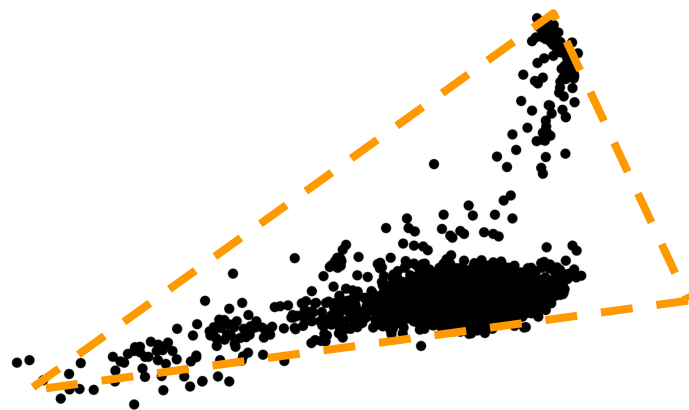


Average cosine-similarity



BERT

.4

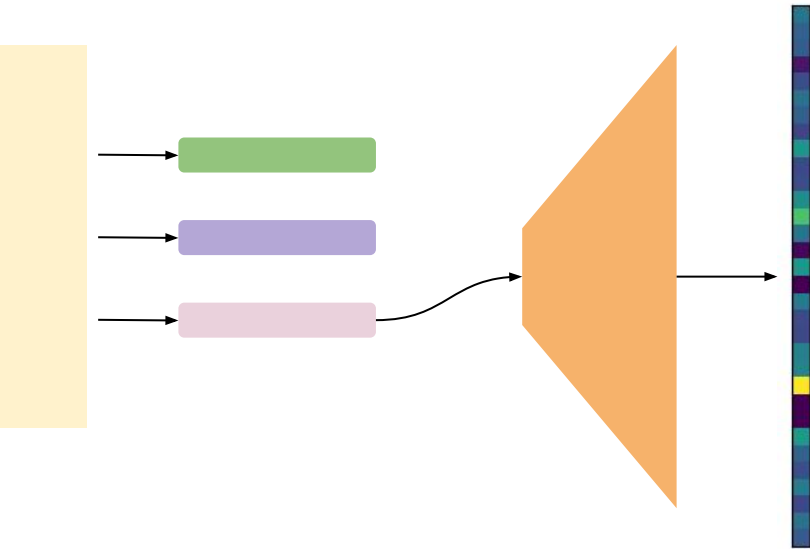


GPT-2

.97 1

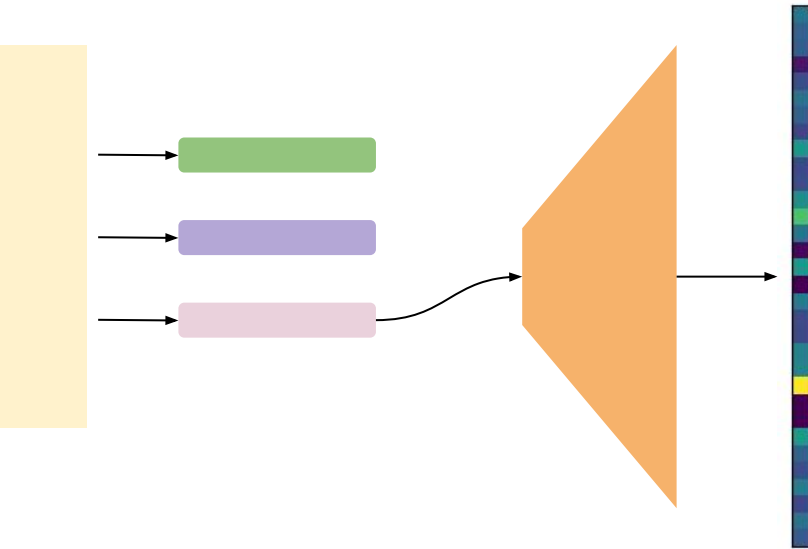
Why?

Example: Once upon a  time



Cross-entropy paradigm.

Example: Once upon a  time

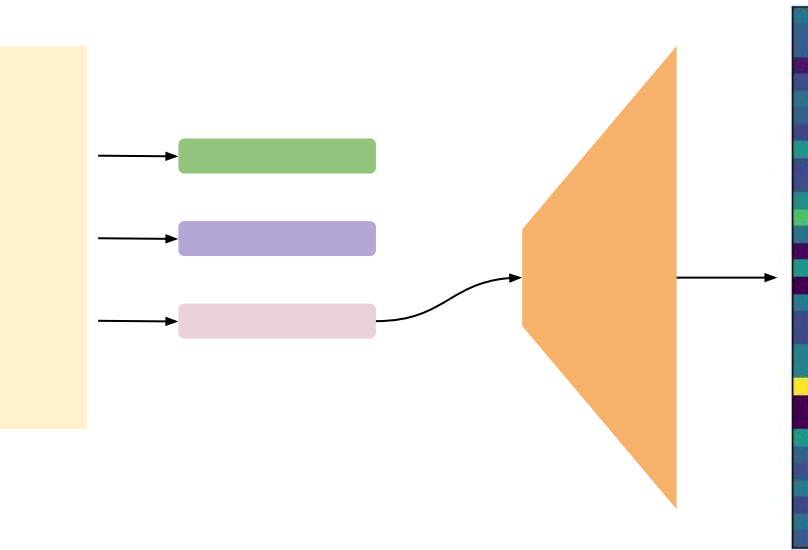


Predict:

- time 

Cross-entropy paradigm.

Example: Once upon a time



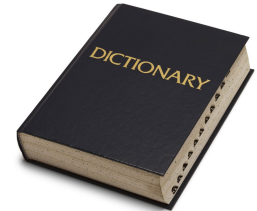
Cross-entropy paradigm.

Predict:

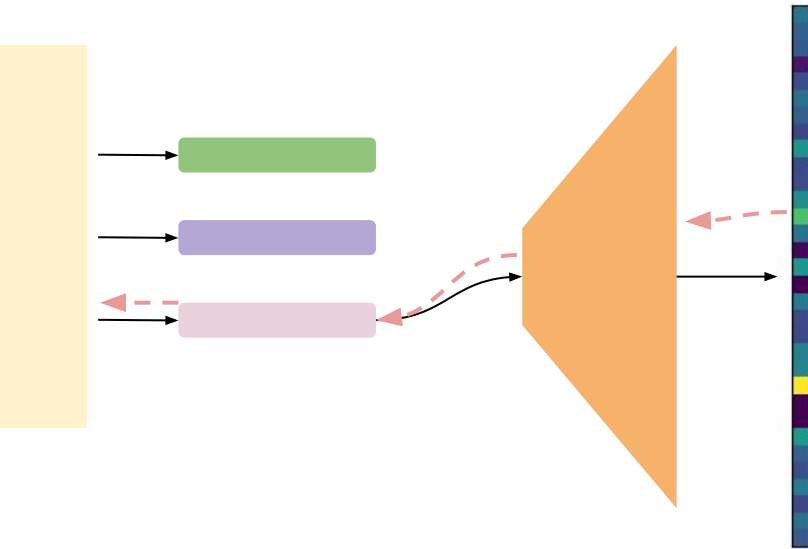
- time 

Don't predict:

- a
- an
- ability
- able
- about
- above
- ...
- zero
- zone
- zoo



Example: Once upon a time



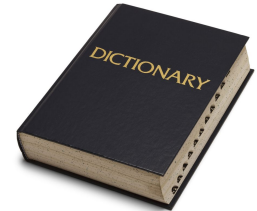
Cross-entropy paradigm.

Predict:

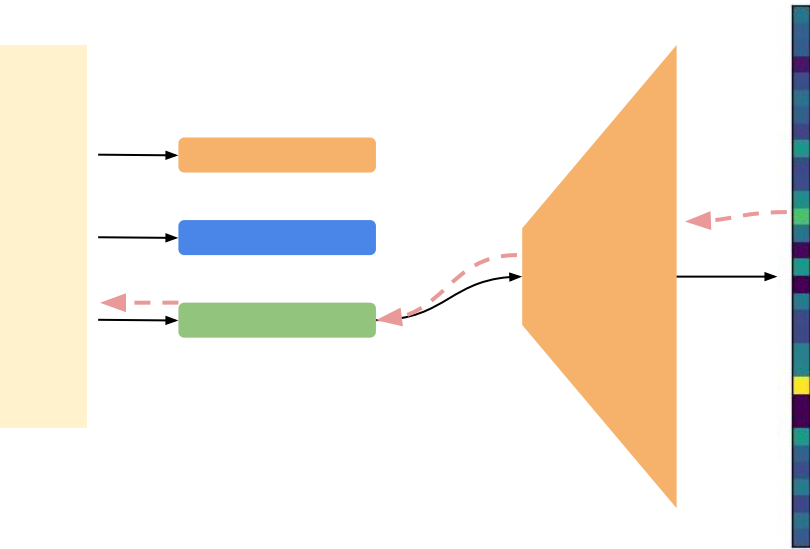
- time 

Don't predict:

- a
- an
- ability
- able
- about
- above
- ...
- zero
- zone
- zoo



Example: I really love 🎯 pizza



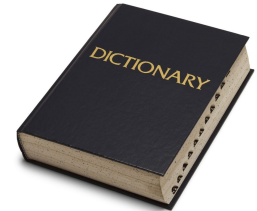
Cross-entropy paradigm.

Predict:

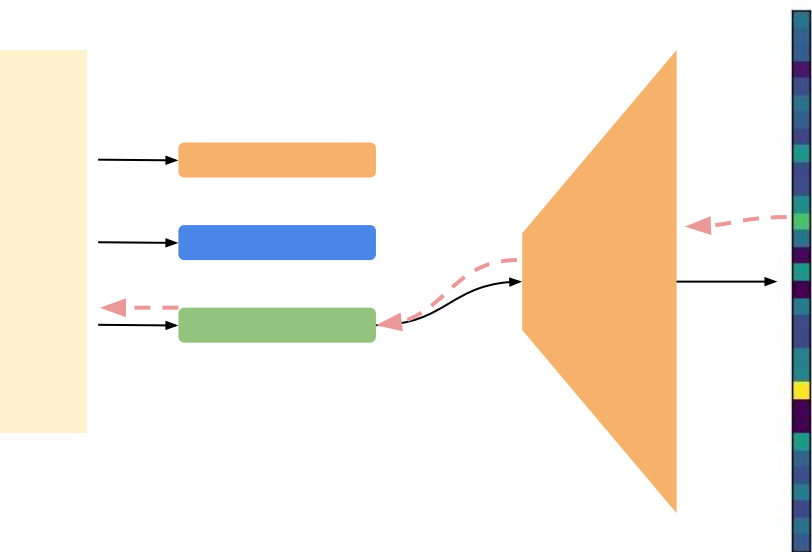
- pizza 🎯

Don't predict:

- a
- an
- ability
- able
- about
- above
- ...
- zero
- zone
- zoo




Example: This is not  about



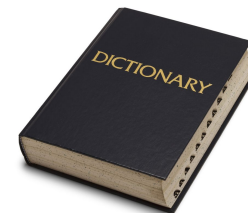
Cross-entropy paradigm.

Predict:

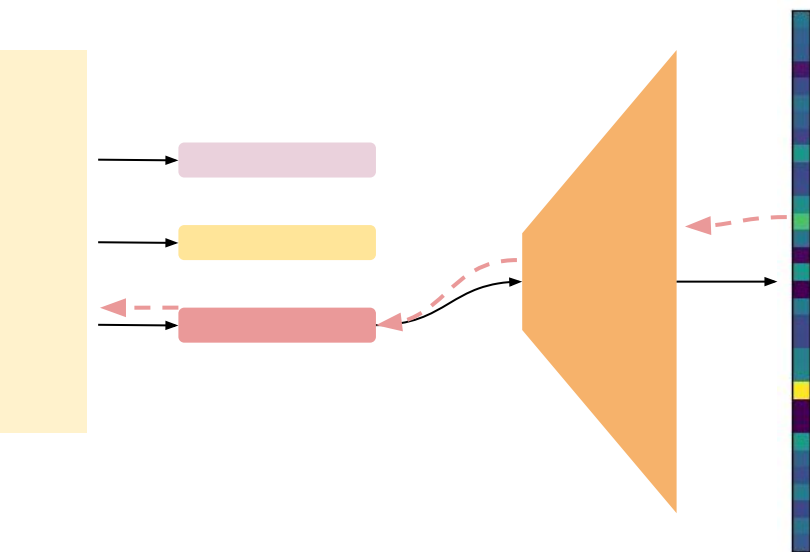
- about 

Don't predict:

- a
- an
- ability
- able
- **about**
- above
- ...
- zero
- zone
- zoo



Example: You begin to  understand



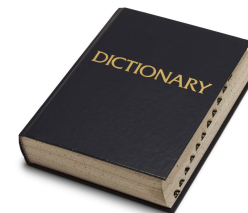
Cross-entropy paradigm.

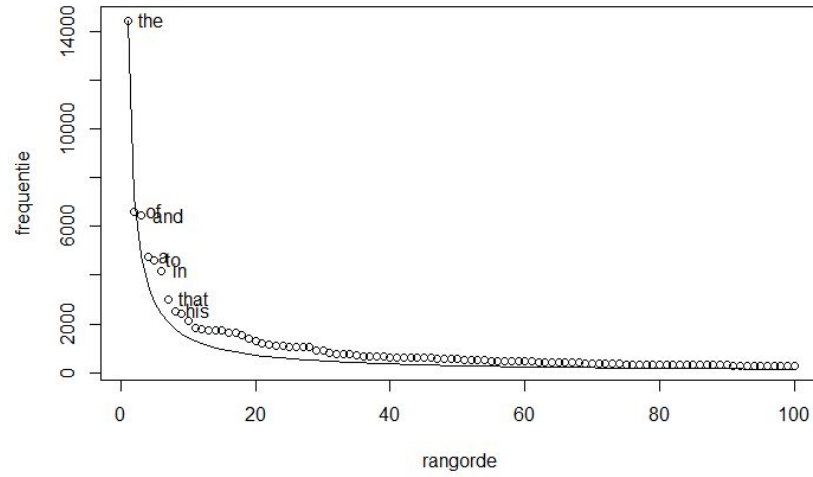
Predict:

- understand 

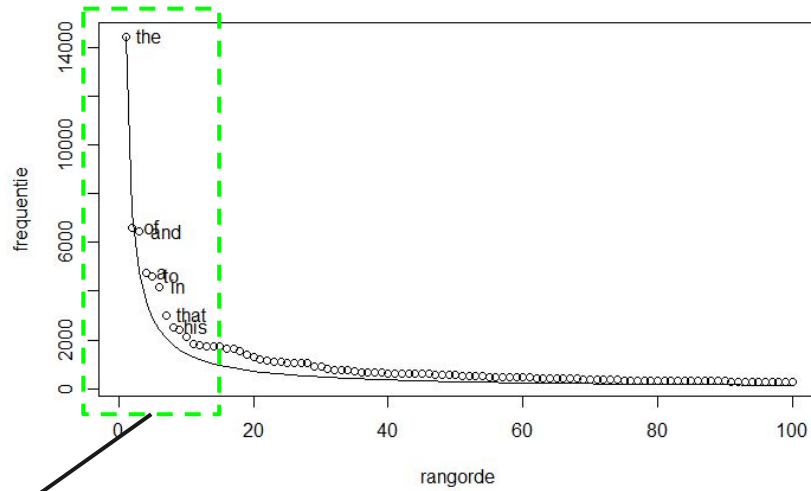
Don't predict:

- a
- an
- ability
- able
- about
- above
- ...
- zero
- zone
- zoo





Zipf's law and limitations.



the

of

and



5%



3%



2%



100%

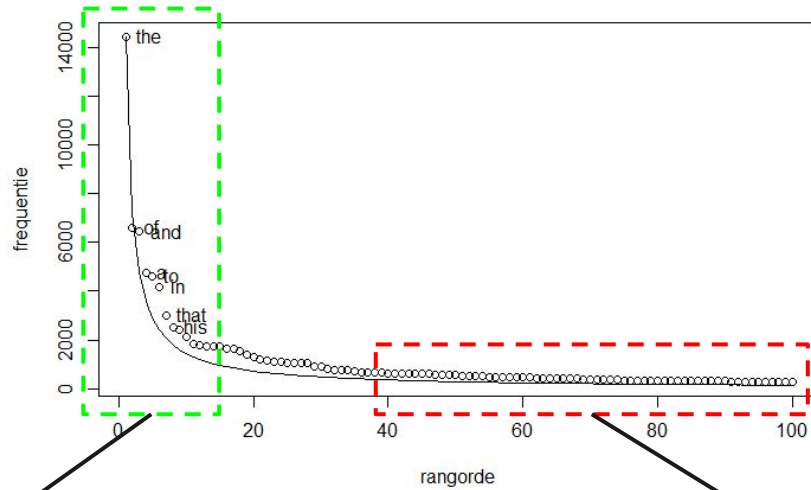


100%



100%

Zipf's law and limitations.

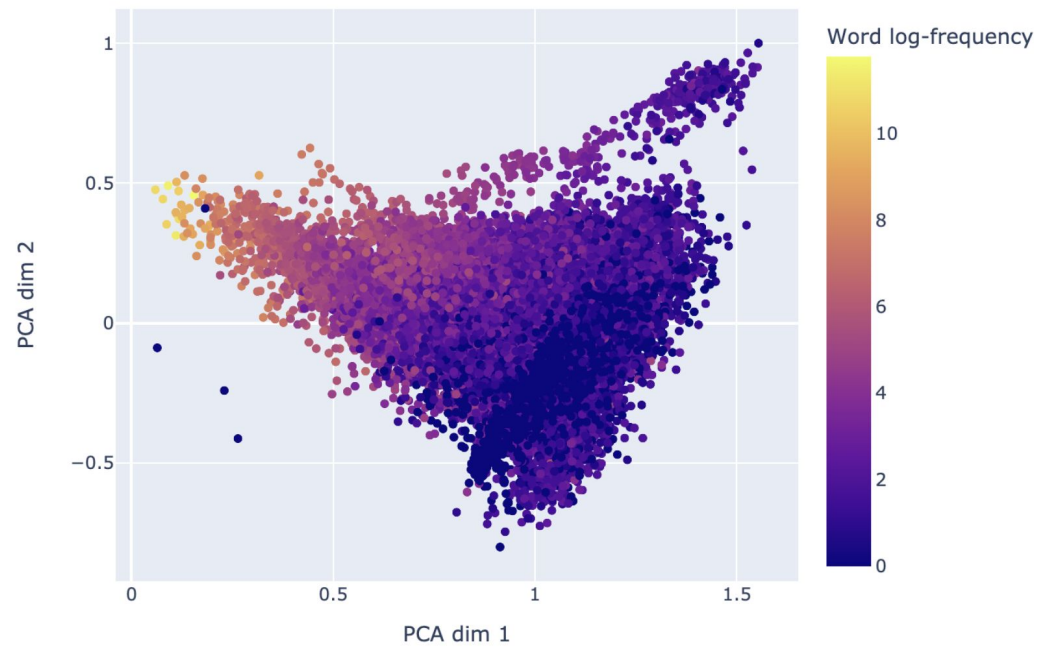


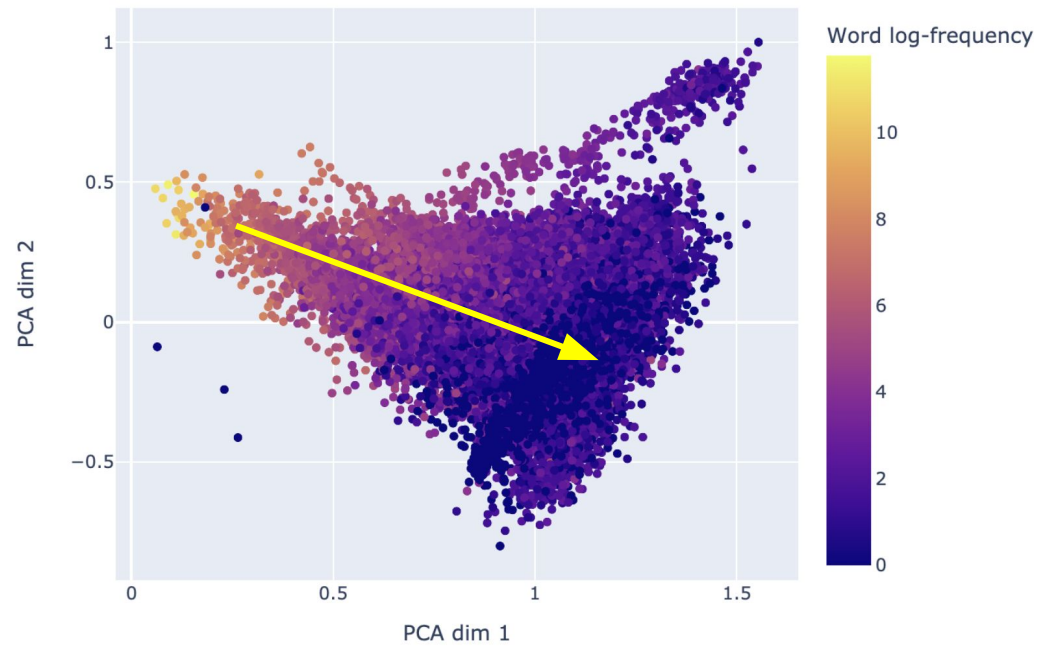
the	of	and	...	Gobbledygook	xhheo
5%	3%	2%		1e-8%	0%
100%	100%	100%		100%	100%

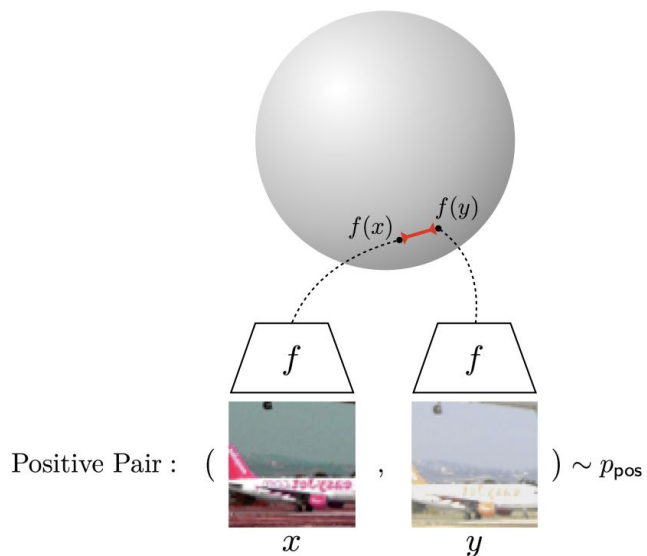
Zipf's law and limitations.

*“...the embeddings of **most words** in the vocabulary will be **pushed towards similar directions negatively correlated with most hidden states...***”

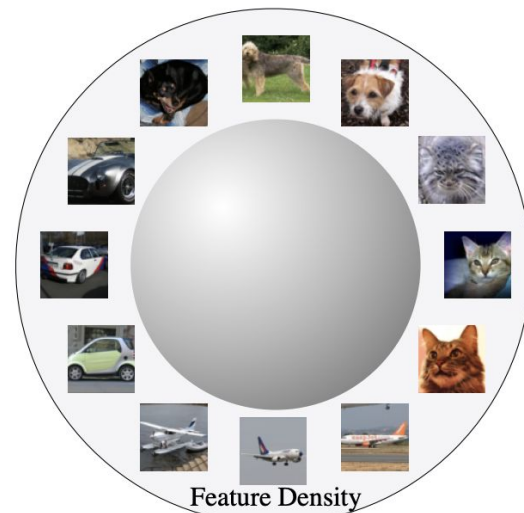
- Jun Gao et. al, 2019







Alignment: Similar samples have similar features.
 (Figure inspired by [Tian et al. \(2019\)](#).)

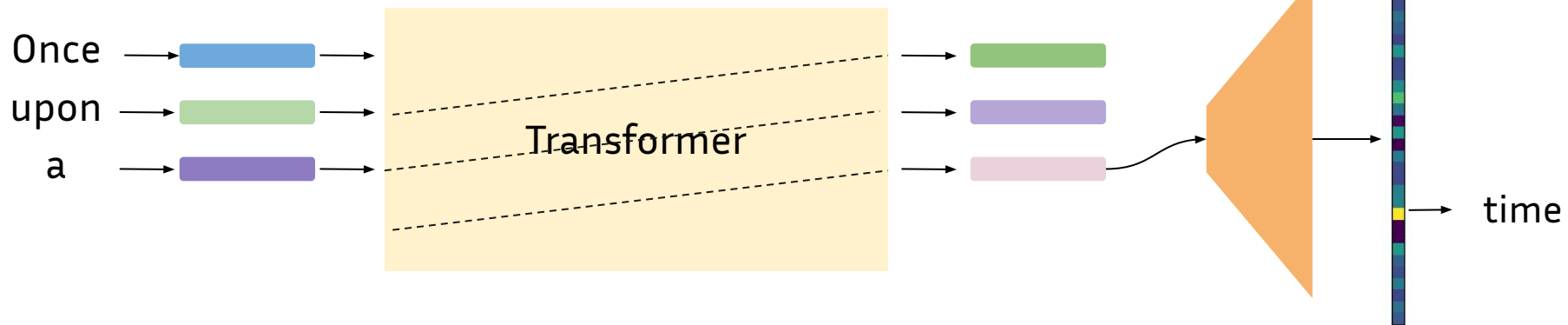


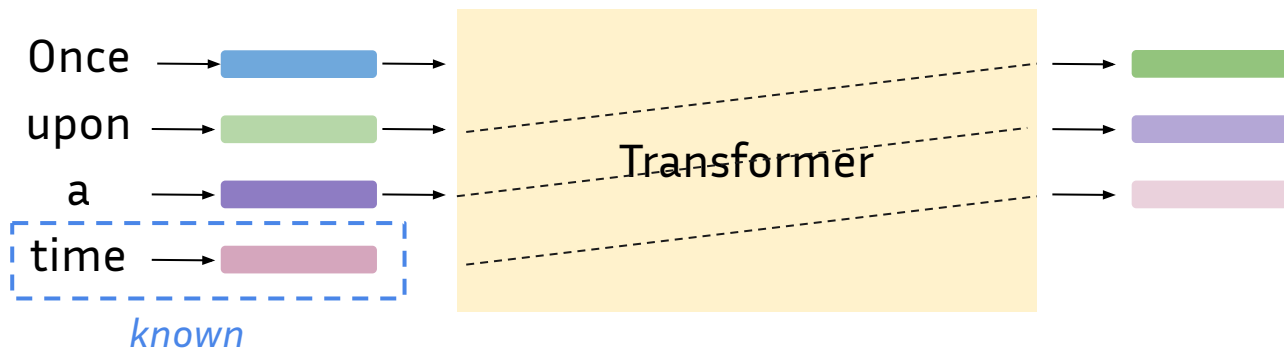
Uniformity: Preserve maximal information.



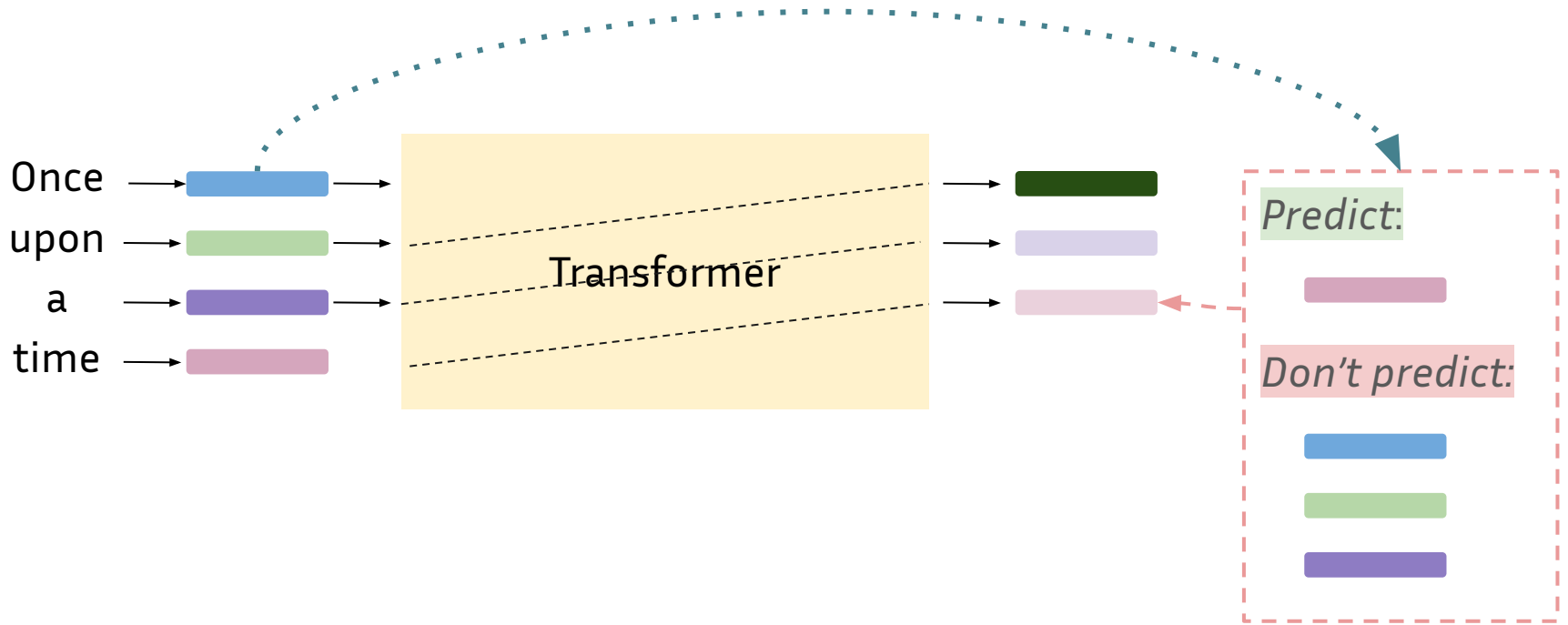


Our fix

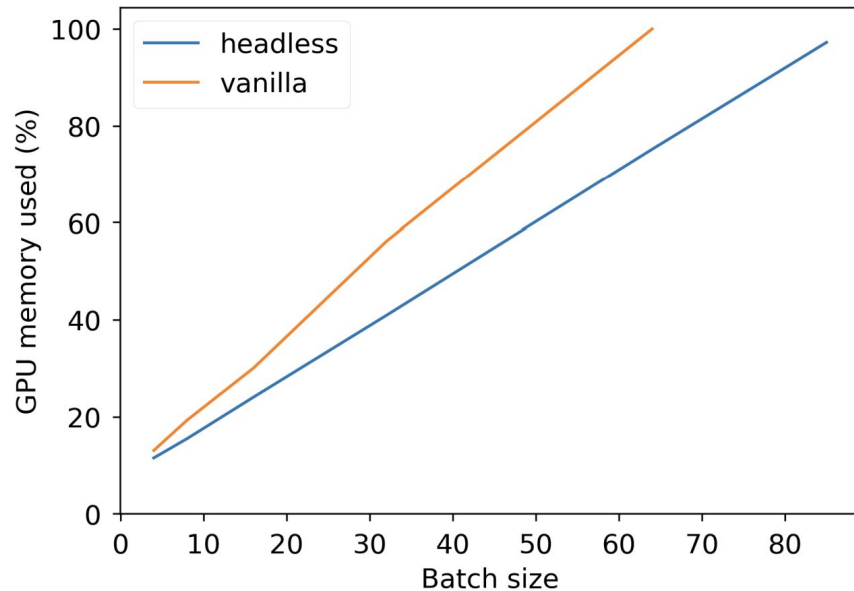
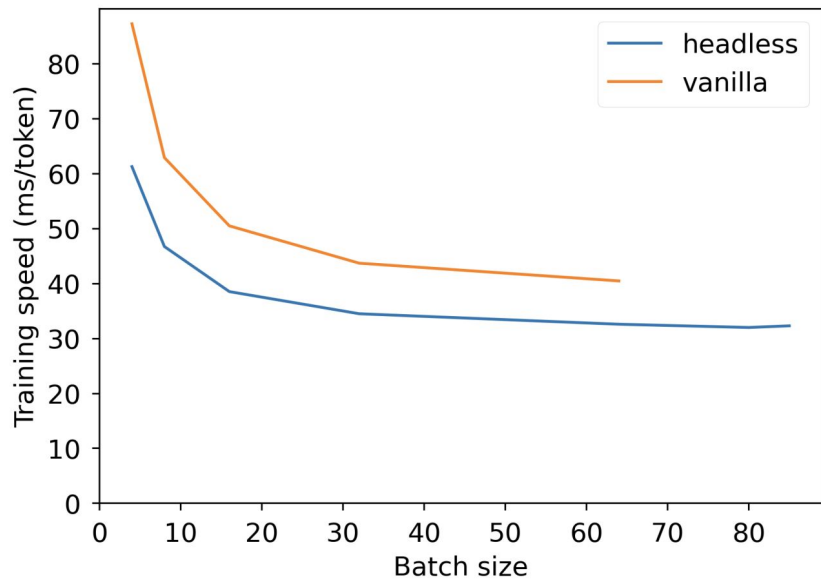




Headless pre-training.

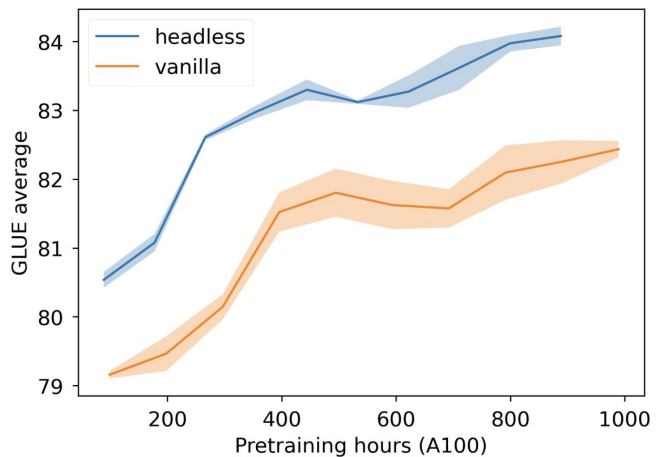


Headless pre-training.

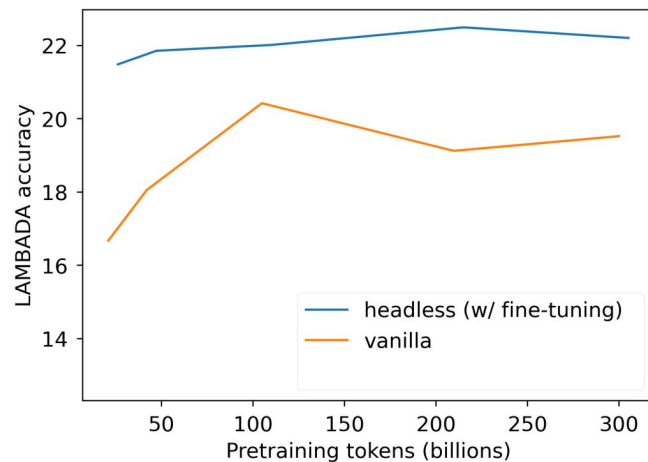


Faster and more memory-friendly.

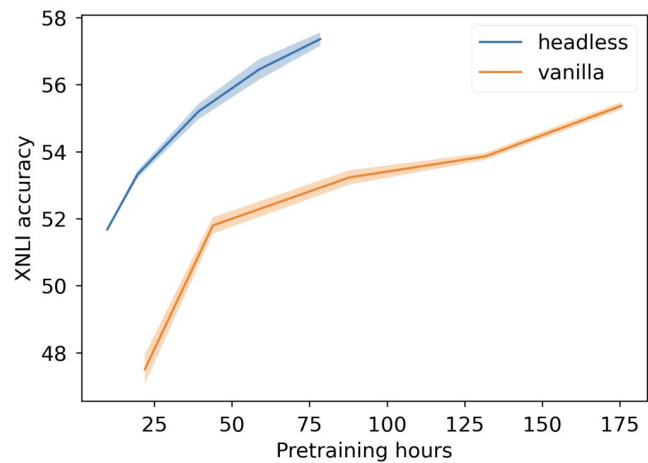
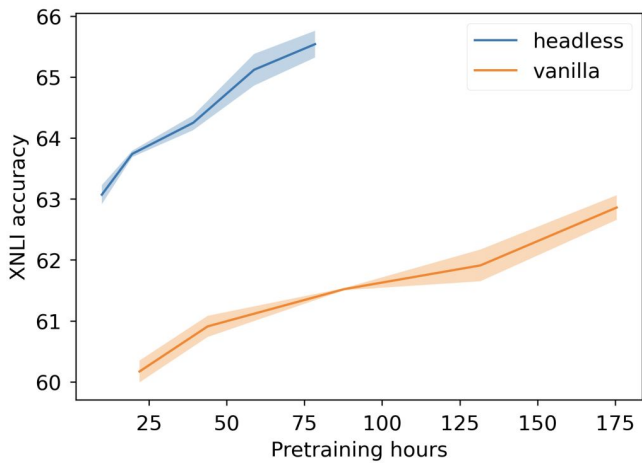
English BERT

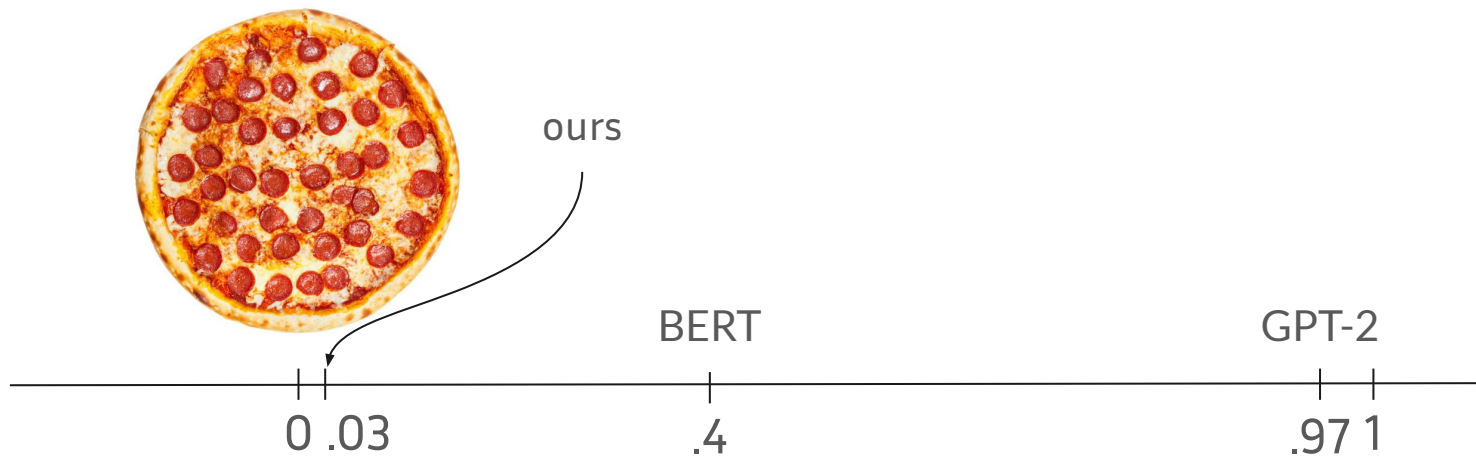


English GPT



Multilingual BERT-small





Average cosine-similarity



Conclusion



Conclusion



Thank you!

