Word Embeddings and LLMs

INF1 CG

Week 9, Lecture 24

Maithilee Kunda

Review: Vector semantics

- Represent the "meaning" of a word (**semantics**) using a list of numbers (**vector**)
- Why did I put "meaning" in quotes??
- In pure vector semantics, there is **no** link between a word and its actual meaning!
- How do we approximate meaning? Fill in the blank:

- Words that have similar meanings will **Be used in similar ways, i.e., if you have two synonyms, the typical word-neighborhoods around these synonyms will be similar!**
- What do we call the word-neighborhood of a given word? **Context**
- What do we call the list of numbers that captures some measurements of a word's context? **Context vector**

Ennugu pilli kante peddadi

Pilli ennugu kante cinnadi

Pilli eluka kante peddadi

Eluka pilli kante cinnadi

Ennugu eluka kante peddadi

Ennugu kukka kante peddadi

- Last time, I showed you sentences in Telugu.
- You were able to "correctly" fill in the blanks.
- Does this demonstrate **understanding**?

- What is missing?

No symbol grounding!

You've got a lexicon, you've got a grammar...

Searle's Chinese Room Argument

- John Searle, philosopher, 1932-2025
- See him explaining it in 1984:

https://www.youtube.com/watch?v=6tzjcnPsZ_w

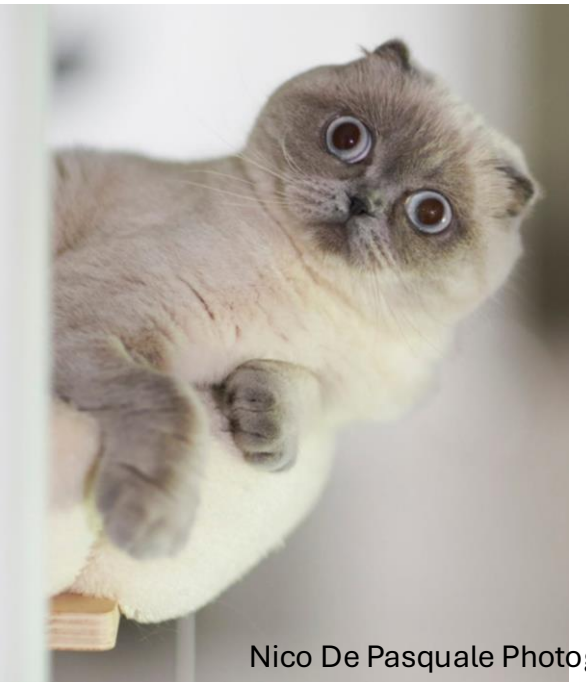
- Imagine I put you in a room and gave you boxes with a bunch of Telugu words and sentences in them, along with instructions for which ones to put together when.
- Then I feed you questions in Telugu, and you follow the instructions to give me back the correct answers, in Telugu.
- **Do you "understand" Telugu?**

Searle had his "emeritus professor" status in the UC Berkeley philosophy department revoked in 2019 following a university investigation of sexual harassment claims against him.

How would you represent the meaning of a word so you can ground your symbols?

- Idea: What if we link words to pictures?

“cat”

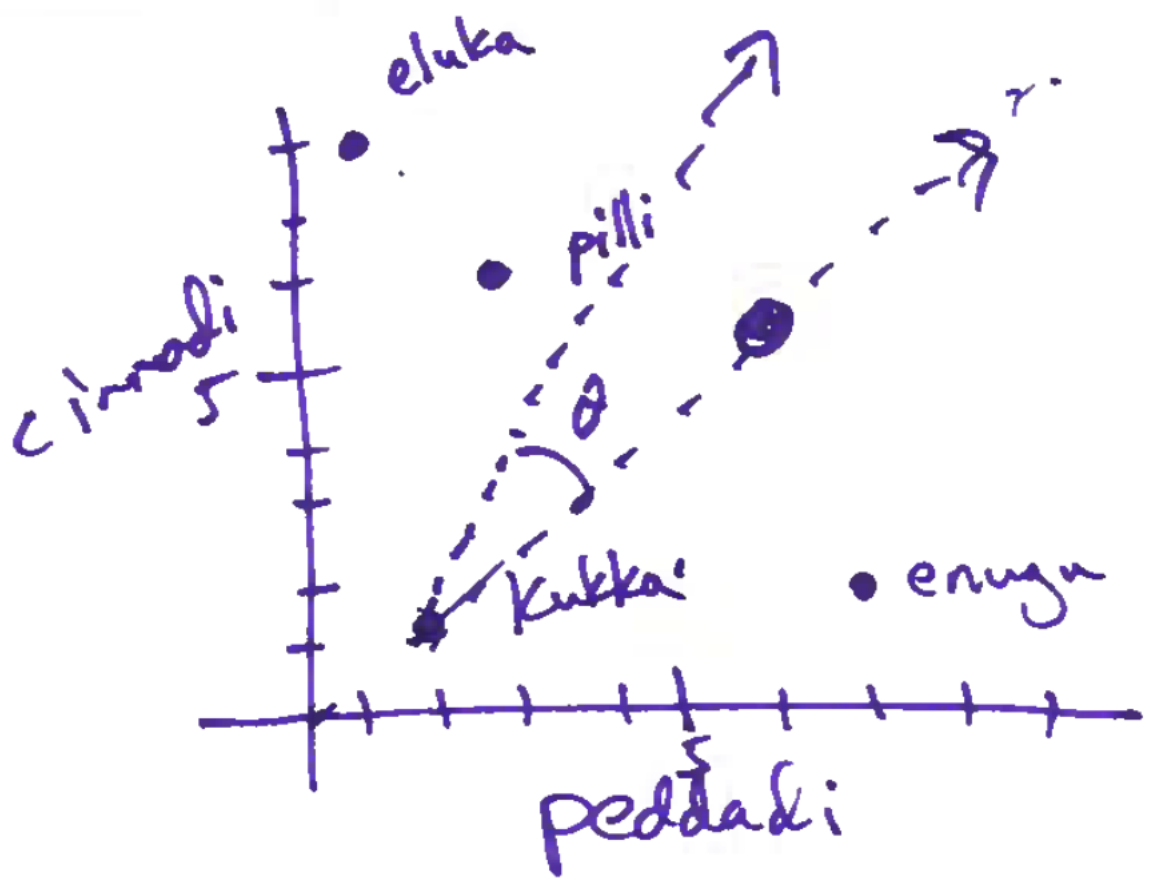


Nico De Pasquale Photography//Getty Images

- Is a picture of a cat a cat?
- **Still no!**
- **What IS a cat???**
- Many things… including a real physical thing in the real physical world.
- Maybe we (humans) ground our symbols by how we experience things in the physical world, with our physical bodies and senses

=> Embodiment or Embodied cognition

But! Forget about bodies for now… back to vector semantics. How far can we get without a body??



- With vector semantics, you can infer relationships between words
- Synonyms
- ranking similarity
- different meanings of the same word (depending on different subspaces of the semantic space)
- Etc.

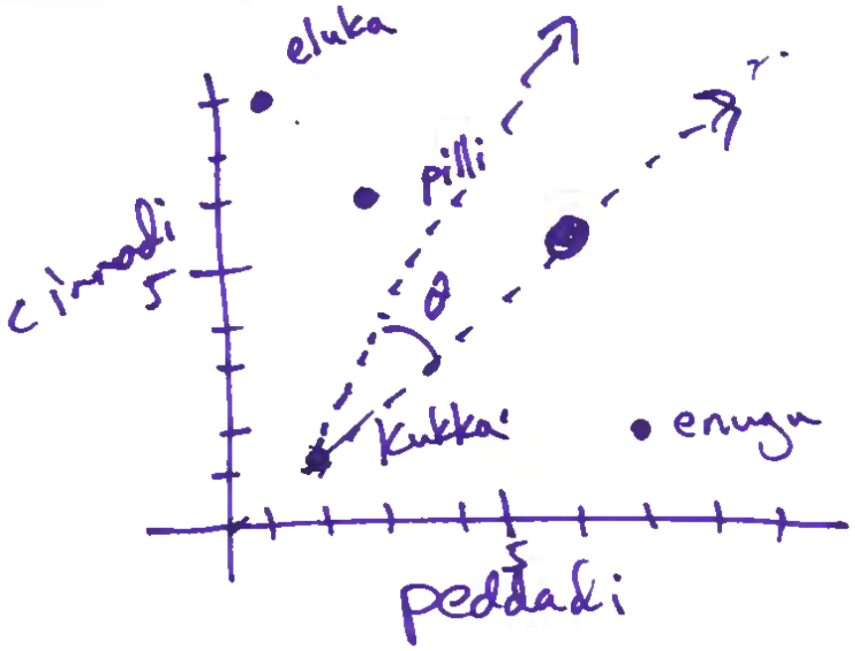
From vector semantics to LLMs: 3 big ideas

1. Learning the embedding

2. Attention / transformers

3. Fine-tuning to build a better chatbot

How did we get these vectors before?



- Count co-occurrences of the main word (“eluka”) near other words (“cinnadi” and “peddadi”)
- As they co-occur within a context (a neighborhood of some given size)
- In some given dataset of language

- With a lexicon of size N, how big is each context vector?
- Length N! This is big. 50,000 words => 50,000 dimensional space
- What are most of the entries in a given vector going to be?
- Zero! Most words don’t occur near most other words.
- We call this a “sparse” vector

Problems with “count the co-occurrences” approach

- Each context vector is big and mostly zeroes (“sparse”)
- Sparse vectors are difficult to work with because it’s mostly just nothing, the amount of “signal” is small

- Also… no relationships between dimensions!
- ”Wheel” co-occurs with “car” and with “automobile”
- But the “car” and “automobile” places in the vector are treated as two completely different counts / two completely different dimensions

- **What to do???**

World Premiere!
Kunda's Alien Bucket Argument

- Eventually… the buckets show some useful organization!
- One bucket for fruits. One bucket for pizza toppings, etc.
- Can use the buckets to generalize! Blueberry muffins -> cherry muffins

World Premiere!
Kunda's Alien Bucket Argument

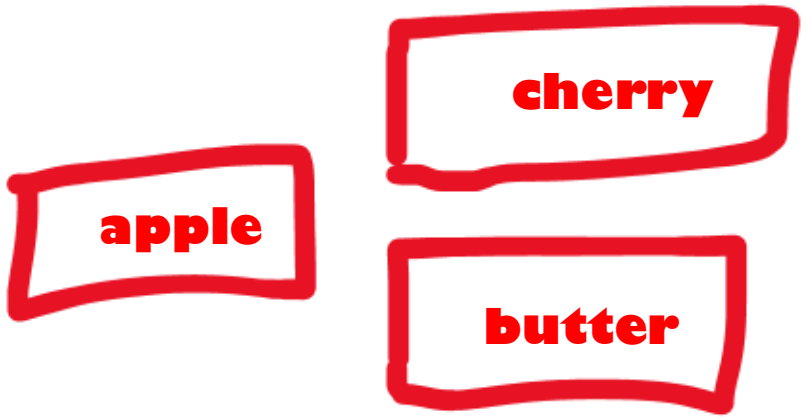
- What happens if there are too few buckets? Like what if there are only TWO buckets???
- What happens if there are too many buckets, like more buckets than words in the whole lexicon? **Buckets too sparse!**
- What happens if the recipes are bad? **Garbage in, garbage out…**

World Premiere!
Kunda's Alien Bucket Argument

- Does the alien know what an apple is?
- Does the alien know what a recipe is?
- Does the alien know what a fruit is?

- This is a learned word embedding!

- This is the training data!



- These are the tokens!
- We can also let the alien rip up words into pieces, or tape words together to make common phrases
- ”Learning” the tokens, i.e., learn the lexicon

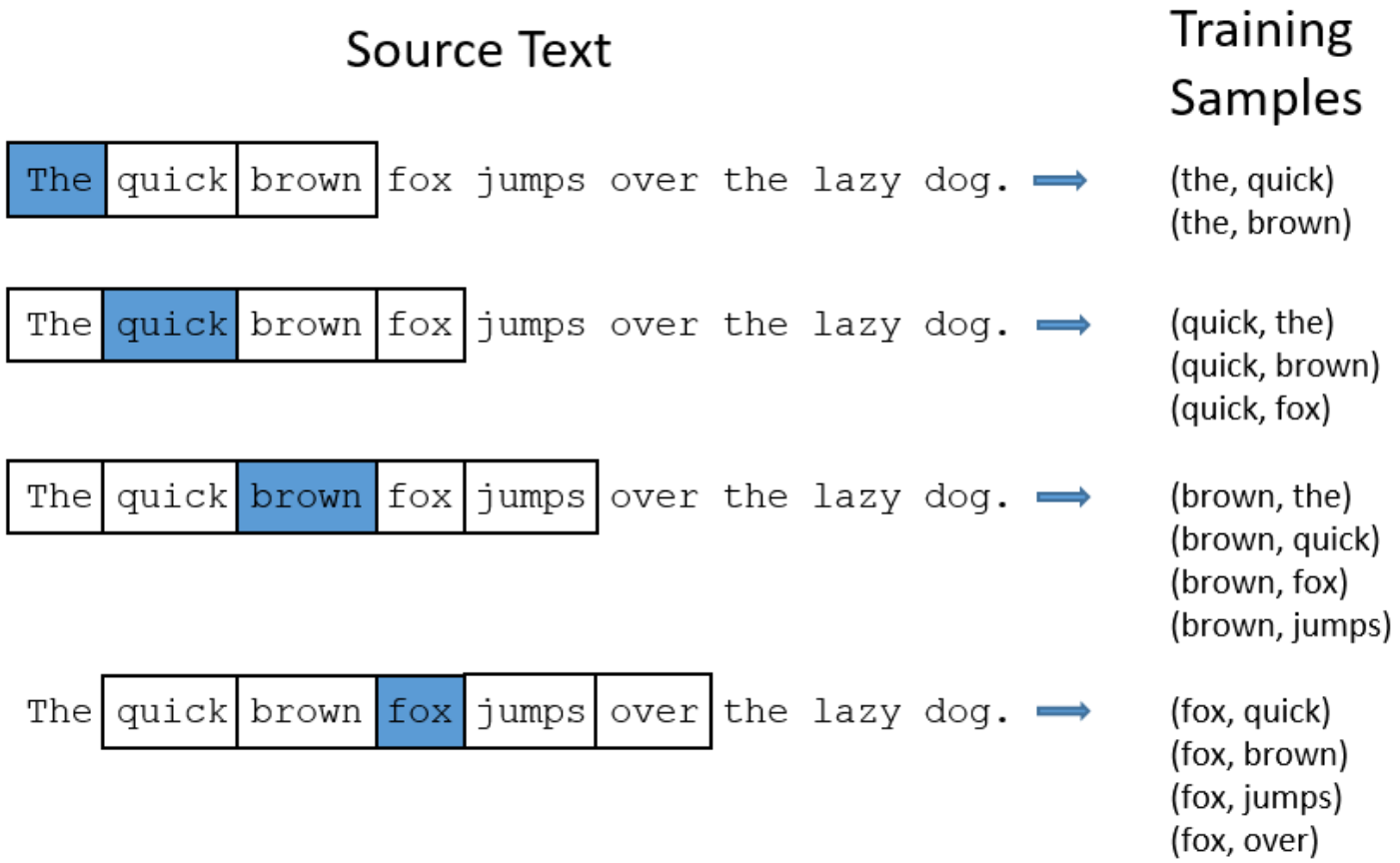
History of these ideas…

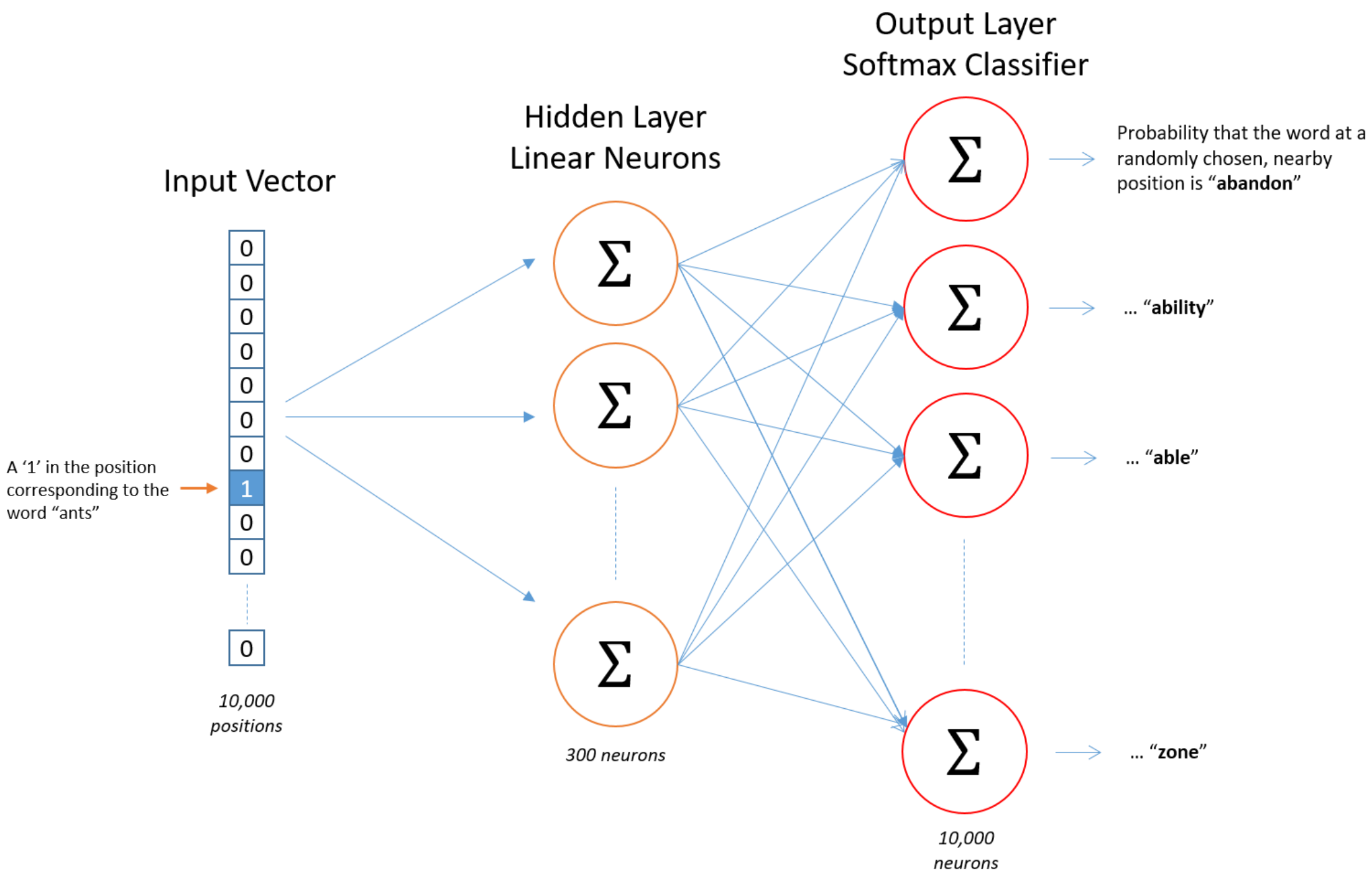
- 1990s – Use context vectors to help with information retrieval (e.g., find two documents that are related)
- 2000s – Use neural networks to learn the embedding instead of using manually defined dimensions
- 2013: Efficient Estimation of Word Representations in Vector Space ---> “Word2Vec” model

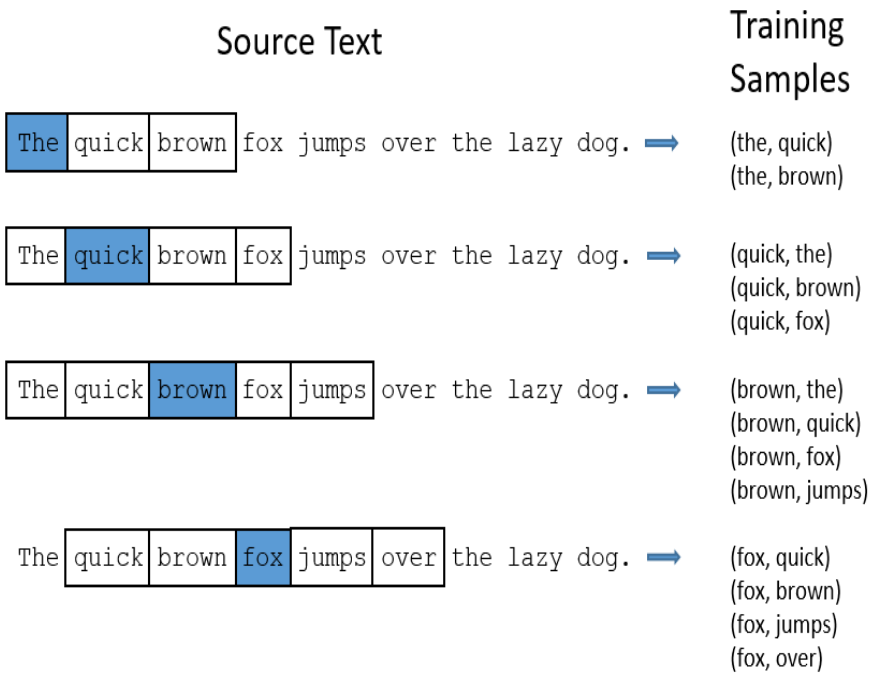
Word2Vec – two training methods, we'll just talk about one: The Skip Gram model

- https://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

- Training task:

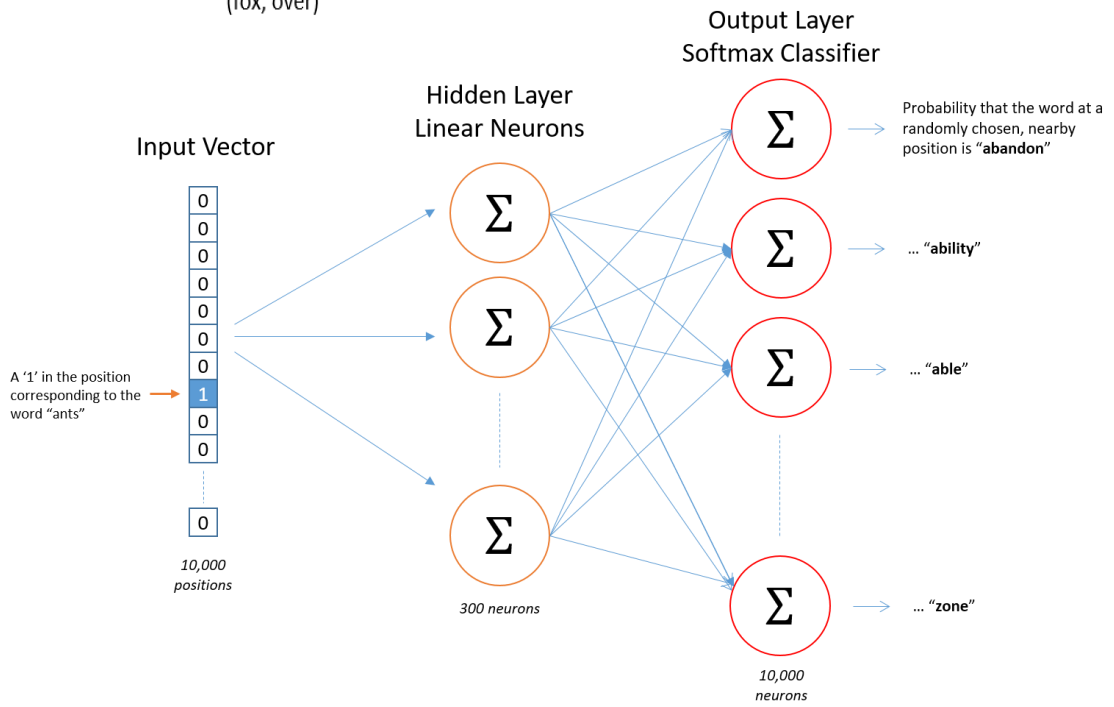



Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
10,000 neurons



Result:

- A really good bucket system! (embedding)
- AND we didn't even need to label the data beforehand!
- "Self-supervised learning" – learn from correct answers and errors, just like supervised learning, but the "correct" answers are already in the data



Embedding to LLM……

- 1990s – Use context vectors to help with information retrieval (e.g., find two documents that are related)
- 2000s – Use neural networks to learn the embedding instead of using manually defined dimensions
- 2013: Efficient Estimation of Word Representations in Vector Space ---> “Word2Vec” model
- 2017: Google paper presenting “transformer” neural network architecture: “Attention is all you need”

The cat left its litter mates and uses the litter

_____

- What is the problem here for word embeddings?

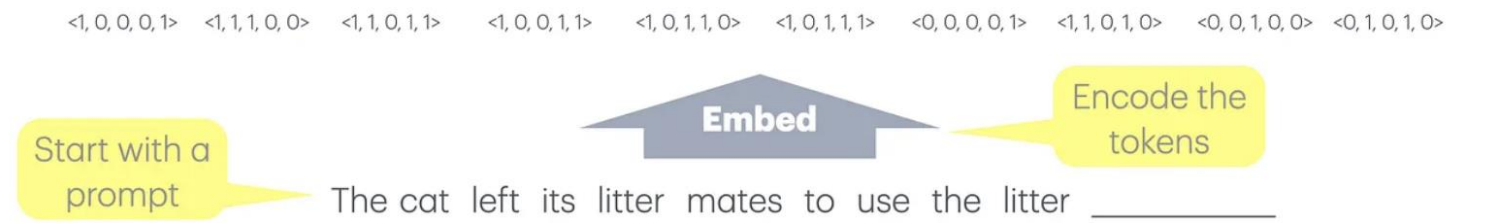
- ”Litter” gets used twice, but it means completely different things!
- But it goes to the SAME embedding.

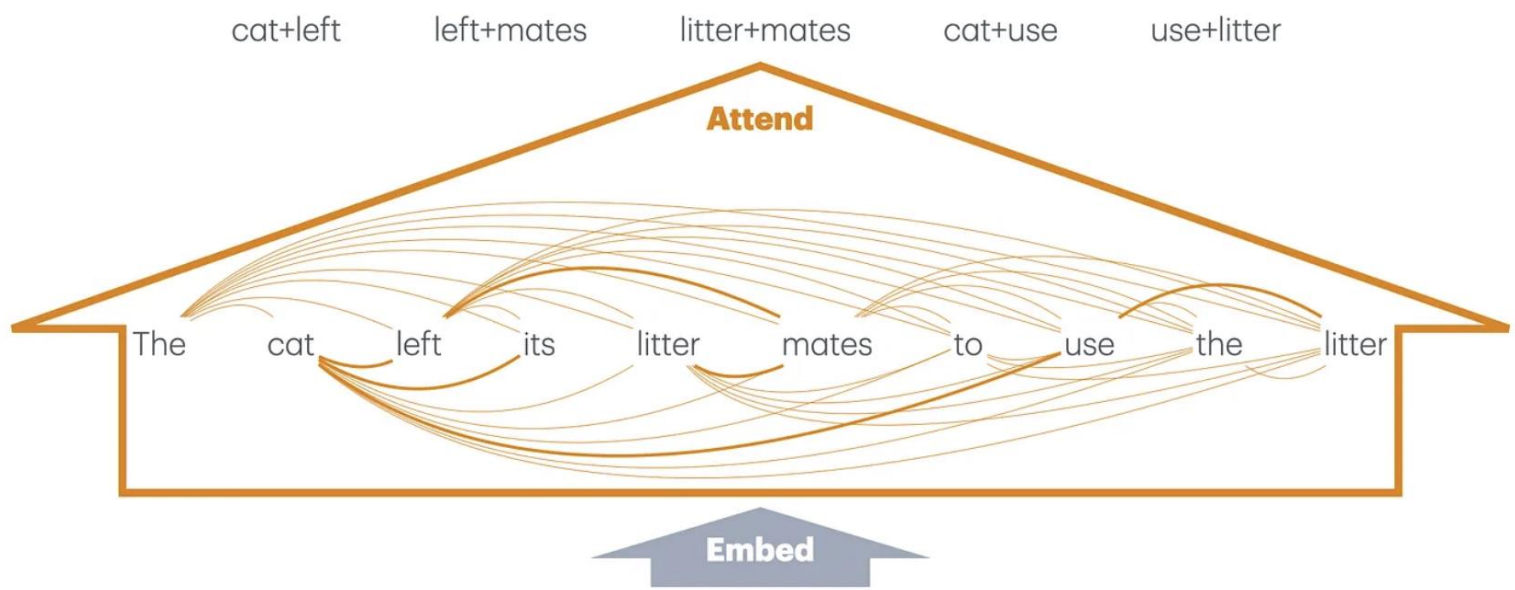
- How do WE know that litter means two different things?
- Context again!

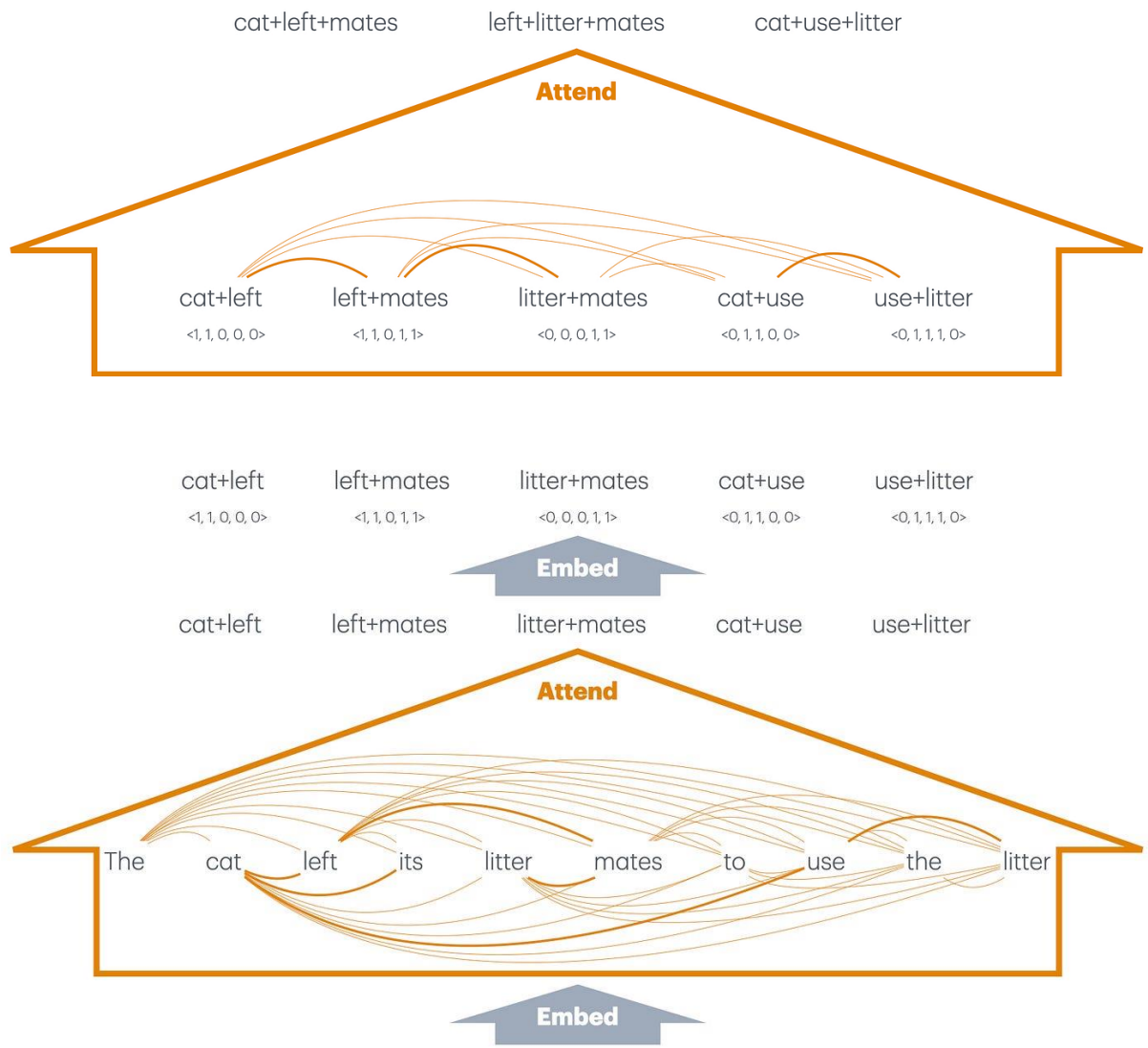
- https://mark-riedl.medium.com/the-intuition-behind-how-large-language-models-work-166cf2fb278a

The cat left its litter mates and uses the litter

_____





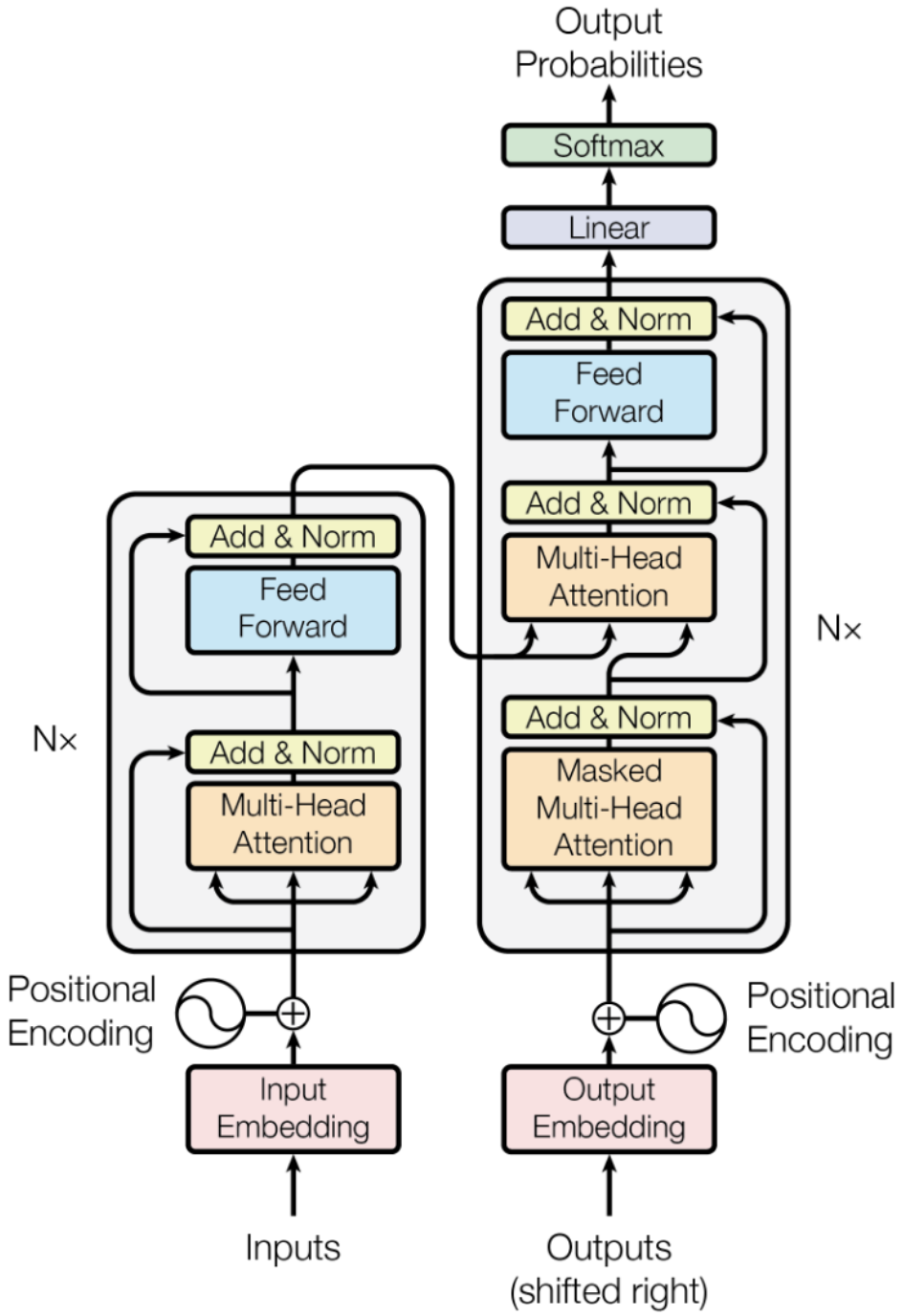


The cat left its litter mates to use the litter _____

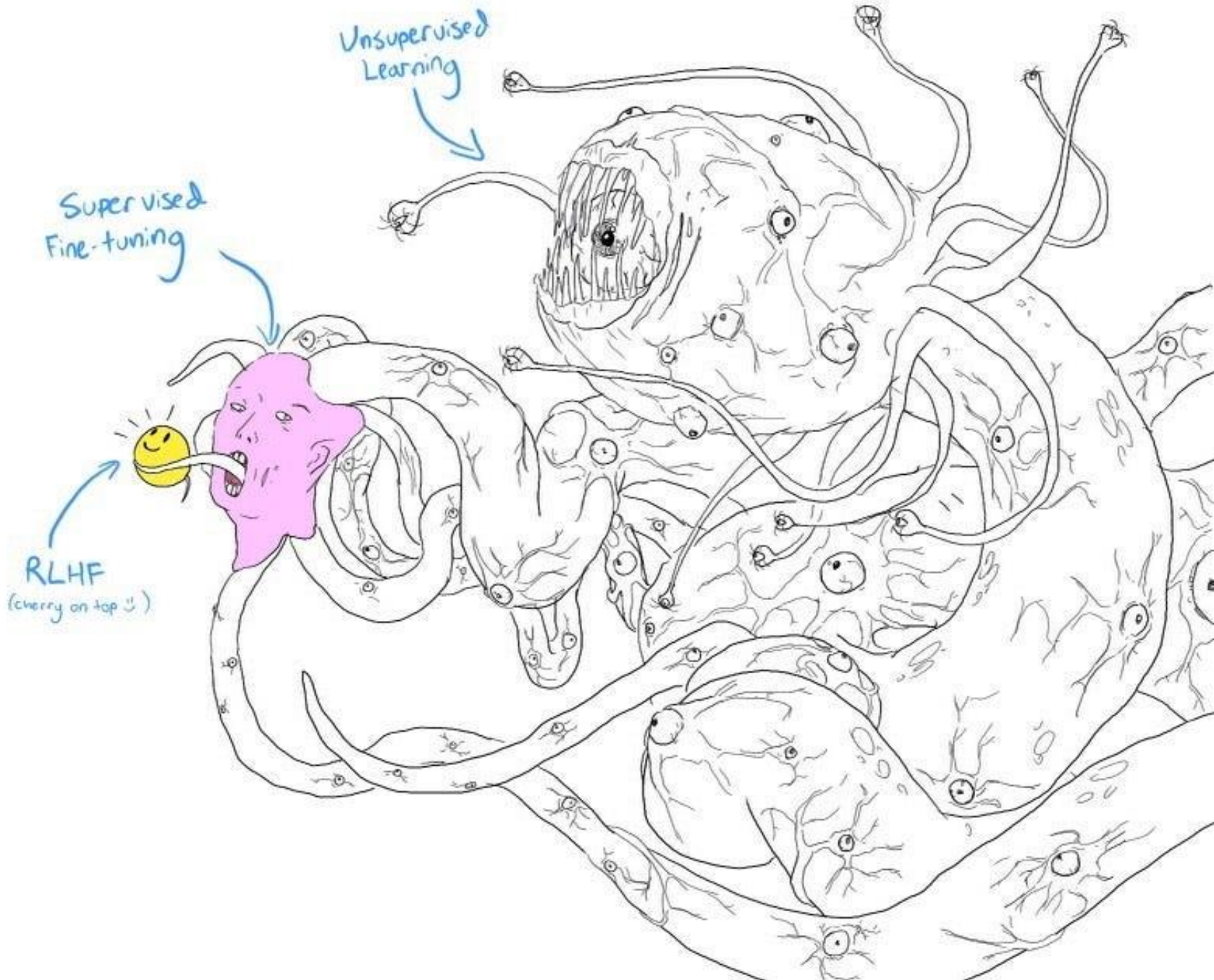
Google's Transformer

- All of these are sets of nested buckets!
- There are convolutional buckets.
- Left: Encoder = words to buckets
- Right: Decoder = buckets to words

- The original Google paper trained the transformer on a translation task (e.g., English to German sentence pairs)



Tomorrow....


Unsupervised Learning
Supervised Fine-tuning
RLHF
(cherry on top :))