kokchun giang

how do ChatGPT work? An introduction to **large language models**

how can we represent text for a computer? Naive approach

- computer doesn't understand text
- it understands numbers (which can be represented in binary 1 & 0

=> need to represent text with numbers

Suppose our vocabulary has these words:

[hej, kanin, fisk, då]

=> hej can be represented with

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

one-hot encoded vector

126000 swedish words

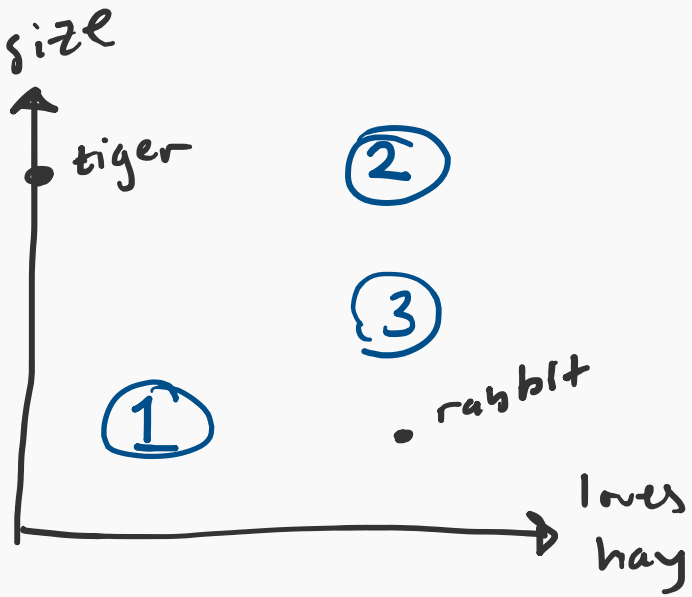
=> get a very sparse vector

but we also want semantic meaning between similar words

how can we represent text for a computer? embeddings

word2vec 2013

→ represent words w. vector embeddings that captures semantic meaning



where does cow & calf go?

however we have larger dimensional embedding vector

to find similar words we use dot product, which gives cosine similarity betw. vectors

attention is all you need 2017 — transformers architecture

goal: predict next word based on previous sequence

ex. hur är $\boxed{\text{läget}}$?
hur mår $\boxed{\text{du}}$?

for this we need to understand context through attention

ex.
I am cool

The ice cream is cool

compute similarities to see which words that determine the context for cool as it has diff. meanings in diff. contexts

With the transformer we can generate text word by word using previous words as context

Ex
I am
I am $\boxed{\text{cool}}$
I am cool. $\boxed{\text{Yo}}$
I am cool. Yo $\boxed{\text{zup?}}$

gpt – general pretrained transformers

trained with unsupervised learning on large corpus of text - pretraining

---

using internets text it can predict most probable next word to generate

add temperature & we get variations & "creativity"

---

finetuned to specific tasks using supervised learning

for example chat in a certain format

rlhf – reinforcement learning with human feedback

we let the model answer a set of questions several times

then let humans score these answers and feed back to the system

the system will try to maximize its scores

Ex. an answer of how to make a bomb will get low scores

model will try predict which type of answers humans like