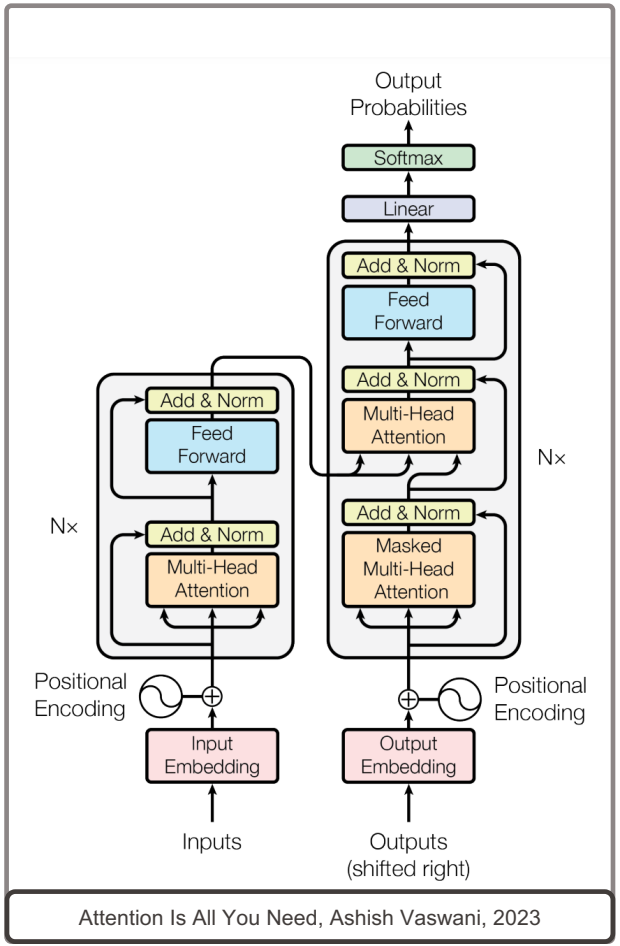
Exercise 5
Transformers

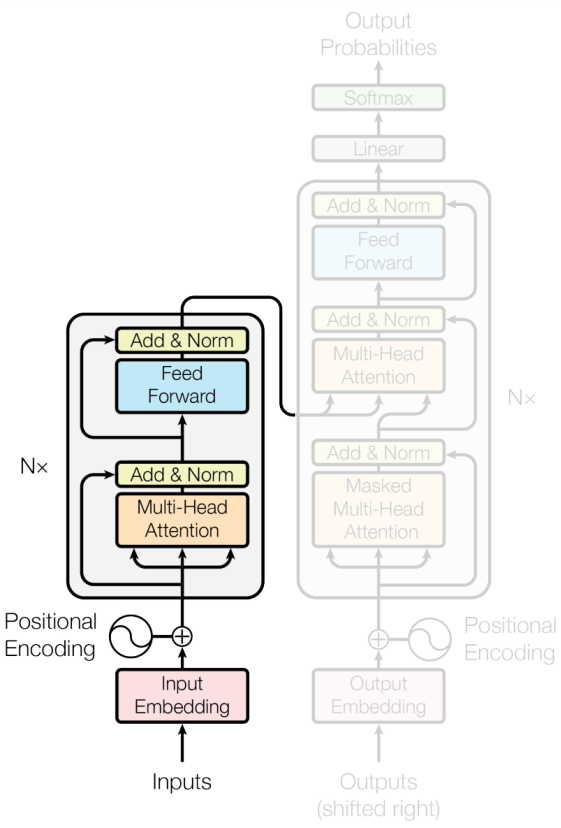
EPFL – 20.10.2023

Transformer - Recap



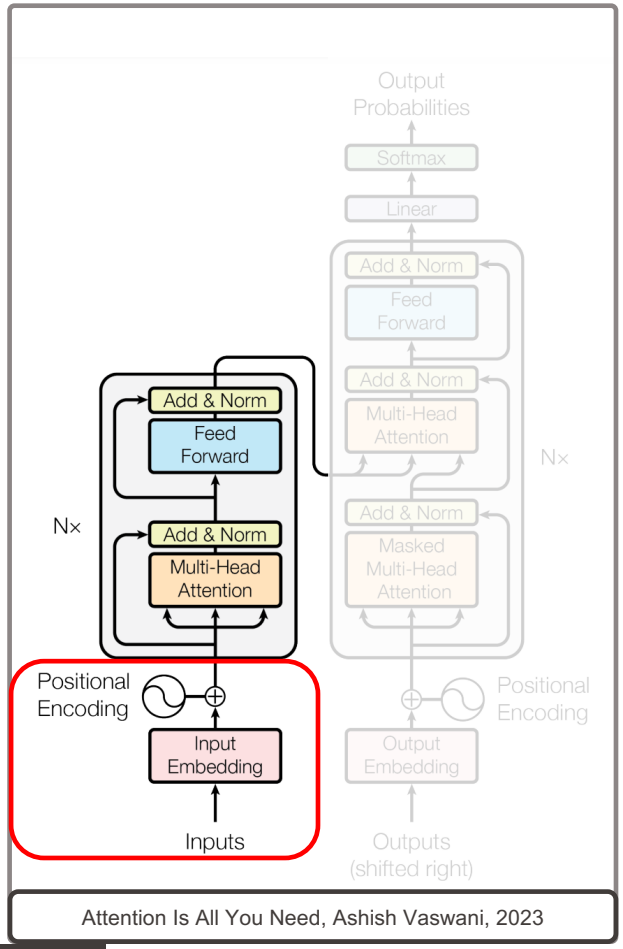
Attention Is All You Need, Ashish Vaswani, 2023

Transformer - Recap



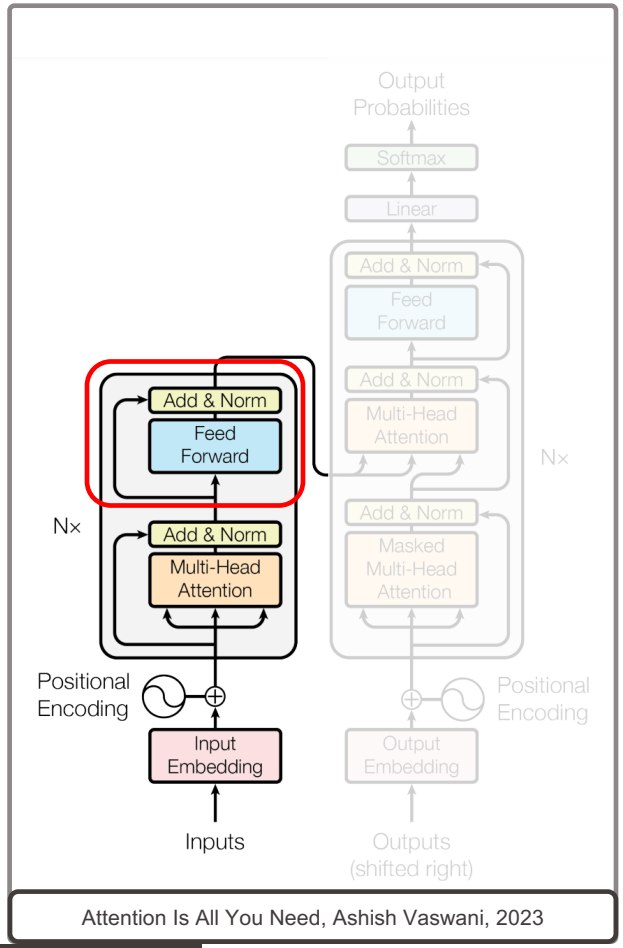
Attention Is All You Need, Ashish Vaswani, 2023

Transformer - Recap



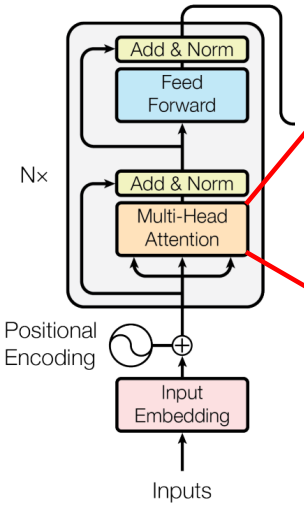
Attention Is All You Need, Ashish Vaswani, 2023

Transformer - Recap

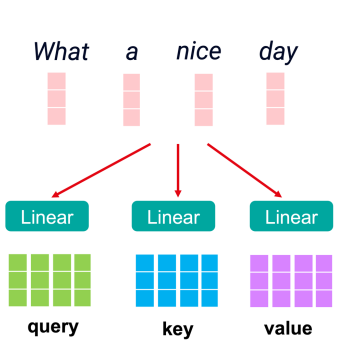


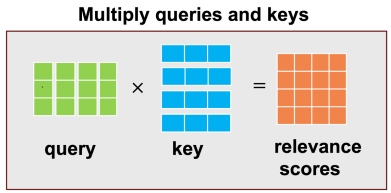
Attention Is All You Need, Ashish Vaswani, 2023

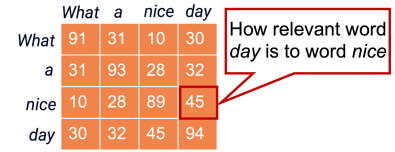
Transformer - Recap



Attention Is All You Need, Ashish Vaswani, 2023

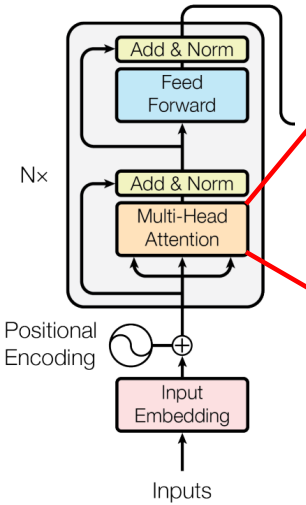




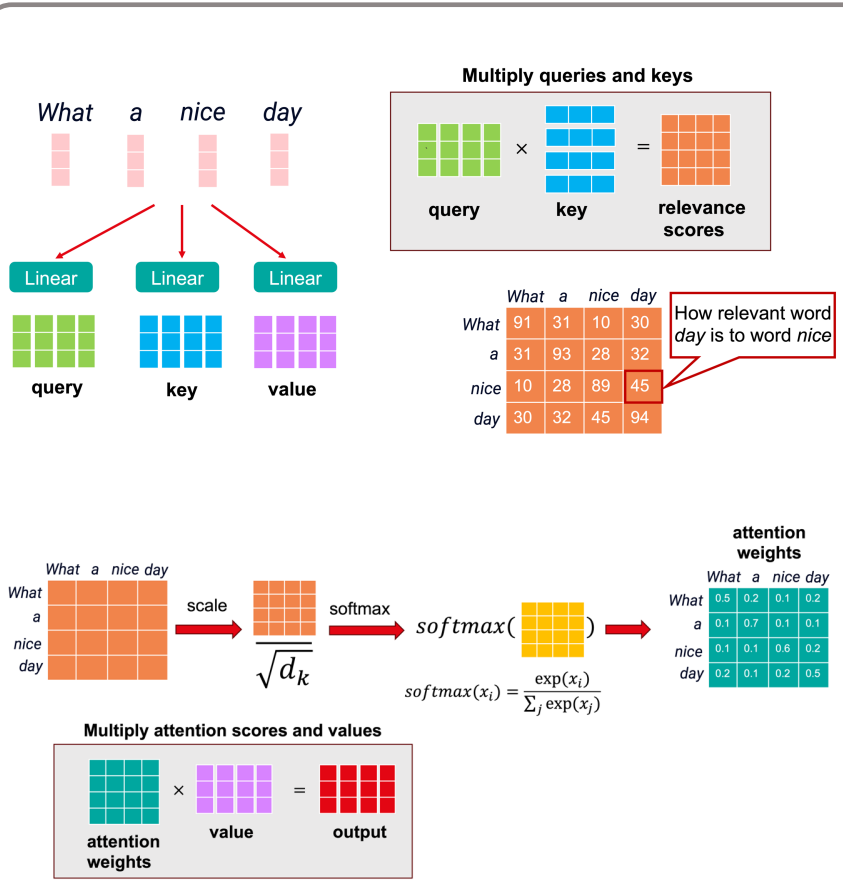


CS502: Lecture 5

Transformer - Recap



Attention Is All You Need, Ashish Vaswani, 2023

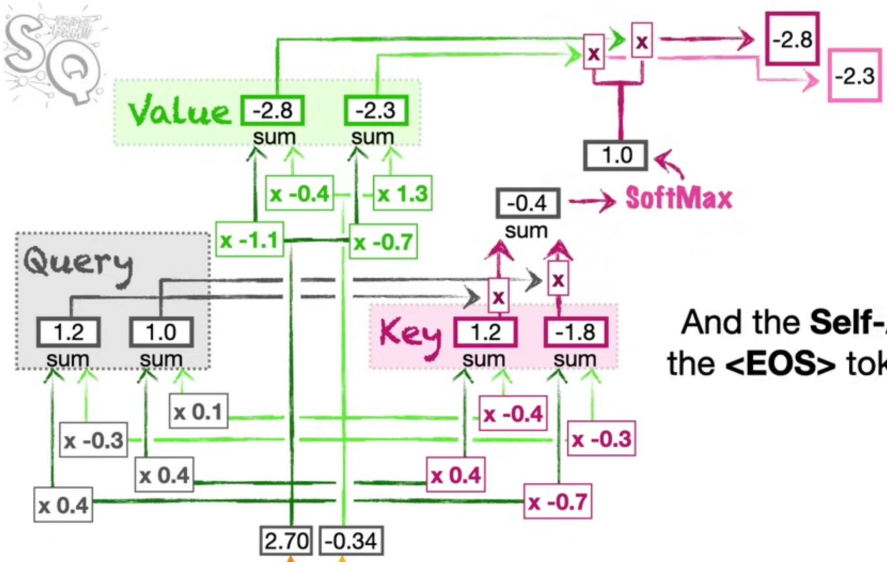


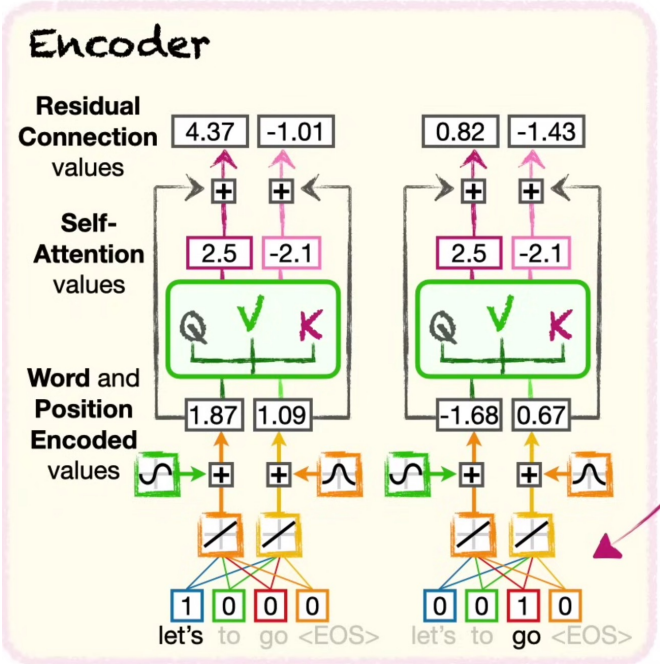
CS502: Lecture 5

Transformer

Tutorial Suggestion:

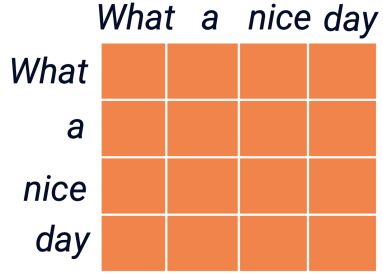
https://www.youtube.com/watch?v=zxQyTK8quyY&t=1087s





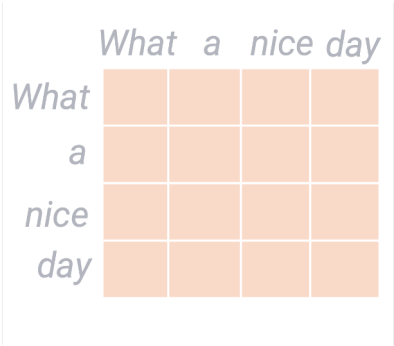
Self-Attention

Quadratic dependency on Sequence Length for Memory and Computation



Self-Attention

Quadratic dependency on Sequence Length for Memory and Computation



Sparse Attention

- use fixed pattern of attention to reduce number of connections

Flash Attention

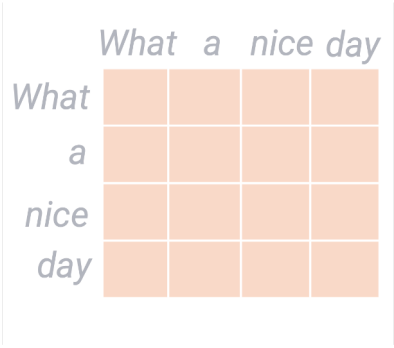
- Split score matrix into blocks
- Recompute attention matrix

Longformer

- local attention for most tokens
- global attention for a few selected tokens

Self-Attention

Quadratic dependency on Sequence Length for Memory and Computation



Sparse Attention

- use fixed pattern of attention to reduce number of connections

Flash Attention

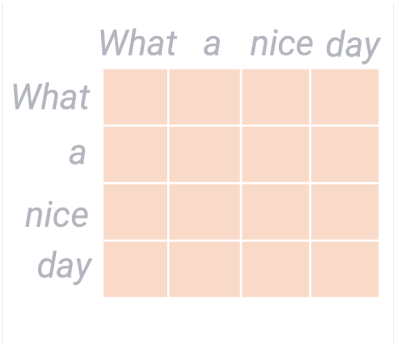
- Split score matrix into blocks
- Recompute attention matrix

Longformer

- local attention for most tokens
- global attention for a few selected tokens

Self-Attention

Quadratic dependency on Sequence Length for Memory and Computation



Sparse Attention

- use fixed pattern of attention to reduce number of connections

Flash Attention

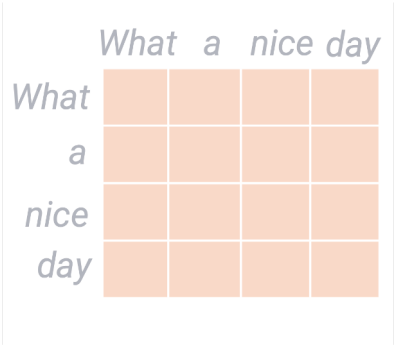
- Split score matrix into blocks
- Recompute attention matrix

Longformer

- local attention for most tokens
- global attention for a few selected tokens

Self-Attention

Quadratic dependency on Sequence Length for Memory and Computation



Sparse Attention

- use fixed pattern of attention to reduce number of connections

Flash Attention

- Split score matrix into blocks
- Recompute attention matrix

Longformer

- local attention for most tokens
- global attention for a few selected tokens

Exercise 5

Debugging Tips - 1

Memory Usage

Which parameters impact the model's memory usage?

Dimensions Mismatch

How are input and output dimensions defined in a transformer?

Training and Evaluation Metrics

Are the metrics behaving as expected?

Debugging Tips - 1

Memory Usage

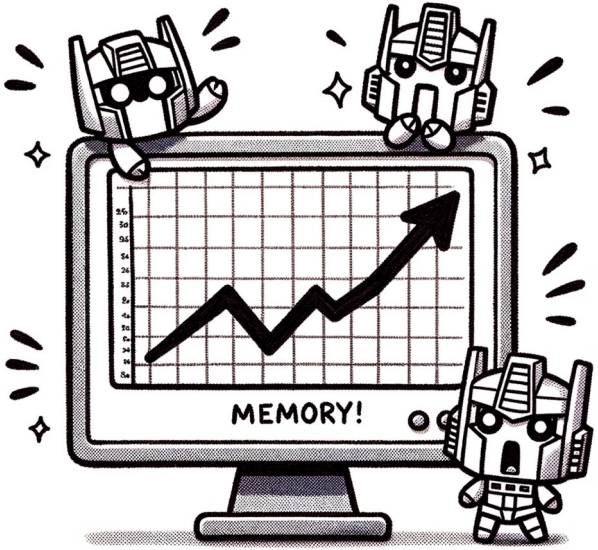
Which parameters impact the model's memory usage?

Dimensions Mismatch

How are input and output dimensions defined in a transformer?

Training and Evaluation Metrics

Are the metrics behaving as expected?



Debugging Tips - 1

Memory Usage

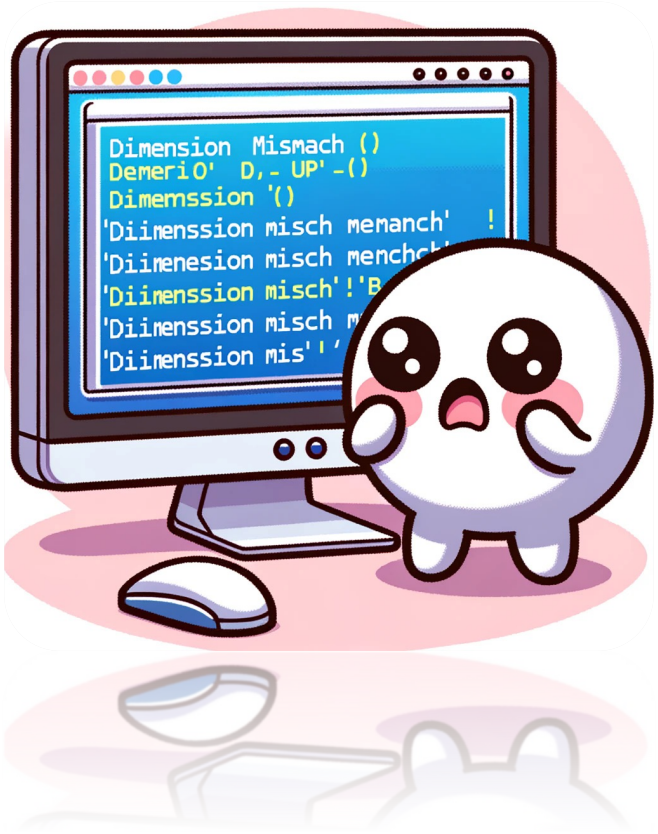
Which parameters impact the model's memory usage?

Dimensions Mismatch

How are input and output dimensions defined in a transformer?

Training and Evaluation Metrics

Are the metrics behaving as expected?



Debugging Tips - 1

Memory Usage

Which parameters impact the model's memory usage?

Dimensions Mismatch

How are input and output dimensions defined in a transformer?

Training and Evaluation Metrics

Are the metrics behaving as expected?

Debugging Tips - 2

Start Small

We provide you a Debug and Run mode

Performance Indicator

Expect a training perplexity around 20 after 50 epochs

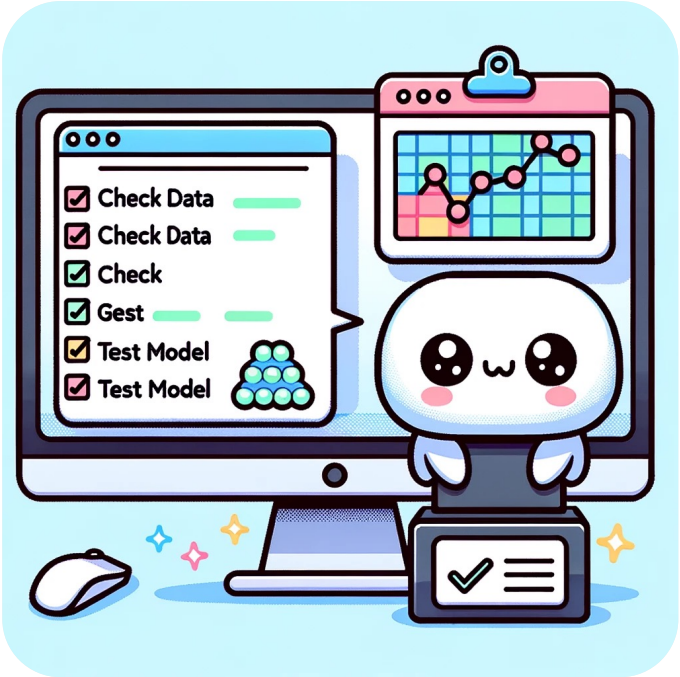
Debugging Tips - 2

Start Small

We provide you a Debug and Run mode

Performance Indicator

Expect a training perplexity around 20 after 50 epochs



Debugging Tips - 2

Start Small

We provide you a Debug and Run mode

Performance Indicator

Expect a training perplexity around 20 after 50 epochs