University of St.Gallen
Institute of Computer Science

University of St.Gallen
Institute for Computer Science in Vorarlberg

Integrating the Attention Mechanism into State Space Models

Tomas Hrycej, Bernhard Bermeitinger, Siegfried Handschuh
University of St.Gallen (HSG), Switzerland
tomas.hrycej@unisg.ch, bernhard.bermeitinger@unisg.ch, siegfried.handschuh@unisg.ch

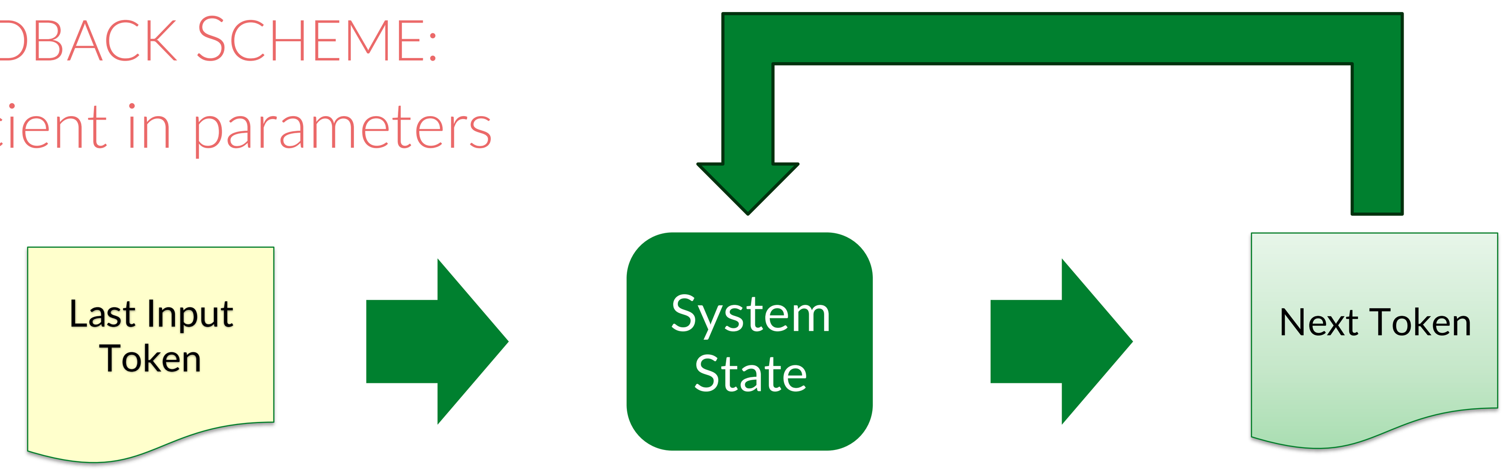
ATTENTION MECHANISM | STATE SPACE MODEL

SIGNAL FLOW

FEEDFORWARD SCHEME: extensive need of parameters

System theory: Finite Impulse Response Model (FIR)

FEEDBACK SCHEME: efficient in parameters



System theory: Infinite Impulse Response Model (IIR)

CONTEXT SOURCE AND REPRESENTATION

CONTEXT SOURCE (Attention Mechanism)

- Complete sequence of token predecessors
- Variable volume growing with the sequence length and position in it
- Quadratic complexity of context relevance

CONTEXT SOURCE (State Space Model)

- Recursively updated state vector
- Constant volume
- Necessity to commit to the state dimension
- Context relevance not yet captured
- Capturing context relevance is the essence of the proposal presented here

WEIGHTING OF THE CONTEXT RELEVANCE IN THE TEXT SEQUENCE

ATTENTION MECHANISM (WITHOUT MULTI-LAYER-PERCEPTRON)

$$y_i = \sum_{j=1}^{i} x_j a_{i,j} W_V W_O$$

context relevance

position independent parameters

STATE SPACE MODEL IN DIAGONAL FORM

$$y_i = \sum_{j=1}^{i} x_i B A^{i-j} C = \sum_{j=1}^{i} x_i \left(\sum_{k=1}^{N} a_k^{i-j} b_k c_k \right)$$

Positional forgetting

position independent parameters

USE OF ATTENTION MECHANISM

CONVENTIONAL CONCEPT

- Relevance determined by attention weights
- They result from the cross-similarities between input embeddings (queries and keys)
- Similarity: $x_i{}' W_Q W_K x_j$ (all i and j)
- Quadratic number of combinations

OUR PROPOSAL

- In addition to positional forgetting: Introducing attention weights
- Resulting from the similarities between input embedding (query) x_i and the actual state vector h_i
- Similarity: $x_i{}' W_S h_i$(only along i)
- Linear complexity

Presented at SDS2025, Zürich, Switzerland