NLP Recent History

Text Features

- Natural Language Processing = Machine Learning with Text

- Step 1 – Identify text-based features
- Step 2 – Label data for a task
- Step 3 – Train a classifier

Text Features – Old School

- Lexical Features
  - Part-of-speech tags (Verb/Noun/Adj)
  - Stems / Lemmas
  - Word Frequency
  - Psycholinguistic Features
  - Terminologies
  - Morphology (affixes)

Text Features – Old School

- Sentential Features
  - Sentence Structure
  - Parse Tree Features
  - Sentence Dependencies
  - POS ratios
  - Punctuation frequency
  - Averages of lexical features

Text Features – Old School

- Document Features
  - Bag-of-words
  - TF-IDF Vectors

Text Features – Word Embeddings

- Represent each word as a vector of numbers

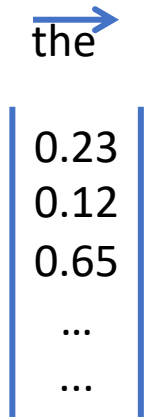
Text Features – Word Embeddings

- Represent each word as a vector of numbers

The cat sat on the mat

Text Features – Word Embeddings

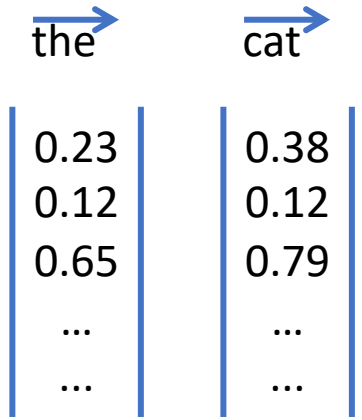
- Represent each word as a vector of numbers



The cat sat on the mat

Text Features – Word Embeddings

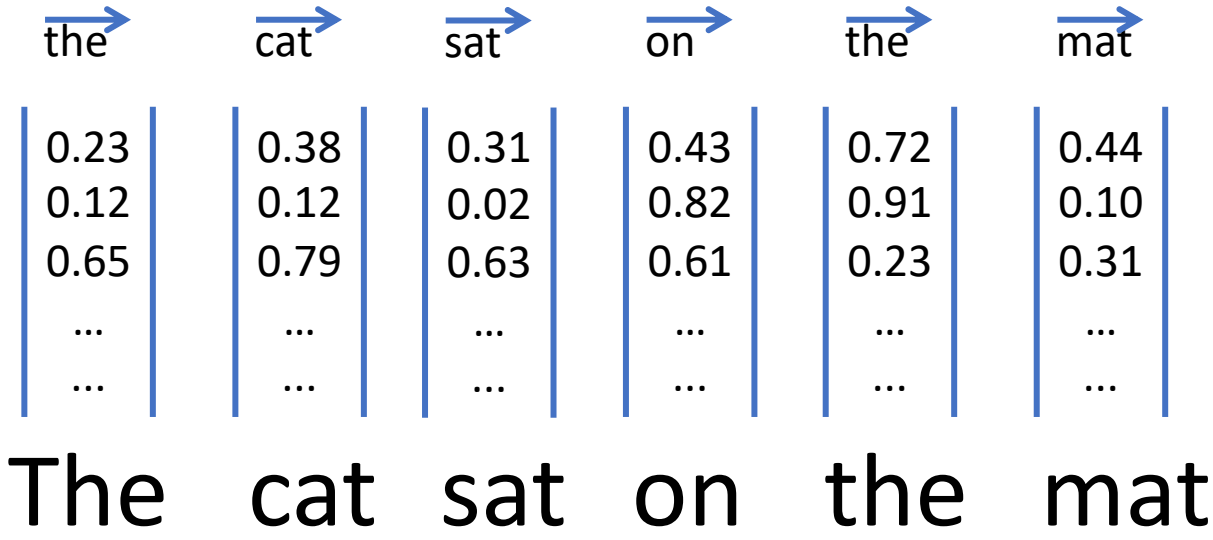
- Represent each word as a vector of numbers



The cat sat on the mat

Text Features – Word Embeddings

- Represent each word as a vector of numbers



Text Features – Count-based Embeddings

- The cat sat on the mat
- Dogs and cats are animals
- Cats like milk
- The dog chased the stick
- Dogs like bones

	cat	sat	mat	dog	chase	stick	animal	bone	milk
cat									
sat									
mat									
dog									
chase									
stick									
animal									
bone									
milk									

Text Features – Count-based Embeddings

- The cat sat on the mat
- Dogs and cats are animals
- Cats like milk
- The dog chased the stick
- Dogs like bones

	cat	sat	mat	dog	chase	stick	animal	bone	milk
cat	3	1	1	1	0	0	1	0	1
sat	1	1	1	0	0	0	0	0	0
mat	1	1	1	0	0	0	0	0	0
dog	1	0	0	3	1	1	1	1	0
chase	0	0	0	1	1	1	0	0	0
stick	0	0	0	1	1	1	0	0	0
animal	1	0	0	1	0	0	1	0	0
bone	0	0	0	1	0	0	0	1	0
milk	1	0	0	1	0	0	0	0	1

Text Features – Count-based Embeddings

- The cat sat on the mat
- Dogs and cats are animals
- Cats like milk
- The dog chased the stick
- Dogs like bones

cat	sat	mat	dog	chase	stick	animal	bone	milk
3	1	1	1	0	0	1	0	1
1	1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
1	0	0	3	1	1	1	1	0
0	0	0	1	1	1	0	0	0
0	0	0	1	1	1	0	0	0
1	0	0	1	0	0	1	0	0
0	0	0	1	0	0	0	1	0
1	0	0	1	0	0	0	0	1

Text Features – Count-based Embeddings

- The cat sat on the mat
- Dogs and cats are animals
- Cats like milk
- The dog chased the stick
- Dogs like bones

cat	sat	mat	dog	chase	stick	animal	bone	milk
3	1	1	1	0	0	1	0	1
1	1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
1	0	0	3	1	1	1	1	0
0	0	0	1	1	1	0	0	0
0	0	0	1	1	1	0	0	0
1	0	0	1	0	0	1	0	0
0	0	0	1	0	0	0	1	0
1	0	0	1	0	0	0	0	1

- Issue
  - Dimensionality = vocab size
- Reduce with PCA..

Text Features – Count-based Embeddings

- Which animal likes milk?

animal
1
0
0
1
0
0
1
0
0

+

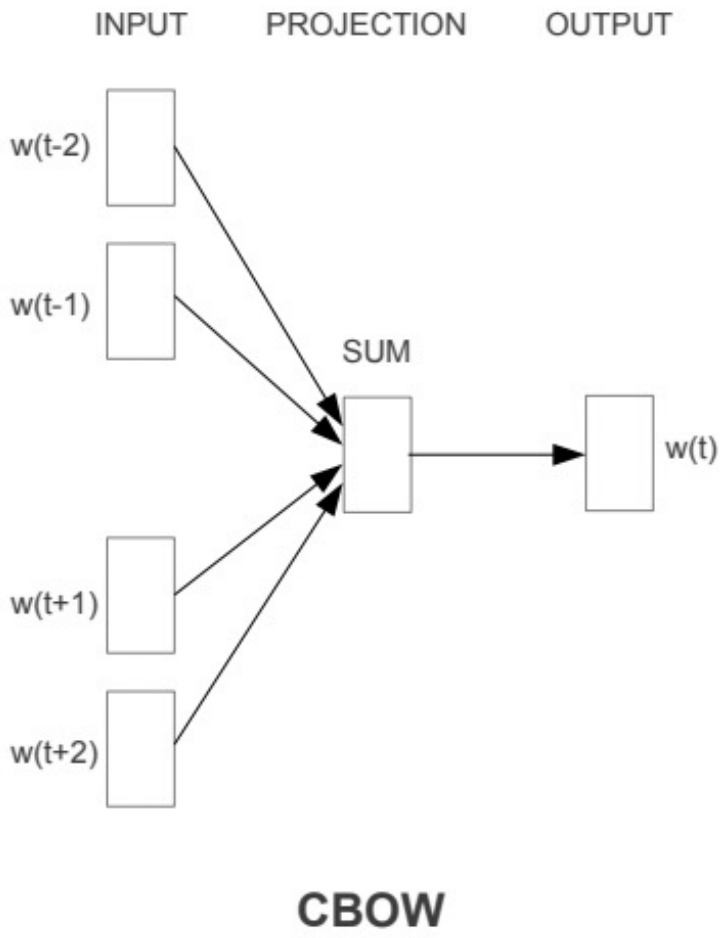
milk
1
0
0
0
0
0
0
0
1

~

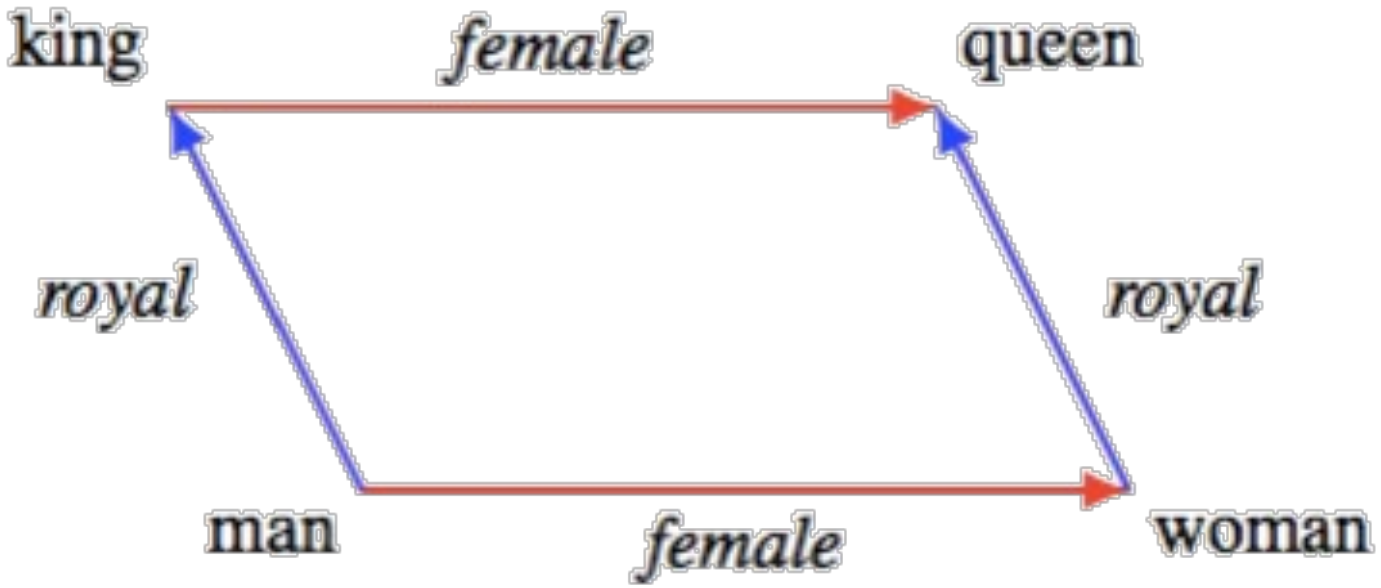
cat
3
1
1
1
0
0
1
0
1

Text Features – Learnt Embeddings (Word2Vec)

- Learn a set of word vectors that can be used to predict surrounding words
- Initialise random vectors and train neural network, updating vector weights with network
- End result is a set of vectors that perform like count-based, but without counting

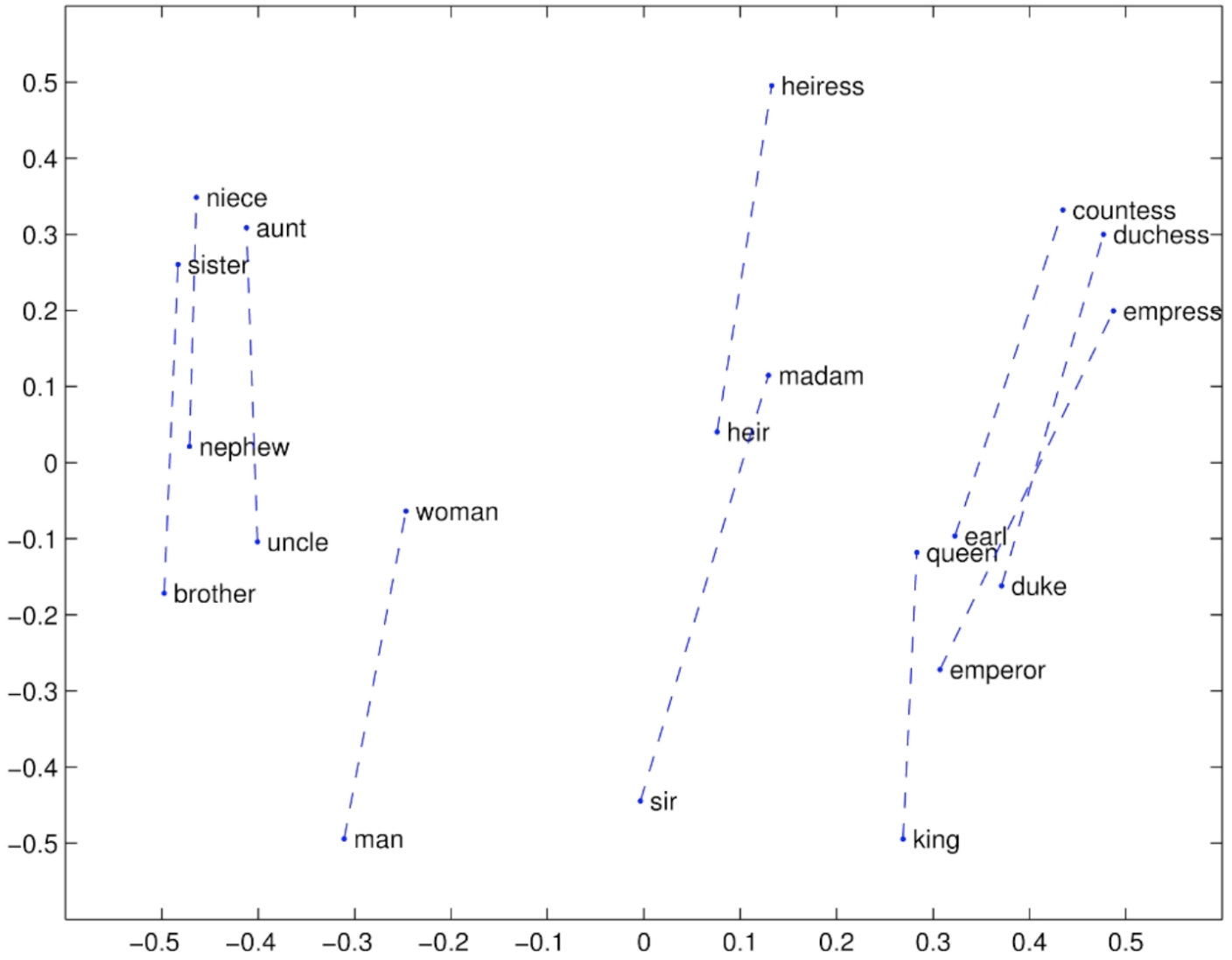


Text Features – Learnt Embeddings (Word2Vec)

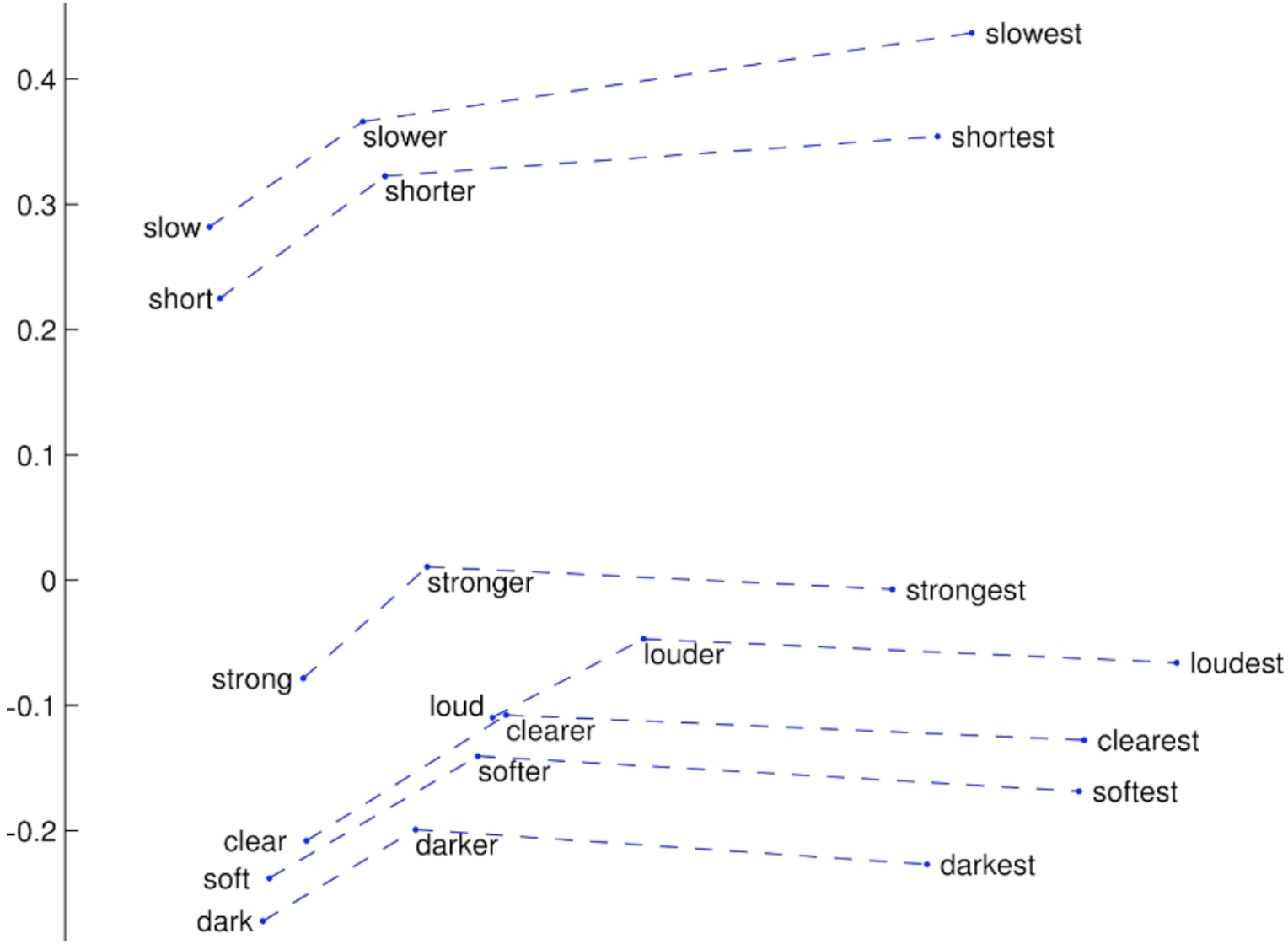


$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

Text Features – Learnt Embeddings (Word2Vec)



Text Features – Learnt Embeddings (Word2Vec)



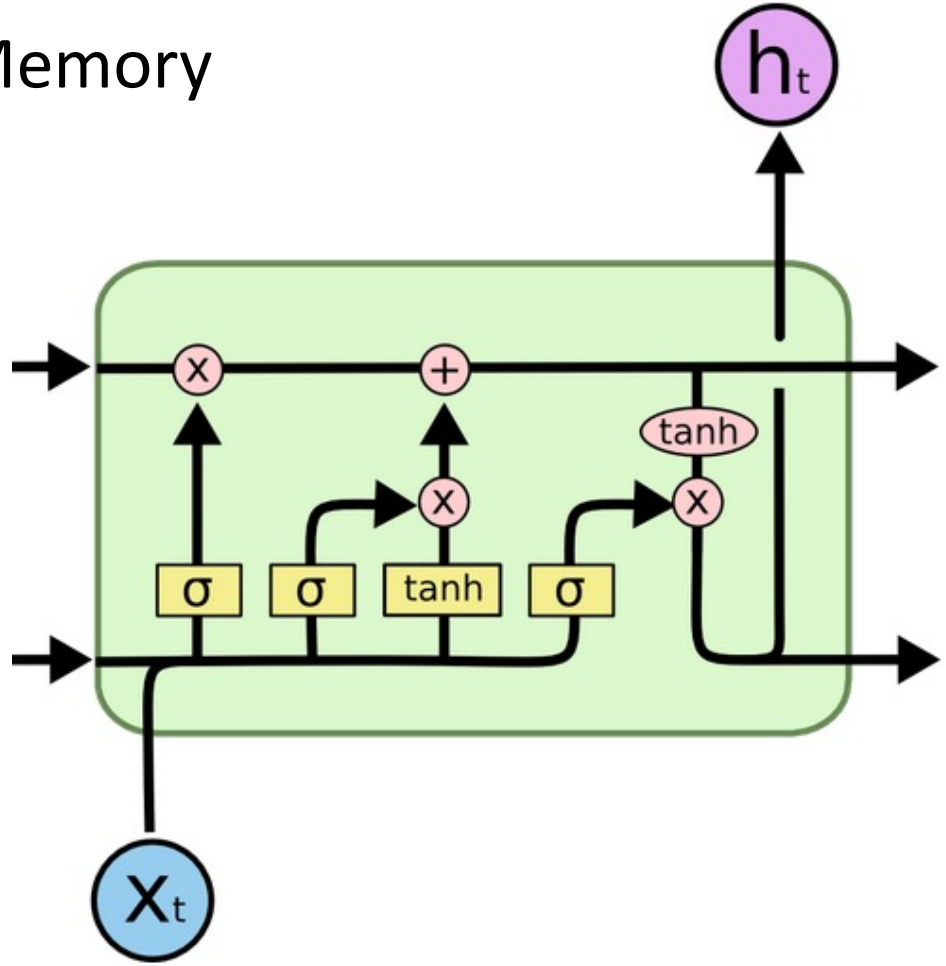
Example Tasks

- Classification
  - Sentiment Analysis
  - Readability
  - Categorisation
- Text Generation
  - Machine Translation
  - Summarisation
  - Data-to-text
- Natural Language Understanding
  - Inference / Reasoning
  - Chatbots / Intelligent Tutoring

http://nlpprogress.com/

Sequence classification

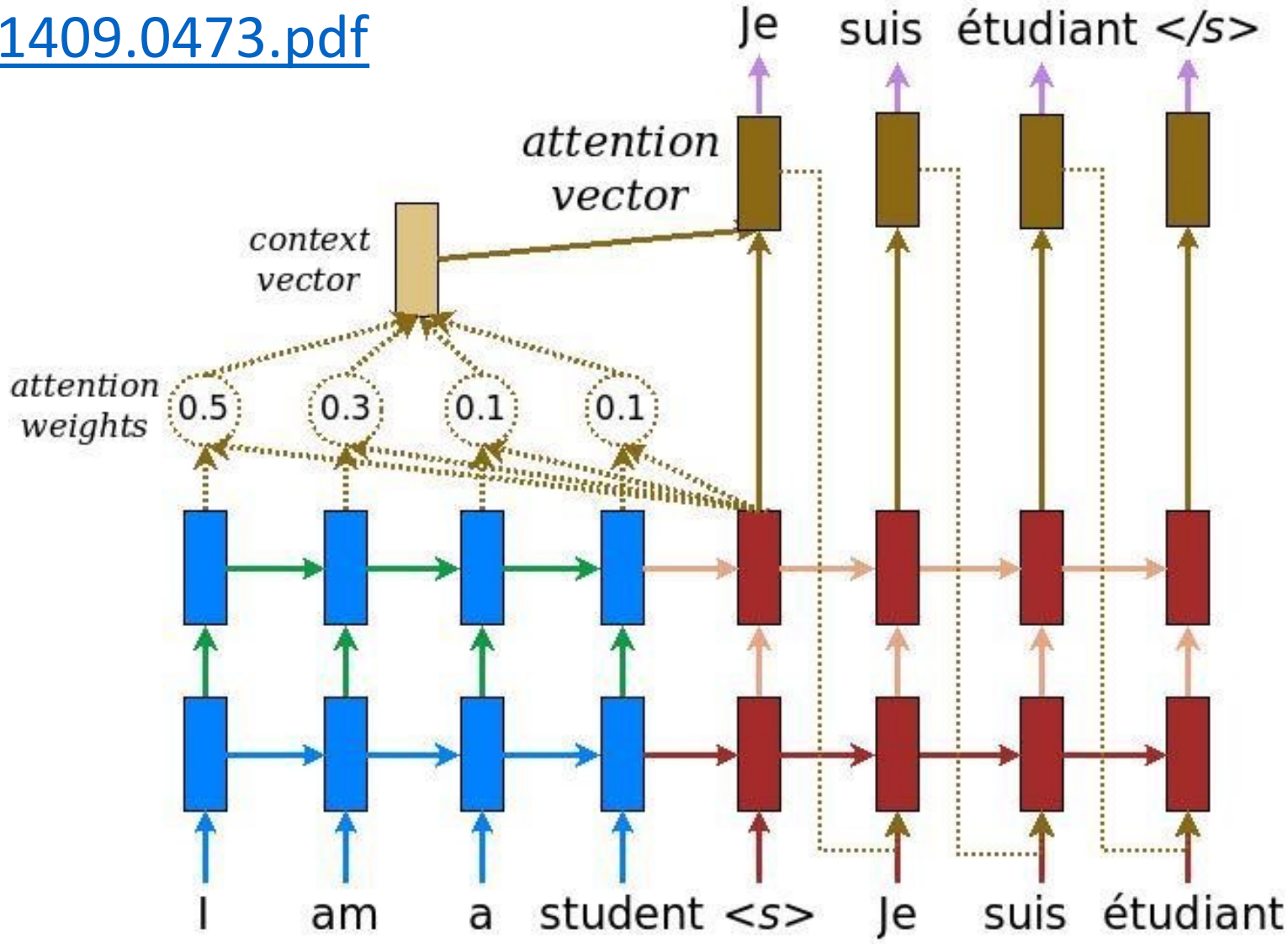
- Long Short Term Memory



Sequence classification

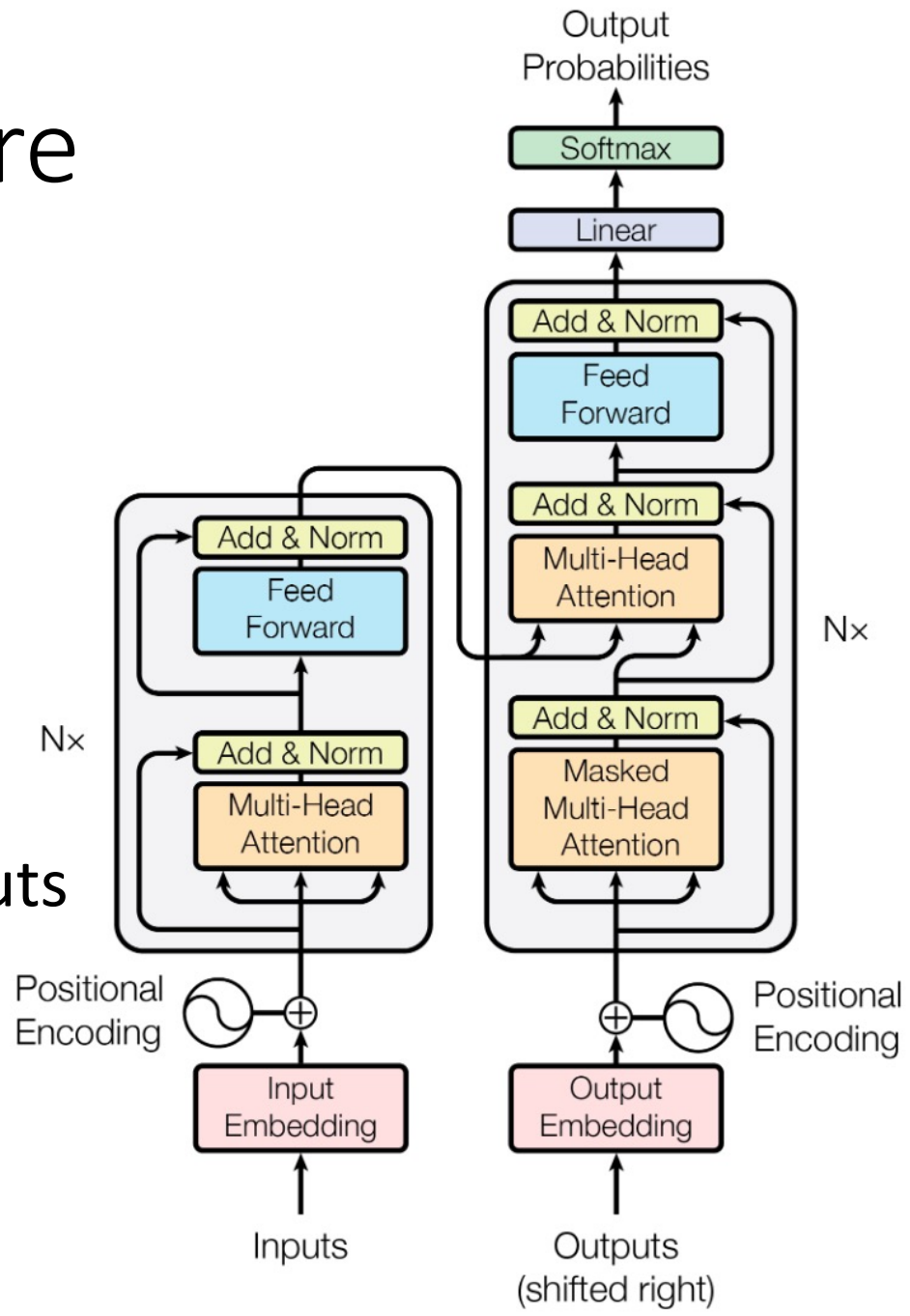
https://arxiv.org/pdf/1409.0473.pdf

- LSTMs + Attention



The Transformer - Architecture

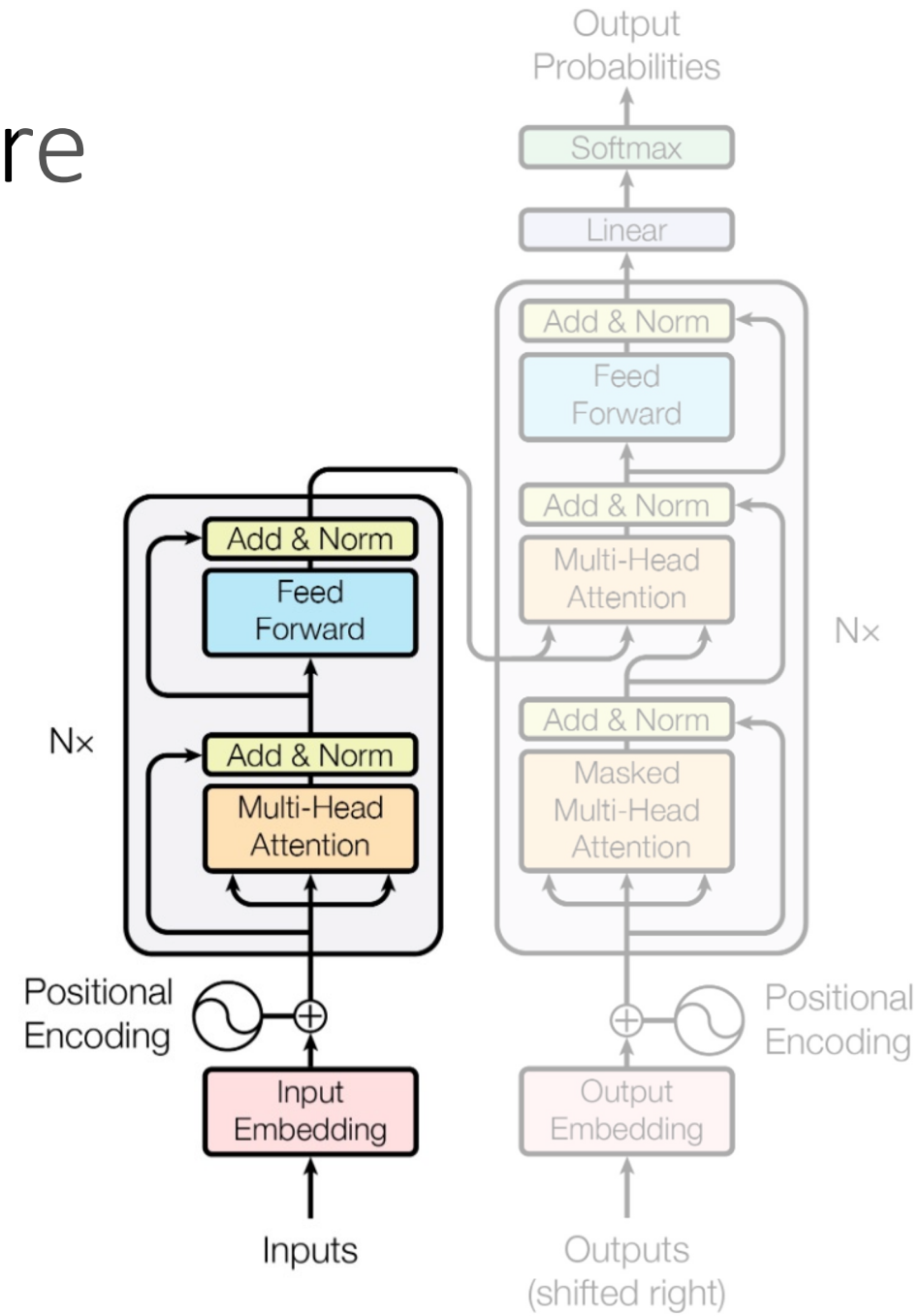
- Encoder-Decoder architecture
  - Feature space acts as an interlingua
- Encoder
  - Inputs (words) are encoded using embeddings
  - Transforms input embeddings to feature space
  - 6 stacked layers attention and feed-forward
- Decoder – transform feature space to outputs
  - Converts output of last encoder layer into a set of probabilities which predict the next word



The Transformer - Architecture

- Encoder Layer

- Creates context vector for input sequence
- N=6 stacked layers

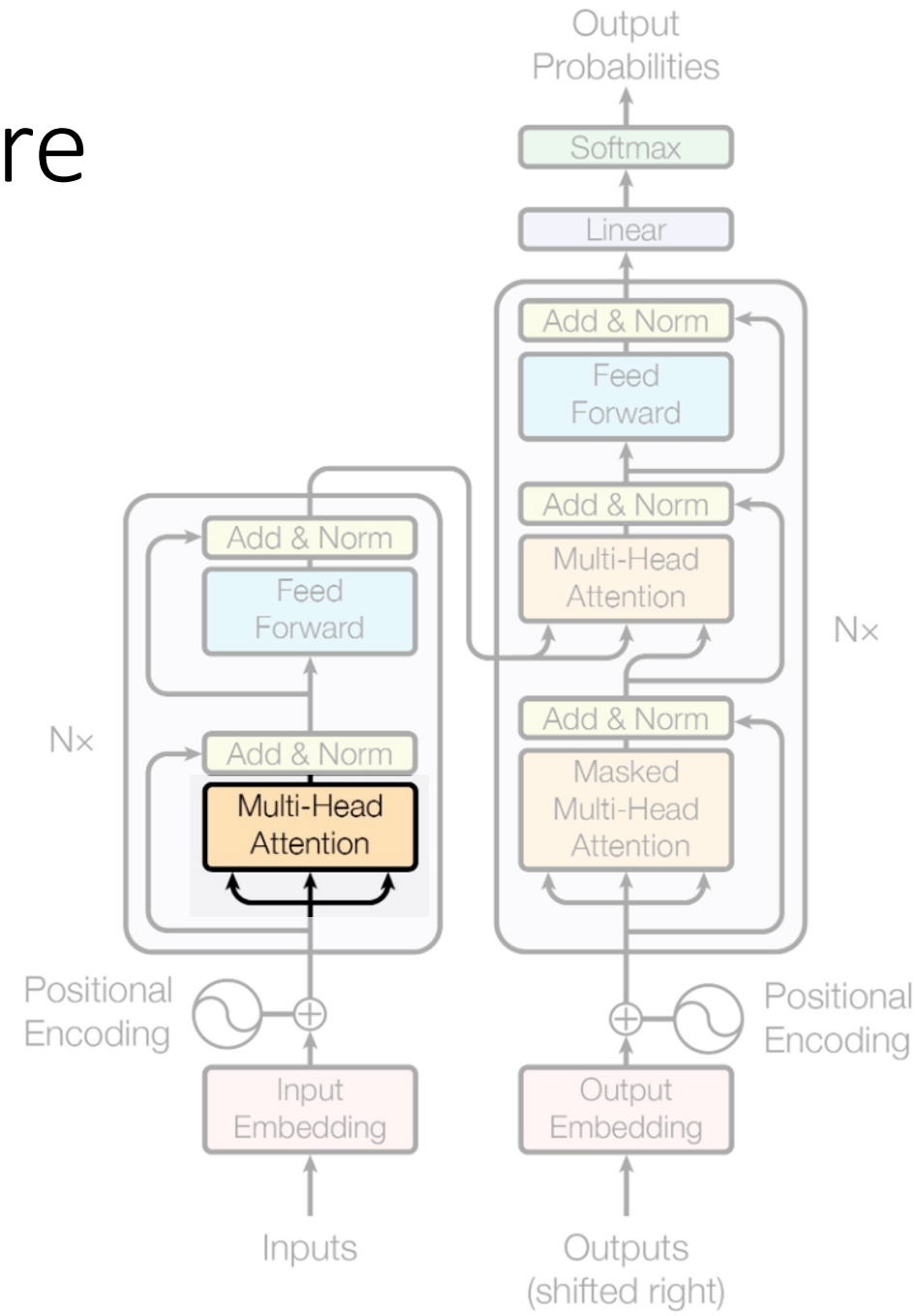


The Transformer - Architecture

- Encoder – multi-headed self attention

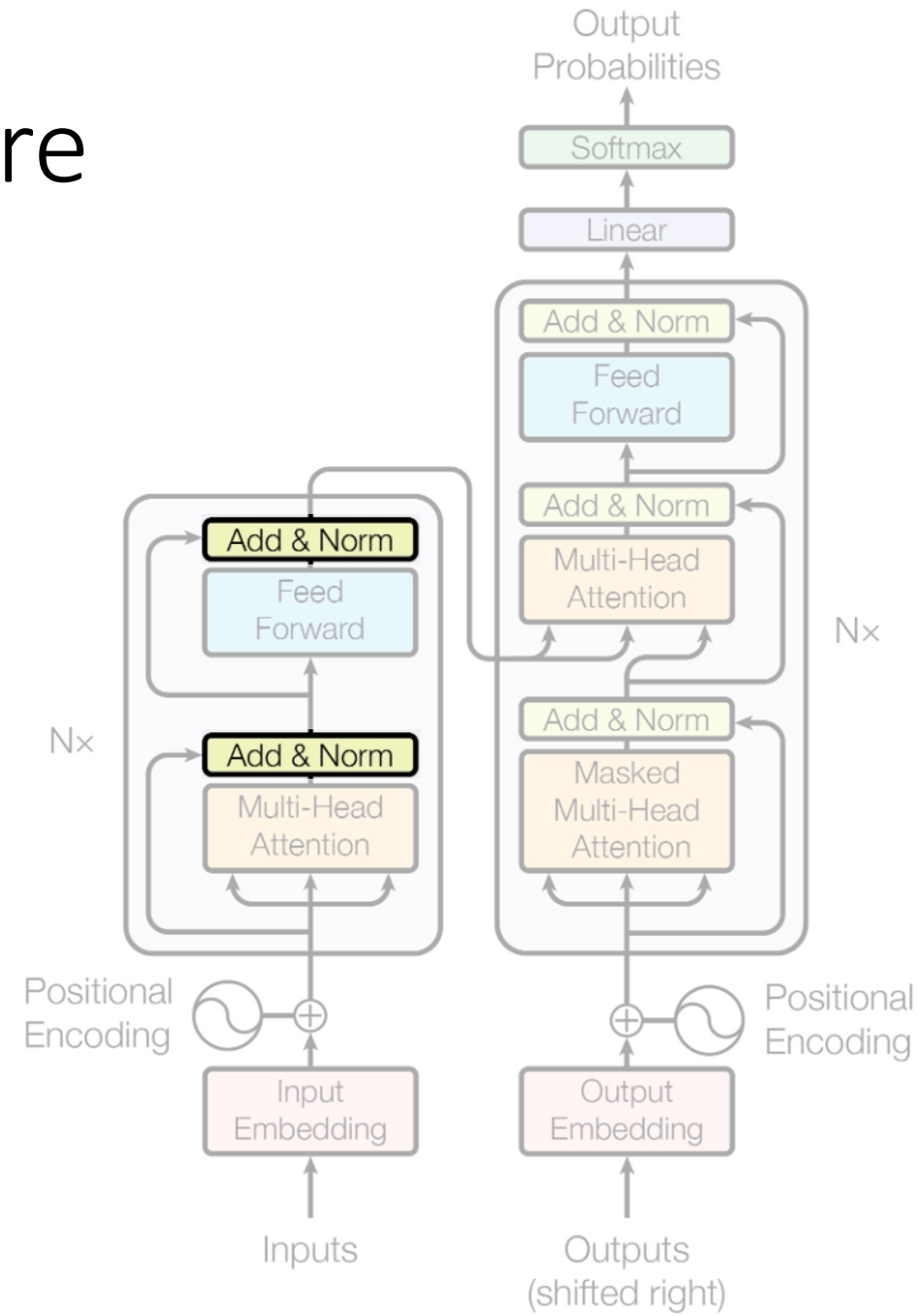
- Self-attention
  - All inputs available to encoder
  - Attention selects relevant parts of input
  - Key, Value, Query (read paper for details)

- Multi-headed
  - Ran in parallel N=8 times
  - Output of each ‘head’ is concatenated



The Transformer - Architecture

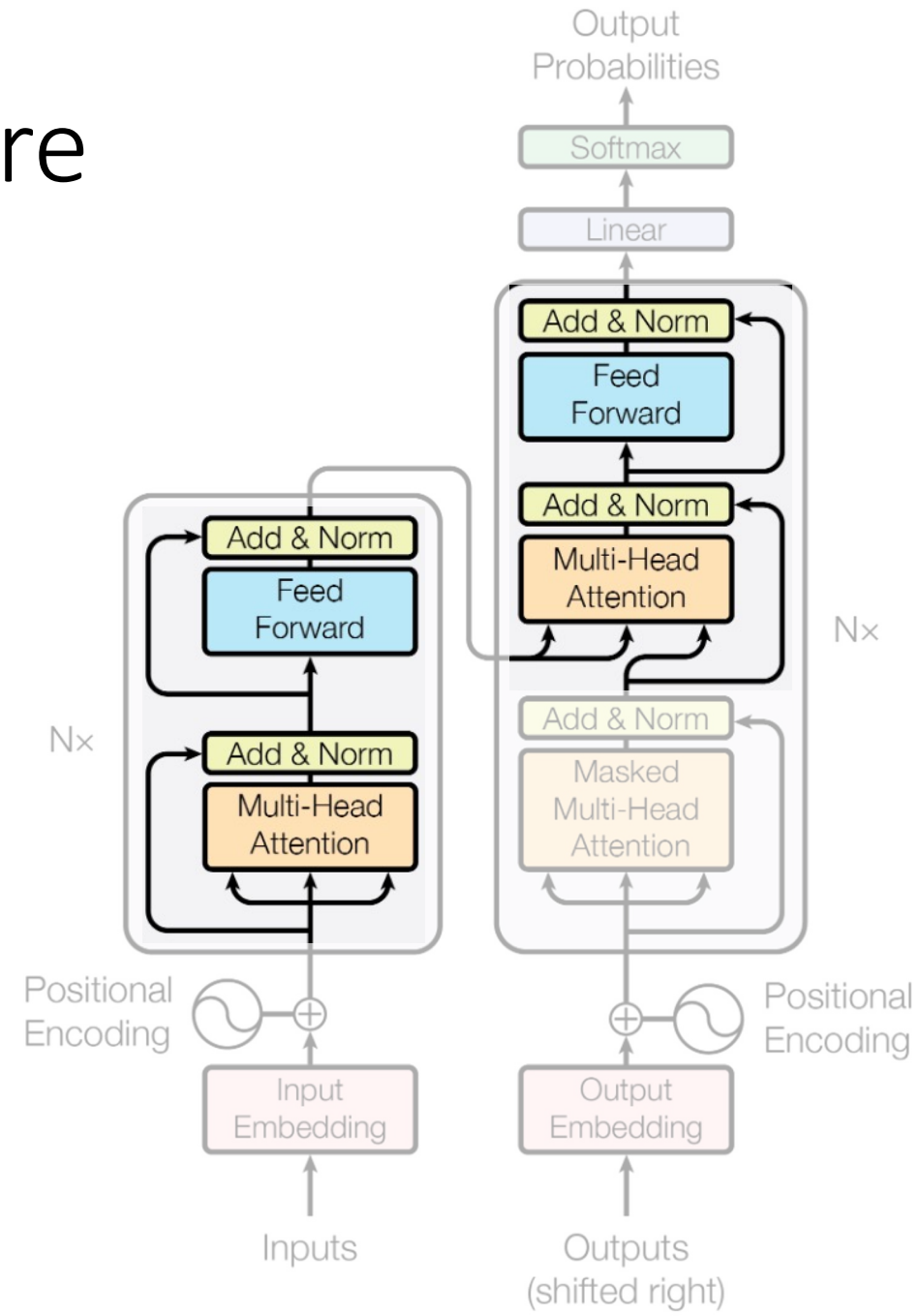
- Encoder - Add and normalize
- Note that the input from before each step is carried through to the next step
- Normalisation prevents exploding gradient



The Transformer - Architecture

- Decoder – Replicates encoder

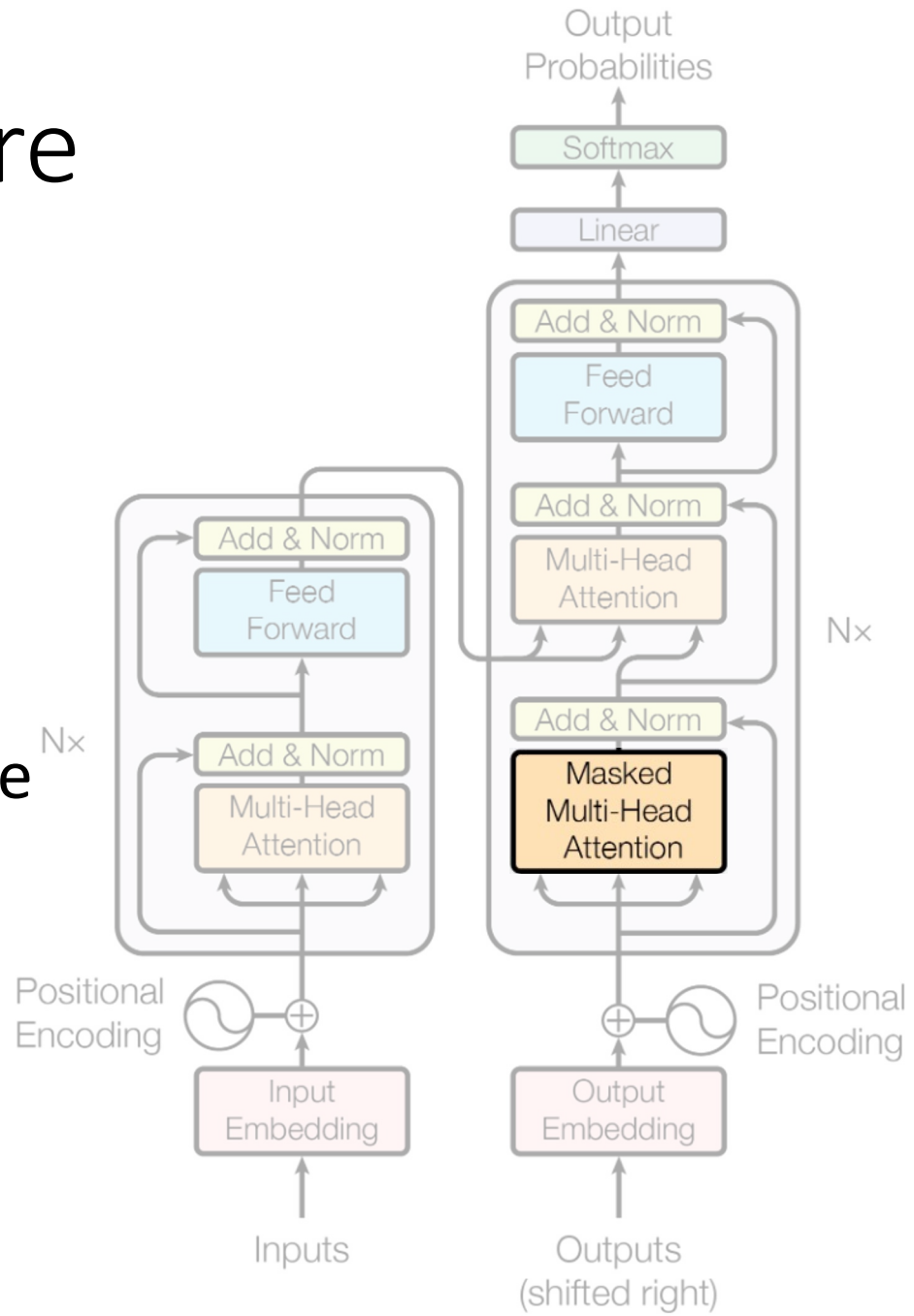
- Multi-head self attention and feed forward network are carried over
  - Applied to the attention weights from encoder
  - Also applied to the last output from decoder



The Transformer - Architecture

- Decoder – Masked Multi-head attention

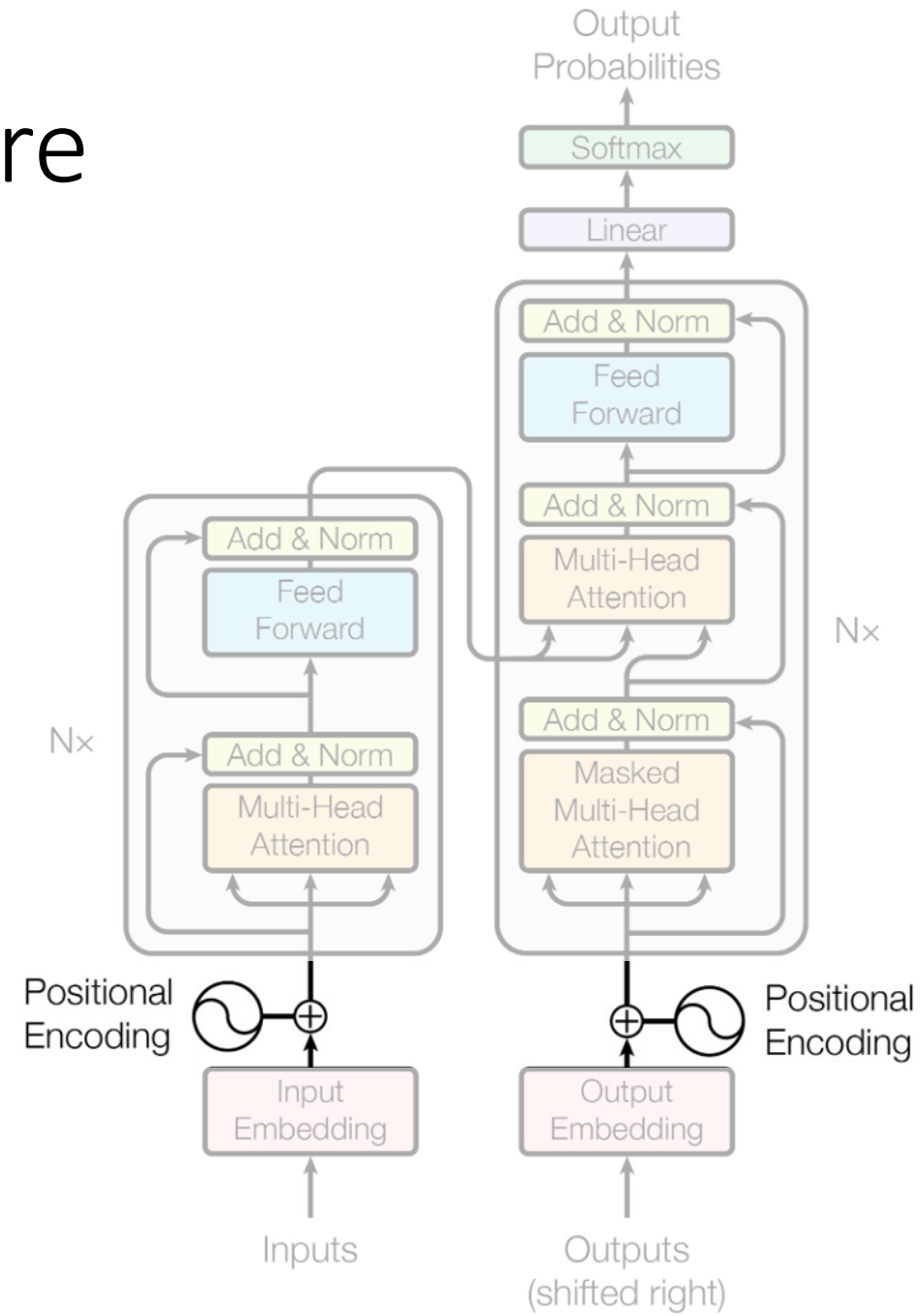
- Masking prevents information from future becoming available at decode time
- Otherwise, self-attention the same as before



The Transformer - Architecture

- Encoder-Decoder – Sinusoidal Positional Encoding

- Encodes whereabout in the sentence the word is likely to have occurred
  - Applies sin/cos function to alternate dimensions of the embedding
  - This is summed with the word embeddings
  - So embedding contains info on meaning + position in sentence

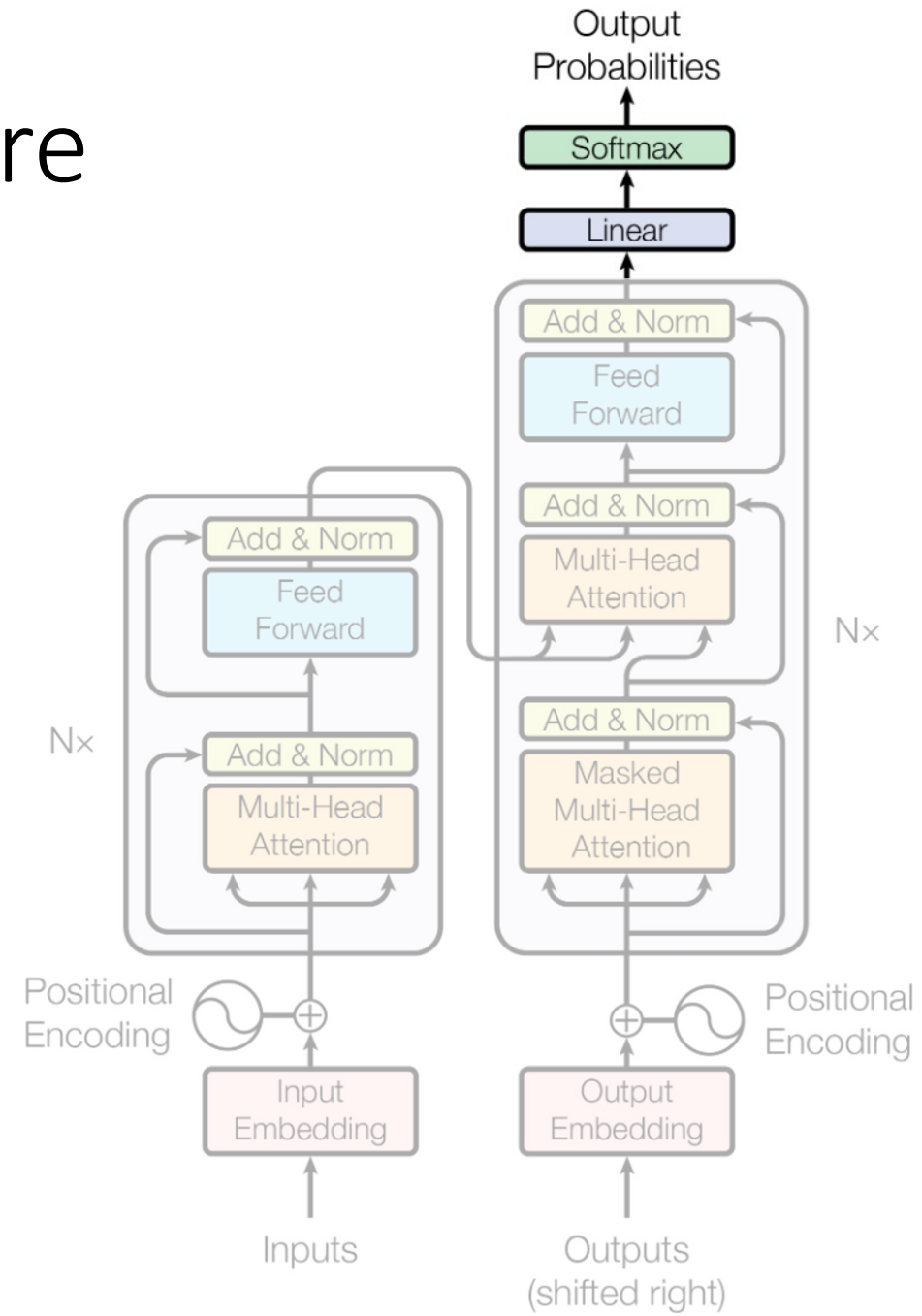


The Transformer - Architecture

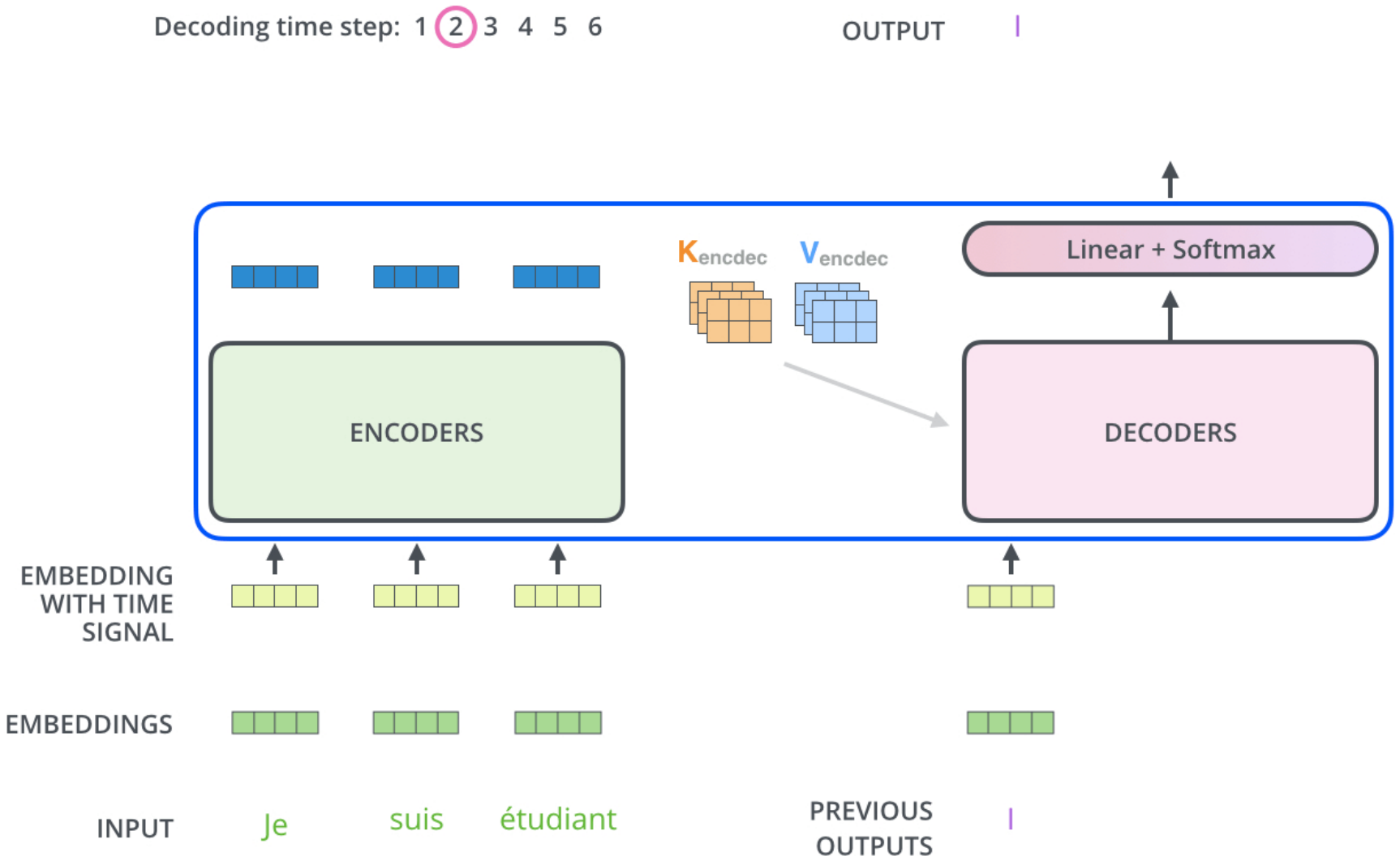
- Output

- Linear layer
  - Fully connected NN
  - Converts k-dim vector to size of output vocab

- Softmax layer
  - Normalises probabilities in 0-1 range.



The Transformer - Visualisation

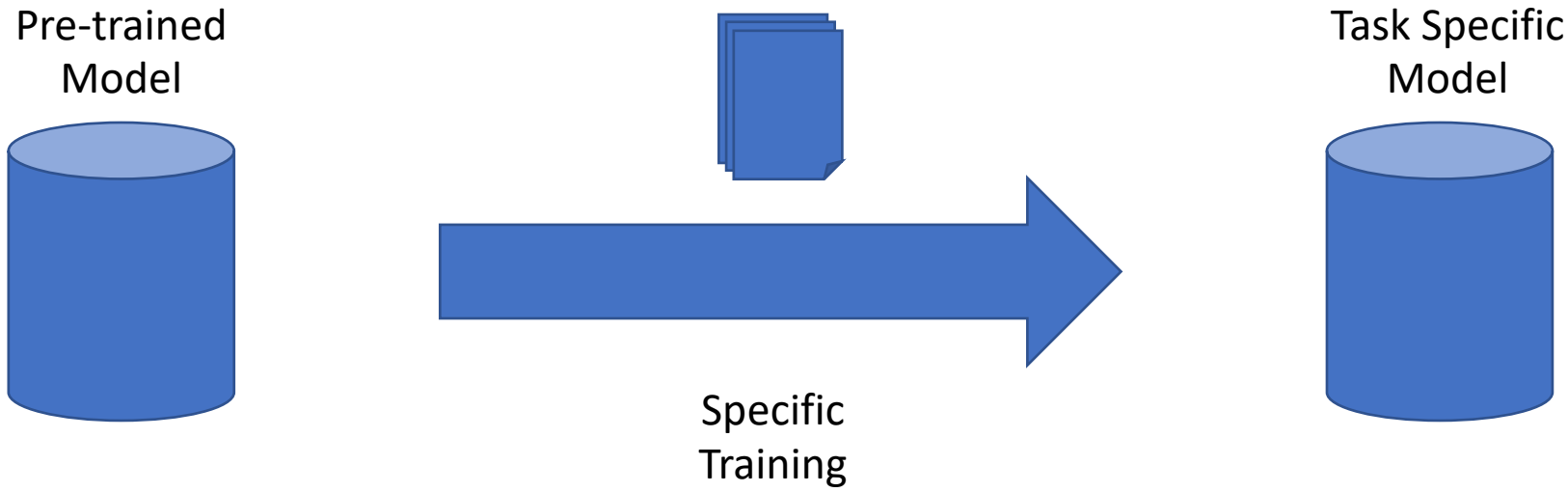


Pre-trained Models

- NLP would be very expensive if we had to retrain models every time
- Fortunately, pre-trained models are available that we can use

Fine-tuning

- Pre-trained models can be adapted to a specific task



Tasks | Libraries | Datasets | Languages | Licenses | Other

Filter Tasks by name

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Models 124,065

Filter by name

Sort: Most Downloads

bert-base-uncased
Updated Nov 16, 2022 • ↓ 25.8M • ♡ 464

Fri, 16 Dec 2022 15:44:21 GMT

gpt2
Updated Dec 16, 2022 • ↓ 14M • ♡ 469

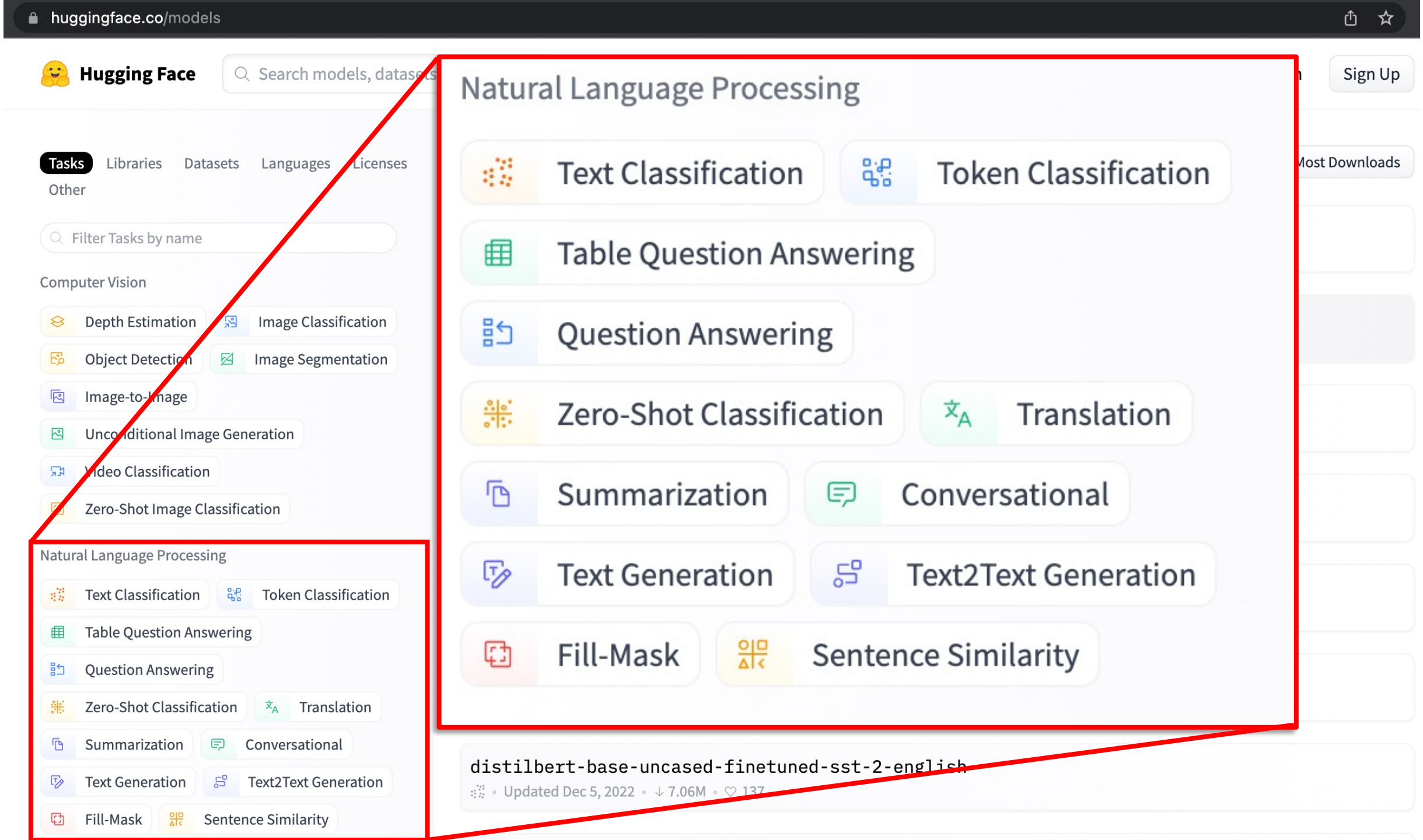
openai/clip-vit-large-patch14
Updated Oct 4, 2022 • ↓ 10.2M • ♡ 160

xlm-roberta-large
Updated Jun 27, 2022 • ↓ 10M • ♡ 64

distilbert-base-uncased
Updated Nov 16, 2022 • ↓ 9.88M • ♡ 127

prajjwal1/bert-tiny
Updated Oct 27, 2021 • ↓ 9.65M • ♡ 25

distilbert-base-uncased-finetuned-sst-2-english
Updated Dec 5, 2022 • ↓ 7.06M • ♡ 137


huggingface.co/models
Hugging Face
Search models, datasets
Sign Up
Tasks
Libraries
Datasets
Languages
Licenses
Other
Most Downloads
Filter Tasks by name
Computer Vision
Depth Estimation
Image Classification
Object Detection
Image Segmentation
Image-to-Image
Unconditional Image Generation
Video Classification
Zero-Shot Image Classification
Natural Language Processing
Text Classification
Token Classification
Table Question Answering
Question Answering
Zero-Shot Classification
Translation
Summarization
Conversational
Text Generation
Text2Text Generation
Fill-Mask
Sentence Similarity
Natural Language Processing
Text Classification
Token Classification
Table Question Answering
Question Answering
Zero-Shot Classification
Translation
Summarization
Conversational
Text Generation
Text2Text Generation
Fill-Mask
Sentence Similarity
distilbert-base-uncased-finetuned-sst-2-english
Updated Dec 5, 2022
↓ 7.06M
♡ 137