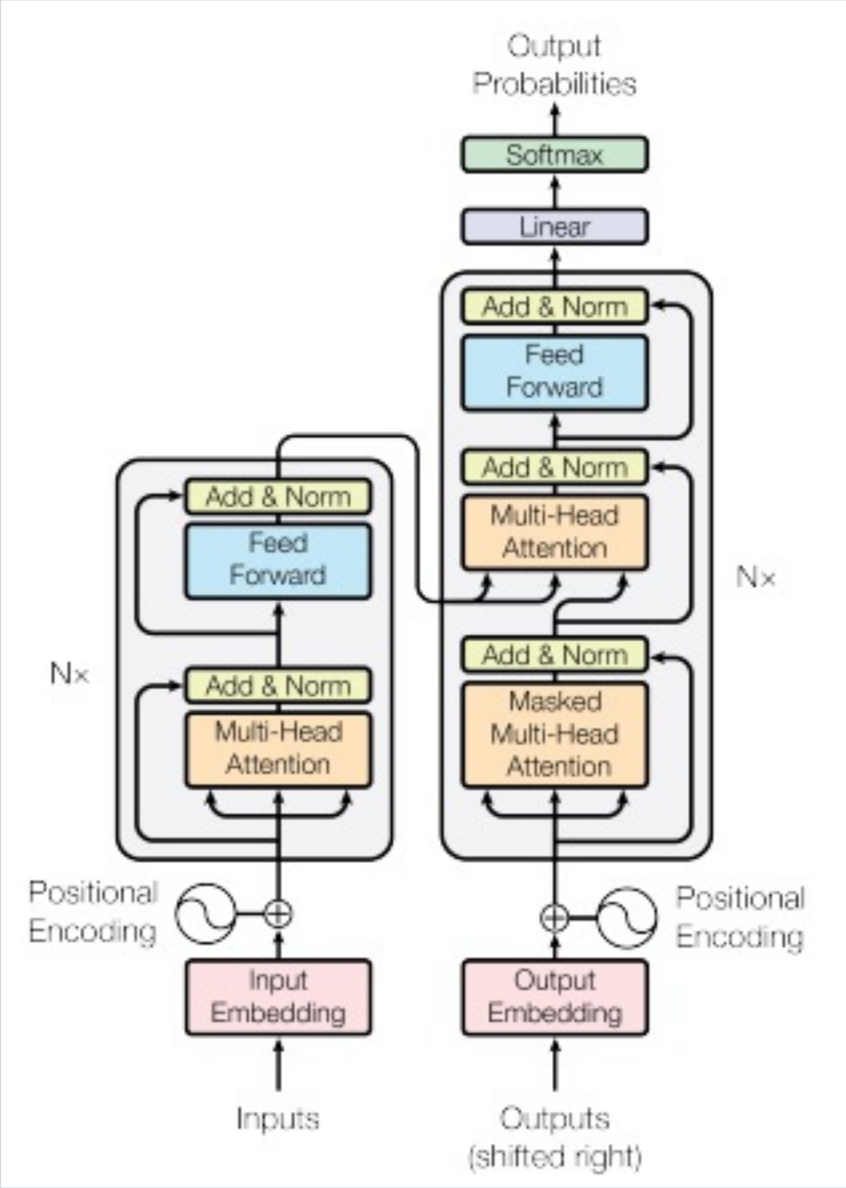
Generative models

NLP in one day

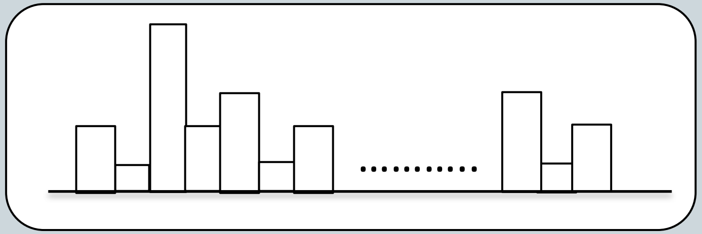
King's College London

NIHR | Maudsley Biomedical Research Centre

A more complex architecture - the Transformer



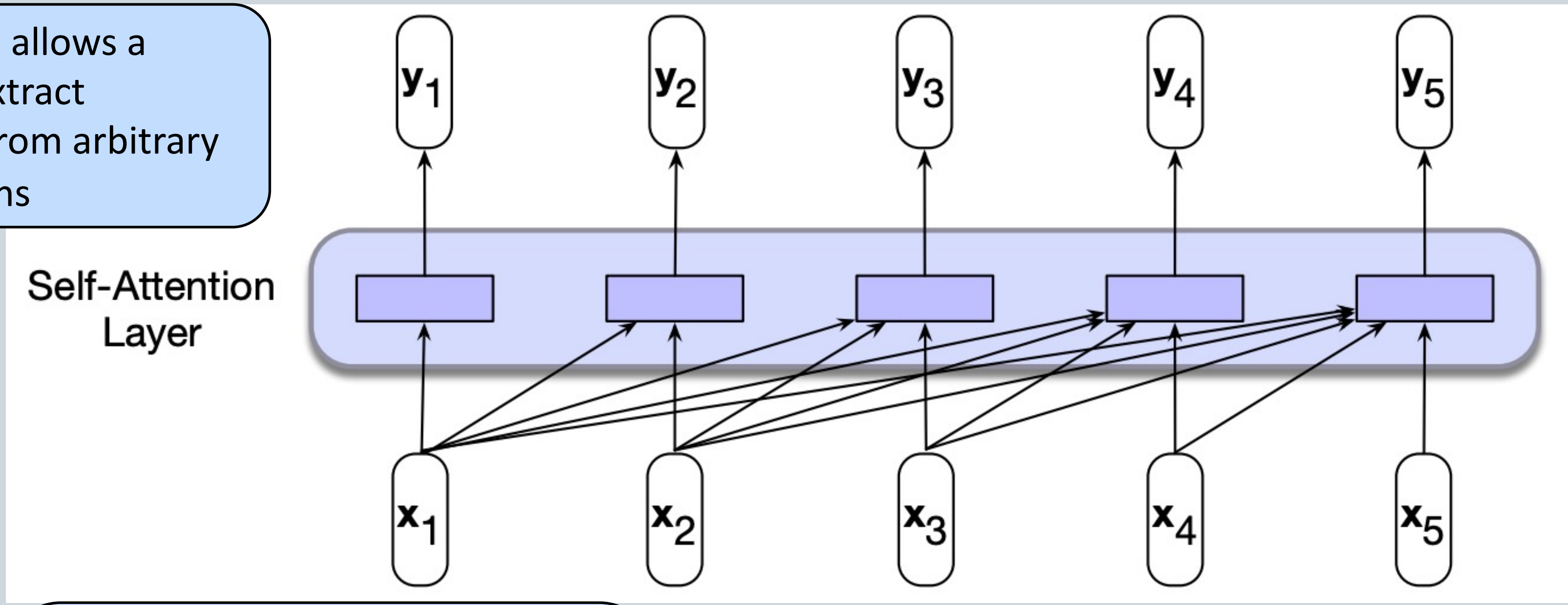
- See "Attention Is All You Need", Vaswani et al, NIPS 2017 (127000 citations)
- Produces probability distributions across the entire vocabulary
- Has proved to be flexible and scalable
- Power comes from number of parameters and size of training corpus
- The encoder and decoder weights are models of language
- Can be reused in other tasks



Output at each step is a probability distribution over our entire vocabulary

Self-attention – encoding word context and word order

Self-attention allows a network to extract information from arbitrary context lengths



Attention compares (scores) each input to the preceding ones, to reveal their relevance

$$\text{score}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$$

Scores are softmaxed and used to weight inputs to predict y

The model learns how to weight different parts of the context

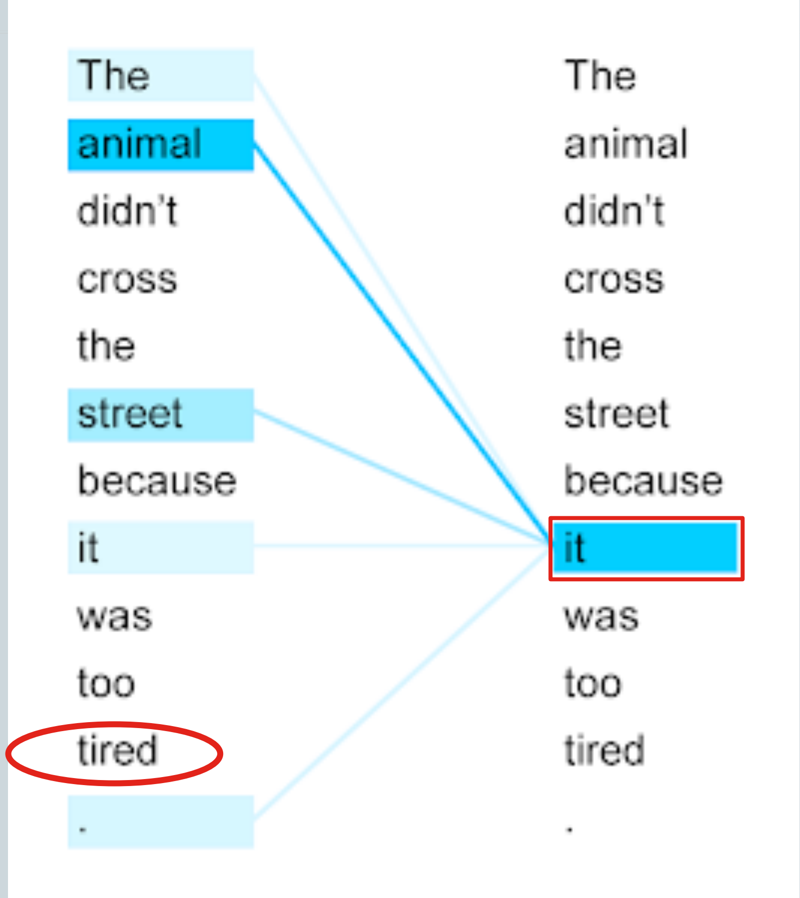
At each input, model has access to all inputs prior to this one.

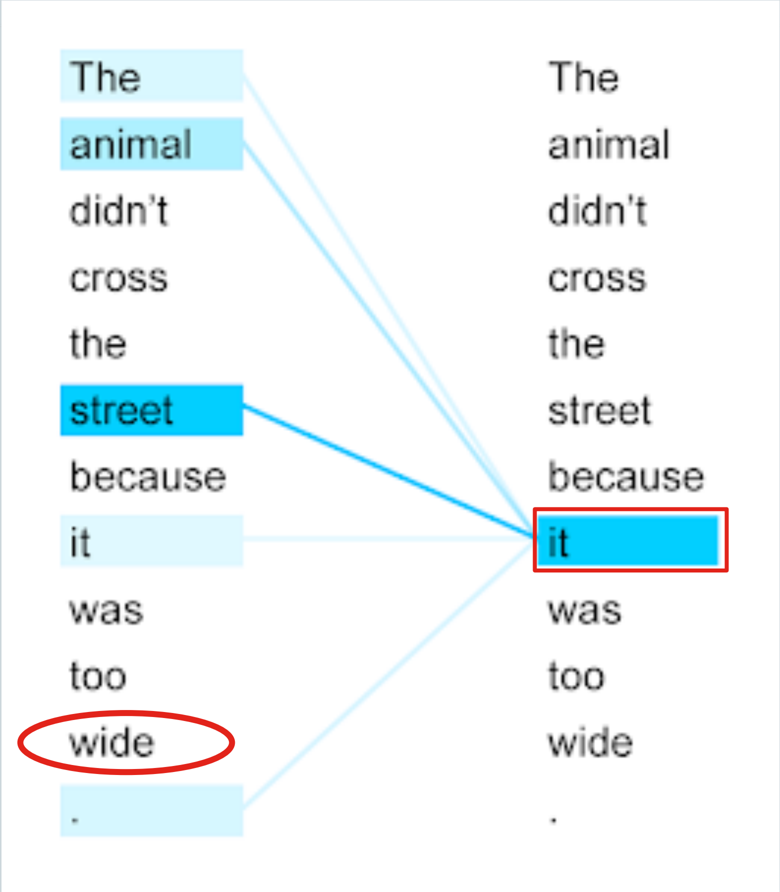
Many models are bi-directional, so access all inputs after this one too

Each input computation is independent of others, so can be parallelized

(Jurafsky and Martin, Fig. 9.15)

Self-attention – example distribution



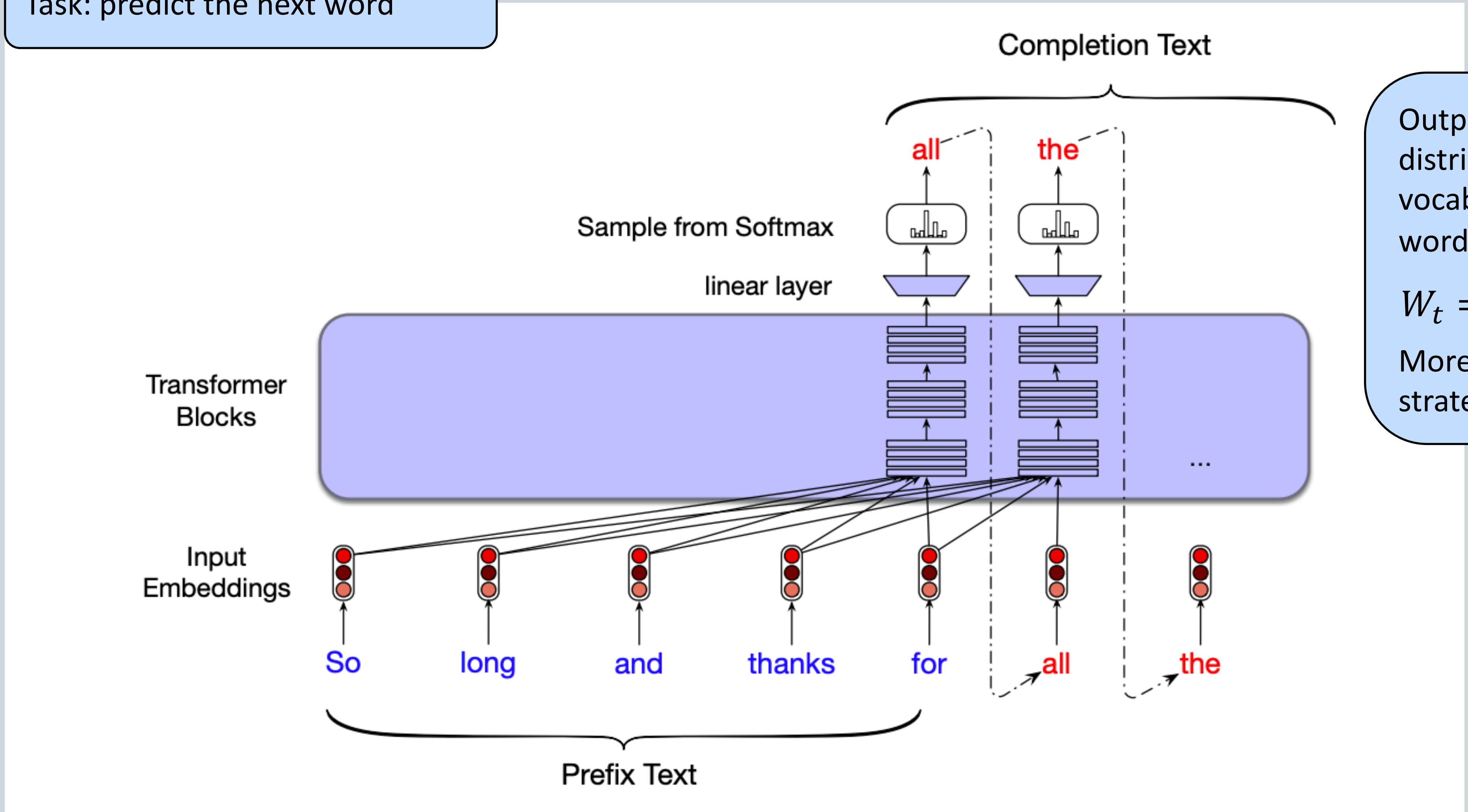


The encoder self-attention distribution for the word "it" from the 5th to the 6th layer of a Transformer trained on English to French translation (one of eight attention heads).

Credit: https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

Completing text

Task: predict the next word



Output at each step is a probability distribution over our entire vocabulary. We could choose the word with the greatest probability:

$$W_t = argmax_W(P(W \mid W_{1:t-1}))$$

More sophisticated sampling strategies are possible

Using text completion for practical NLP tasks

- Sentence classification:

  select the classification with the highest probability:

  `P(positive|The sentiment of the sentence “Such a good movie!” is:)`

  `P(negative|The sentiment of the sentence “Such a good movie!” is:)`

- Question answering:

  `P(w|Q: Who wrote the book “The Origin of Species”? A:)`

  Choose the next most probable word and ask:

  `P(w|Q: Who wrote the book “The Origin of Species”? A: Charles)`

Large Language Models

	BERT	GPT	GPT2	GPT3	GPT4
Model layer	Encoder	Decoder	Decoder	Decoder	Decoder
Pre-training task	MLM, NSP	Text generation	+ task conditioning	+ in-context patterns	
Training data	3.3 billion words	7000 books	40 GB	45 TB	1 PB ?
Context window	512	512	1024	2048	8000 – 32000 ?
Parameters	110 M	117 M	1.5 B	175 B	1 T ?
Suitability	Sequence tasks	Generation	Generation	Generation, adaptable	Generation, adaptable
Availability	Open	Open	Open	Limited, API	Limited, API

Cost →

← Efficiency

← Performance? Biases? →

Thank you

angus.roberts@kcl.ac.uk
https://www.kcl.ac.uk/people/angus-roberts