KI Service Zentrum
by Hasso-Plattner-Institut
HPI

Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt

Time Series Forecasting

3.4 Forecasting with Transformers

Mario Tormo Romero

Design IT.
Create Knowledge.

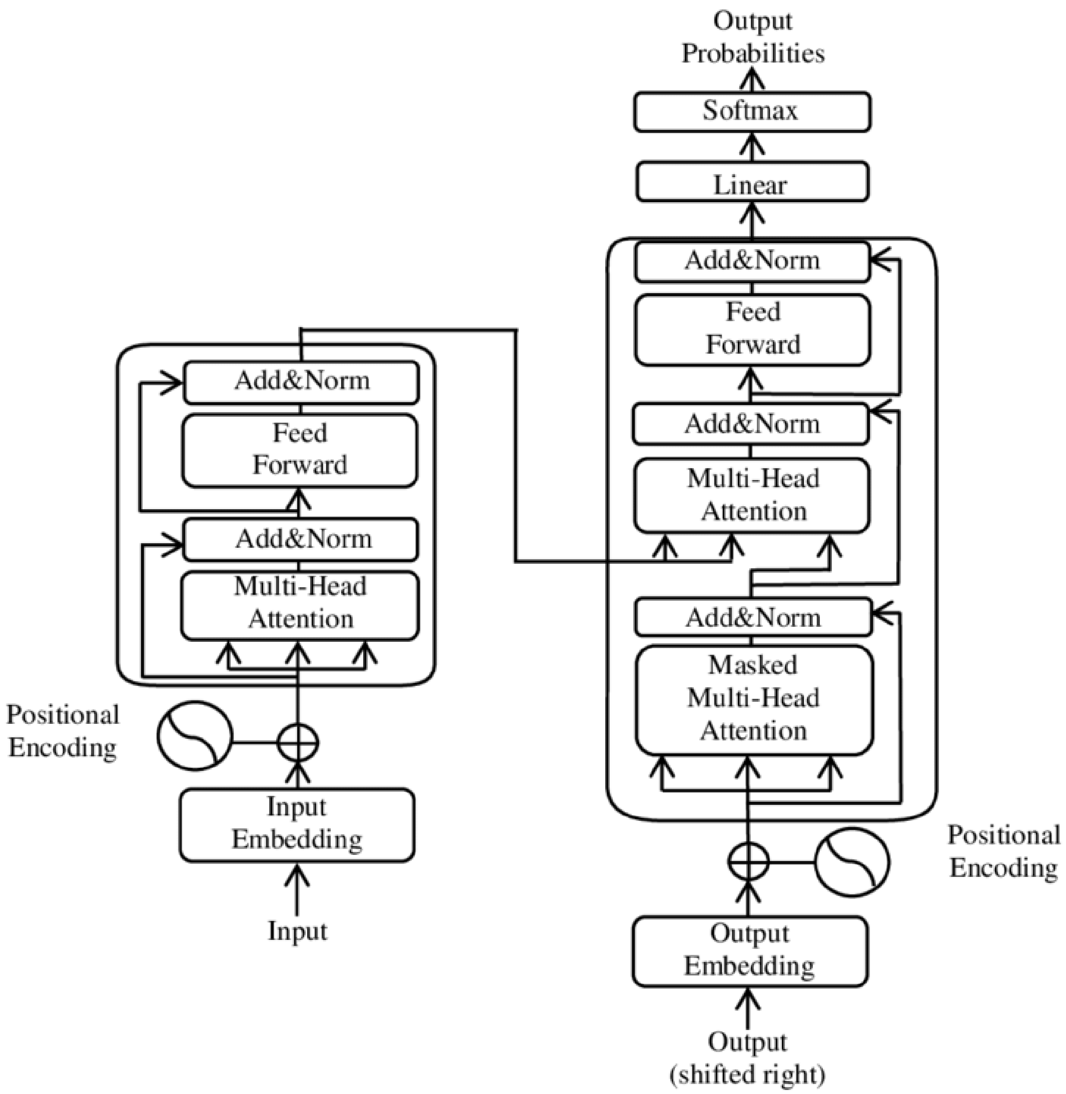
www.hpi.de

What we`ll cover in this video

- Why Transformers are relevant for time series forecasting
- Core concepts of Transformer architecture
- The self-attention mechanism explained
- How to adapt Transformers for sequential numerical data
- Strengths, limitations, and best use cases

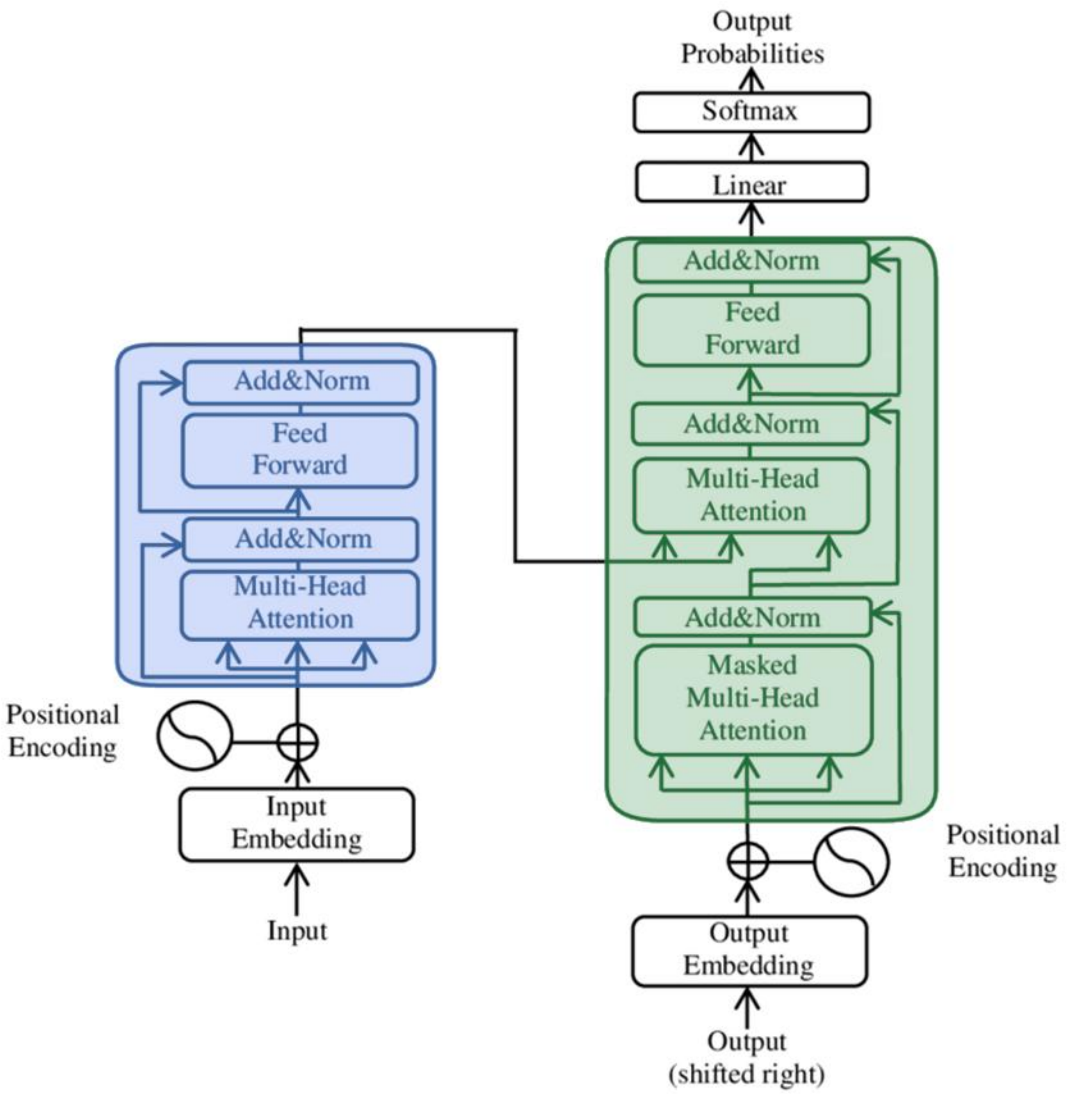
Why Transformers for Forecasting?

- Originally developed for Natural Language Processing (NLP), now adapted across many domains
- Ability to model long-range dependencies without relying on recurrence
- Highly parallelizable — enabling faster training compared to traditional RNNs
- Proven success in large-scale and complex forecasting problems



How Transformers Work (Core Concepts)

- Encoder–decoder structure (encoder usually optional for forecasting)
- Self-attention mechanism to relate all time steps
- Positional encoding to preserve sequence order
- Feed-forward layers for feature transformation



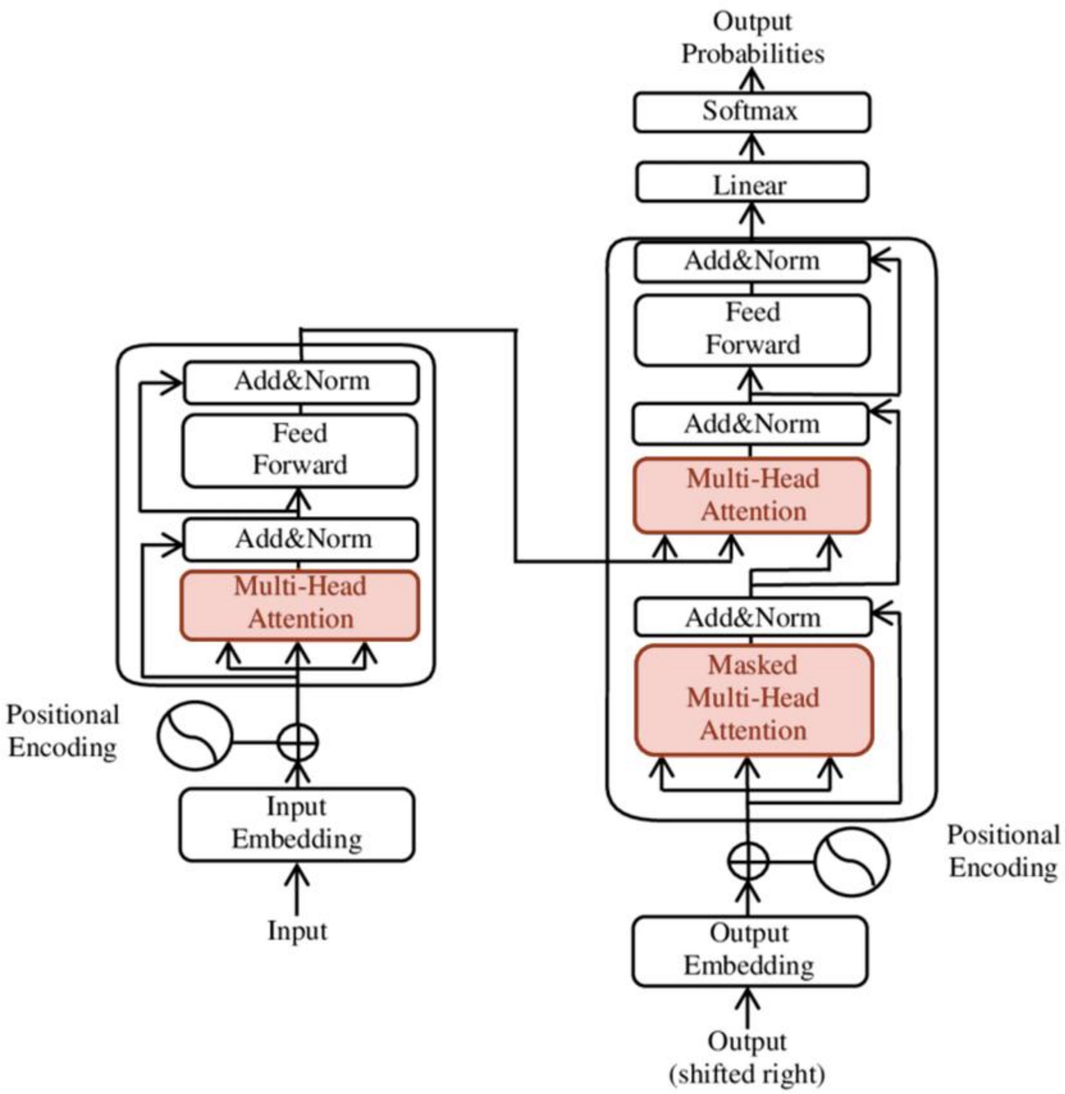
The Self-attention Nechanism

- Key idea: each time step can attend to any other

- Query, Key, and Value matrices

- Scaled dot-product attention

- Multi-head attention for capturing diverse relationships



Adapting Transformers for Time Series

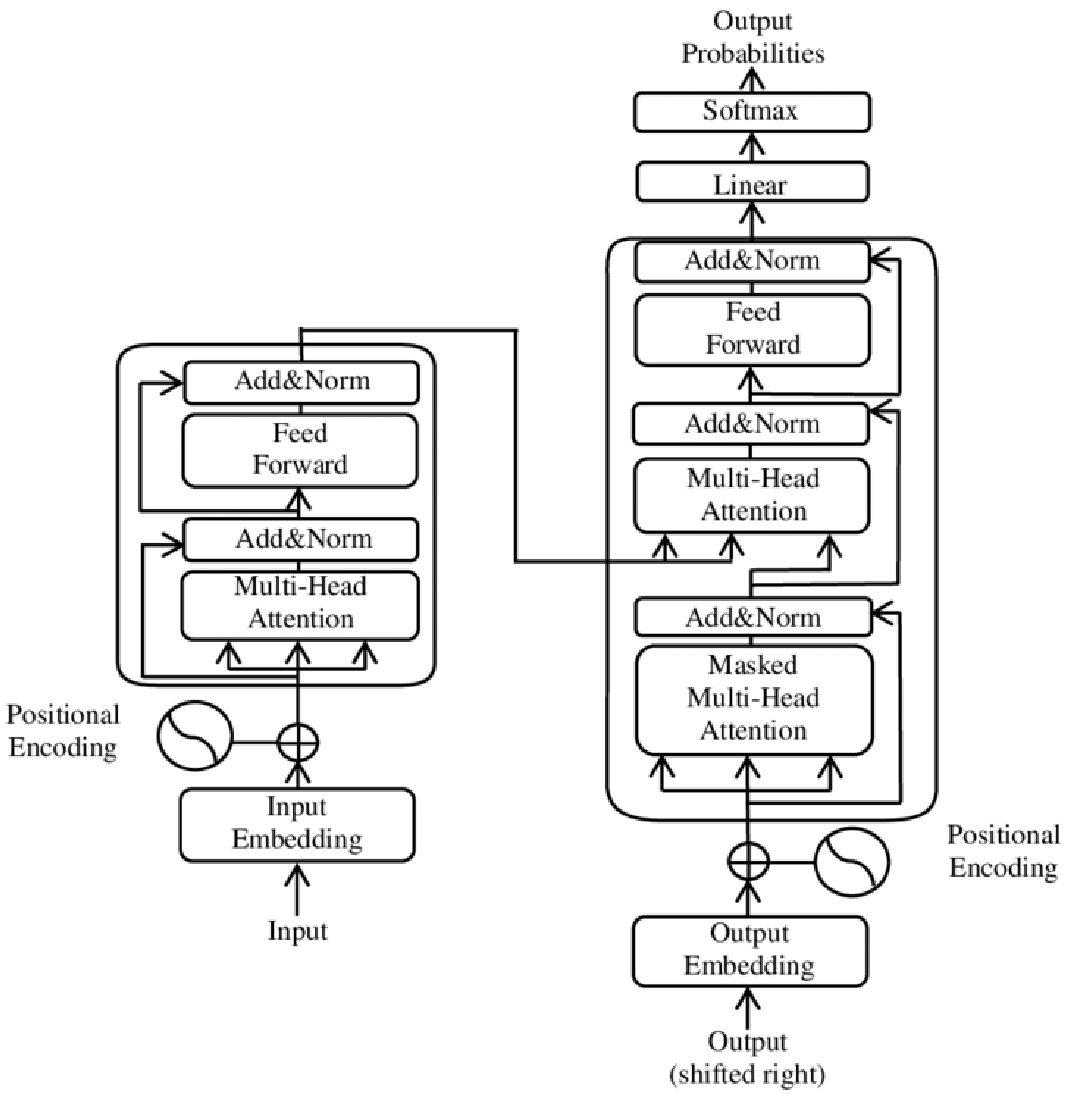
- Differences from NLP:
  - Numeric time series inputs (uni- or multivariate)
  - Time-based positional encodings
  - Often decoder-only structure for forecasting
- Regression output layer instead of classification

Architecture	Use in Forecasting
Decoder-only	Commonly used for autoregressive, step-by-step forecasting (e.g., GPT-style models)
Encoder-decoder	Popular for multivariate and sequence-to-sequence forecasting (e.g., Temporal Fusion Transformer)
Encoder-only	Rarely used standalone; mainly for feature extraction or representation learning in forecasting pipelines

Strengths and Limitations

- Strengths:
  - Captures long-range dependencies
  - Flexible for multi-feature sequences
  - Highly scalable

- Limitations:
  - Requires large datasets
  - Computationally intensive
  - Lower interpretability



What we‘ve learnt

- Transformers are powerful models for complex time series forecasting
- The self-attention mechanism enables global context understanding
- Adaptations are needed to handle numerical sequential data effectively
- Choice of architecture depends on forecasting task requirements
- Best suited for scenarios with ample data and computational resources