Advanced Deep Learning - Transformer Models

Multi-Head Attention Mechanism

About the Instructor

Charles E. Gormley

Areas of Study: Data Science & Economics

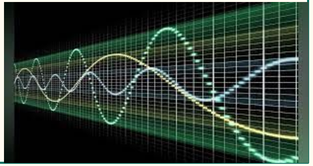
Career

-Economics Researcher

-Machine Learning Engineer - Fintech & Medtech Companies

-Software Engineer - Major Finance Company

-Machine Learning Engineer - Consulting



Agenda

- PreRequisites
- Overview of Transformer Model
- Attention
- Attention Mecahnism
    - Linear Function
    - Query, Key & Value
- Multiple Heads - Why?
    - Attention Filter
- Recap
- What’s Next →

PreRequisites

Mathematics & Modeling

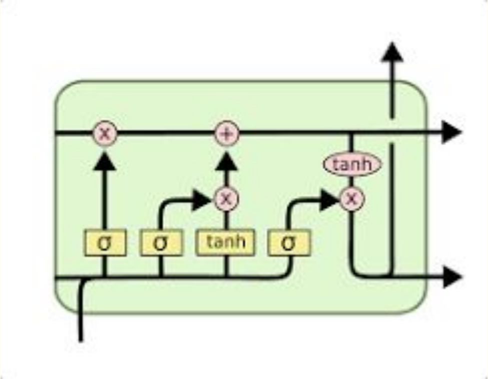
- Linear Algebra: Matrix Multiplication
- Linear Algebra: Scaling Matrices
- Multi-Layer Perceptrons
- Activation Functions
- AutoEncoders
- Positional Encoding
- Tokenization
- Positional Embeddings
- Back-Propagation

Programming

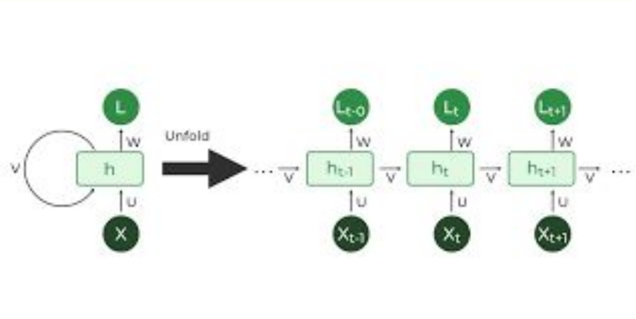
- Basic Python 3.x
- Object Oriented Programming
- Numpy
- SkLearn
- PyTorch & Keras

Overview of Transformer Models: Definition

What is a Transformer Model: Neural Network with extensive ability to handle long range dependencies in sequential data & their capacity for parallelization.



LSTM



RNN

Parallelization

Overview of Transformer Models: Industry Use Cases

GPT Algorithms

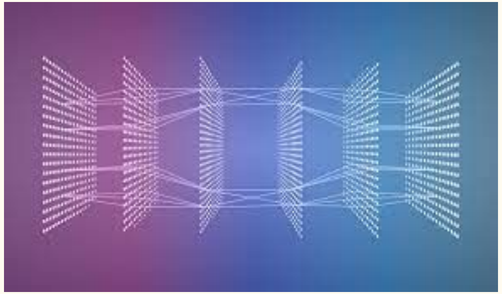
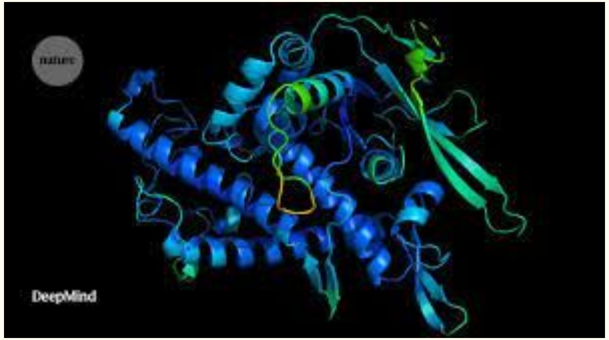


Image Transformers (VIT)

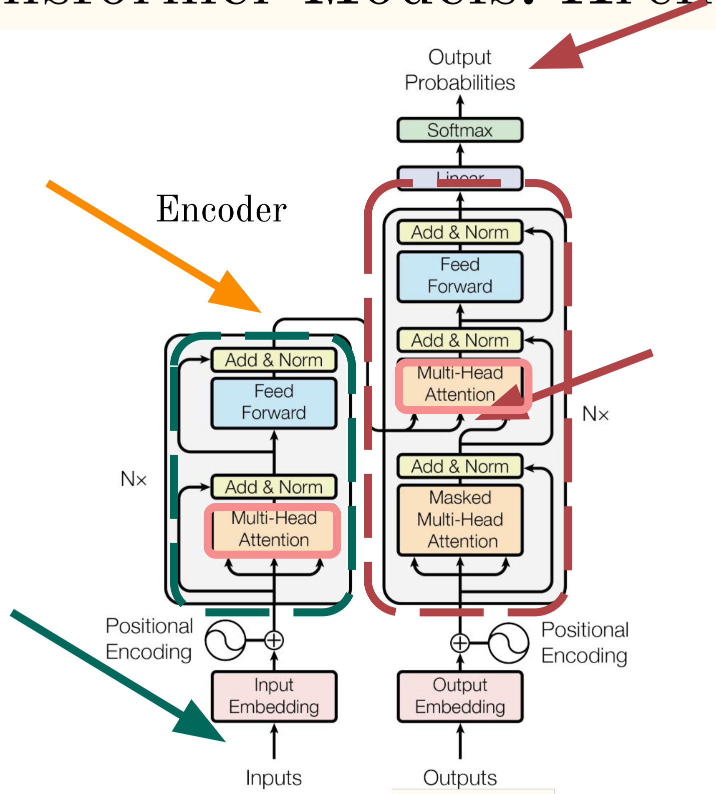


Protein Folding

Overview of Transformer Models: Architecture

BERT

Encoder



Encoder

Decoder

GPT

Decoder

What is Attention?

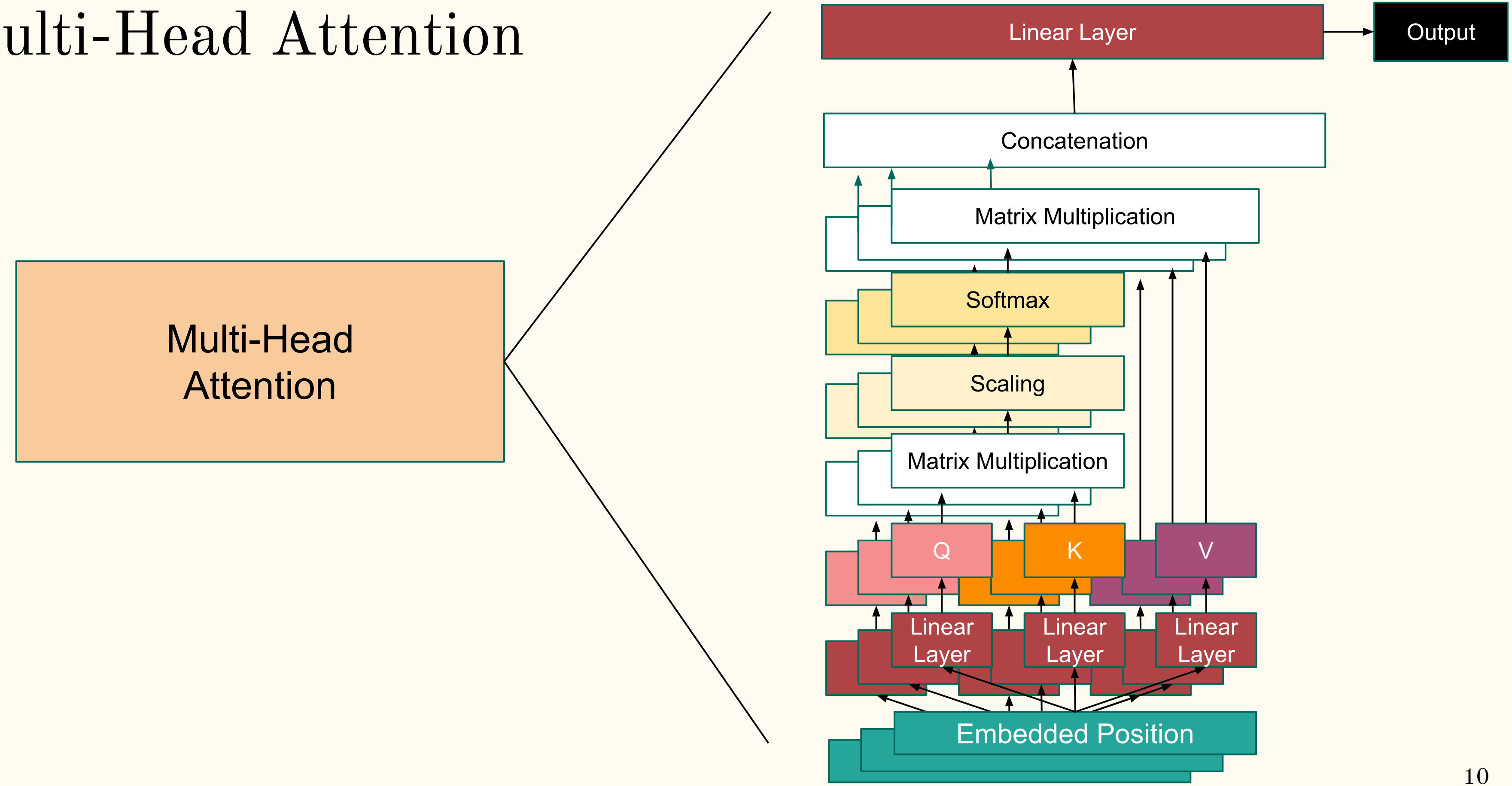
*“Winter is coming, we know what’s coming with it. We can learn to live with the **wildlings** or we can add them to the army of the dead”*
~Jon Snow

What is Self-Attention

What is the meaning of the following sentence:

"The man who passes the sentence should swing the sword" ~ Ned Stark

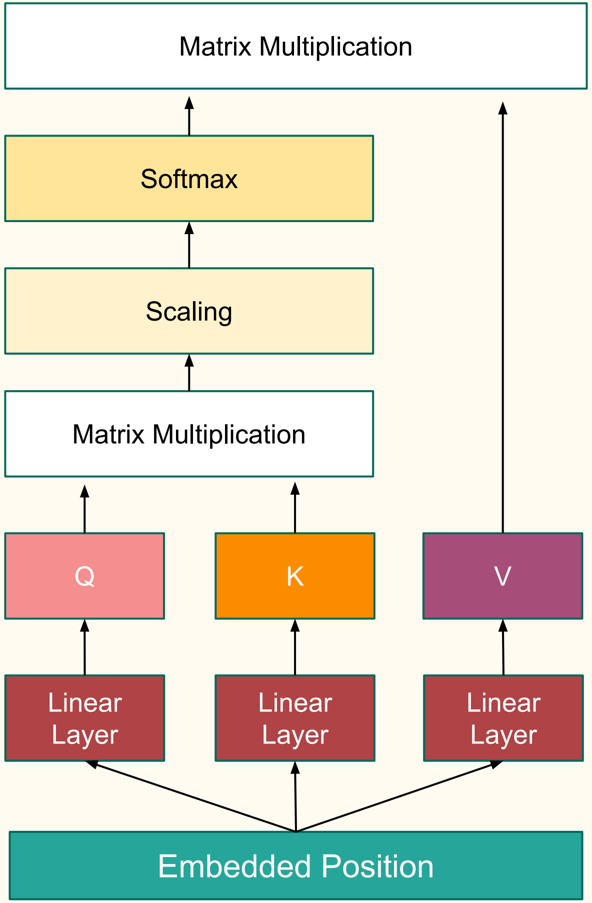
Multi-Head Attention



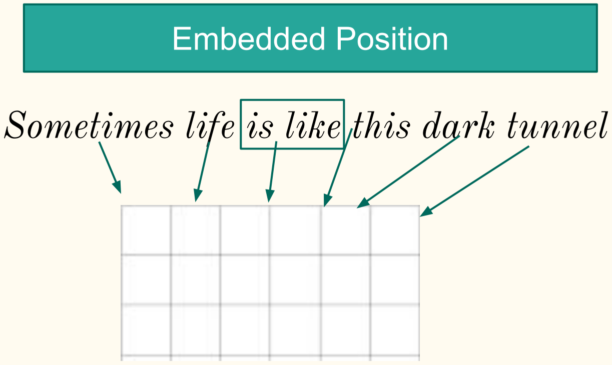
Single Attention Mechanism

Sections

1. Embedded Position Layer
2. Linear Layers
3. Query, Key, Value Matrices
4. Query x Key - MatMul
5. Scaler
6. Softmax - Normalization



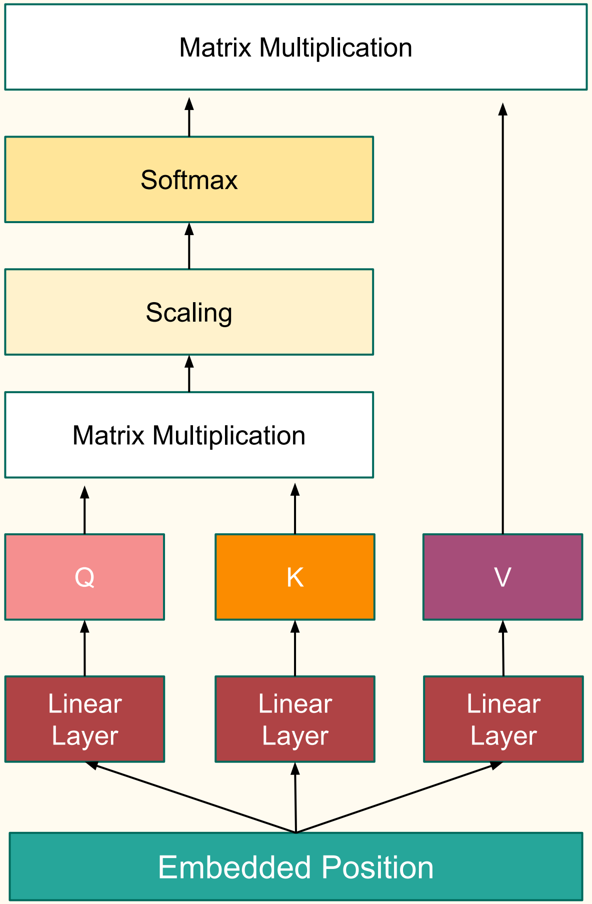
Single Attention Mechanism



Embedded Position

Sometimes life is like this dark tunnel

6x3

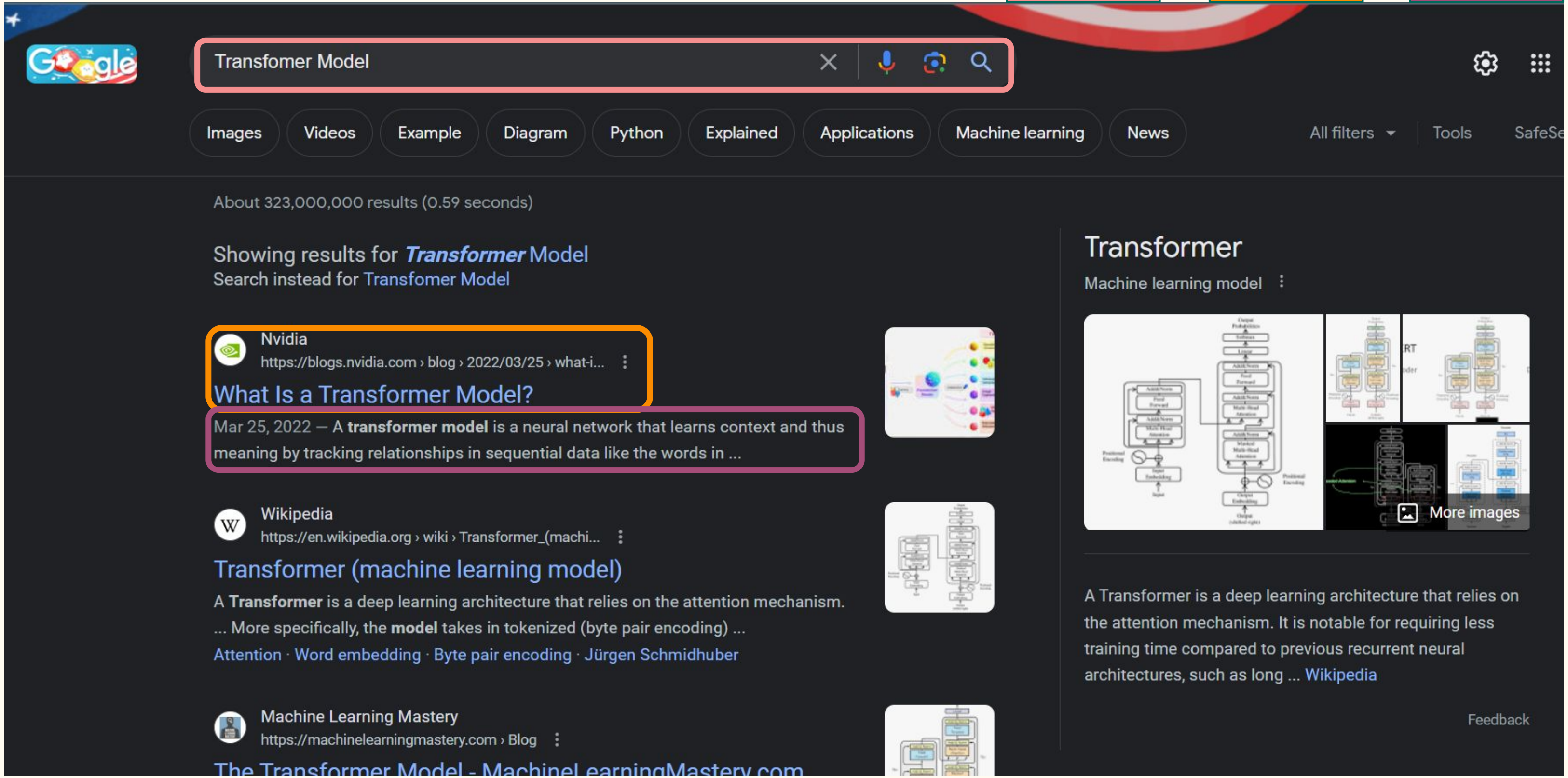


What is Query, Key, Value

Q

K

V



Transfomer Model

About 323,000,000 results (0.59 seconds)

Showing results for *Transformer* Model
Search instead for Transfomer Model

Nvidia
https://blogs.nvidia.com › blog › 2022/03/25 › what-i...

What Is a Transformer Model?

Mar 25, 2022 — A **transformer model** is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in ...

Wikipedia
https://en.wikipedia.org › wiki › Transformer_(machi...

Transformer (machine learning model)

A **Transformer** is a deep learning architecture that relies on the attention mechanism. ... More specifically, the **model** takes in tokenized (byte pair encoding) ...
Attention · Word embedding · Byte pair encoding · Jürgen Schmidhuber

Machine Learning Mastery
https://machinelearningmastery.com › Blog

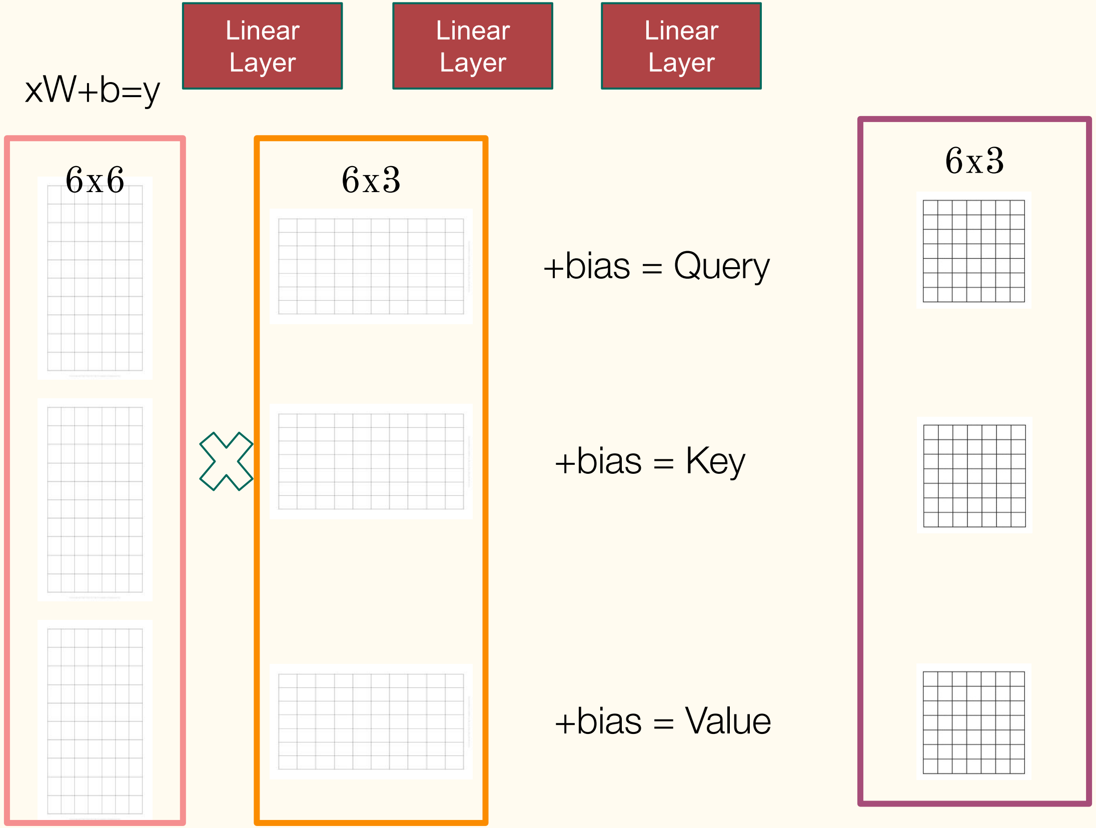
The Transformer Model - MachineLearningMastery.com

Transformer

Machine learning model

A Transformer is a deep learning architecture that relies on the attention mechanism. It is notable for requiring less training time compared to previous recurrent neural architectures, such as long ... Wikipedia

Single Attention Mechanism



xW+b=y

Linear Layer

Linear Layer

Linear Layer

6x6

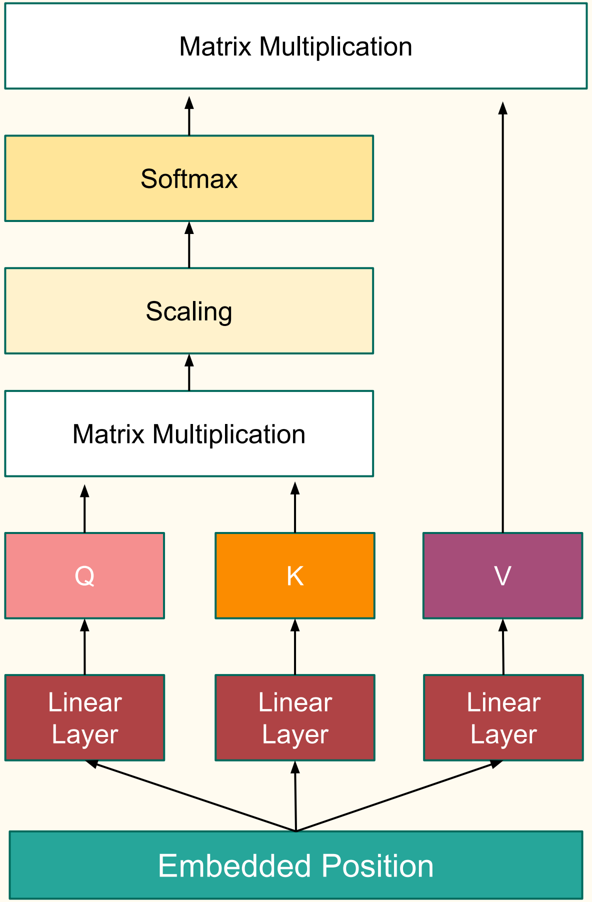
6x3

+bias = Query

+bias = Key

+bias = Value

6x3



Matrix Multiplication

Softmax

Scaling

Matrix Multiplication

Q

K

V

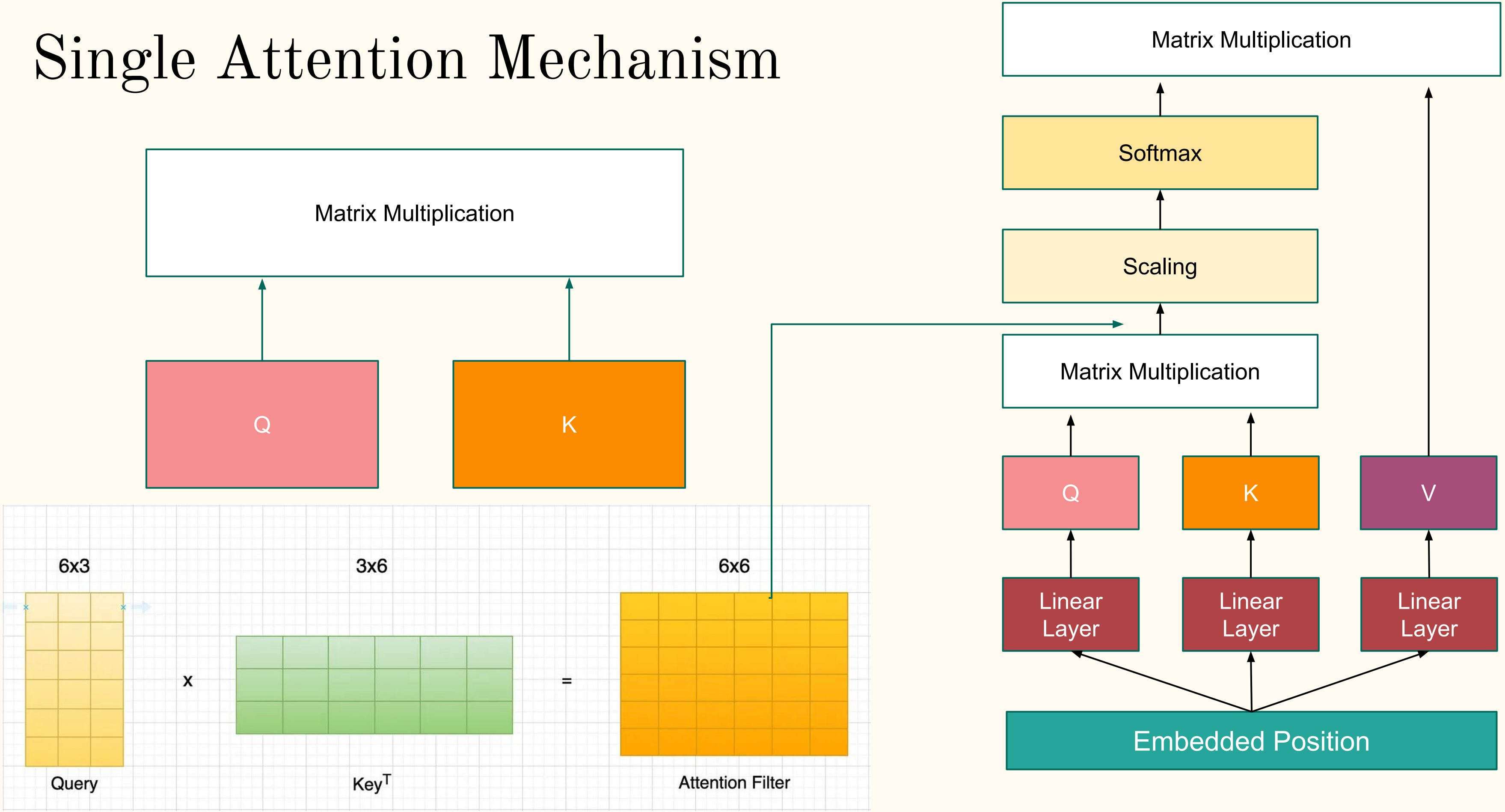
Linear Layer

Linear Layer

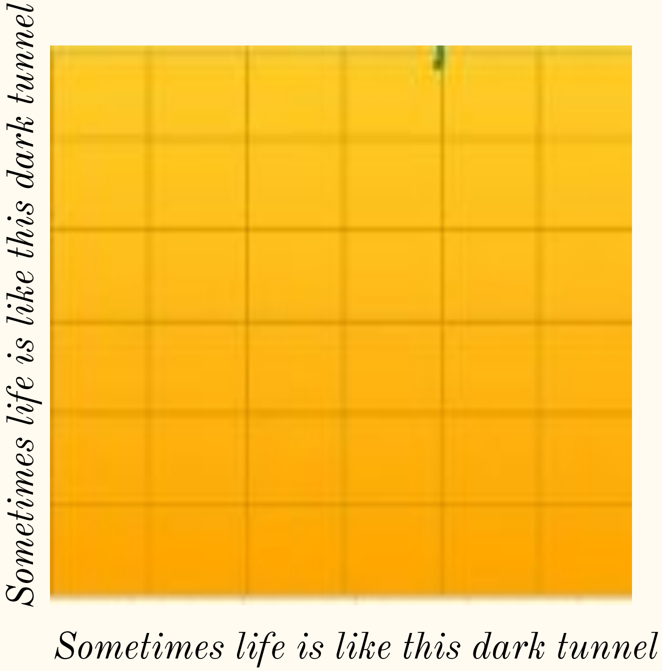
Linear Layer

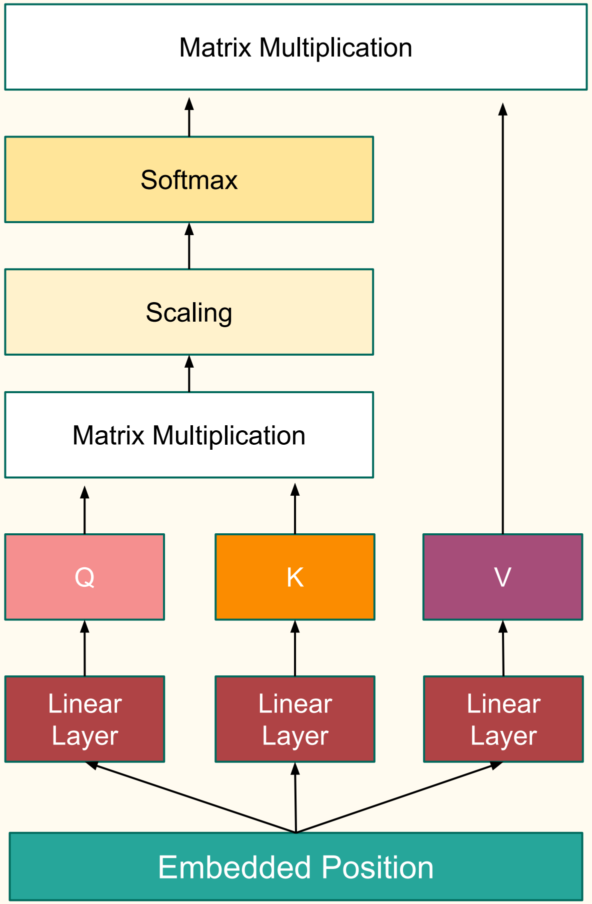
Embedded Position

Single Attention Mechanism

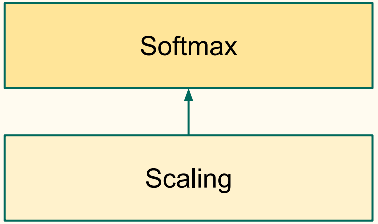


Single Attention Mechanism -Attention Filter





Single Attention Mechanism



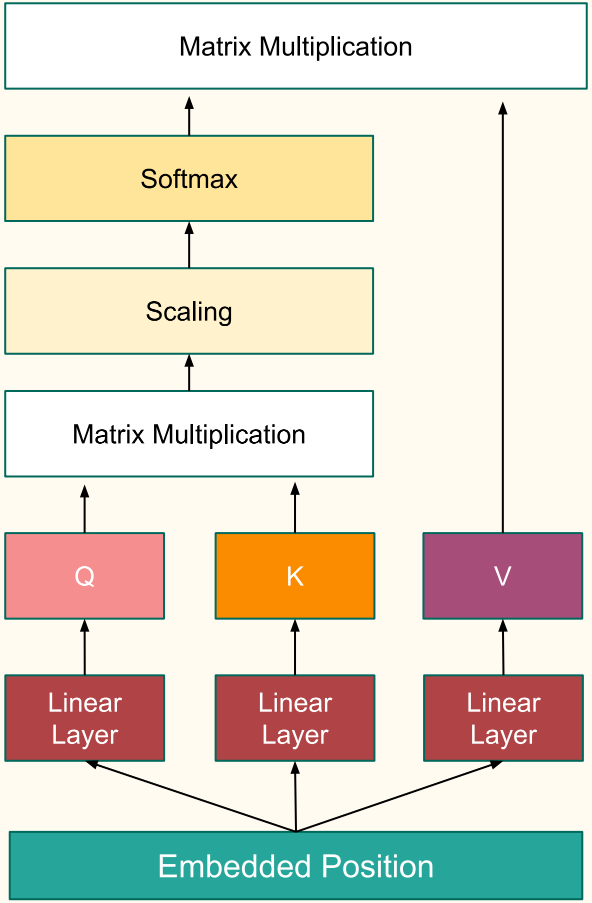
Scaling

✖ 1 / $6^{1/2}$

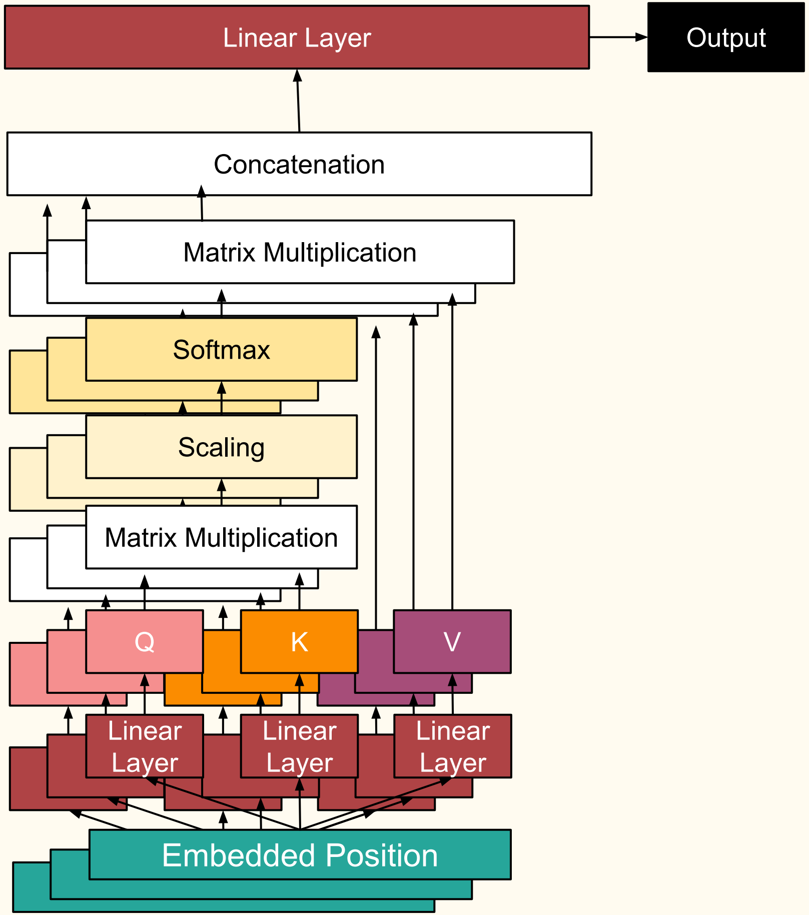
6 = Key Vector Length

Softmax

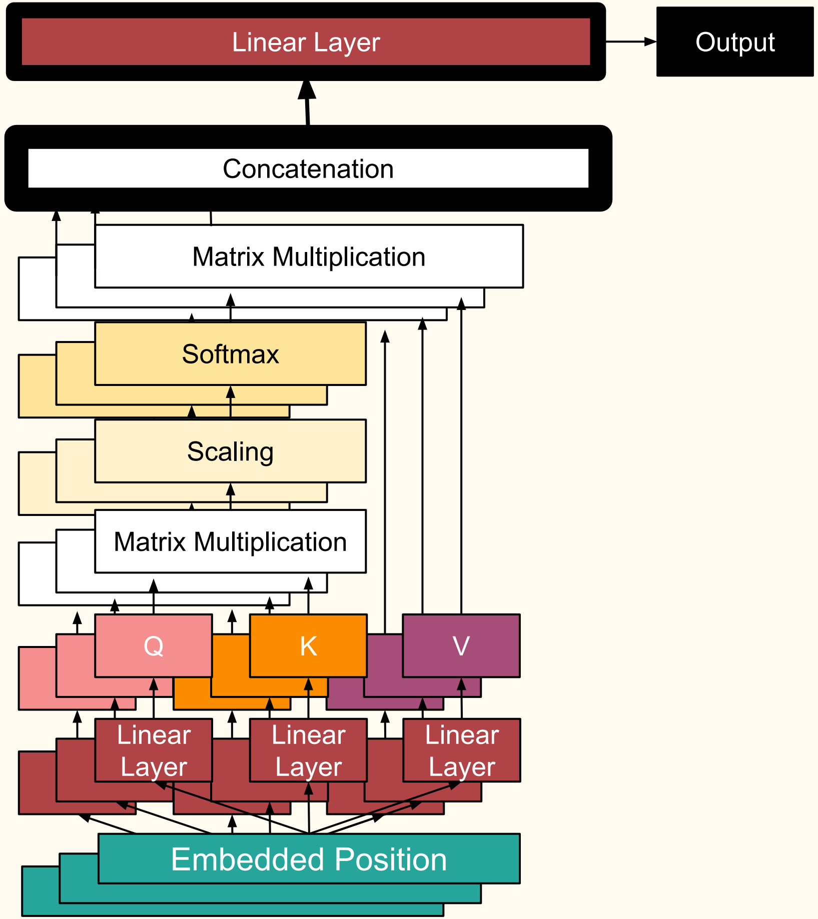
Helps Normalize Values between 0→1



Multi-Head - Self-Attention Filters



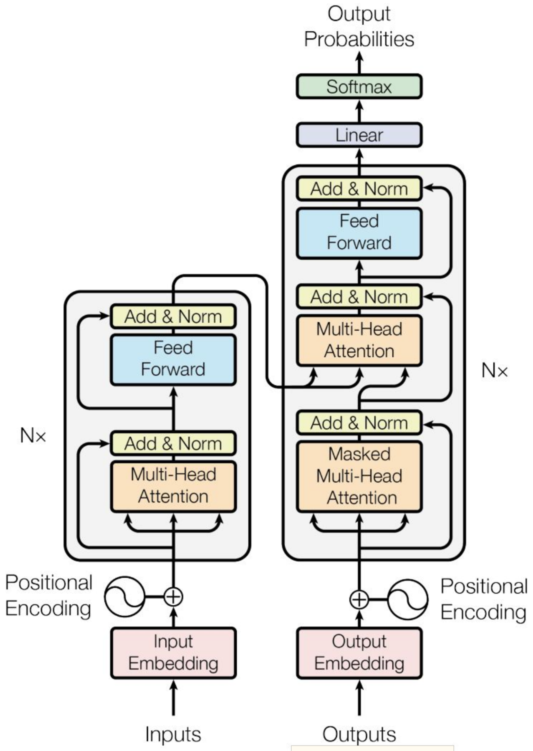
Multi-Head Attention



What’s Next..

BERT

Encoder



GPT

Decoder

What We Covered Today

- ➢ Overview of Transformer Model
- ➢ What is Attention
- ➢ Attention Mechanisms
- ➢ Multiple Head Attention