Netzwerkarchitekturen

Teil 3 – Encoder-Decoder-Architektur

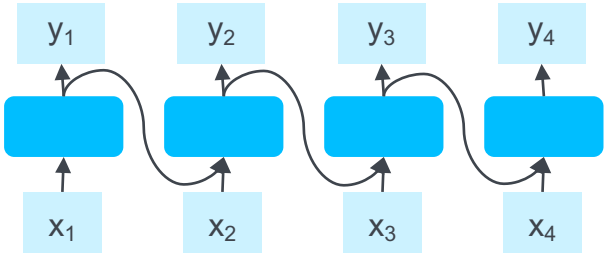
INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

INNOVET
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Sequenz-zu-Sequenz-Modelle

- „sequence-to-sequence“, seq2seq
- RNNs sind einfache Sequenz-zu-Sequenz-Modelle:
  generieren für eine Input-Sequenz eine Output-Sequenz



I	would	like	to	eat	ice-cream

Ich	möchte	gerne	Eis	essen

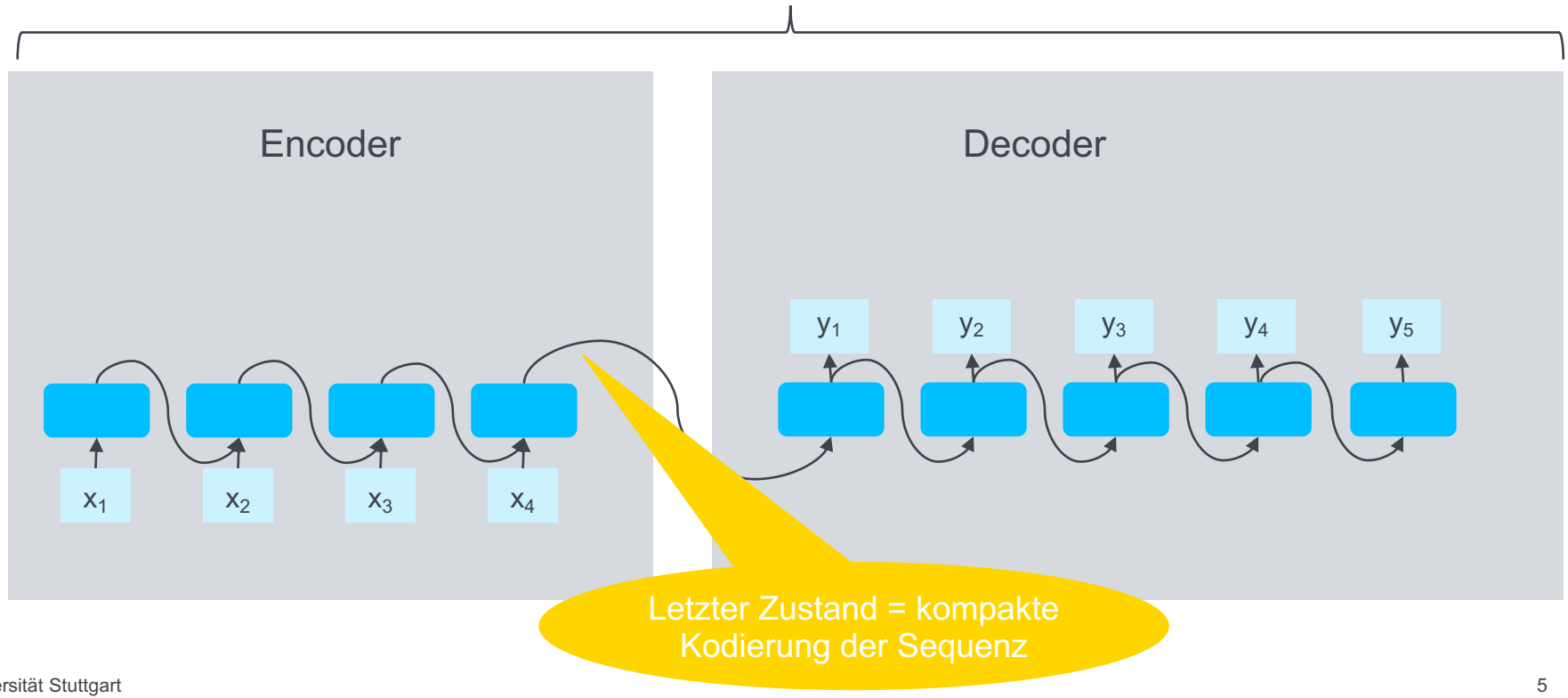
In Anwendungen aber häufig ungleiche Länge, z.B. maschinelle Übersetzung

Sequenz-zu-Sequenz-Probleme sind Fragestellungen, bei denen längere Folgen von Daten in andere, nicht genau gleich lange Folgen von Daten umgewandelt werden müssen.

Typische solche Probleme sind Spracherkennung, maschinelle Übersetzung, Textextraktion (die Zusammenfassung von Texten), Chatbots, automatische Erzeugung von Untertiteln und viele mehr.

Kombination zweier RNNs

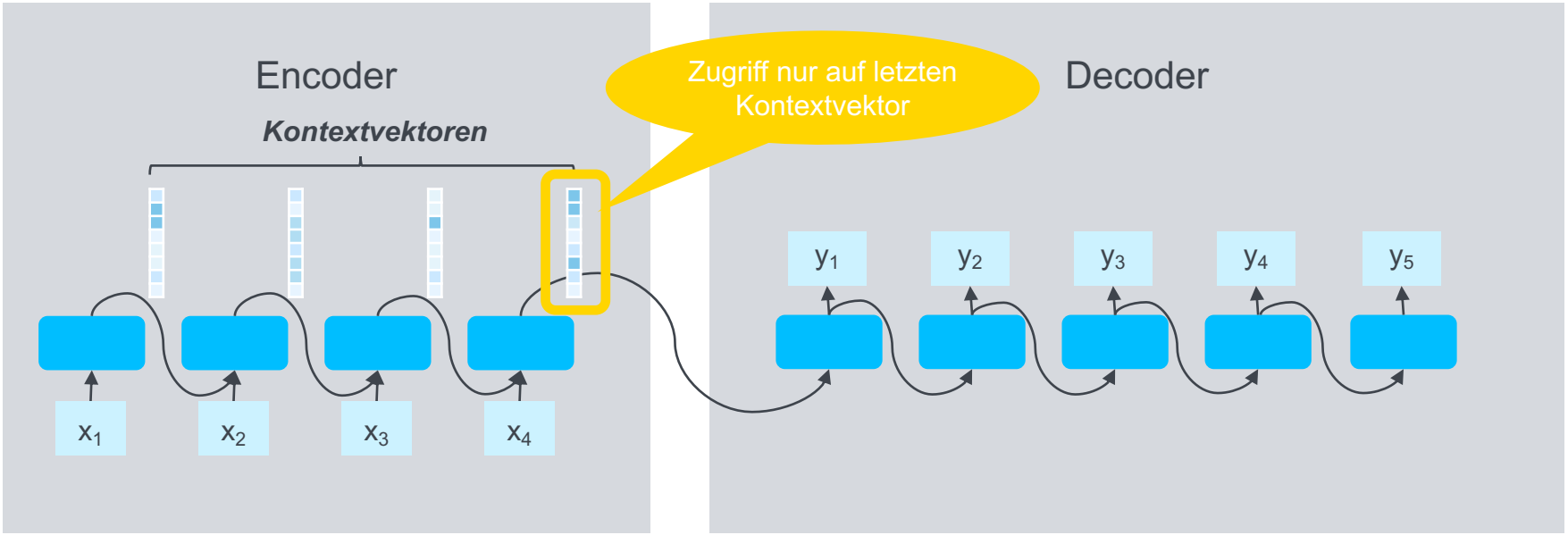
Encoder-Decoder-Architektur



Die Encoder-Decoder-Architektur erlaubt die Modellierung von Sequenz-zu-Sequenz-Problemen. Als Bausteine können sowohl klassische RNNs als auch LSTMs verwendet werden.

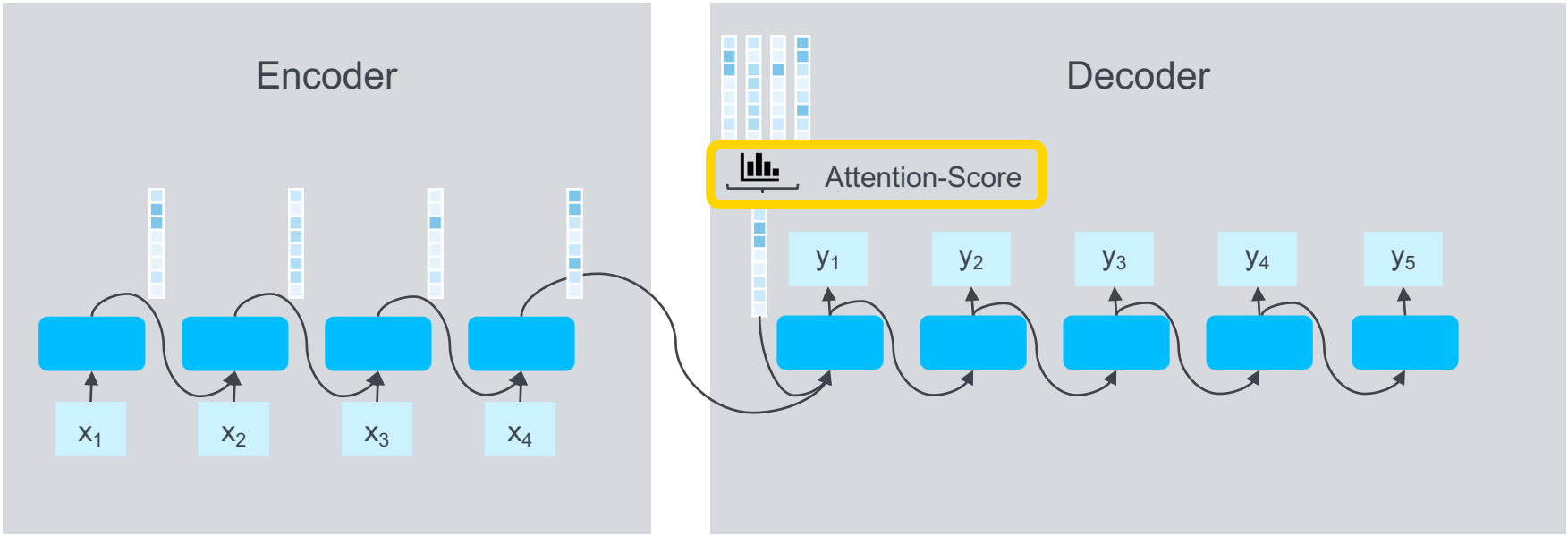
Encoder und Decoder werden als ein gemeinsames Netzwerk trainiert.

Encoder-Decoder Architektur

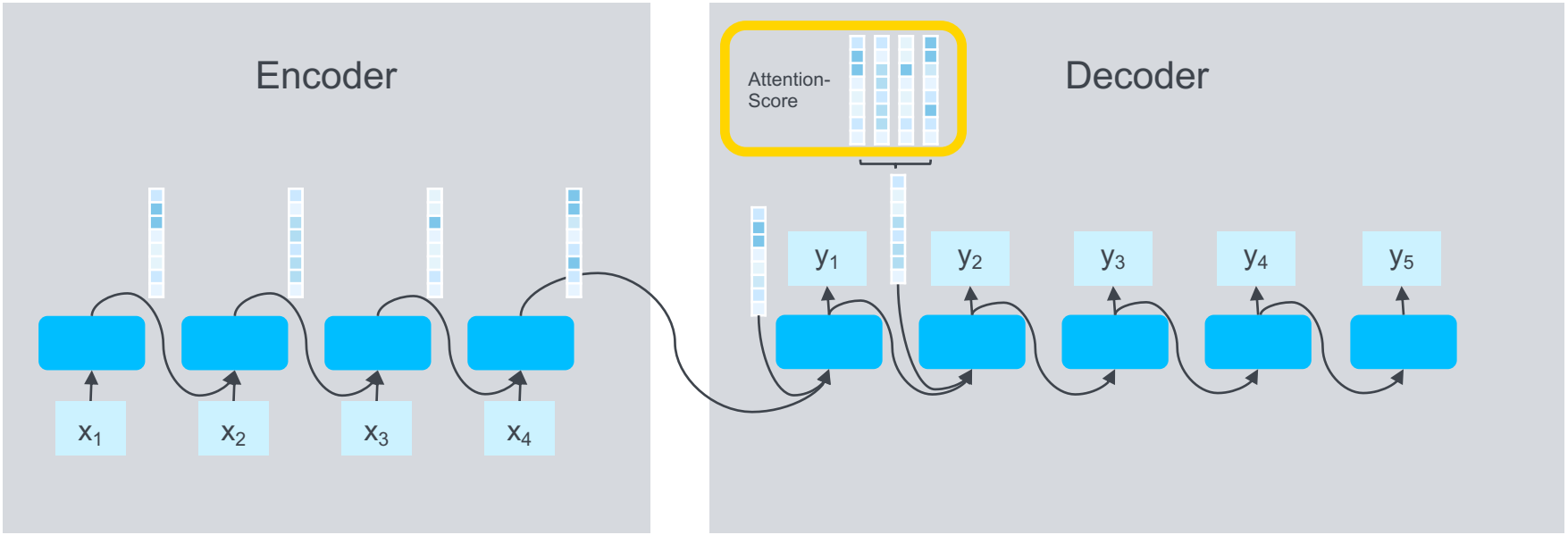


mit Attention

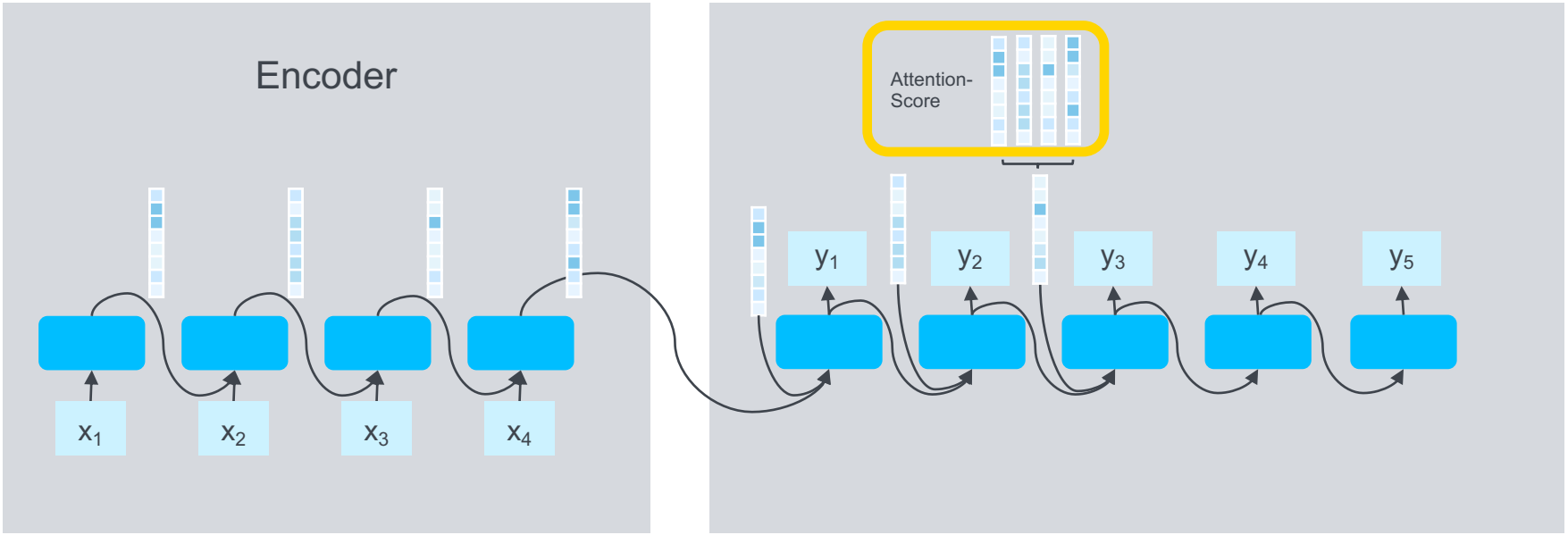
Encoder-Decoder Architektur mit Attention



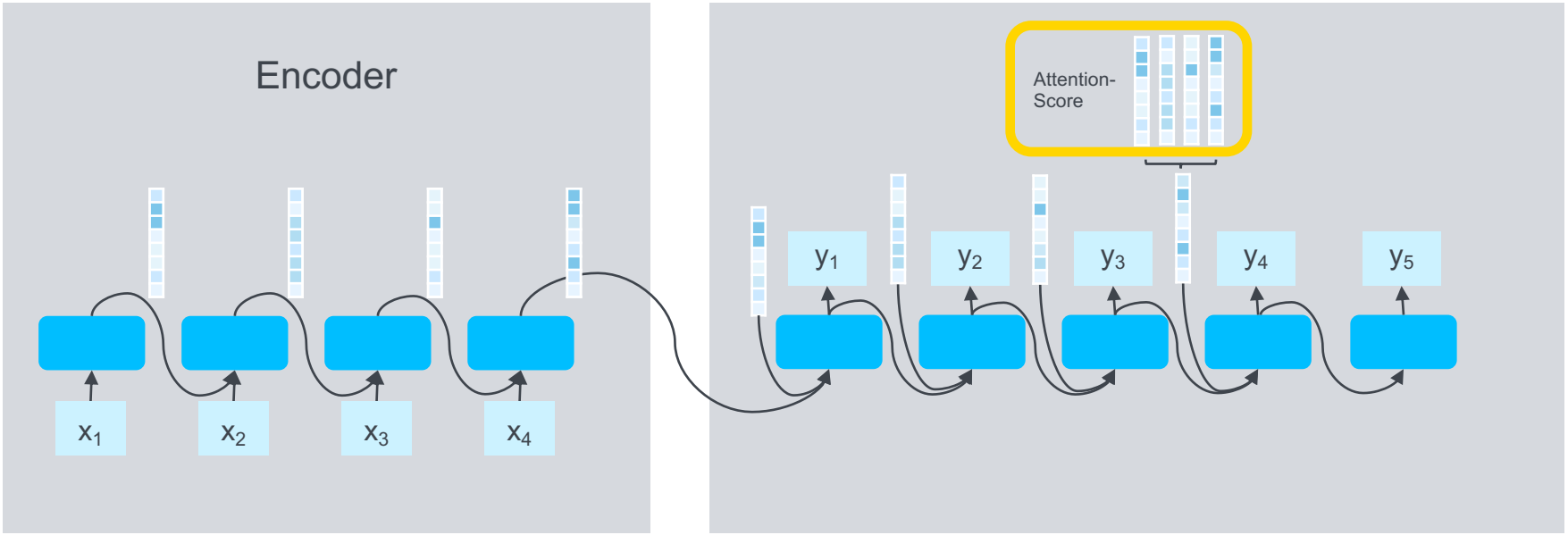
Encoder-Decoder Architektur mit Attention



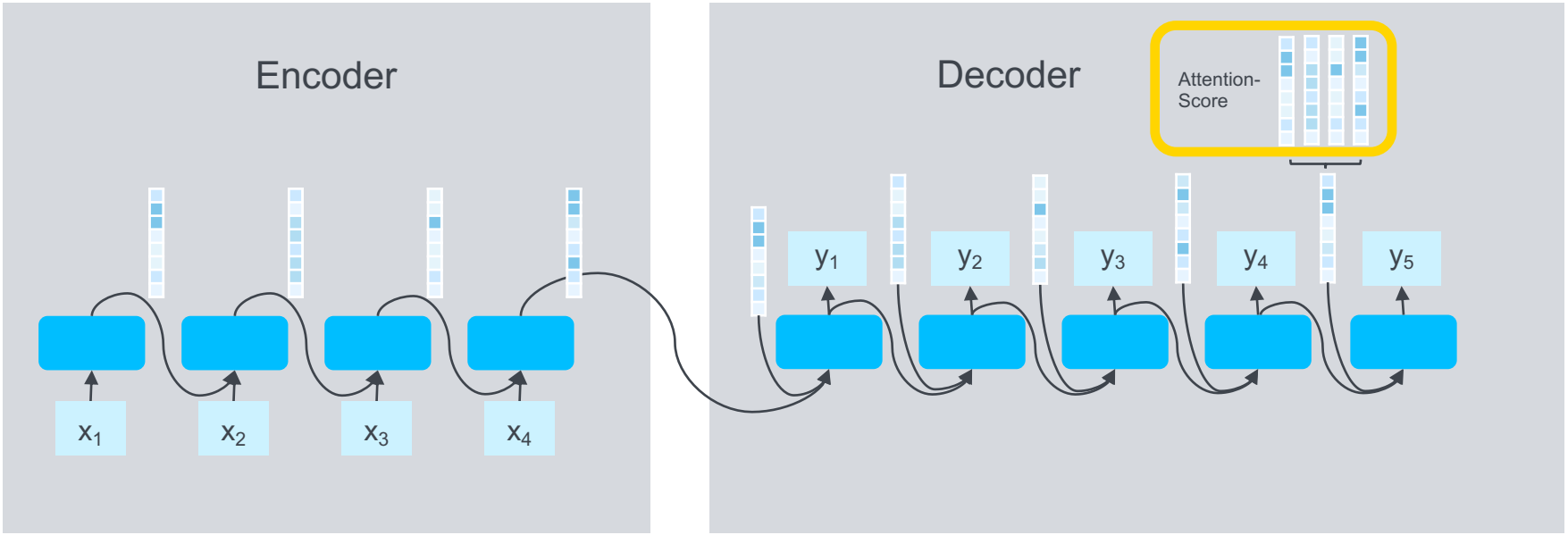
Encoder-Decoder Architektur mit Attention



Encoder-Decoder Architektur mit Attention



Encoder-Decoder Architektur mit Attention



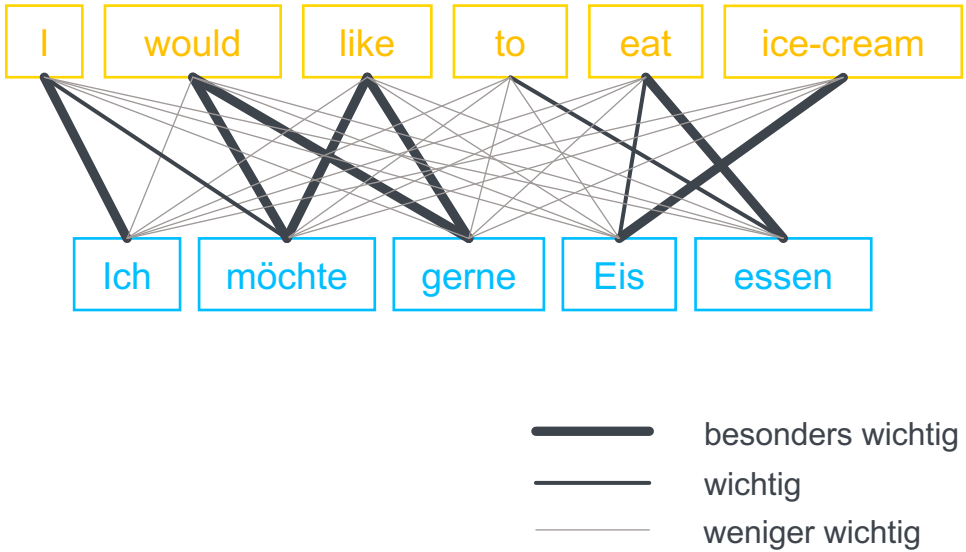
Encoder-Decoder-Architekturen mit Attention erlauben dem Decoder Zugriff auf alle verborgenen Zustände des Encoders.

Sie berechnen in jedem Decoder-Schritt einen Attention-Score für alle verborgenen Zustände des Encoders. Die Zustände werden dann mit Hilfe des Attention-Scores gewichtet und zu einem einzigen Vektor addiert. Dieser wird als zusätzliches Input verwendet.

Effekt der Attention-Scores

- Elemente in der Input-Sequenz sind für die Elemente in der Output-Sequenz unterschiedlich wichtig
- Netzwerk soll sich in jedem Decoder-Schritt auf die Informationen konzentrieren, die an dieser Stelle am wichtigsten sind

➡ Attention-Score



Attention-Scores modellieren, dass die Elemente in der Input-Sequenz für die Elemente in der Output-Sequenz verschieden wichtig sind.

Die Encoder-Decoder-Architektur kann mithilfe von Attention erweitert werden. Dadurch verbessert sich die Performanz vor allem für Probleme, bei denen bestimmte Elemente im Input mit bestimmten Elementen im Output zu tun haben.

Multiple Choice: Encoder-Decoder-Architektur - Netzwerkarchitekturen (Teil 3)

Drag the Words: Encoder-Decoder-Architektur - Netzwerkarchitekturen (Teil 3)

INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München

GEFÖRDERT VOM
INNOVET
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Lizenzbestimmungen

“Netzwerkarchitekturen – Teil 3: Encoder-Decoder-Architektur" von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

 CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.de)
(Namensnennung 4.0 International)

Quellenverzeichnis

Titelfoto: Lars Kienle (https://unsplash.com/@larskienle), „fibre optic cable rack“, auf Unsplash (https://unsplash.com/de/fotos/IlxX7xnbRF8), lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.