Friedrich-Alexander-Universität
Technische Fakultät

Machine Learning
Data Analytics

FAU

Transformer

Richard Dirauf, M.Sc.
Machine Learning and Data Analytics (MaD) Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
MLTS Exercise, 30.01.2025

MLTS Exercise – Organization

~~Introduction (31.10.2024)~~

~~Bayesian Linear Regression (07.11.2024)~~

~~Bayesian Linear Regression (14.11.2024)~~

~~Kalman Filter (21.11.2024)~~

~~Kalman Filter (28.11.2024)~~

~~Dynamic Time Warping (05.12.2024)~~

~~Dynamic Time Warping (12.12.2024)~~

~~No exercise planned (19.12.2024)~~

Holiday

~~RNN + LSTM (09.01.2025)~~

~~RNN + LSTM (16.01.2025)~~

~~RNN + LSTM (23.01.2025)~~

Transformers (30.01.2025)

Transformers (06.02.2025)

Recap: Recurrent Neural Networks

Machine Learning
Data Analytics

FAU

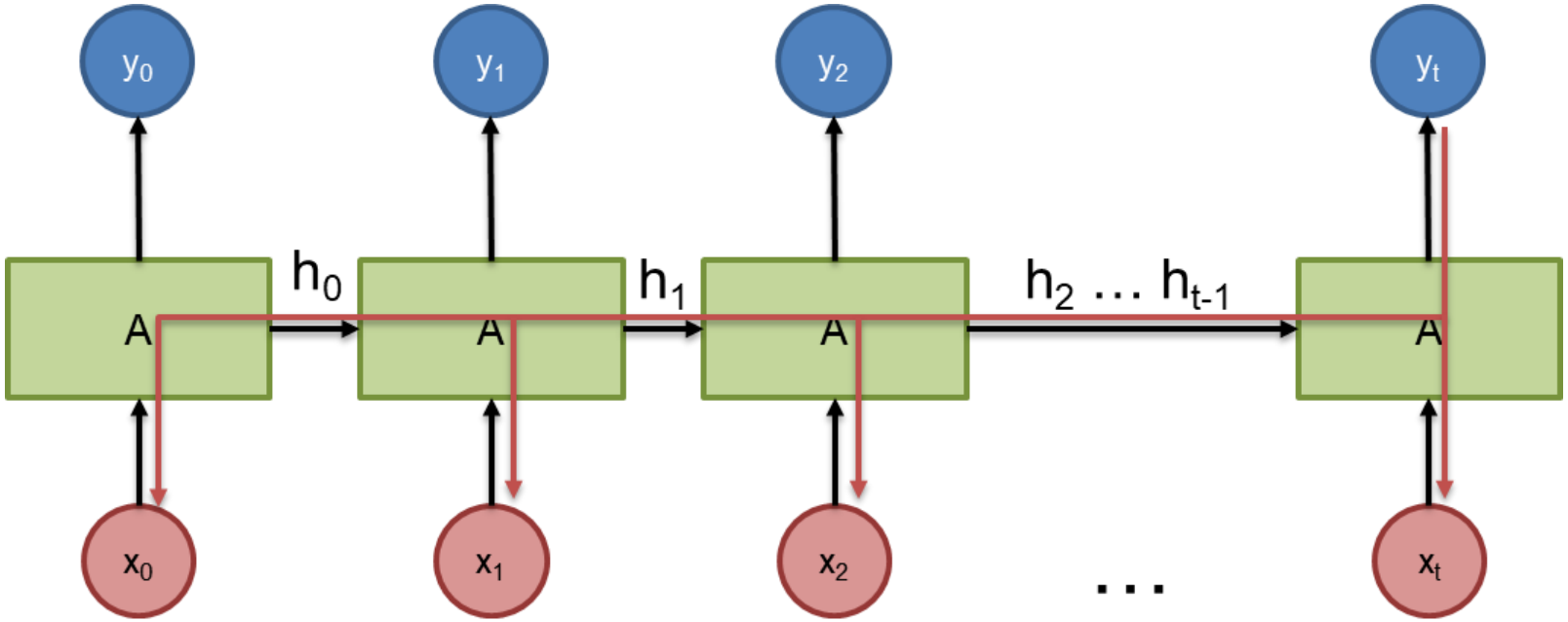
RNN/LSTM can model:

- Sequences of variable lengths
- Long term dependencies (using LSTMs)

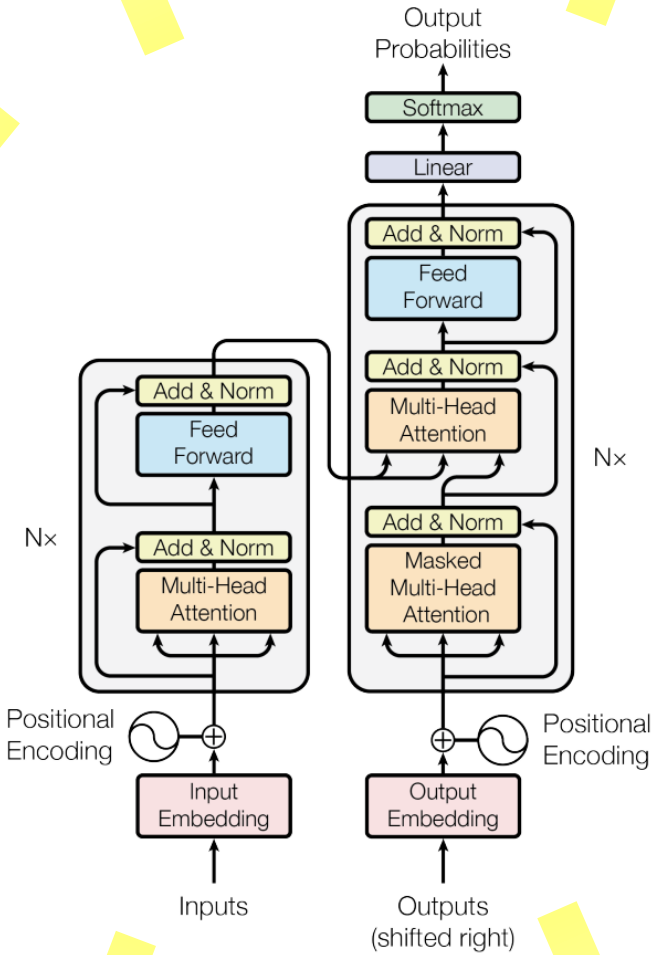
However:

- Non-parallelism -> Long training time
- Large memory usage
- Difficult to train (vanishing/exploding gradients)

Recap: Recurrent Neural Networks

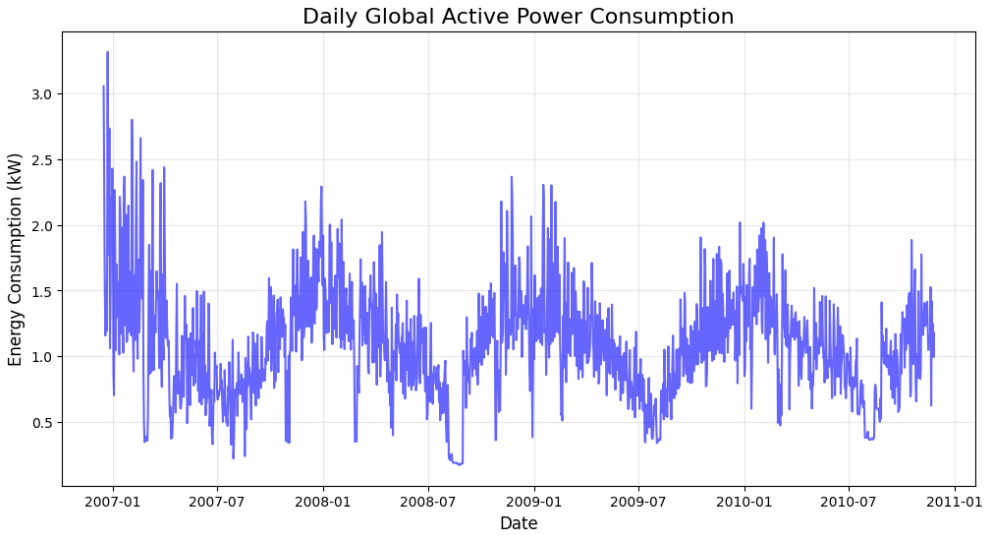


Transformer to the Rescue



Input - Sequences

Time Series



Event Sequence

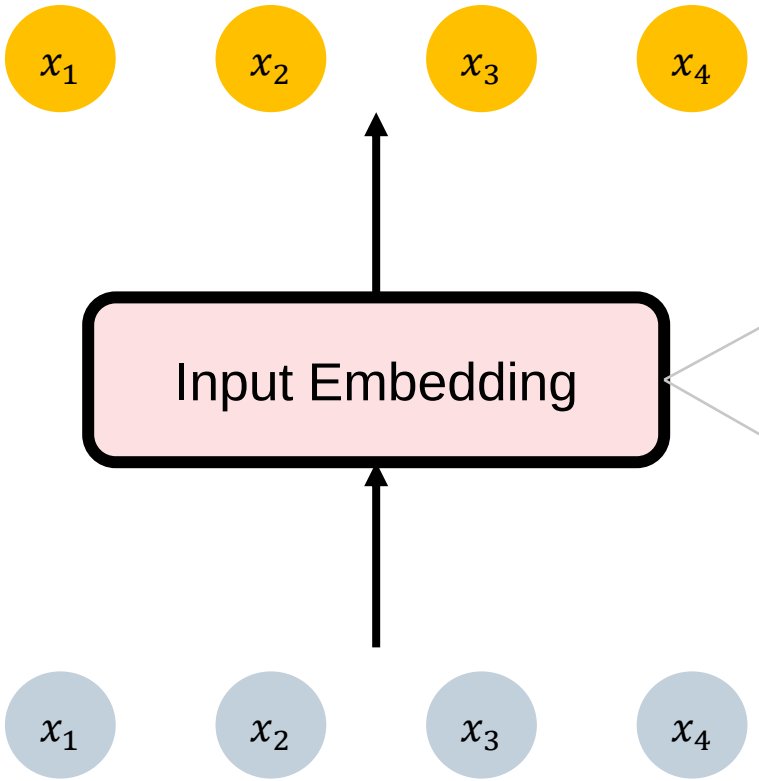
x_1 x_2 x_3 x_4

Text

Harry Potter invented a new spell

Video

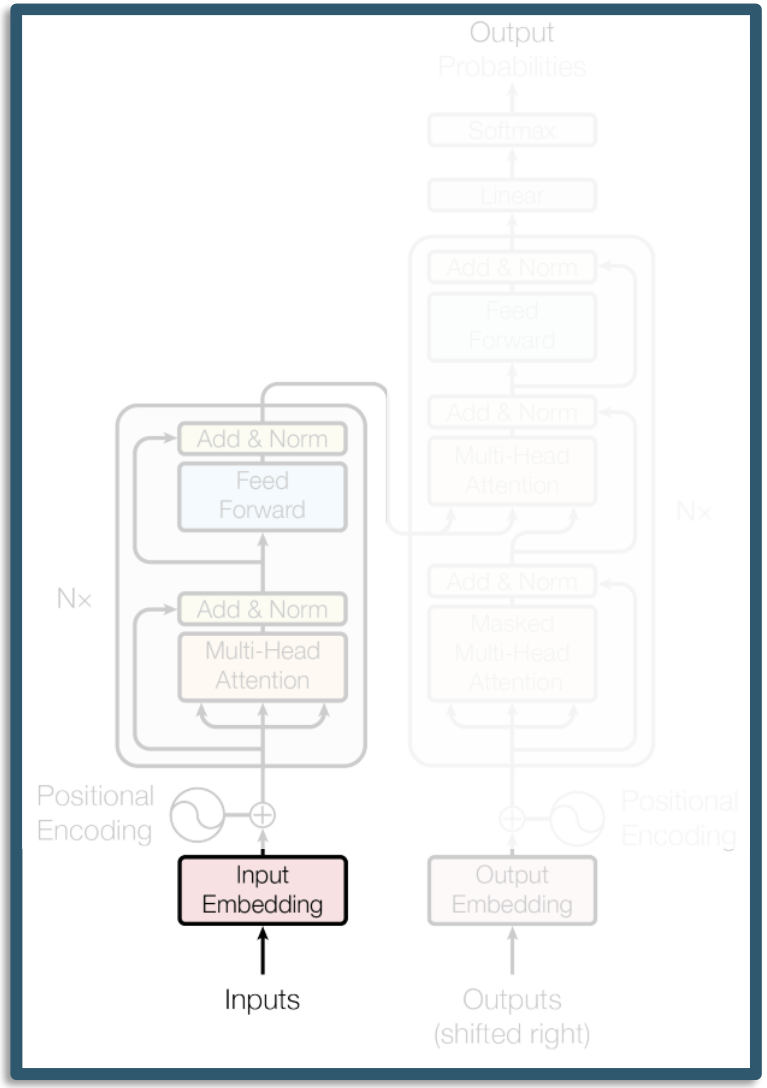
Transformer – Input Embedding



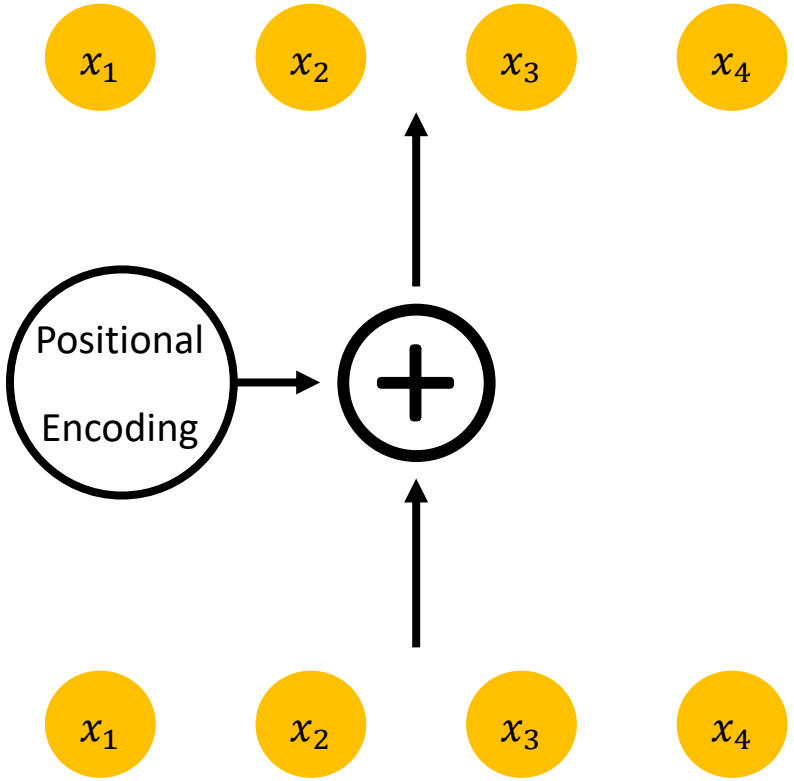
Latent Space Representation

- Feature Extractor
- Linear Layer
- CNN
- Transformer

Inputs

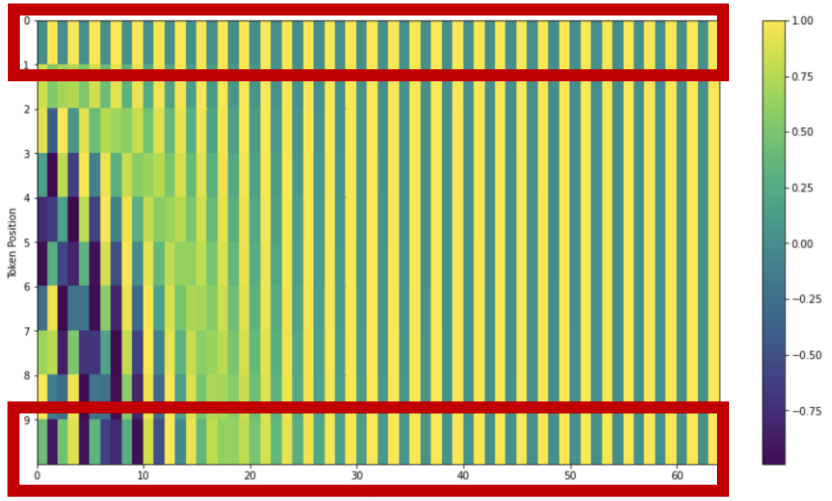


Transformer – Positional Encoding

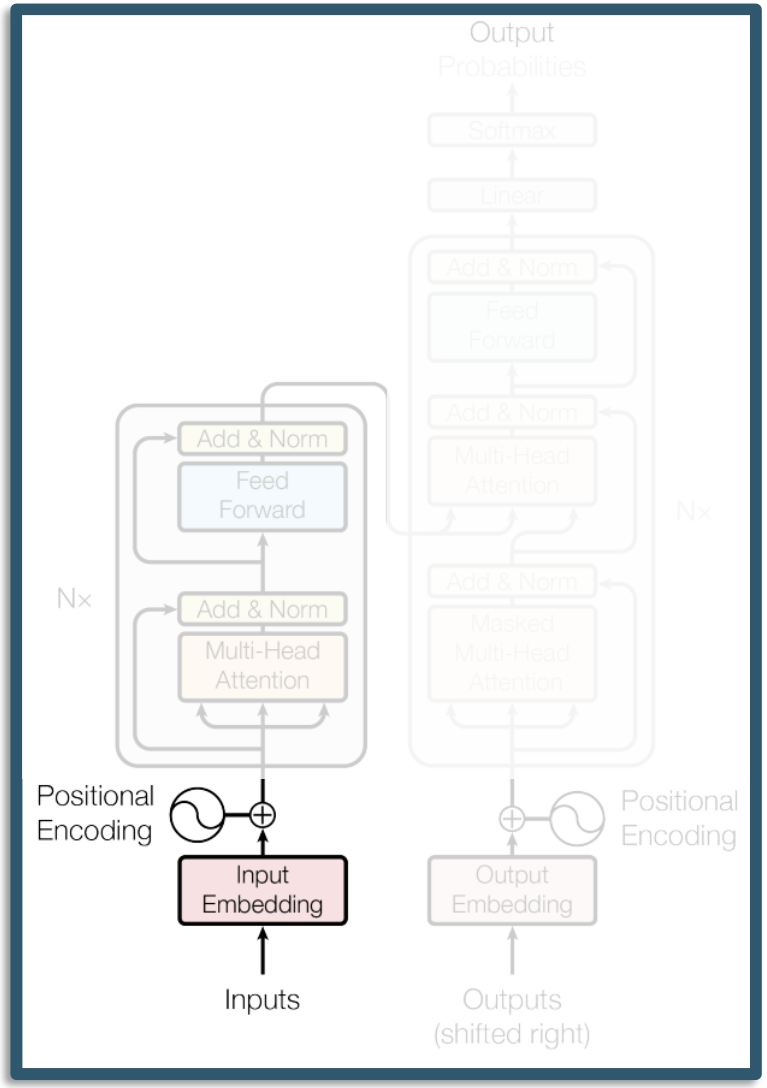


Latent Space Representation

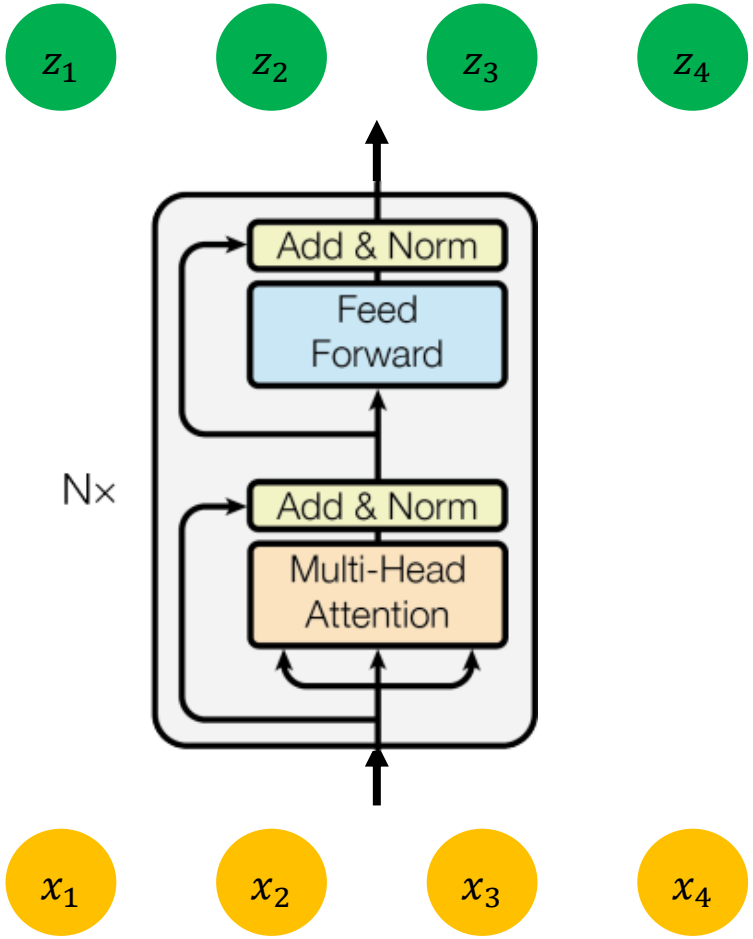
PE of the 1st input



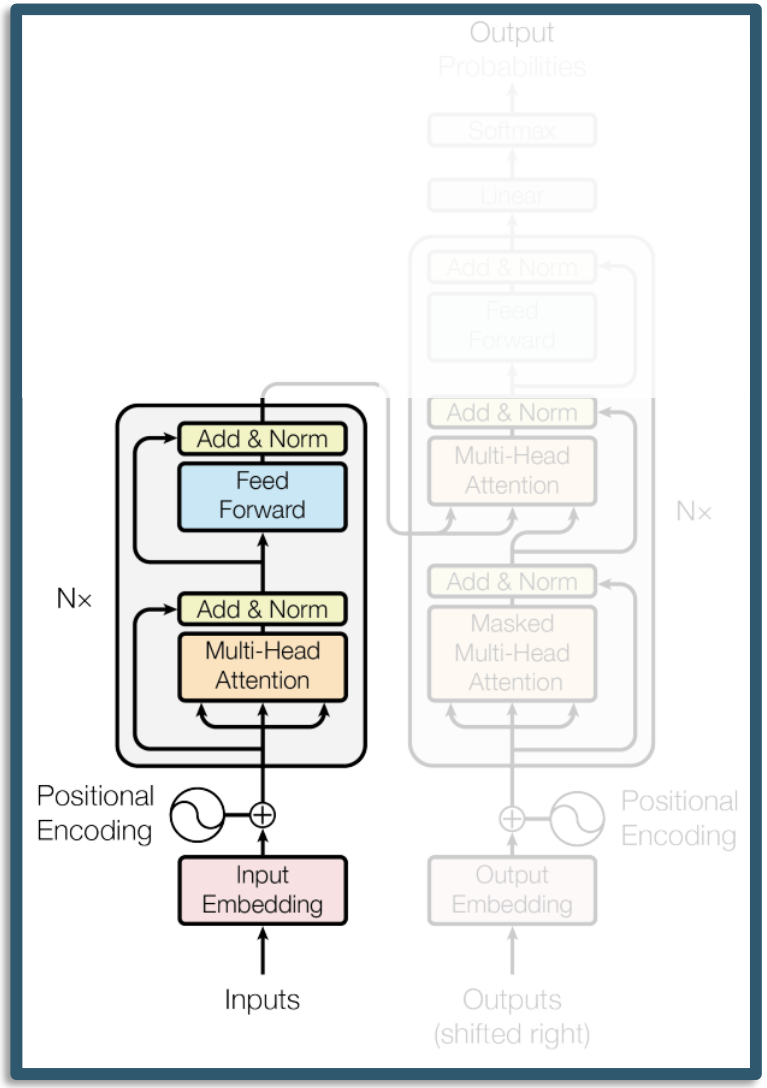
PE of the 10th input



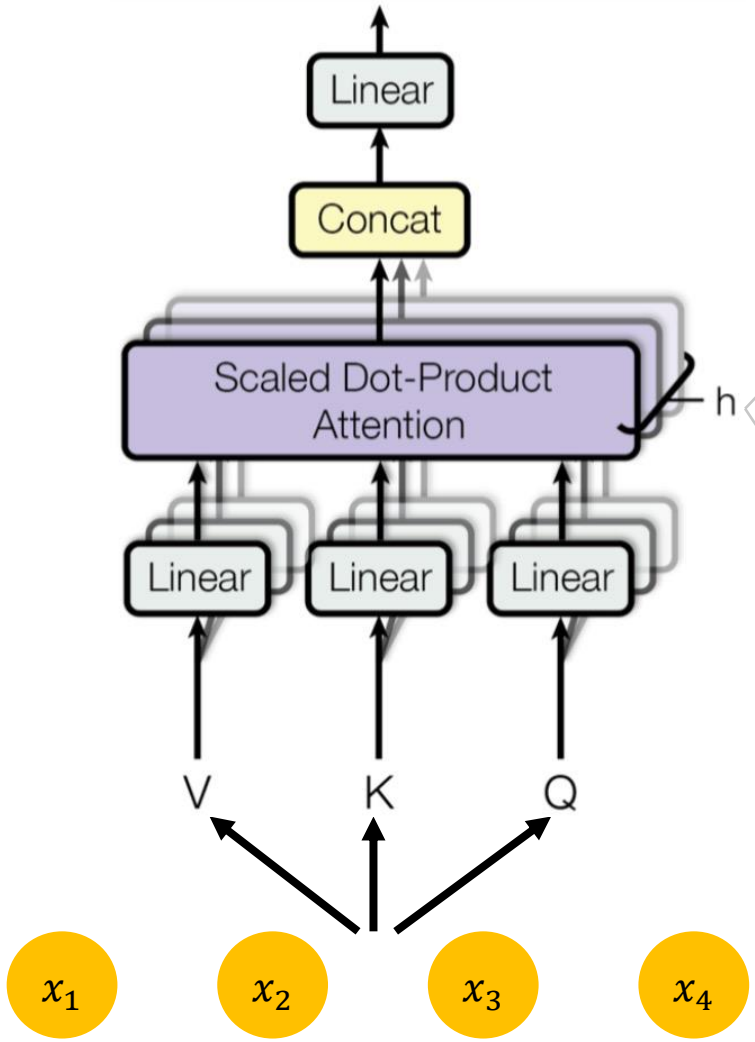
Transformer – Encoder



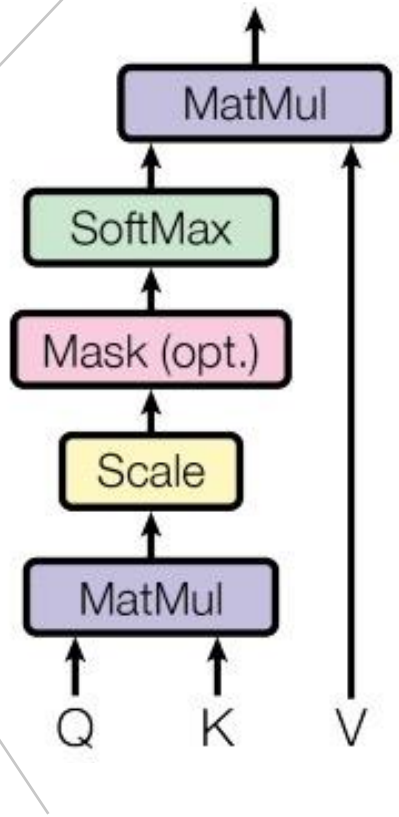
Latent Space Representation

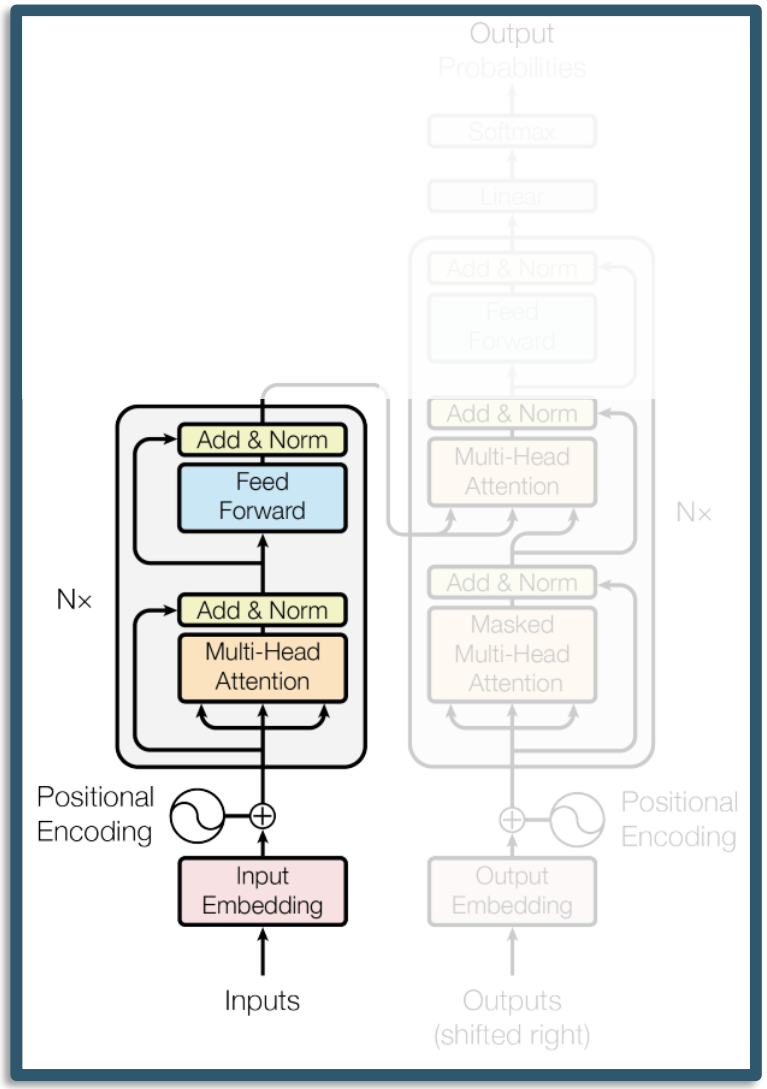


Transformer – Multi-Head Attention

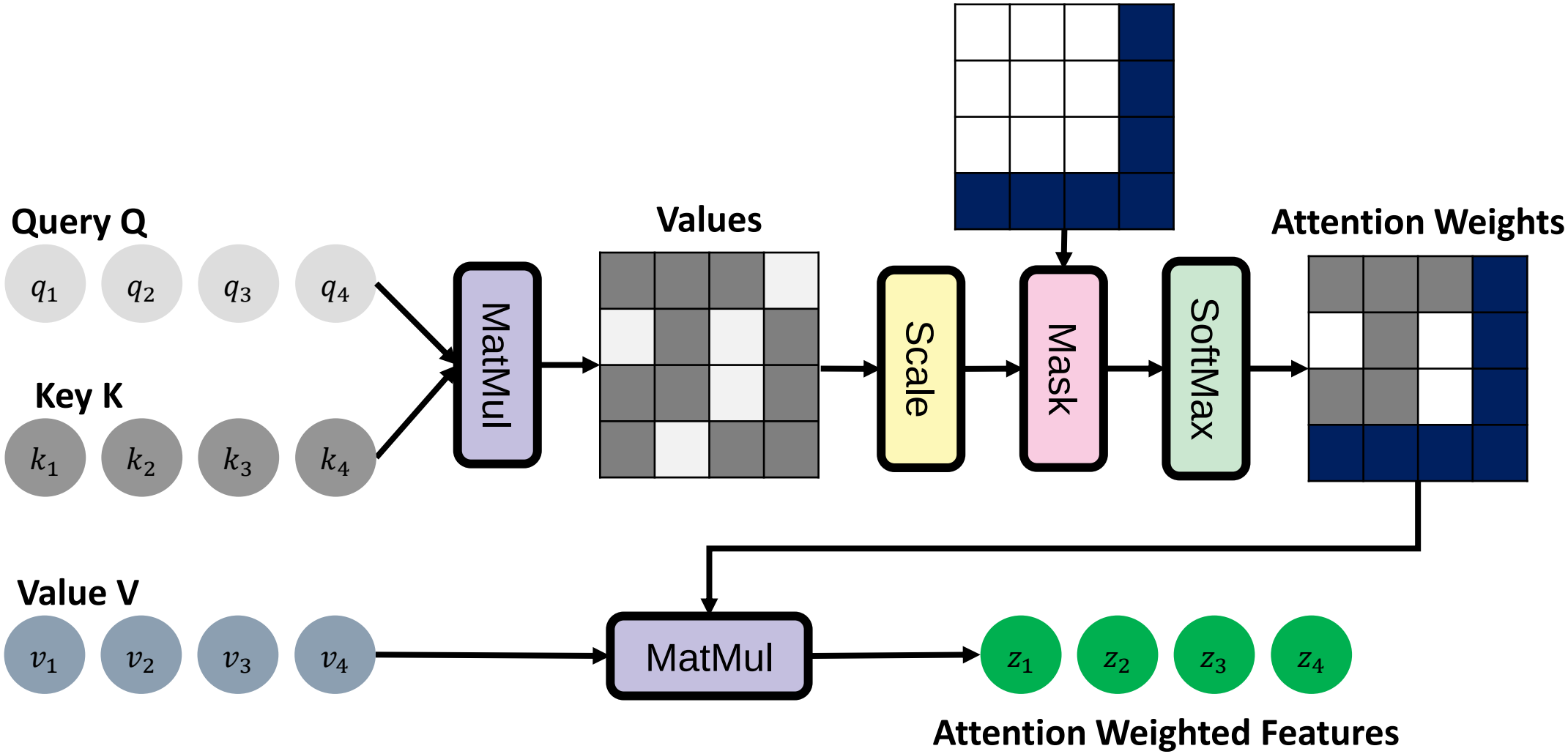


Scaled Dot Product Attention

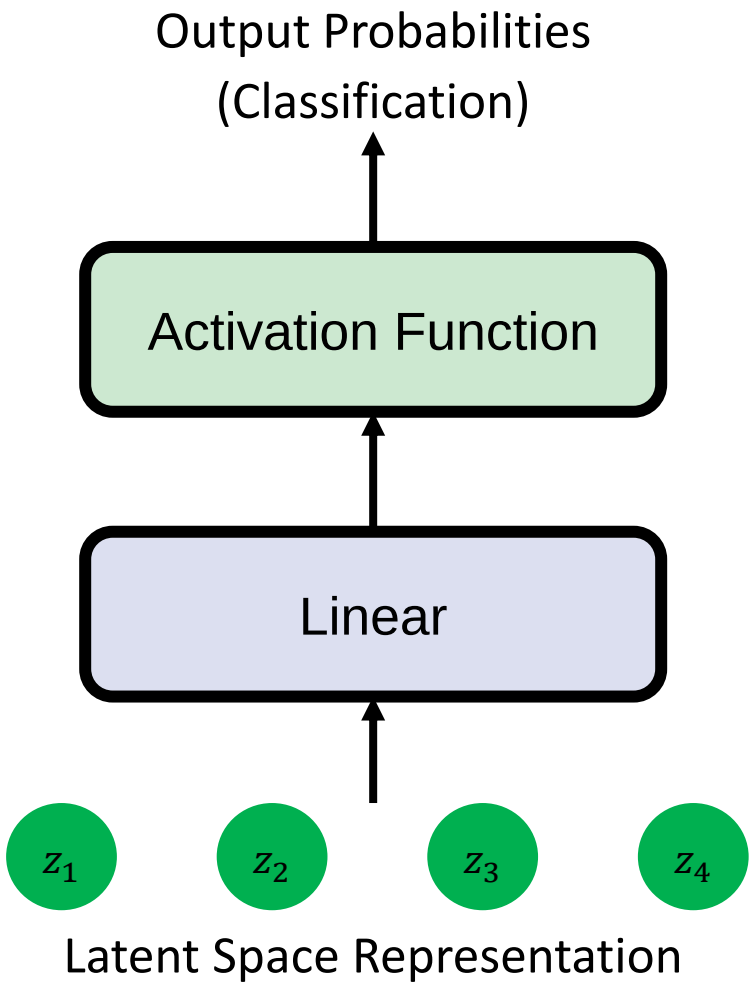




Transformer – Scaled Dot Product Attention



Transformer – Encoder-Only Model

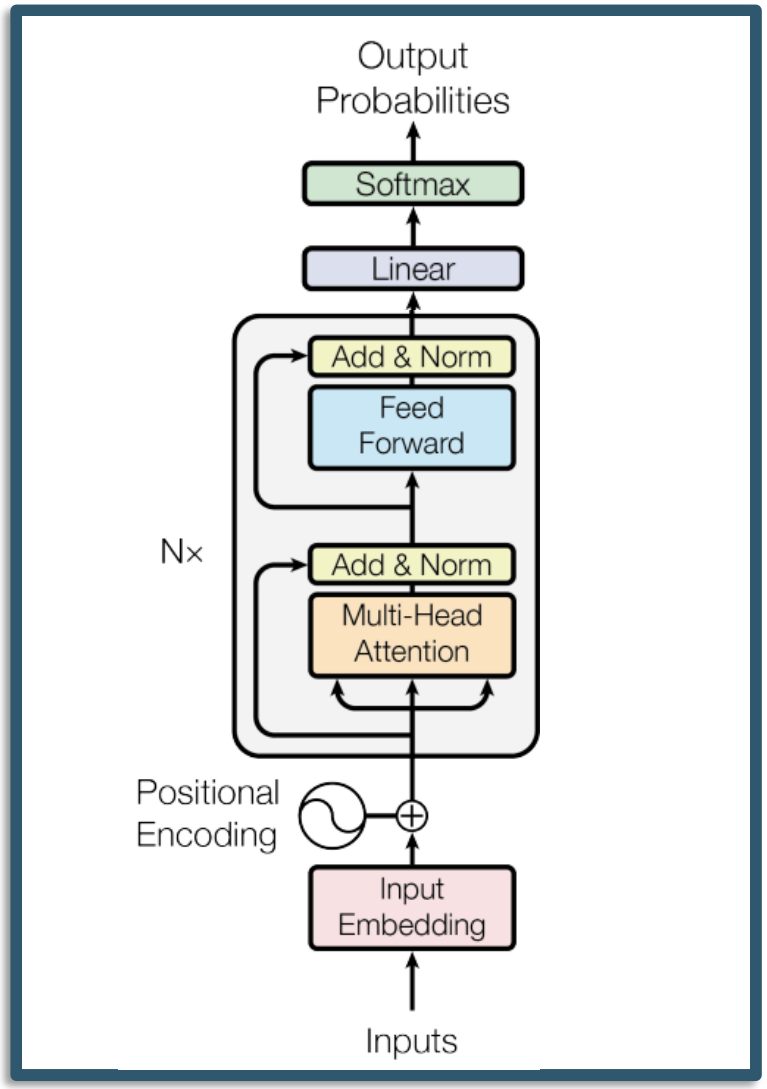


Task: understand / encode the input

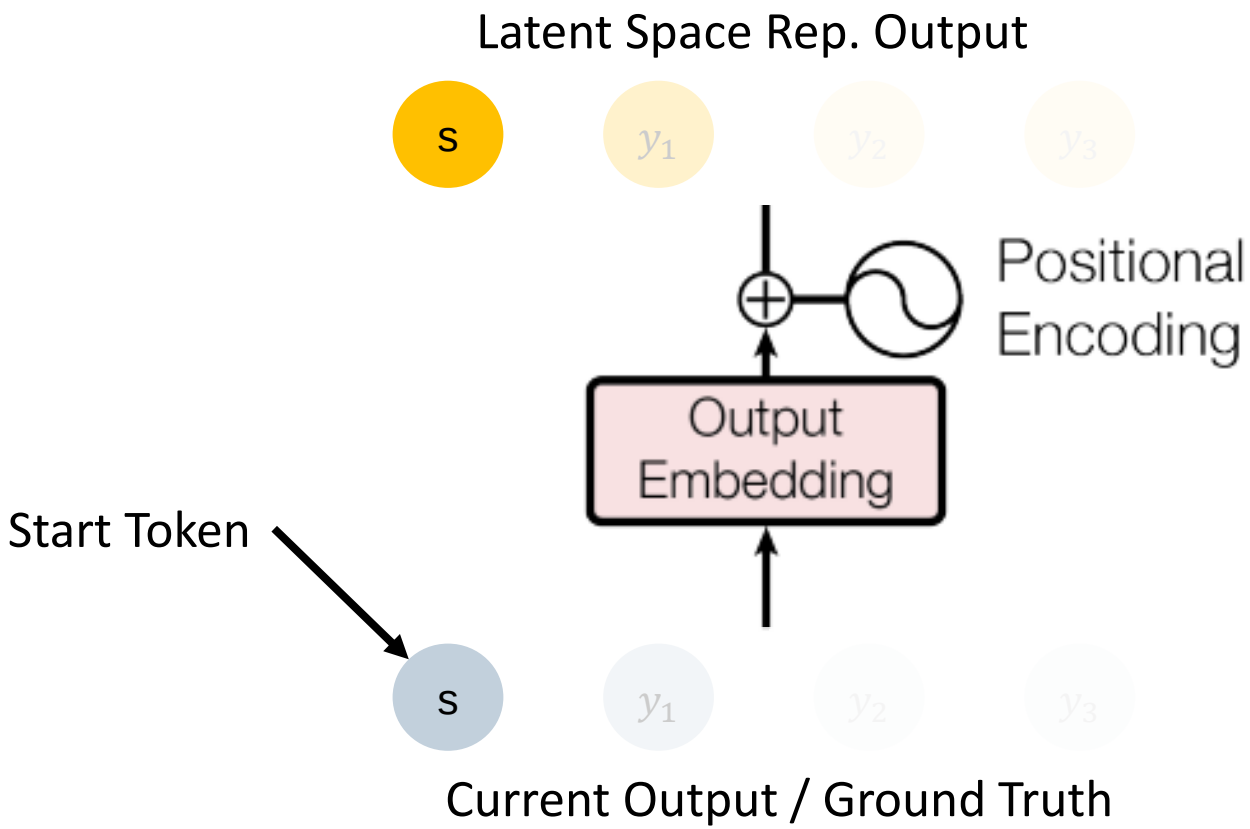
Fixed-length contextual representation for input token / entire sequence

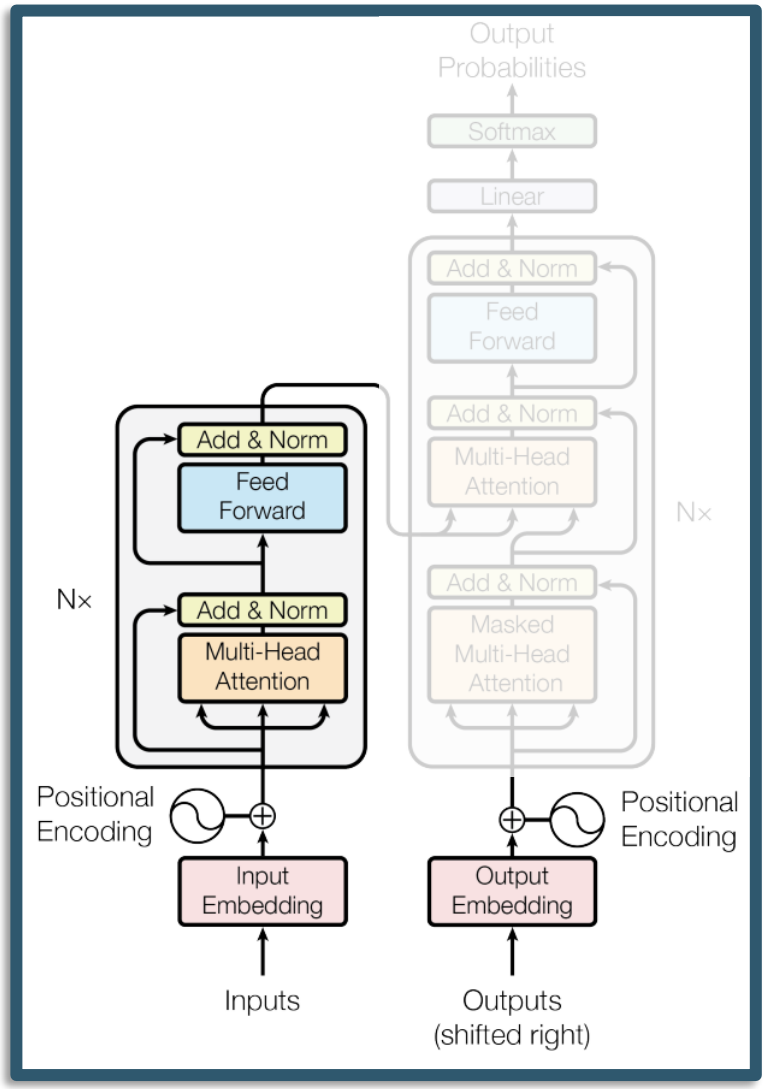
Examples:

- Classification
- Similarity
- Anomaly Detection

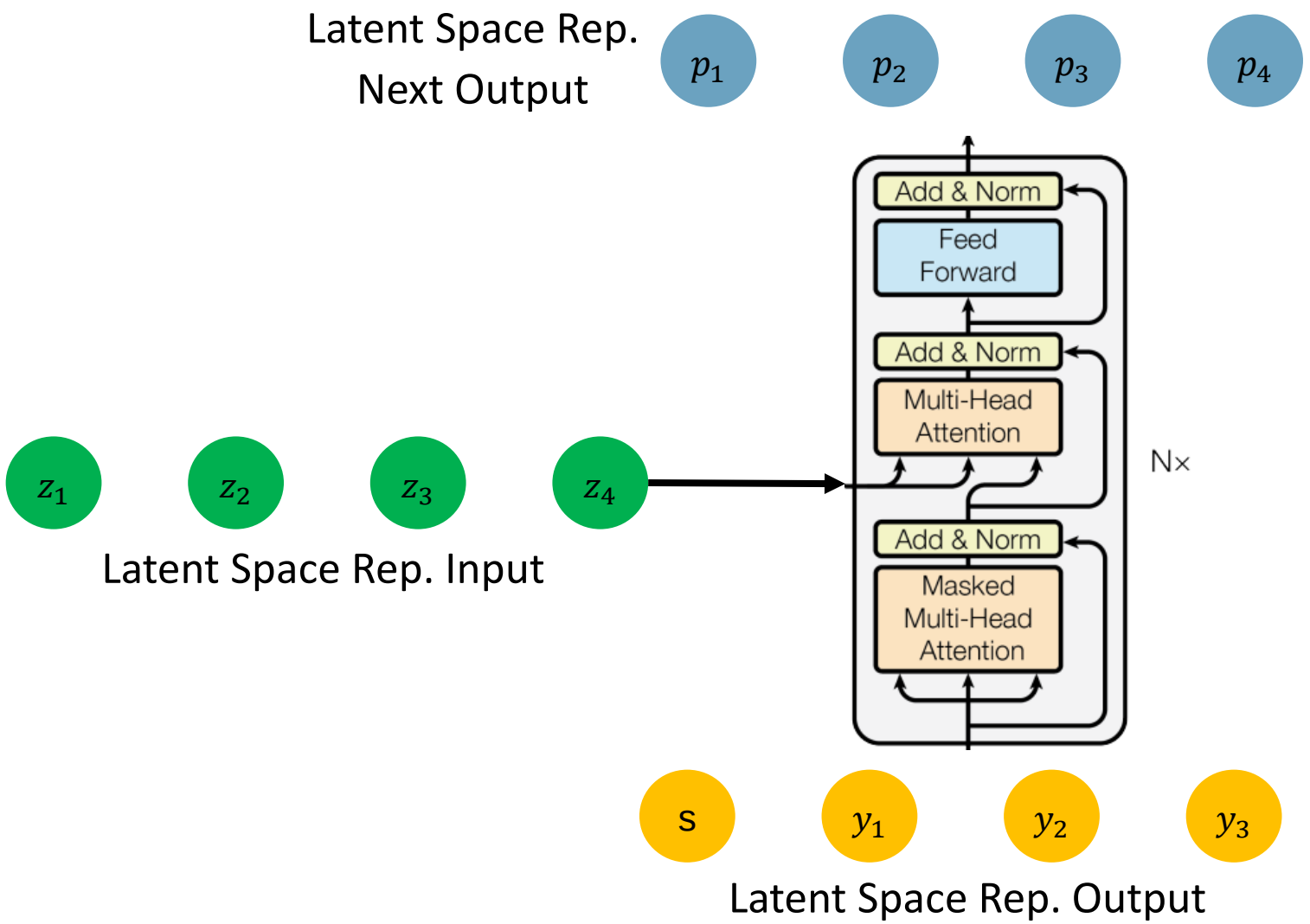


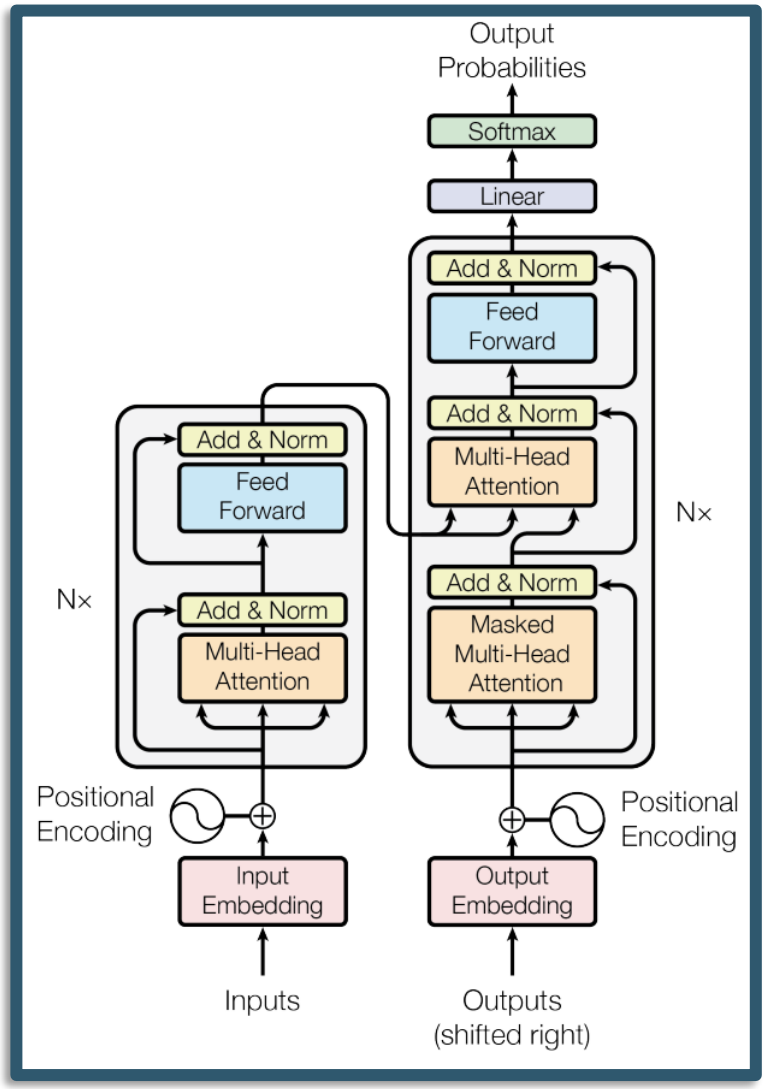
Transformer – Decoder



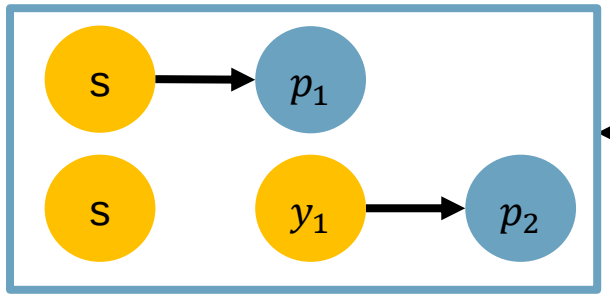


Transformer – Decoder

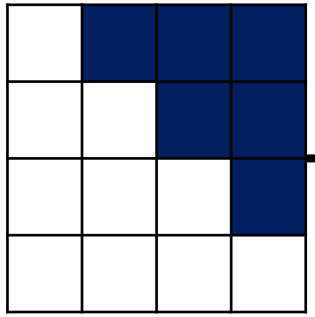


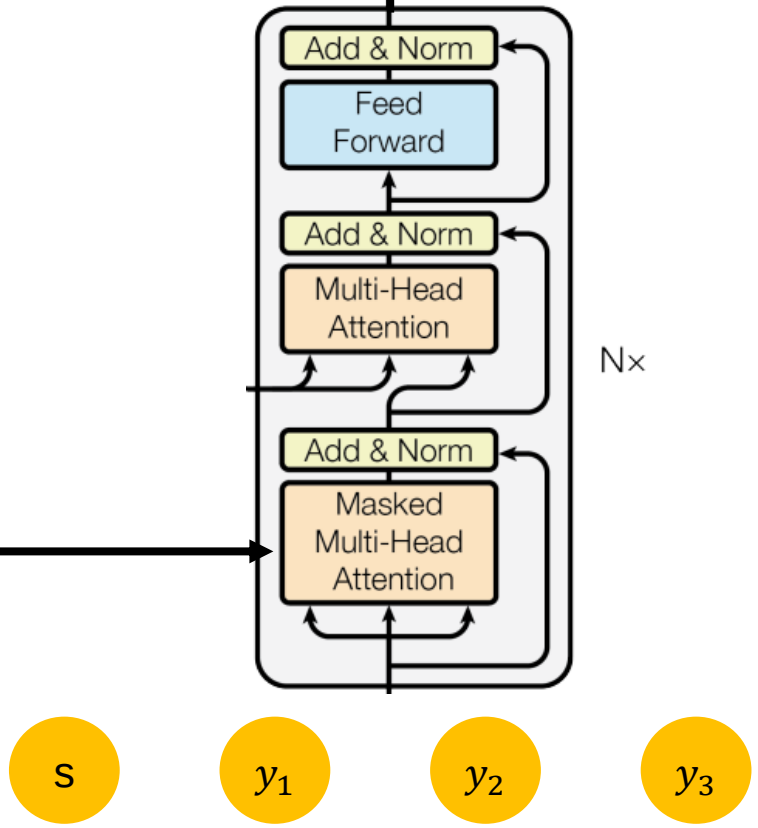


Transformer – Decoder - Training

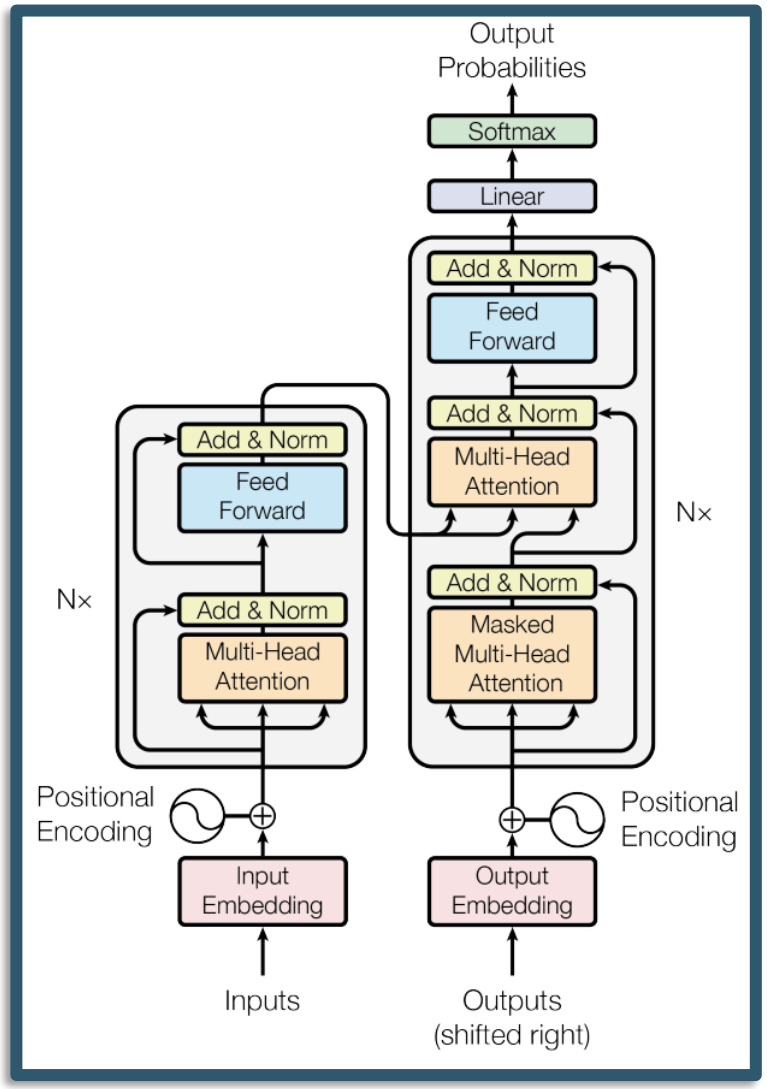


Look Ahead Mask

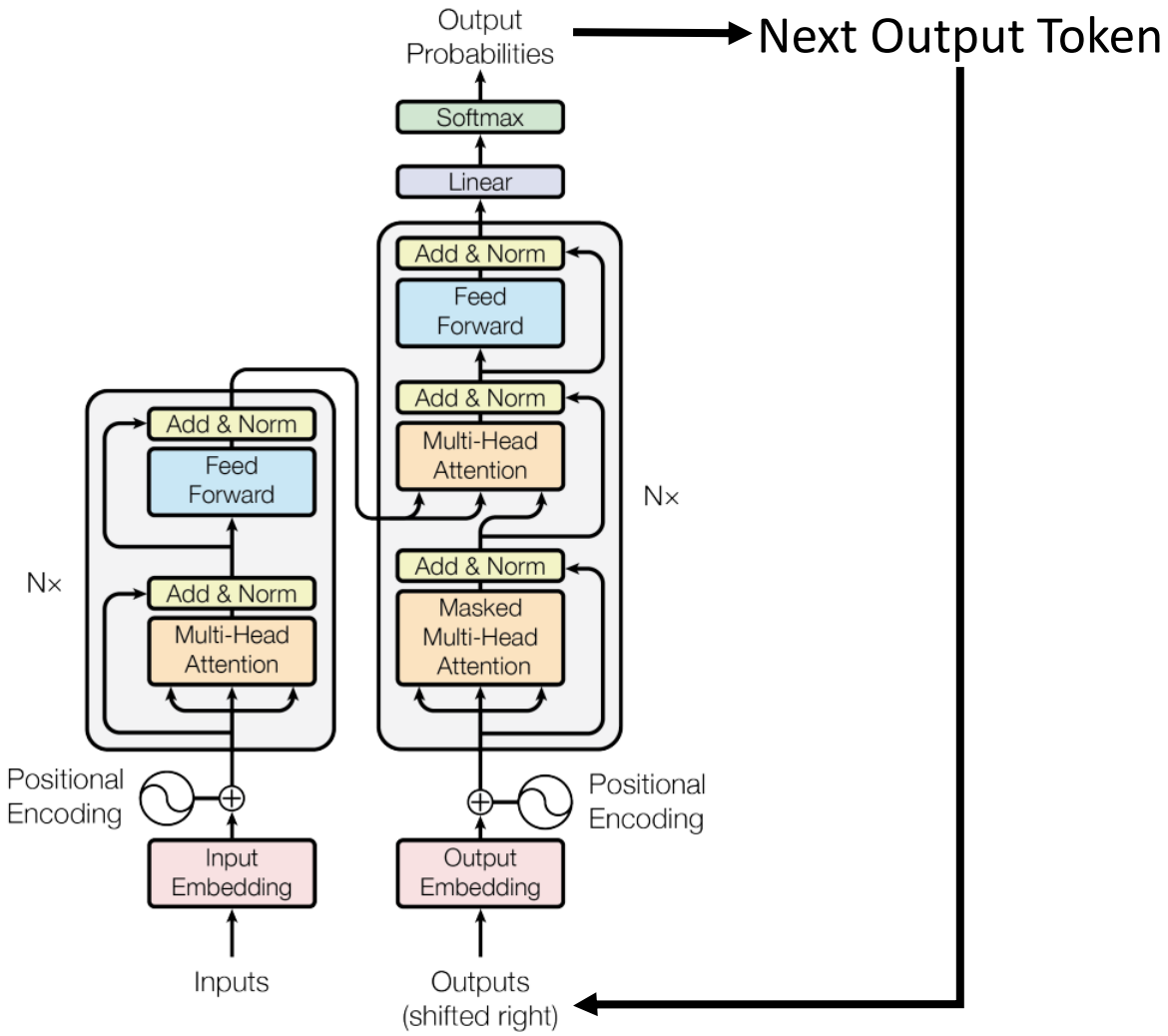




Latent Space Rep. Ground Truth



Transformer – Encoder-Decoder Model



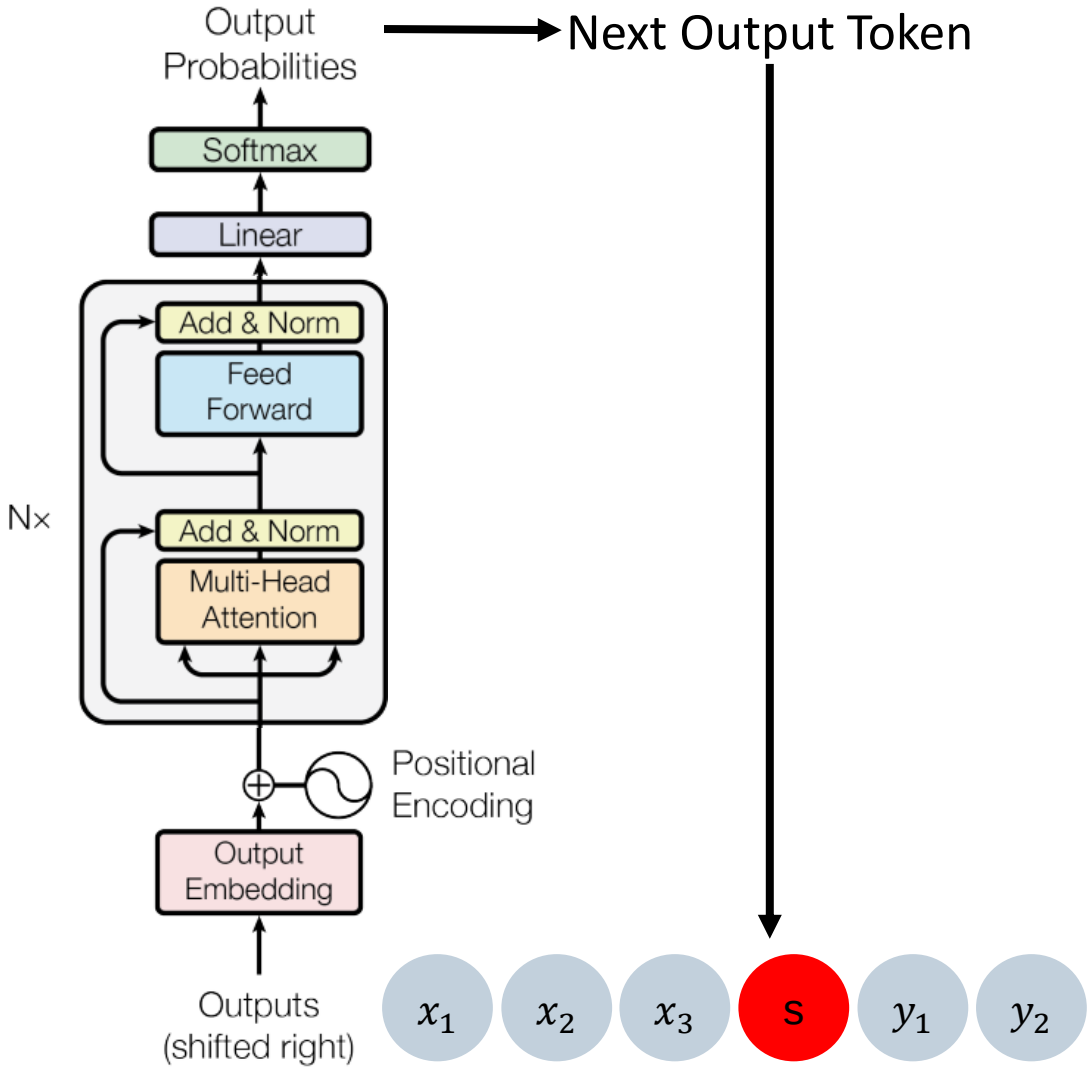
Task: understand / encode the input + generate sequence of output tokens

End of generation: fixed length / end token

Examples:

- Machine Translation
- Multivariate time series forecasting

Transformer – Decoder-Only Model



Task: generate sequence of output tokens

End of generation: fixed length / end token

Examples:

- Text generation
- Time series forecasting

Transformers: pros and cons

Pros:

- Long dependencies
- Multi-head attention can learn complex dependencies

Cons:

- Quadratic time and memory complexity
- Training is insidious
- (Still limited research on time series data.)

Any questions?