CS 480/680 Introduction to Machine Learning

Lecture 16

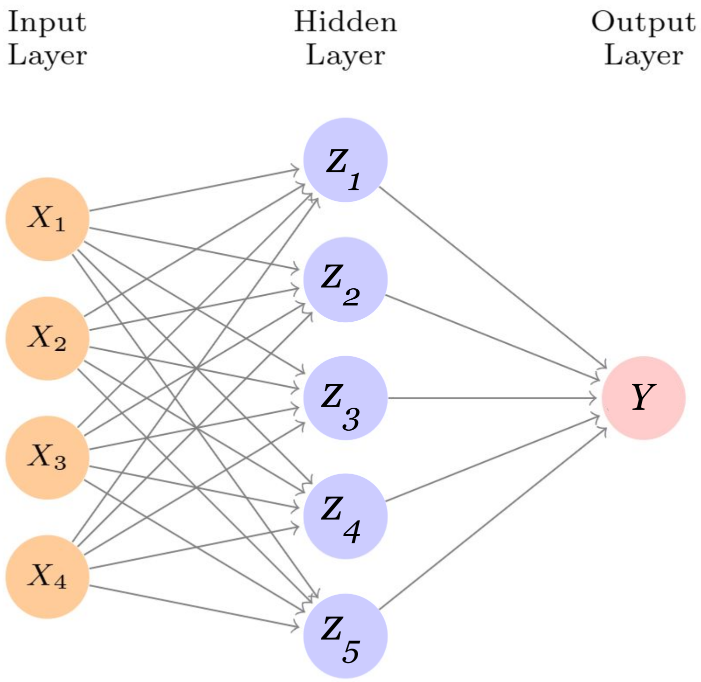
Attention Mechanisms

Kathryn Simone

12 November 2024

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Hidden activations depend on a linear combination of inputs



$$\mathbf{Z} = \varphi(\mathbf{XW})$$

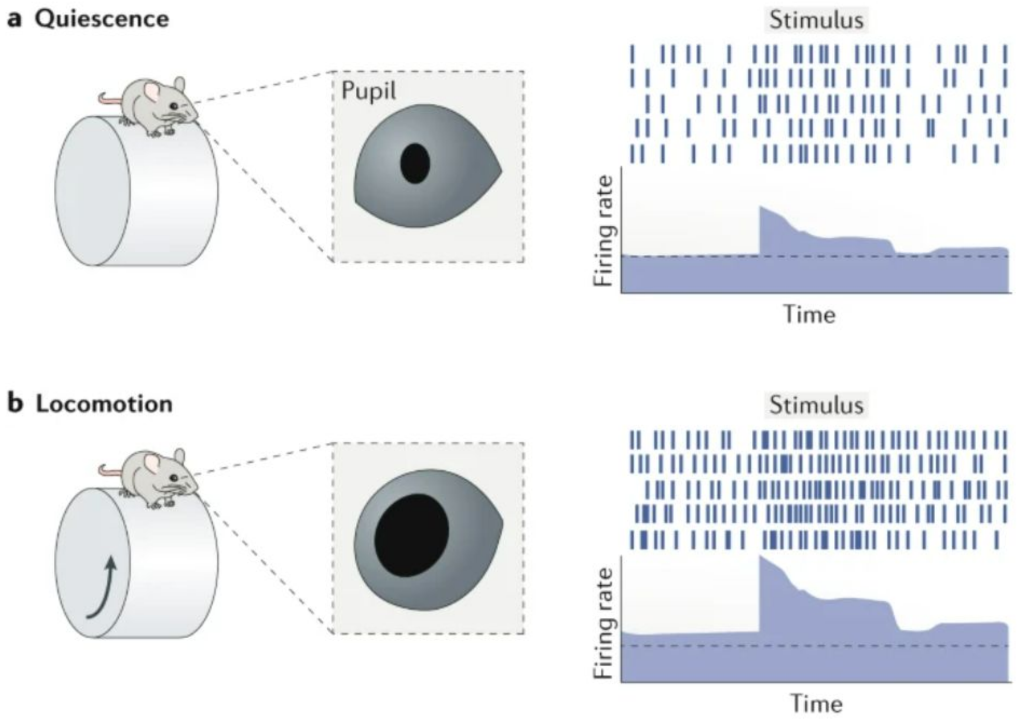
$\mathbf{X} \in \mathbb{R}^{\mathrm{n \times d}}$ (hidden) feature vectors

$\mathbf{W} \in \mathbb{R}^{\mathrm{d \times m}}$ learnable weights

$\mathbf{Z} \in \mathbb{R}^{\mathrm{n \times m}}$ outputs

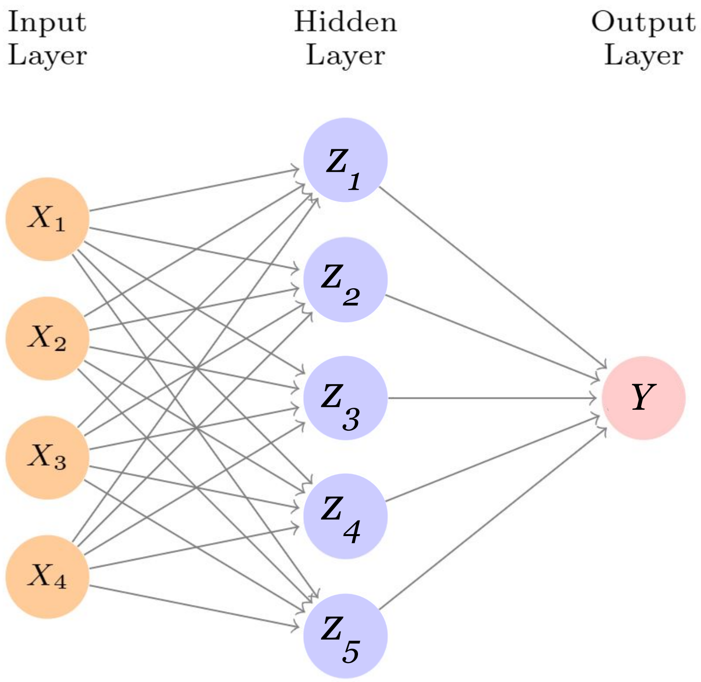
Adapted from ESL Section 10.1

The brain uses the context to modulate gain of inputs



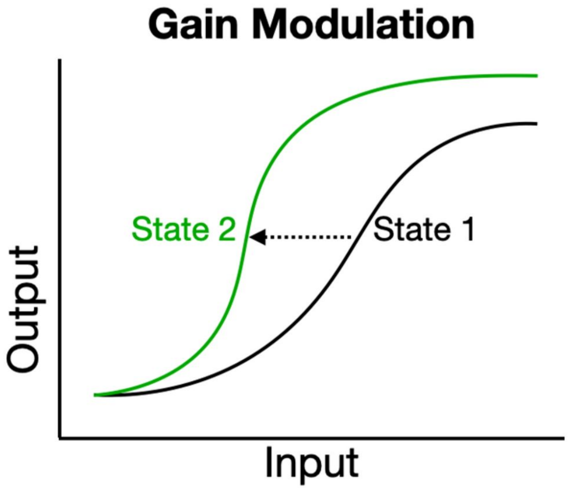
Left: Ferguson and Cardin, 2020
Right: Corbetta, Patel, and Schulman 2008

What if we allowed the weights to depend on the input?

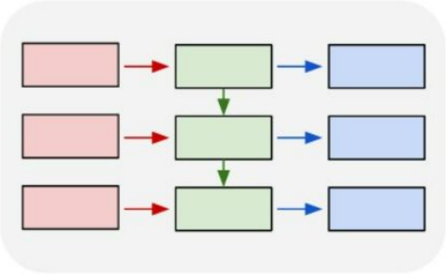


$$\mathbf{W} = \psi(\mathbf{X})$$

$$\mathbf{Z} = \varphi(\mathbf{X}\psi(\mathbf{X}))$$



Attention is particularly useful for language tasks



$$f_\theta : \mathbb{R}^{\mathrm{TD}} \rightarrow \mathbb{R}^{\mathrm{T'C}}$$

Application: Machine Translation

<| French → {Reprise de la session,
Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre
dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances.,
Comme vous avez pu le constater, le grand "bogue de l'an 2000" ne s'est pas produit. En
revanche, les citoyens d'un certain nombre de nos pays ont été
victimes de catastrophes naturelles qui ont vraiment été terribles.},
English → {Resumption of the session, I declare resumed the session of the European Parliament
adjourned on Friday 17 December 1999, and I would like once again to wish
you a happy new year in the hope that you enjoyed a pleasant festive period.,
Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people
in a number of countries suffered a series of natural disasters that truly were dreadful.} |>

Key questions

I. How can we apply attention to sequence tasks?

II. Can we get attention without sequential processing?

III. What are the limitations of these models?

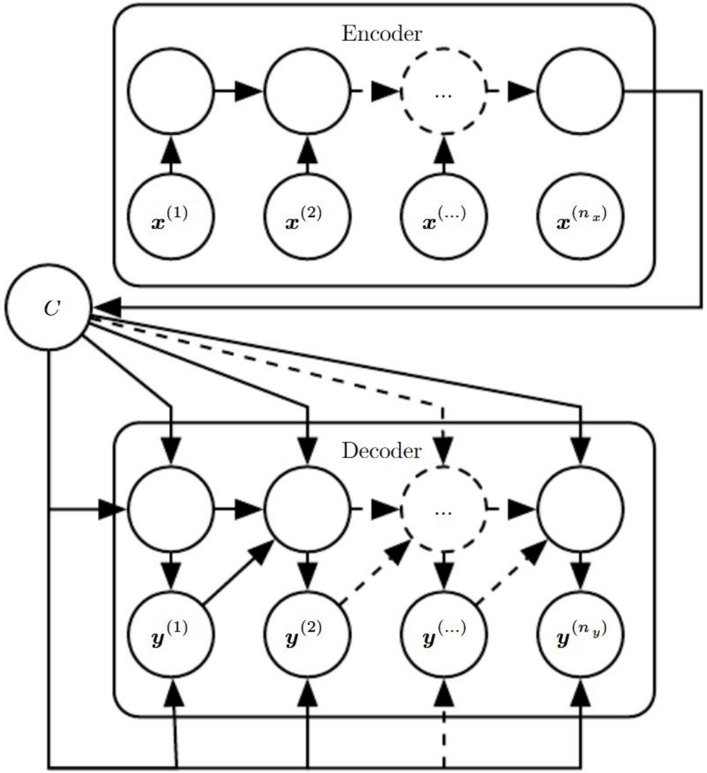
Key questions

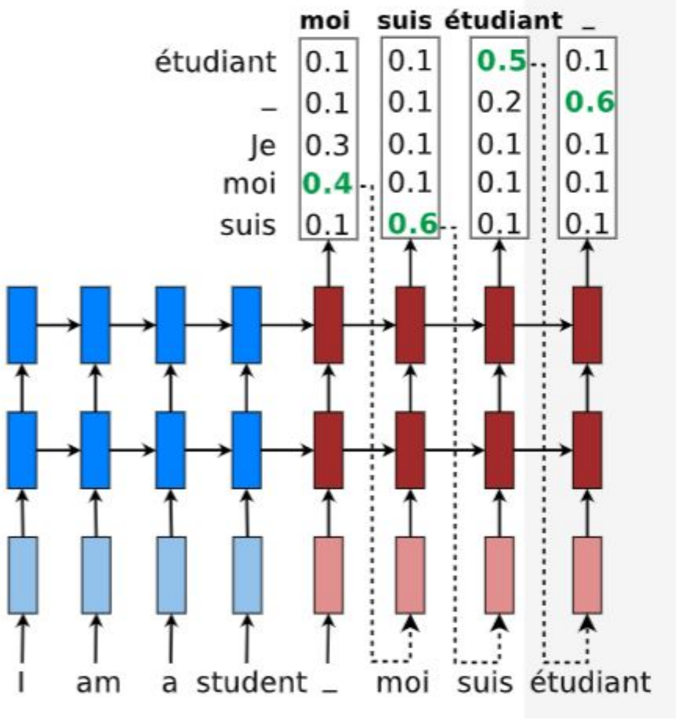
I. **How can we apply attention to sequence tasks?**

II. Can we get attention without sequential processing?

III. What are the limitations of these models?

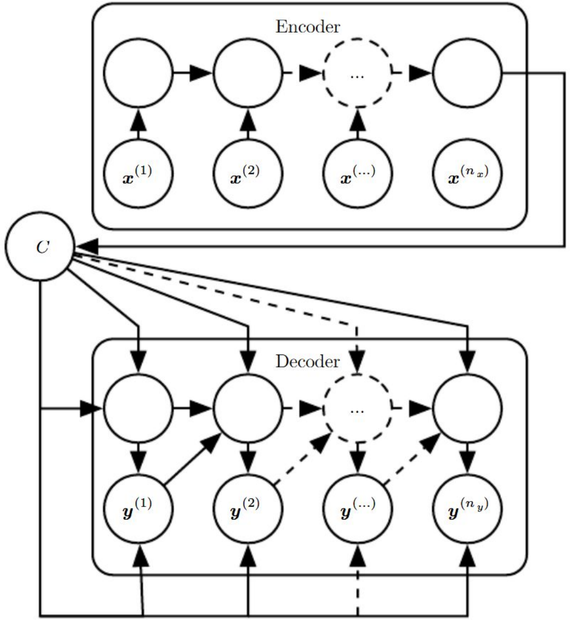
Encoder/Decoder architecture for language translation

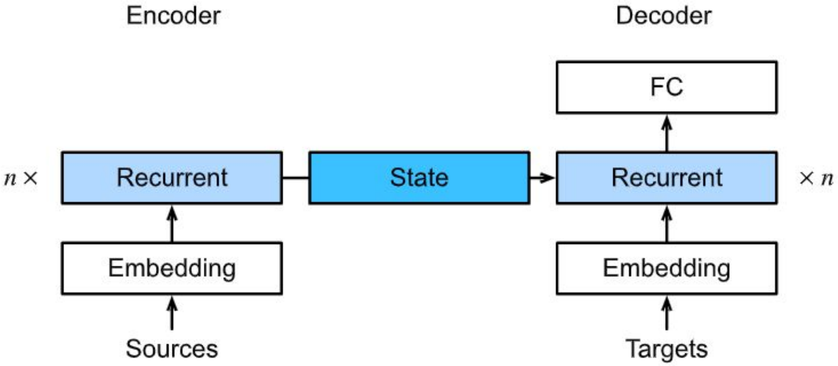




Deep Learning, Section 10.4

The Encoder/Decoder has an information bottleneck



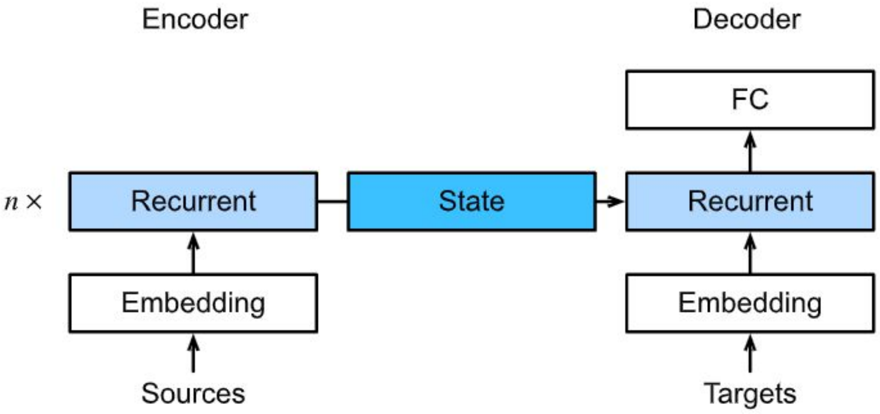


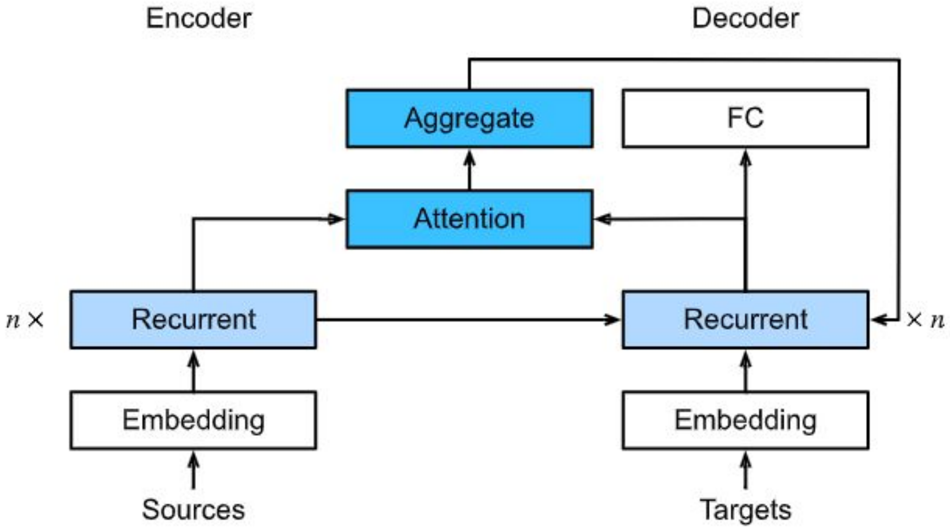
An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

Left: Deep Learning Section 10
Bottom right: Bahdanau, Cho, and Bengio 2015

Encoder/Decoder with Attention





Dive into Deep Learning, Section 11.4

A context vector avoids the bottleneck

Alignment scores:

$$e_{t,i} = a(s_{t-1}, h_i)$$

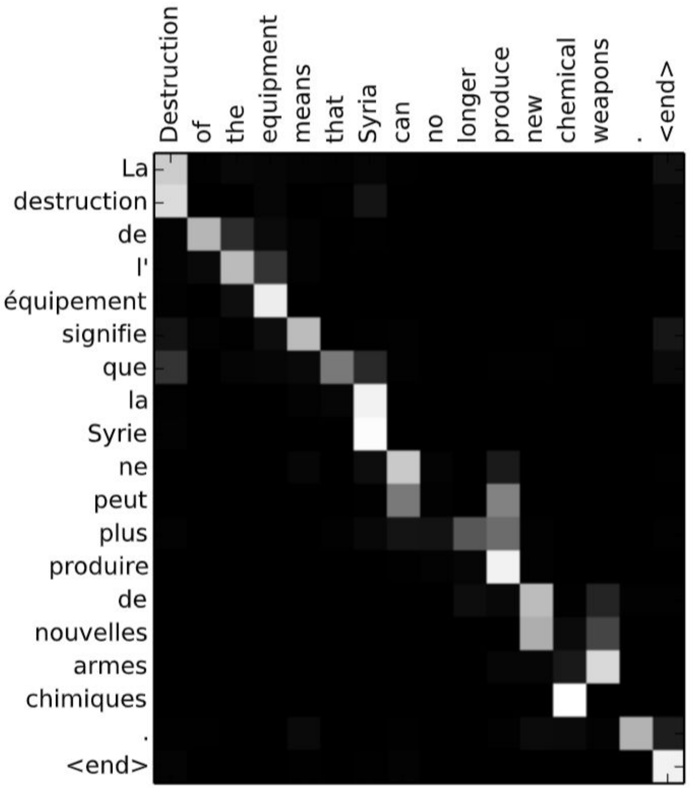
h_i : encoded hidden states
s_{t-1} : previous decoder output

Context vector presented as input to the decoder:

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} h_i$$

Weights:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{T} \exp(e_{t,k})}$$



Right: Bahdanau, Cho, and Bengio 2015

A general definition of an attention function

Suppose we have a database of m (key,value) pairs $\mathcal{D} = \{(k_1, v_1), \ldots, (k_m, v_m)\}$, and some query $\mathbf{q}$. The *attention* over $\mathcal{D}$ is

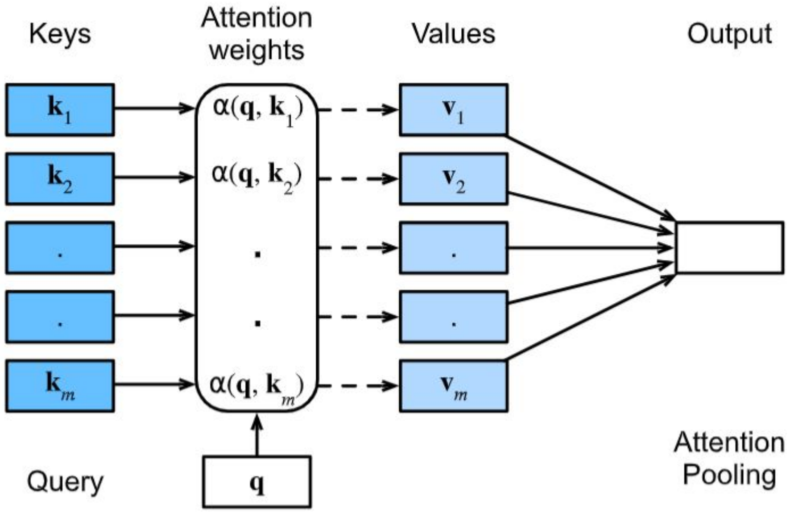
$$\text{Attention}(\mathbf{q}, \mathcal{D}) = \sum_{i=1}^{m} \alpha(\mathbf{q}, \mathbf{k}_i)\mathbf{v}_i,$$

where $\alpha(\mathbf{q}, \mathbf{k}_i) \in \mathbb{R}$ $(i = 1, \ldots, m)$ are scalar attention weights.

Special cases:

Exactly one of the weights $\alpha(\mathbf{q}, \mathbf{k}_i)$ is 1, while all others are 0.

Uniform weighting, where $\alpha(\mathbf{q}, \mathbf{k}_i) = \frac{1}{m} \forall i$.



Dive into Deep Learning, Section 11.1

Some attention functions

Additive Attention:

$$a(\mathbf{q}, \mathbf{k}) = \mathbf{w}_v^\top \tanh(\mathbf{W}_q\mathbf{q} + \mathbf{W}_k\mathbf{k}) \in \mathbb{R}$$

$$\text{where } \mathbf{W}_q \in \mathbb{R}^{h \times q}, \mathbf{W}_k \in \mathbb{R}^{h \times k}$$

Scaled Dot-Product Attention

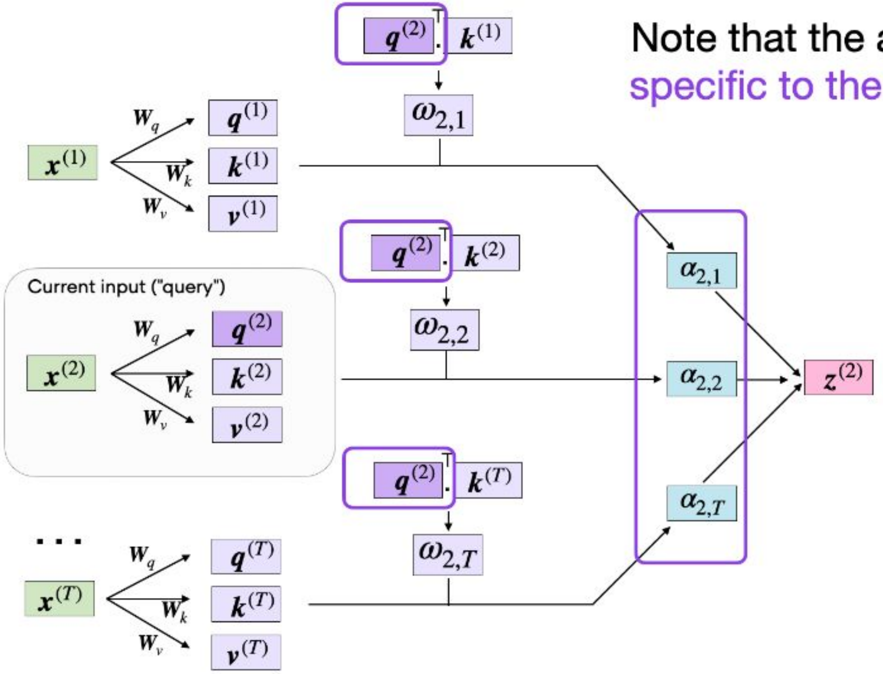
$$a(\mathbf{q}, \mathbf{k}) = \frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}} \in \mathbb{R}$$

For a minibatch of data:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \in \mathbb{R}^{n \times v}$$

$$\text{where } \mathbf{Q} \in \mathbb{R}^{n \times d}, \mathbf{K} \in \mathbb{R}^{m \times d}, \text{ and } \mathbf{V} \in \mathbb{R}^{m \times v}$$

Implementing attention in a neural network



Note that the attention scores are specific to the current input token

where $z^{(2)} = \sum_{j=1}^{T} \boxed{\alpha_{2,j}} \, v^{(j)}$

https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Key questions

I. How can we apply attention to sequence tasks?

II. **Can we get attention without sequential processing?**

III. What are the limitations of these models?

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Vaswani et al, 2018

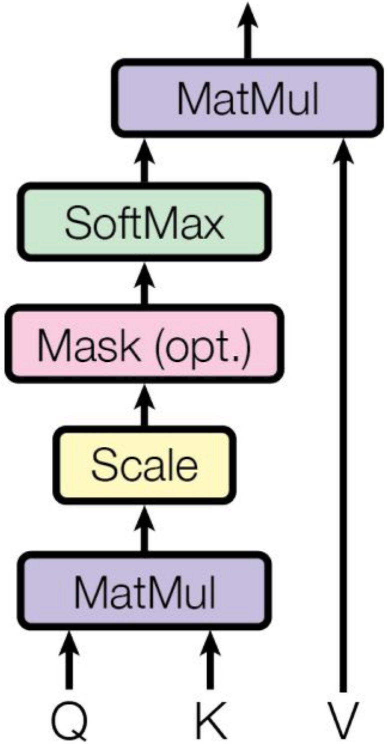
Self-attention models the relationships within a sequence

Given a sequence of input tokens $\{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^{\mathrm{d}}$, self-attention generates a sequence

$$y_i = \mathrm{Attn}(x_i, (x_1, x_1), \ldots, (x_n, x_n))$$

where the query is x_i, and the keys and values are all the (valid) inputs $x_1, \ldots, x_n$.

- Encoder and Decoder have self-attention
- Advantages:
  - Parallelizability
  - Shorter paths for signals to “travel” between tokens

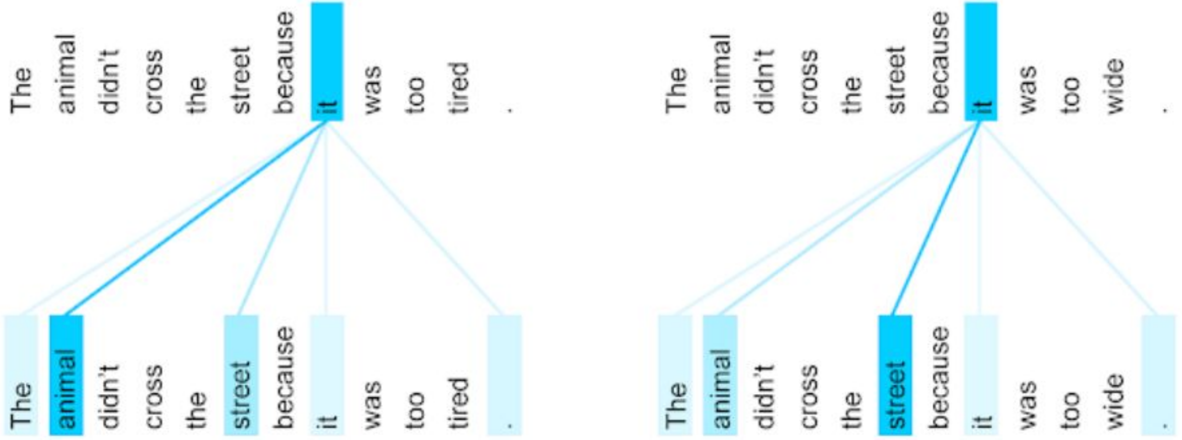


Vaswani et al, 2018

Self-attention improves representation of context

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.



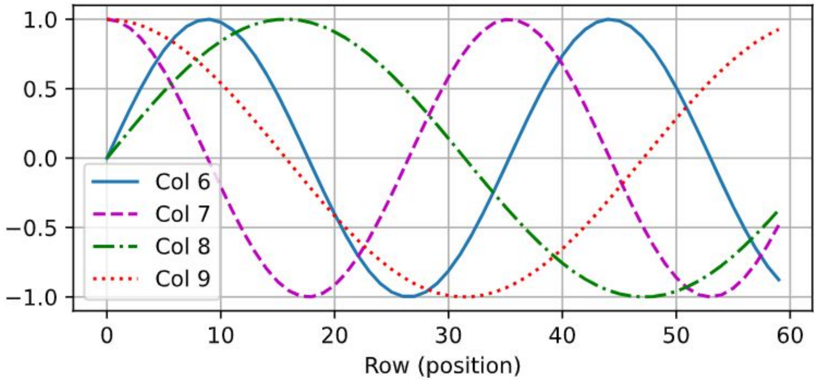
https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

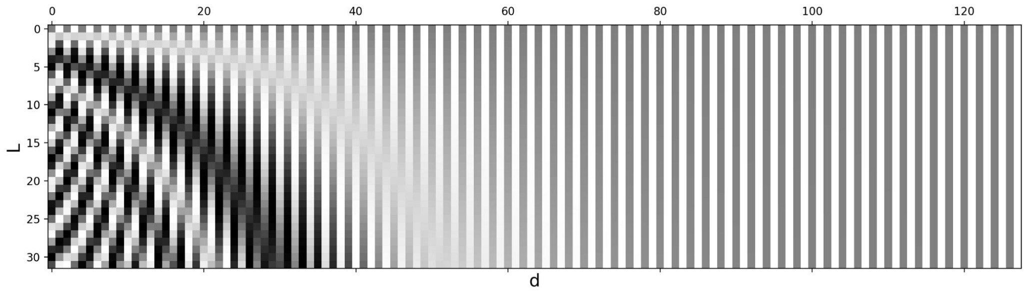
Positional encoding provides order information

$$\mathrm{POS}(\mathrm{Embed}(\mathbf{X})) = \mathbf{X} + \mathbf{P}$$

$$\mathbf{P}_{i,j} = \begin{cases} \sin\left(\frac{i}{C^{2j/d}}\right) & \text{if } j = 2\delta \\ \cos\left(\frac{i}{C^{2j/d}}\right) & \text{if } j = 2\delta + 1 \end{cases}$$

where C is the maximum sequence length.





https://lilianweng.github.io/posts/2020-04-07-the-transformer-family

Example: Compute positional embedding

$$\mathbf{P}_{i,j} = \begin{cases} \sin\left(\frac{i}{C^{2j/d}}\right) & \text{if } j = 2\delta \\ \cos\left(\frac{i}{C^{2j/d}}\right) & \text{if } j = 2\delta + 1 \end{cases}$$

Compute row $i = 5$ for an embedding dimension of $d = 3$,
given max sequence length $C = 10$

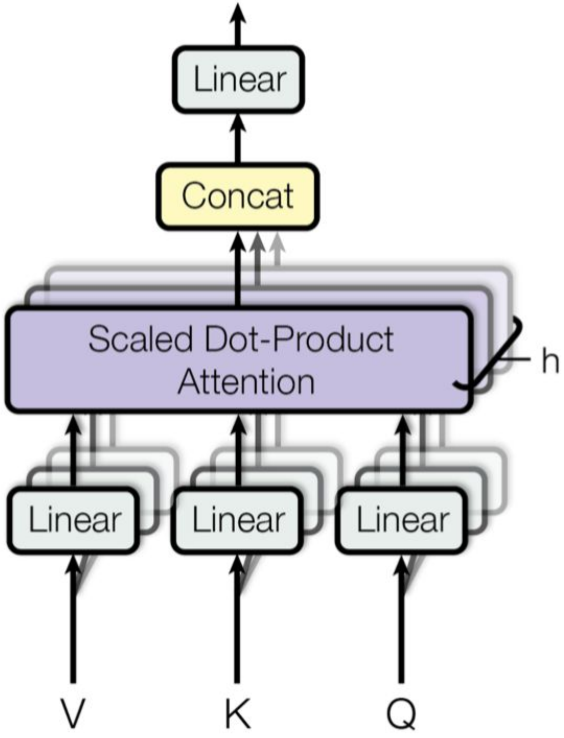
Multi-head attention captures different notions of similarity

The cat chased the mouse down the dark alley.

Semantic Similarity?
The cat chased the mouse down the dark alley.

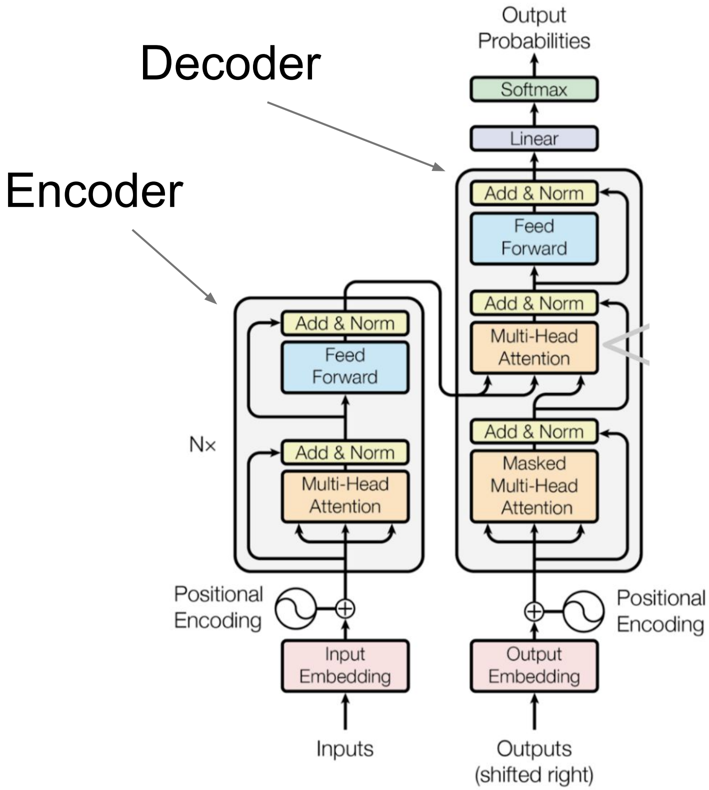
Positional Similarity?
The cat chased the mouse down the dark alley.

Subject/Verb Similarity?
The cat chased the mouse down the dark alley.



https://lilianweng.github.io/posts/2020-04-07-the-transformer-family

The Transformer Neural Network Architecture



https://lilianweng.github.io/posts/2023-01-27-the-transformer-family-v2/

Complete this “autoregressive task” :)

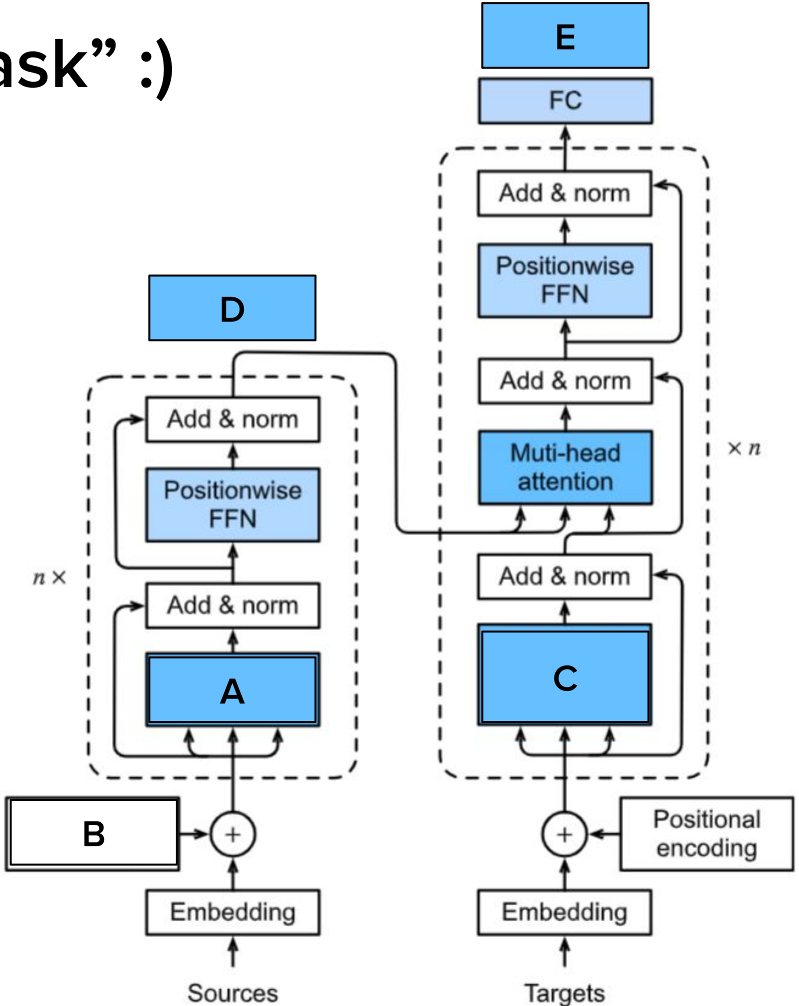
Generate a rich and nuanced attention-based representation

Provide ordering information to the model

Ensure attention is directed at past inputs only

Generate the output sequence

Combines information across the sequence



GPT: Generative Pre-training Transformer

1. **“Unsupervised” Pre-Training: Predict the next word**

   The quick brown fox jumped over the lazy ______.

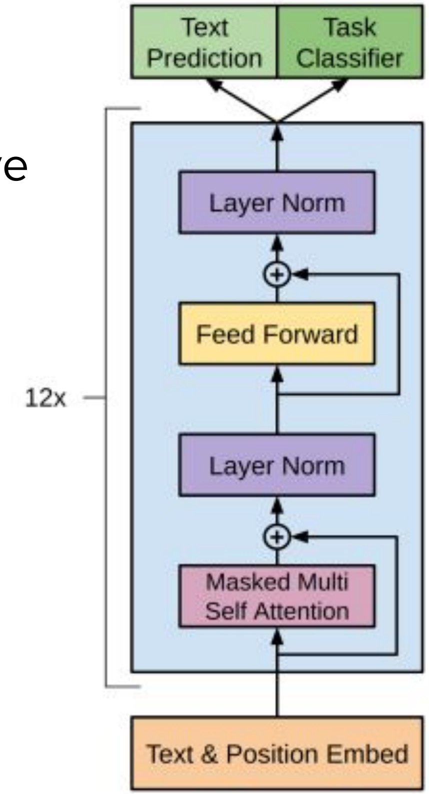
   Train the decoder only with “Language modelling” objective

$$\mathcal{L}_{\mathrm{LM}} = -\sum_{x \in \mathcal{D}_U} \sum_t \log p(x_t | x_{1:t-1})$$

2. **Supervised Fine-Tuning: Predict the label**

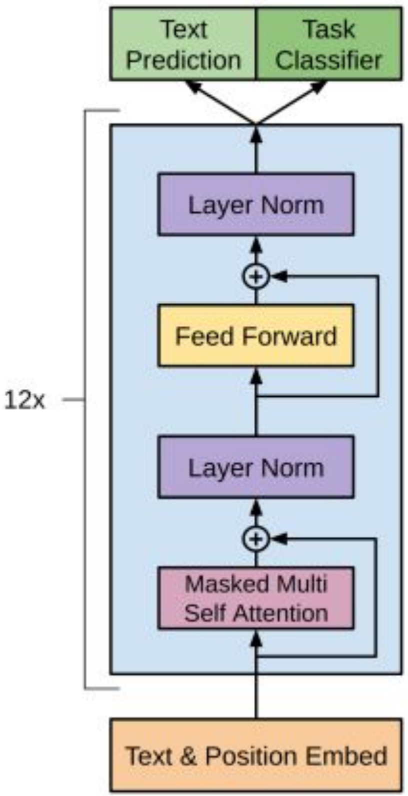
$$\mathcal{L}_{\mathrm{cls}} = -\sum_{(x,y) \in \mathcal{D}_L} \log p(y|x)$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{LM}}$$



Radford et al, 2018

Scaling up GPTs



GPT-1 (2018)

- 117M parameters
- Trained on: 600b words, 40Gb of data

GPT-2 (2019)

- 1.5B parameters
- Trained on WebText: 41Gb of data; no labelled data!

GPT-3 (2020)

- 175B parameters
- 570Gb of data
- Reinforcement learning with human feedback

GPT-4

- 1.7T parameters

GPT-2: Performs tasks without supervised training

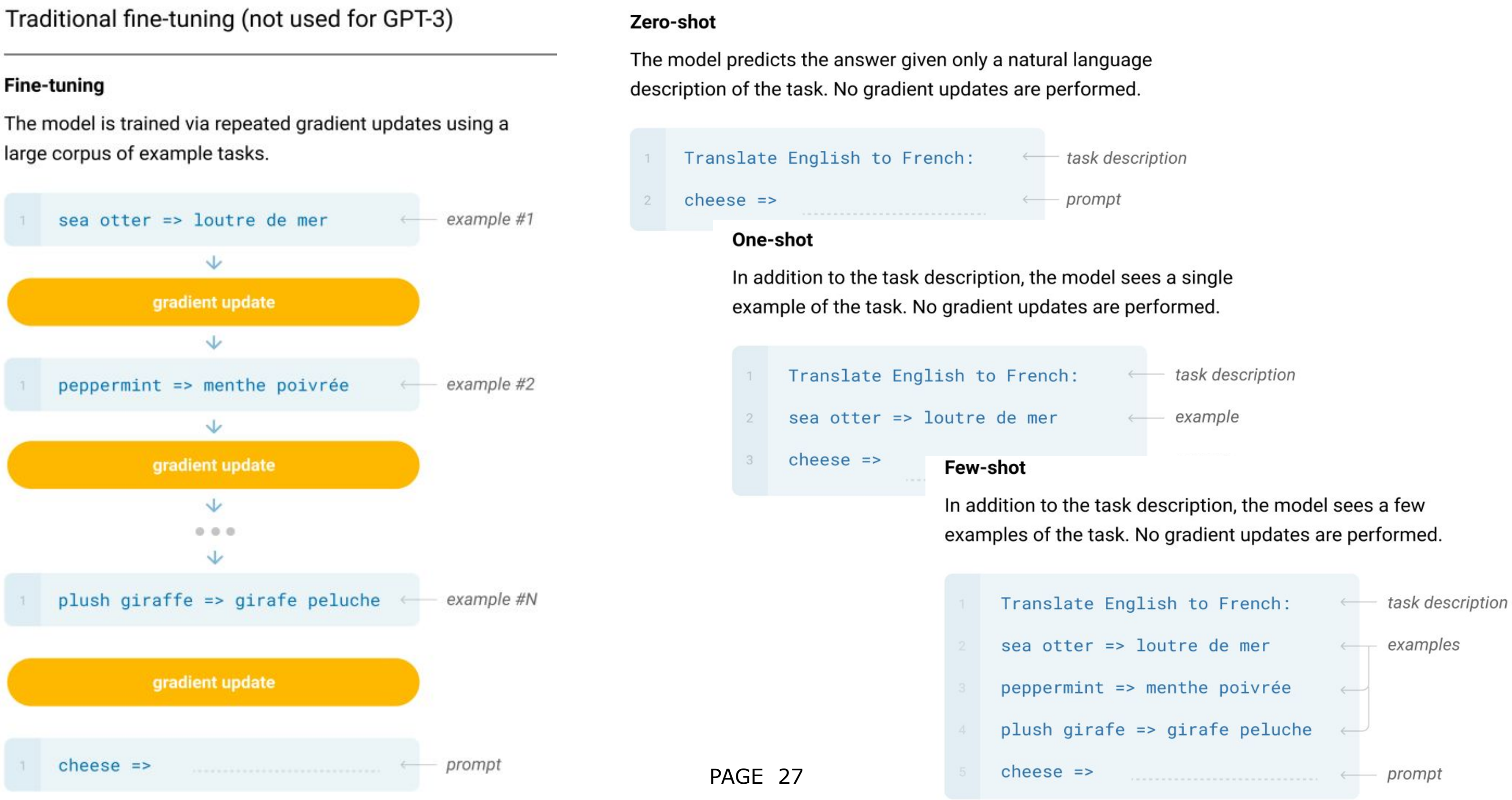
Language Models are Unsupervised Multitask Learners

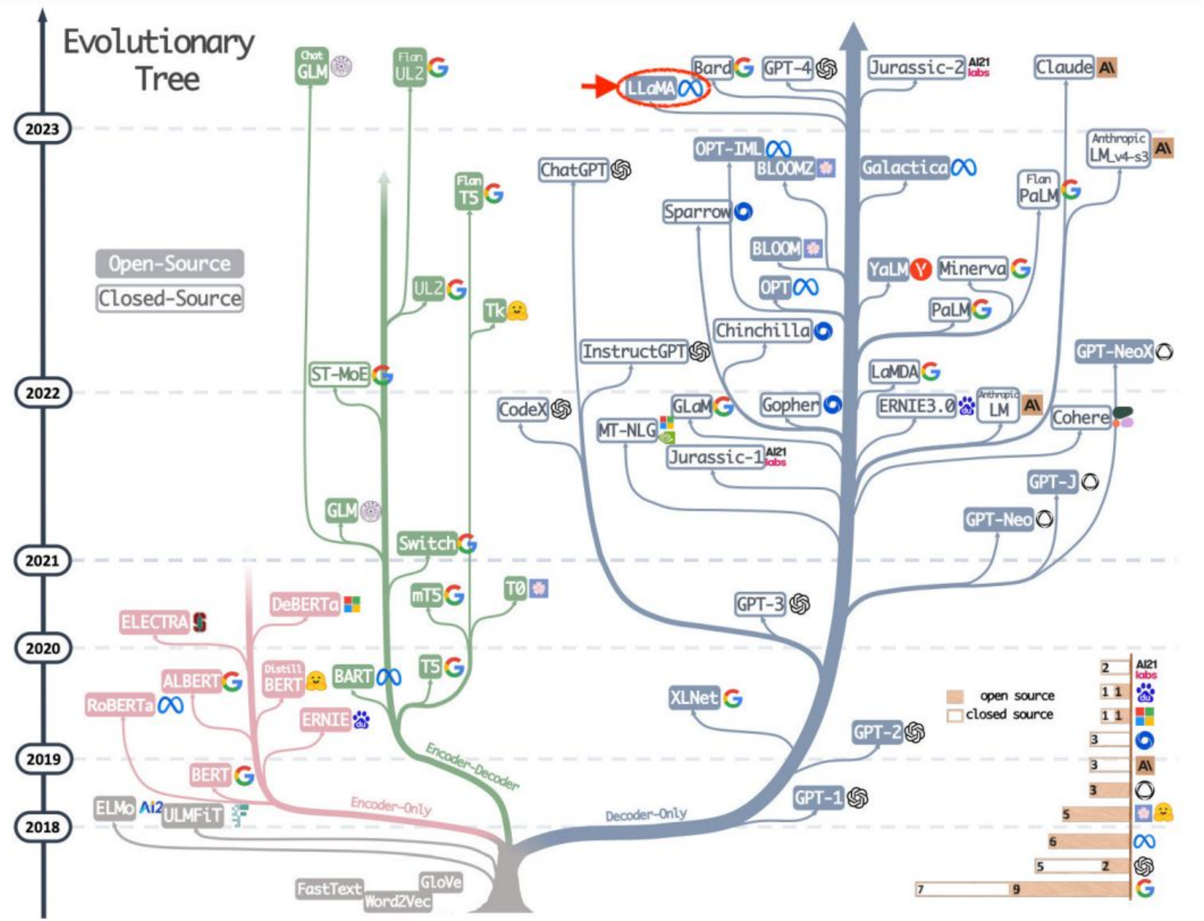
Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
Who is the head of the department of homeland security 2017?	John Kelly	✓	47.0%
What is the name given to the common currency to the european union?	Euro	✓	46.8%
What was the emperor name in star wars?	Palpatine	✓	46.5%
Do you have to have a gun permit to shoot at a range?	No	✓	46.4%
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓	45.7%
Nuclear power plant that blew up in russia?	Chernobyl	✓	45.7%
Who played john connor in the original terminator?	Arnold Schwarzenegger	✗	45.2%

Table 5. The 30 most confident answers generated by GPT-2 on the development set of Natural Questions sorted by their probability according to GPT-2. None of these questions appear in WebText according to the procedure described in Section 4.

Radford et al, 2019

GPT-3: In-context learning





Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
LLaMA
Bard
GPT-4
Jurassic-2
AI21 labs
Claude
OPT-IML
BLOOMZ
Galactica
Anthropic LM v4-s3
ChatGPT
Flan T5
Sparrow
Flan PaLM
BLOOM
YaLM
Minerva
UL2
OPT
Tk
PaLM
Chinchilla
InstructGPT
GPT-NeoX
ST-MoE
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GLM
GPT-Neo
Switch
mT5
T0
GPT-3
ELECTRA
DeBERTa
Distill BERT
BART
T5
ALBERT
RoBERTa
XLNet
ERNIE
GPT-2
open source
closed source
BERT
Encoder-Decoder
Encoder-Only
Decoder-Only
GPT-1
ELMo
Ai2
ULMFiT
FastText
Word2Vec
GloVe
2
1 1
1 1
3
3
3
5
6
5 2
7 9

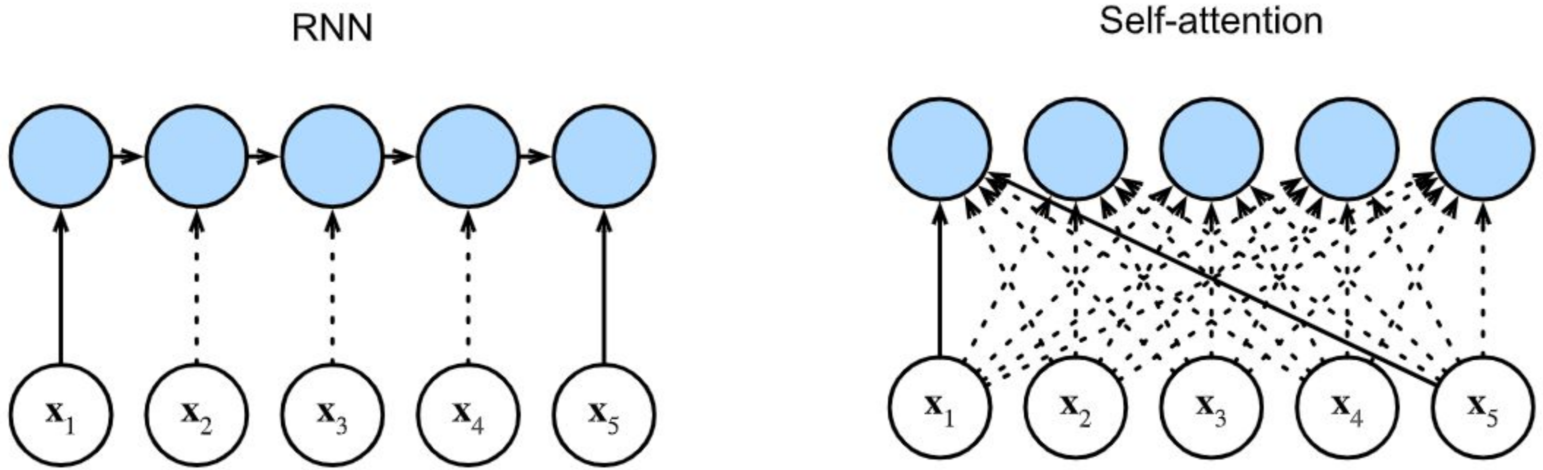
Key questions

I. How can we apply attention to sequence tasks?

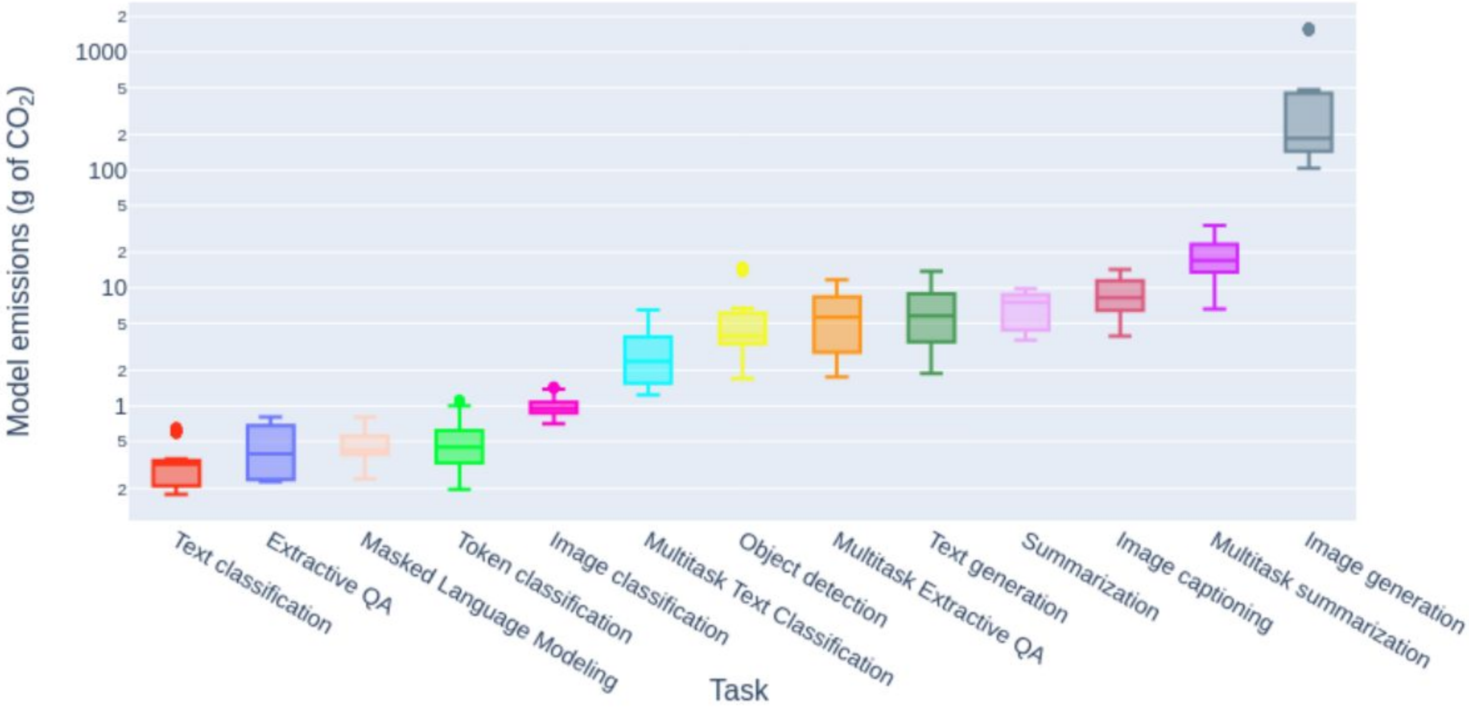
II. Can we get attention without sequential processing?

III. **What are the limitations of these models?**

Self-attention is computationally expensive



Layer type	Complexity	Sequential ops.	Max. path length
Self-attention	$O(n^2 d)$	$O(1)$	$O(1)$
Recurrent	$O(nd^2)$	$O(n)$	$O(n)$



Model emissions (g of CO_2)
2
1000
5
2
100
5
2
10
5
2
1
5
2
Text classification
Extractive QA
Masked Language Modeling
Token classification
Image classification
Multitask Text Classification
Object detection
Multitask Extractive QA
Text generation
Summarization
Image captioning
Multitask summarization
Image generation
Task

Discussion: Other limitations or ethical issues?

Now that we're at the end of the lecture, you should be able to...

- ★ Explain the operation of the **encoder/decoder architecture** of an RNN.
- ★ Describe the **challenges of dealing with sequential data** using RNNs, and the key **advantages offered by attention** as compared to RNNs.
- ★ **Compute attention** for a short sequence given a queries, keys, and values.
- ★ Differentiate between **attention, self-attention, masked attention, and multi-head attention**.
- ★ Label the **key components of the transformer architecture**.
- ★ Describe the **generative pre-training paradigm** as used in GPT-1 with reference to the appropriate objectives.
- ★ Recall significant **applications of the transformer** in language modeling.
- ★ Discriminate between **fine-tuning** and **zero-, one-**, and **few-shot generation**.
- ★ Defend the **scaleability issues of transformer-based attention models** and .