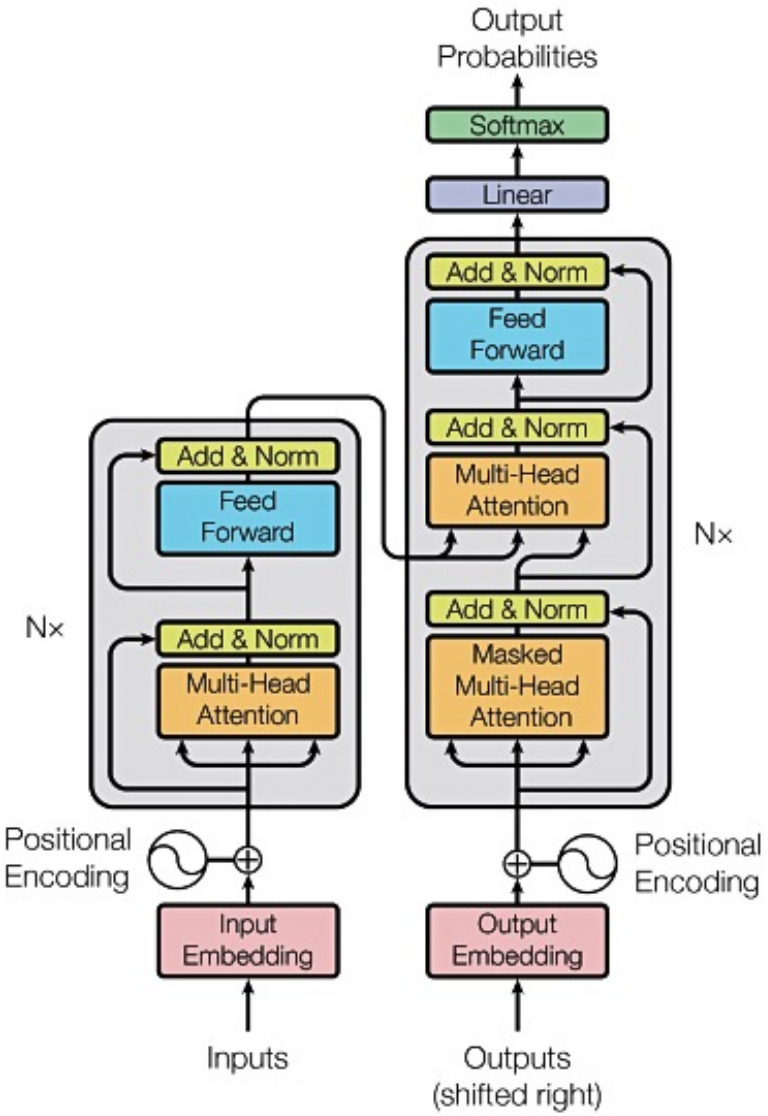


Transformers

Background (1)

- The **RNN** and **LSTM** neural models were designed to process language and perform tasks like classification, summarization, translation, and sentiment detection
  - RNN: Recurrent Neural Network
  - LSTM: Long Short Term Memory
- In both models, layers get the next input word and have access to some previous words, allowing it to use the word's left context
- They used **word embeddings** where each word is encoded as a vector of 100-300 real numbers representing its meaning

Learning word meaning?

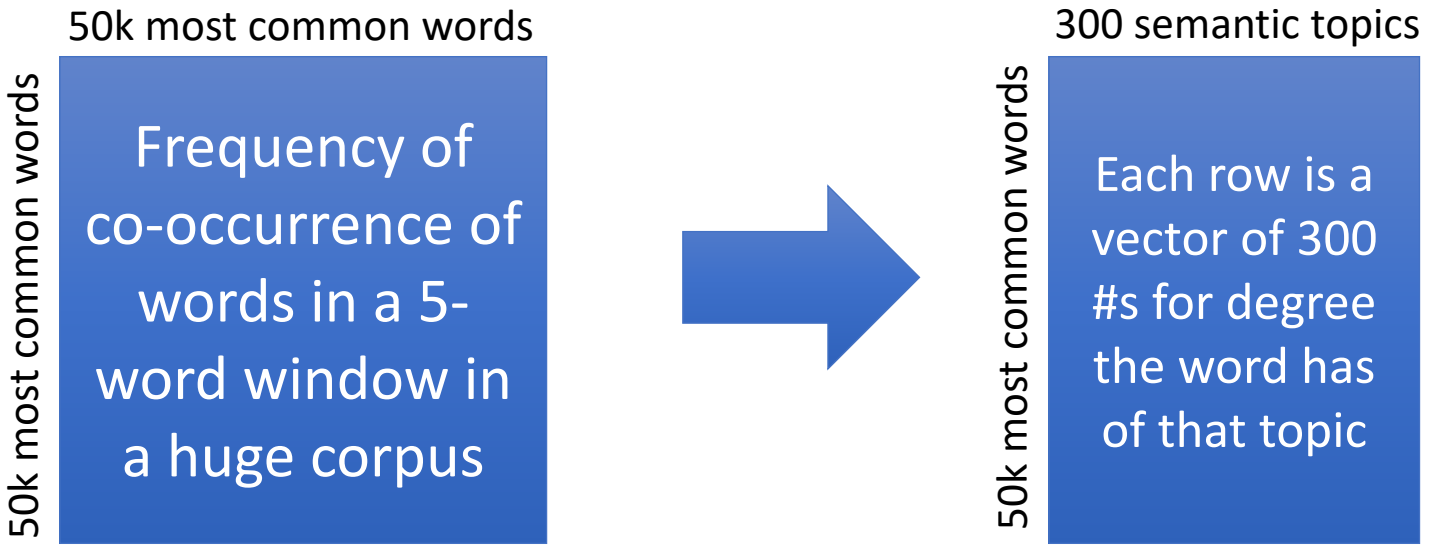
- How can we learn what a word means?
- The linguist John Rupert Firth famously wrote in 1957

  "You shall know a word by the company it keeps"

- A way to recognize that two words have similar meanings is to note that they occur in similar contexts
  - E.g., physician & doctor, nurse & doctor, love & hate

Word Embeddings

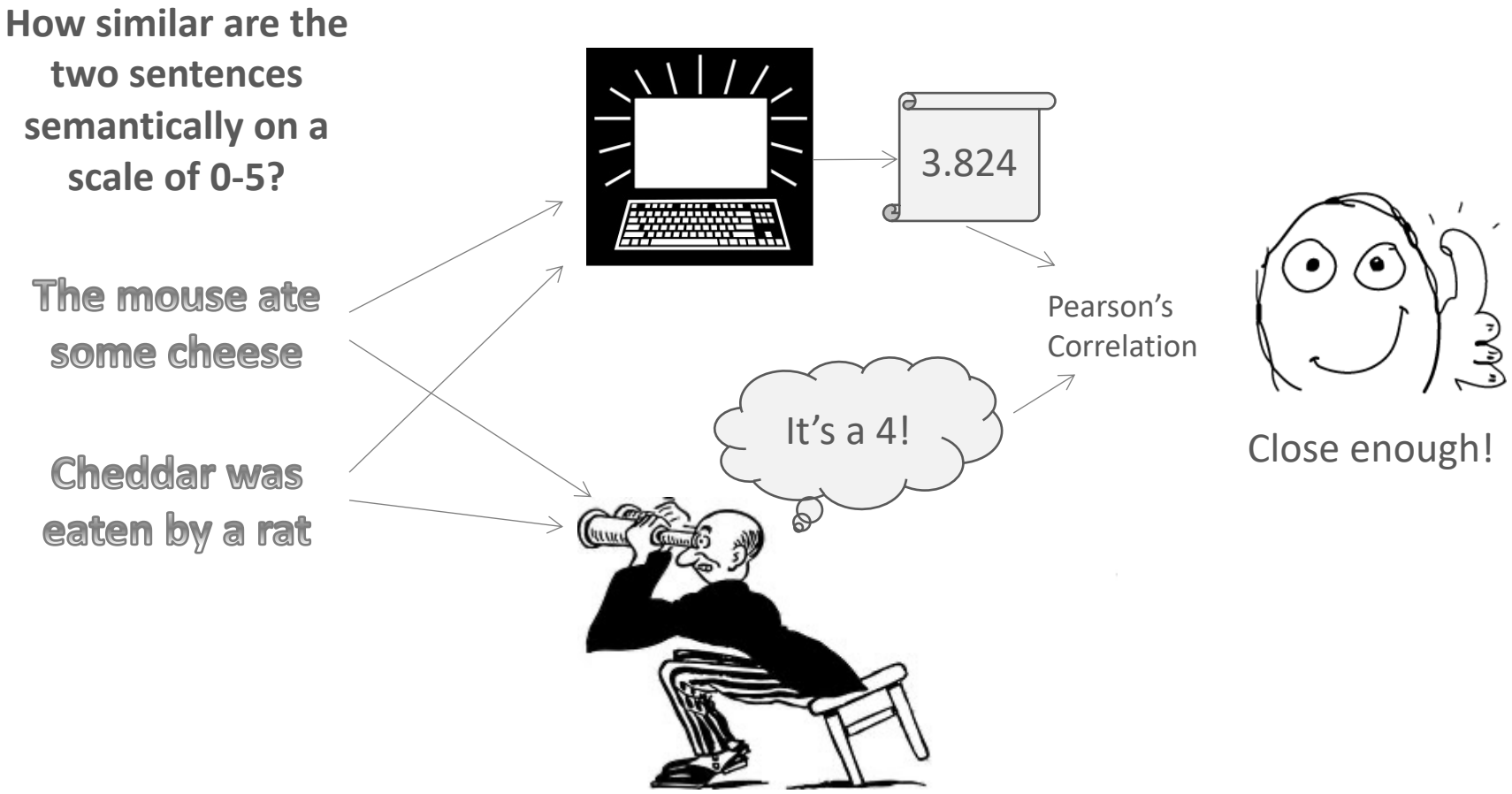
- Latent Semantic Analysis (LSA) learns a vector (e.g., 300 reals 0..1) for each unique word in a corpus to represent its meaning
  - LSA also used for document topic modelling
  - An example of dimentionality reduction that uses Principal component analysis, which does a linear mapping of the data to a lower-dimensional space



The **semantic similarity** of two words is the dot produce of their vectors, e.g.

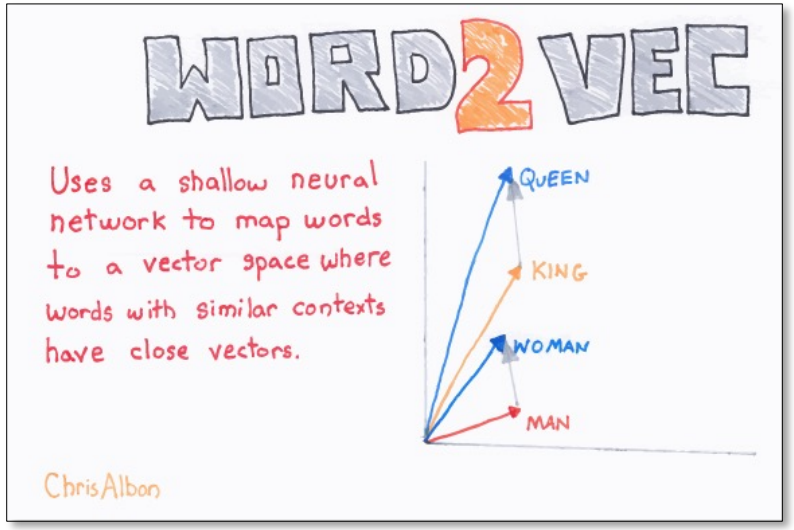
- dog ◦ cat = 0.8
- dog ◦ hound = 0.7
- dog ◦ ape 0.4

Sentence similarity

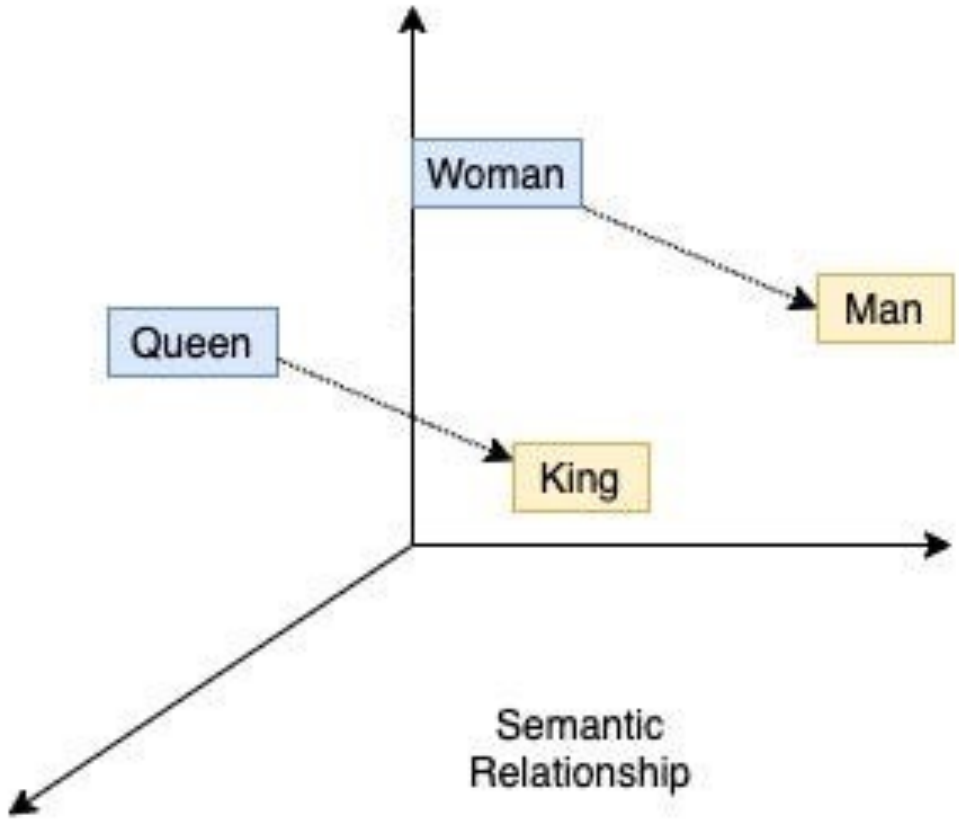


We used this approach in 2013 to win in a sentence similarity task

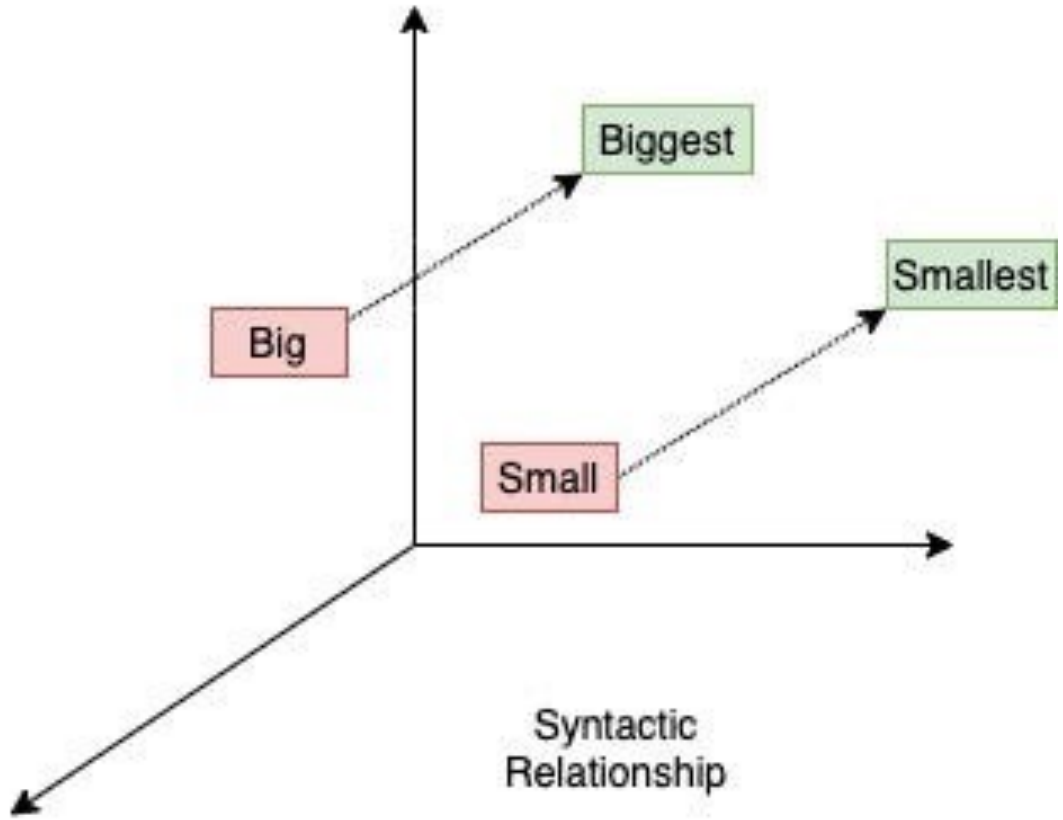
Word2Vec



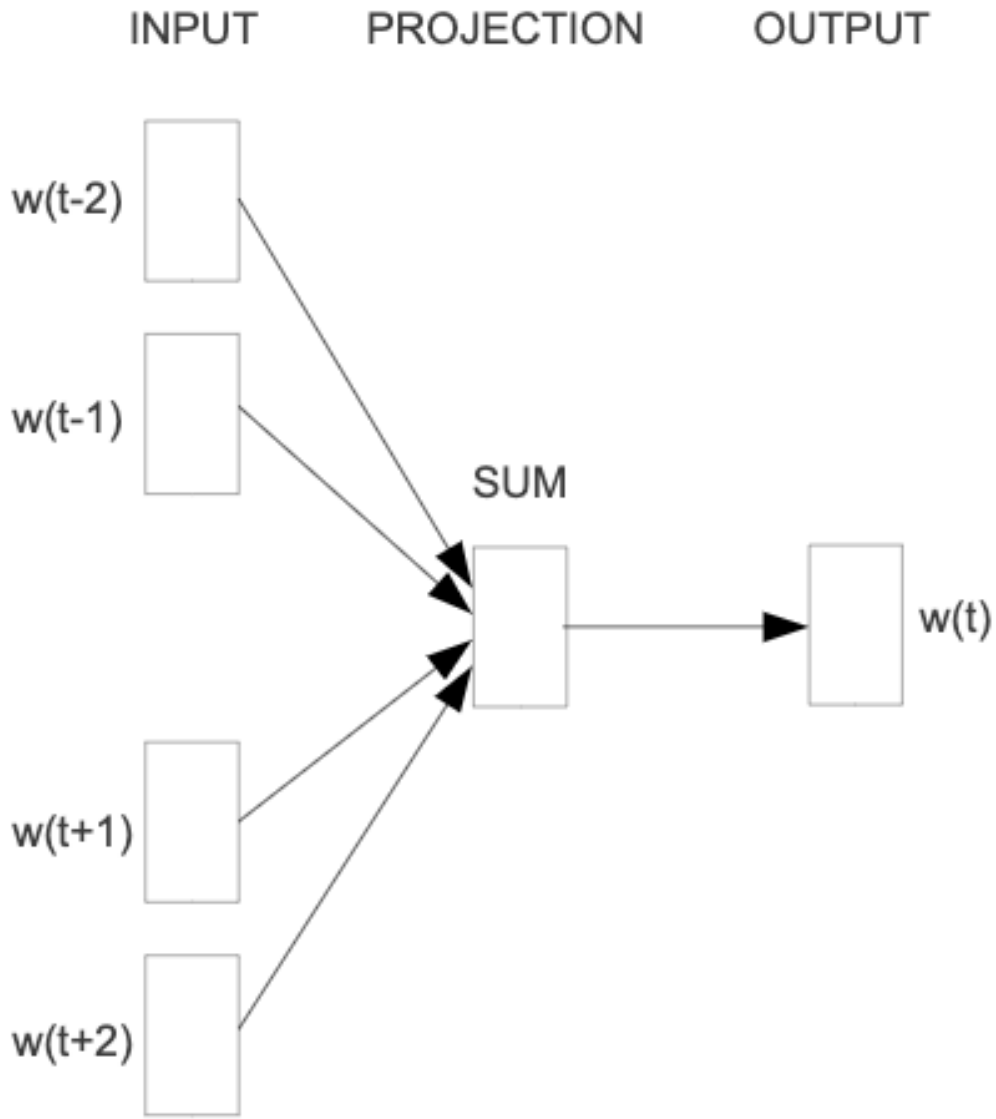
- Developed by Google also in 2013 using a neural network approach
- Two models: CBOW and skip grams
- Trained on a much larger corpus from the Web
- Models can be downloaded and are still used today
  - E.g., the spaCy NLP system uses word2vec to measure similarity for language understanding tasks



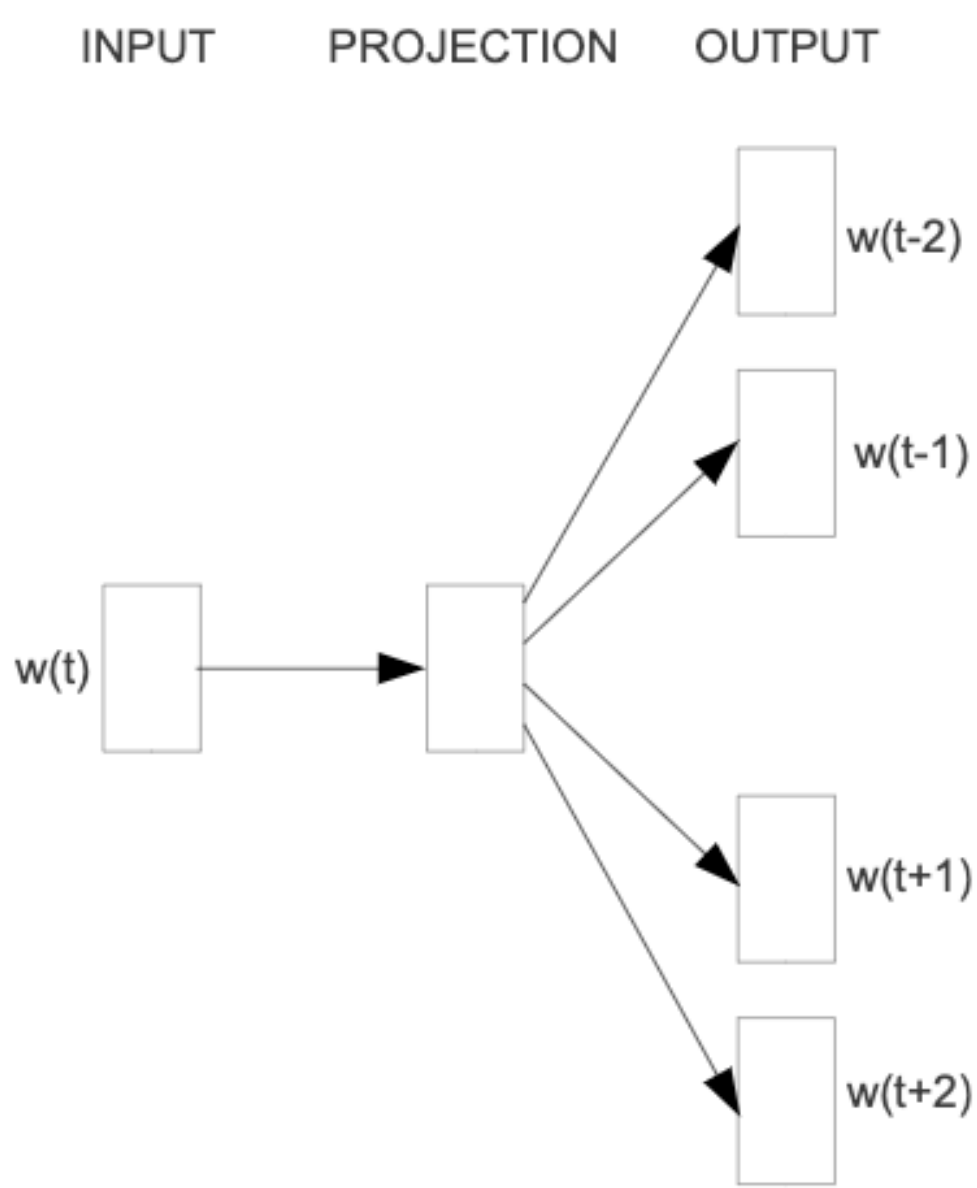
Semantic
Relationship



Syntactic
Relationship

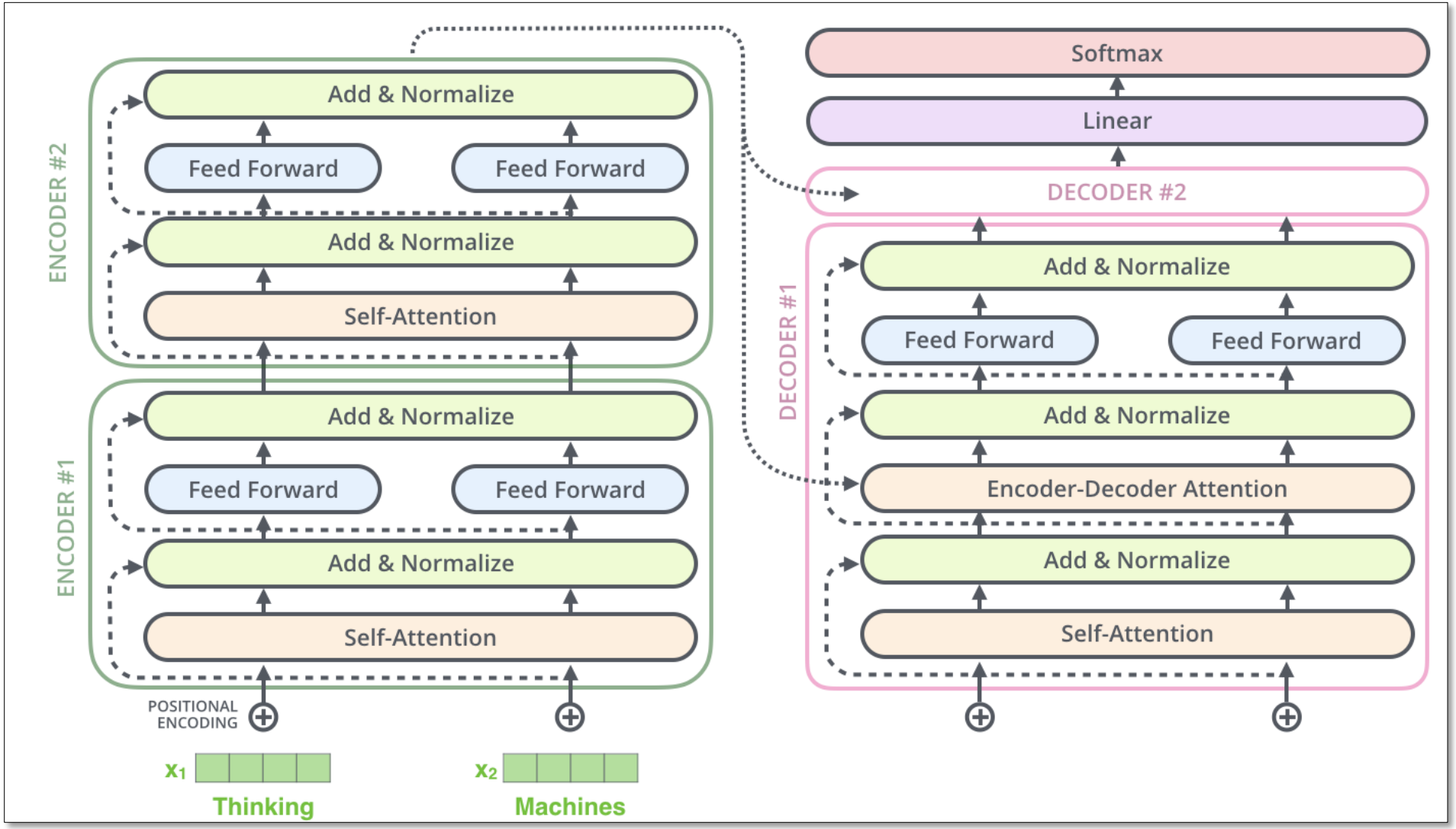


CBOW



Skip-gram

Transformer model



Encoder (e.g., BERT)

Decoder (e.g., GPT)