TECHNICAL APPLICATIONS AND DATA MANAGEMENT. WS 2021/2022.

VORLESUNG 9

16.11.2021

MÜNCHEN

STUDIENGANG DIGITAL MANAGEMENT.

AGENDA

1. Wahl Projektarbeit AI
2. Sequentielle Daten
3. Rekurrente Neuronale Netze
4. Natural Language Processing
5. Case Study

WAS HABEN WIR BIS JETZT GEMACHT?

ROADMAP	WAS HABEN WIR GEMACHT?
Vorlesung 1	Workflow Data Management, Datentypen und Datenqualität
Vorlesung 2	Einführung Data Science und Data Science Workflow, Grundlagen Data Management
Vorlesung 3	Grundlagen Stochastik: Wahrscheinlichkeitsrechnung, deskriptive und explorative Statistik
Vorlesung 4	Statistische Inferenz, lineare Regression
Vorlesung 5	Einführung Machine Learning, Unüberwachtes Lernen
Vorlesung 6	Überwachtes Lernen
Vorlesung 7	Neuronale Netze und Convolutional neural networks
Vorlesung 8	Aufgabenstellung Data Science, Case Study CNN: Malaria

1. WAHL PROJEKTARBEIT AI

WAS IST IM RAHMEN PROJEKTARBEIT ZU TUN? 1. TEIL: SCHULTERBLICK AI IN KW50.

- Zu erstellen ist eine Word-Datei à 4 Seiten mit:
  - Problem statement: „Welches Problem wollen wir lösen? Wieso ist es ein Problem? Was ist der Nutzen einer Lösung?"
  - Metriken zur Evaluation Ergebnisse
  - Vorgehensweise Lösungsansatz anhand Data Science Workflow: „Wie gehen wir es an?"
  - Aufteilung Projektarbeit: wer in der Gruppe macht was?
  - Aktueller Status: gibt's Probleme? Wie kommen Sie voran?

- Vorstellung durch die Gruppen in der Vorlesung („Amazon"- Ansatz)

- Gemeinsame Diskussion

Templates finden Sie auf der Homepage des Kurses unter Materialien

WAS IST IM RAHMEN PROJEKTARBEIT ZU TUN? 2. TEIL: ABSCHLUSSPRÄSENTATION IN KW01

- Powerpoint-Präsentation: jede Gruppe präsentiert ihre Ergebnisse mit gesamthaft 30 Minuten (jeder ca. 10 Minuten)

- Schriftliche Ausarbeitung je Teilnehmer à 12-15 Seiten:
  - Projektübersicht
  - Vorgehensweise anhand Data Science Workflow (Get Data, Explore the Data, Model the Data, Visualise Results):
    - eingesetzte Verfahren
    - Implementierung Datenaufbereitung, Datenmodellierung und Datenvisualisierung
  - Ergebnisse: Visualisierung Ergebnisse und Bewertung anhand Metriken
  - Reflektion: Was lief gut? Was lief schlecht?
  - Future Work: „Was würden wir als nächste Schritte machen?“
  - Literaturverzeichnis

Templates finden Sie auf der Homepage des Kurses unter Materialien

WAHL PROJEKTARBEIT: FOLGENDE THEMEN STEHEN FÜR DEN AI-ANTEIL ZUR AUSWAHL.

Bilderanalyse: Einsatz und Vergleich von Basic CNN und Transfer Learning inkl. Parametertuning

1. Erkennen von Lungenentzündung (https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia)
2. Verkehrszeichen erkennen (https://www.kaggle.com/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign)
3. Tierartenerkennung: Erkennung von Katzen oder Spinnen (https://www.kaggle.com/c/dog-breed-identification zeigt es am Beispiel Hunde)
4. Erkennen von Müdigkeit (https://www.kaggle.com/serenaraju/yawn-eye-dataset-new)

Text Analyse: Einsatz und Vergleich von Basic RNN und LSTM, Attention oder Bert inkl. Parametertuning

1. Fake News Detection/ Classifier mit Deep Learning (https://www.kaggle.com/hassanamin/textdb3)
2. Sentiment Analysis of Movies (https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews)
3. Börsenkursvorhersage (https://www.kaggle.com/faressayah/stock-market-analysis-prediction-using-lstm)

2. SEQUENTIELLE DATEN

WELCHE DATENTYPEN HABEN WIR BIS JETZT FÜR MACHINE LEARNING EINGESETZT?

Statistische Datensätze

	PassengerClass	survived	name	sex	age	SiblingSpousesPresent	ParentsChildrenPresent	ticket	fare	cabin	embarked	boat	body	HomeDestination
1289	3	False	Wiklund, Mr. Karl Johan	male	21.0	1	0	3101266	6.4958	0	Southampton	0	0	unknown
1290	3	True	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	0	Southampton	0	0	unknown
1291	3	False	Willer, Mr. Aaron ("Abi Weller")	male	NaN	0	0	3410	8.7125	0	Southampton	0	0	unknown
1292	3	False	Willey, Mr. Edward	male	NaN	0	0	S.O./P.P. 751	7.5500	0	Southampton	0	0	unknown
1293	3	False	Williams, Mr. Howard Hugh "Harry"	male	NaN	0	0	A/5 2466	8.0500	0	Southampton	0	0	unknown
1294	3	False	Williams, Mr. Leslie	male	28.5	0	0	54636	16.1000	0	Southampton	0	14	unknown
1295	3	False	Windelov, Mr. Einar	male	21.0	0	0	SOTON/OQ 3101317	7.2500	0	Southampton	0	0	unknown
1296	3	False	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	0	Southampton	0	131	unknown
1297	3	False	Wiseman, Mr. Phillippe	male	NaN	0	0	A/4. 34244	7.2500	0	Southampton	0	0	unknown
1298	3	False	Wittevrongel, Mr. Camille	male	36.0	0	0	345771	9.5000	0	Southampton	0	0	unknown
1299	3	False	Yasbeck, Mr. Antoni	male	27.0	1	0	2659	14.4542	0	Cherbourg	C	0	unknown
1300	3	True	Yasbeck, Mrs. Antoni (Selini Alexander)	female	15.0	1	0	2659	14.4542	0	Cherbourg	0	0	unknown
1301	3	False	Youseff, Mr. Gerious	male	45.5	0	0	2628	7.2250	0	Cherbourg	0	312	unknown
1302	3	False	Yousif, Mr. Wazli	male	NaN	0	0	2647	7.2250	0	Cherbourg	0	0	unknown
1303	3	False	Youssef, Mr. Gerious	male	NaN	0	0	2627	14.4583	0	Cherbourg	0	0	unknown
1304	3	False	Zabour, Miss. Hileni	female	14.5	1	0	2665	14.4542	0	Cherbourg	0	328	unknown
1305	3	False	Zabour, Miss. Thamine	female	NaN	1	0	2665	14.4542	0	Cherbourg	0	0	unknown
1306	3	False	Zakarian, Mr. Mapriededer	male	26.5	0	0	2656	7.2250	0	Cherbourg	0	304	unknown
1307	3	False	Zakarian, Mr. Ortin	male	27.0	0	0	2670	7.2250	0	Cherbourg	0	0	unknown
1308	3	False	Zimmerman, Mr. Leo	male	29.0	0	0	315082	7.8750	0	Southampton	0	0	unknown

Einsatz von unsupervised und supervised Verfahren für das Lernen von Features aus den Daten

Bilder:

Lernen Features eines Bildes per CNN, dann Einsetzen in Supervised Learning Verfahren.

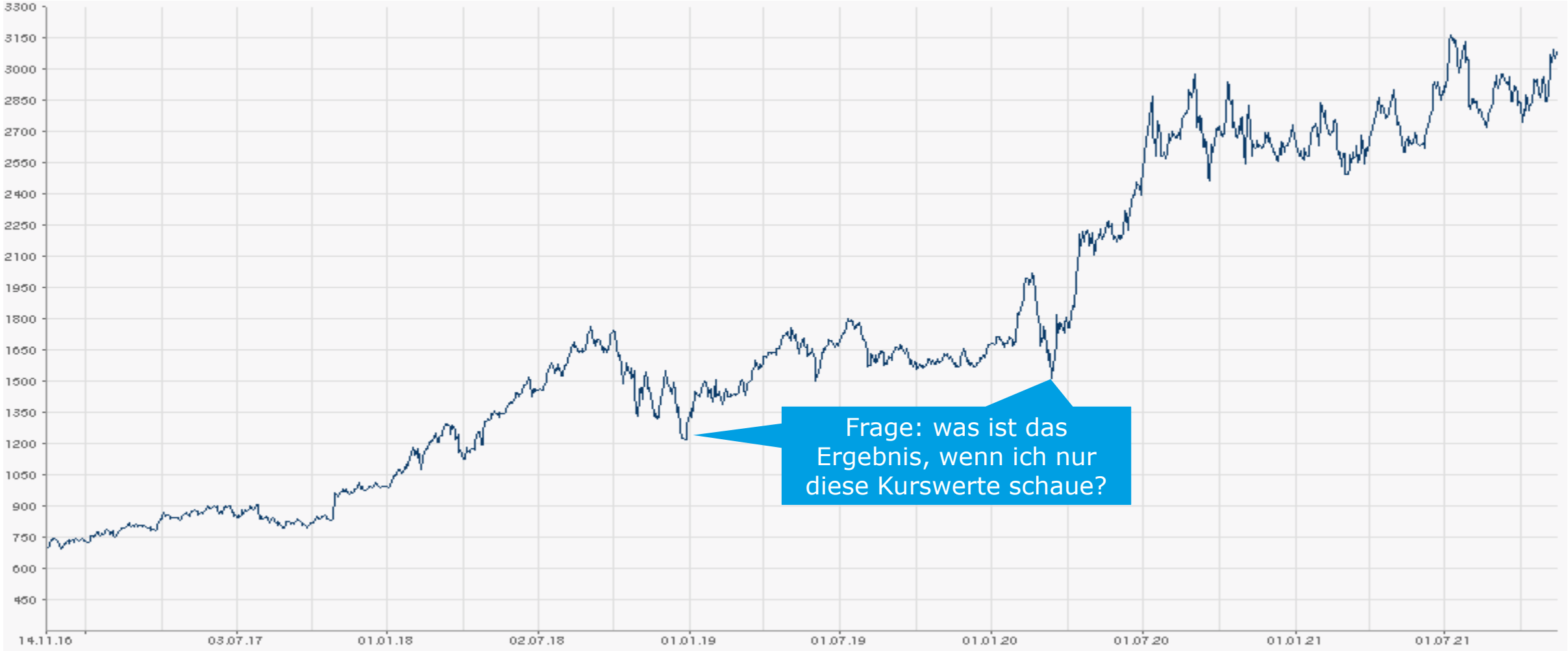
Hierbei handelt es sich um statische, sich über den zeitlichen Verlauf nicht ändernde, Daten

ÜBERSICHT SEQUENTIELLE, ZEITABHÄNGIGE DATEN.

- Zeitreihen: Aktienkurse, Messungen, Temperaturkurven, …
- Video: Folge von einzelnen Bildern (Frames per Second)
- Texte
- Sprache

Bei sequentiellen Daten haben vorherige Daten Einfluß auf die aktuellen und nachfolgenden Daten

ZEITREIHEN: BEISPIEL AKTIENKURS AMAZON.

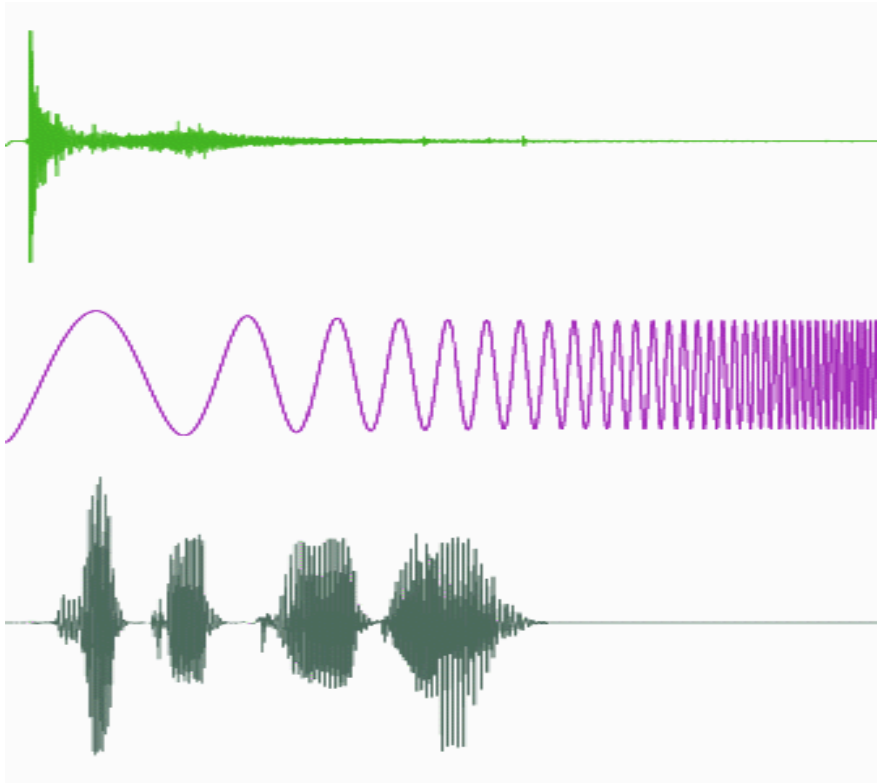


Quellen: Aktienkurs: https://www.finanzen.net/

VIDEO: SEQUENZ EINZELNER BILDER (FRAMES PER SECONDS)

Quelle Video: Castles in the Sky (2010), abrufbar unter Link

SPRACHE: SEQUENZ VON TÖNEN



Gewehrschuss.

Sinusschwingung mit sinkender Periodendauer.

gesprochene Wort Wikipedia.

Frage: wieso brauch ich hier Sequenzen?
Klingt doch gleich?

Quelle: Link

TEXTE: SEQUENZ VON WÖRTERN

- Der Bauer schlägt das Pferd und dann die Dame.
- Am Fuß des Berges.
- Schmerzt des Dichters Ferse? Schmerzen des Dichters Verse?
- „Der ist nicht ganz dicht!“ – „Der Eimer oder er?“
- „Das war ja eine ganz große Leistung, toll!“

Frage: Was passiert, wenn ich die Sätze nicht als Sequenz betrachte, sondern Wort für Wort?

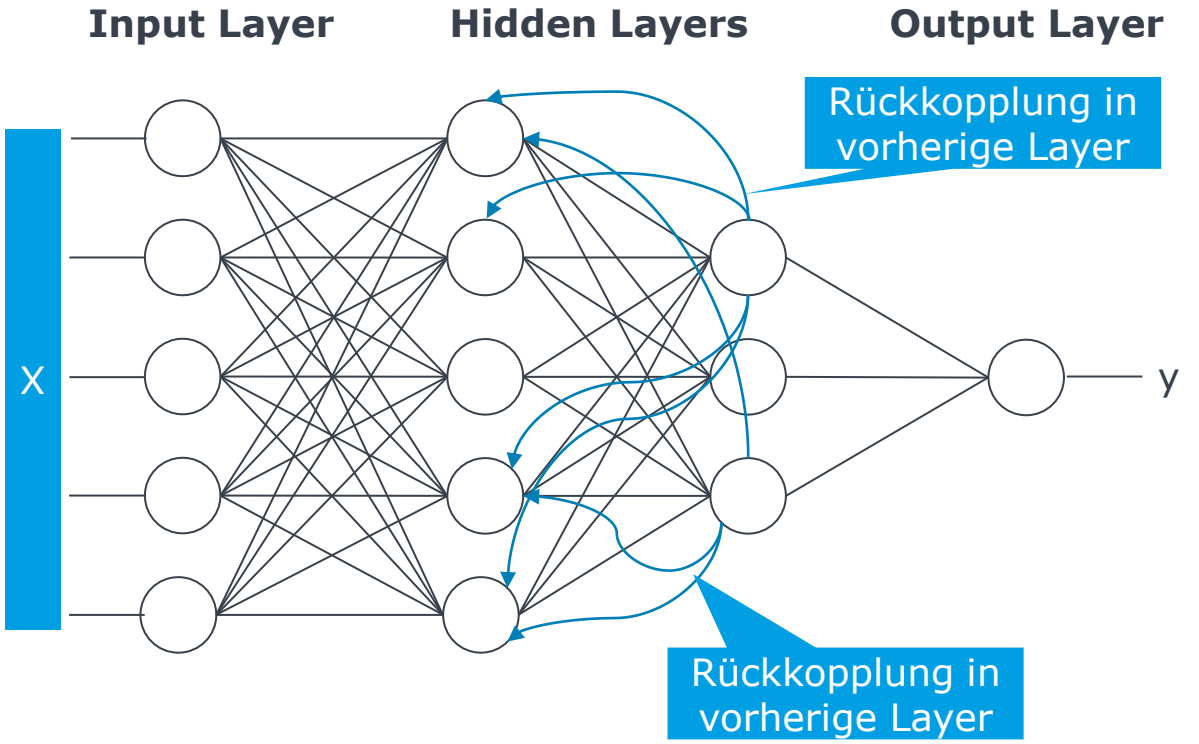
SEQUENTIELLE DATEN

- Bei sequentiellen Daten können die vorherigen Daten Einfluß auf die aktuellen und nachgelagerten Daten haben.
- Für eine Verarbeitung solcher Daten mittels Machine Learning muß das Modell zeitliche Bedingungen abbilden können.
- Das Modell muß also auf aktuellen Daten, aber auch auf vorherige Daten zugreifen.
- Die bisher betrachteten Modelle bilden dies nicht ab, wir brauchen also Modelle mit Rückkopplung oder „Gedächtnis".

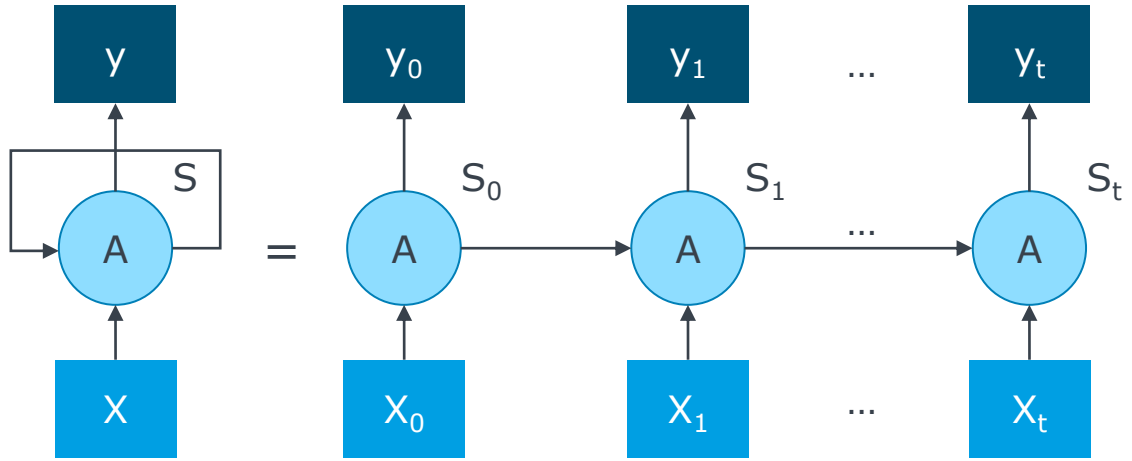
Rekurrente Netzwerke werden sehr häufig für die Verarbeitung von sequentiellen Daten eingesetzt.

3. REKURRENTE NEURONALE NETZE

STRUKTUR REKURRENTE NEURONALE NETZE.



Detaillierung Struktur RNN über Zeitschritte t („Querschnittsbild")

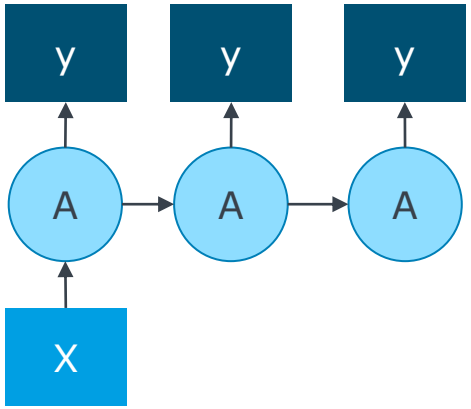


- Eingaben X_i: Sequenz wird auf Inputs aufgeteilt.
- Hidden State S_i: „Kurzzeitgedächtnis". Abhängig vom Input x_i sowie vorherigen Zustände S_i. Überträgt auch (gewichtet) Informationen zum nächsten Schritt.
- A: Neurales Netzwerk inklusive Rückkopplung Ergebnisse.
- Ausgaben y_i: Ausgaben/ Prädiktion des Netzwerks.

Quelle: rechtes Bild: Link und Link
Rekurrente Netzwerke definiert in: Elman, J.: "Finding structure in time", Cognitive Science, 14(2), 179–211, 1990.

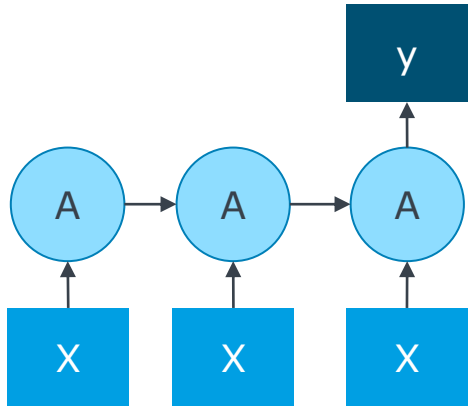
(GROB-)STRUKTUR RNN ABHÄNGIG VOM EINSATZGEBIET.

Ein Input, viele Outputs



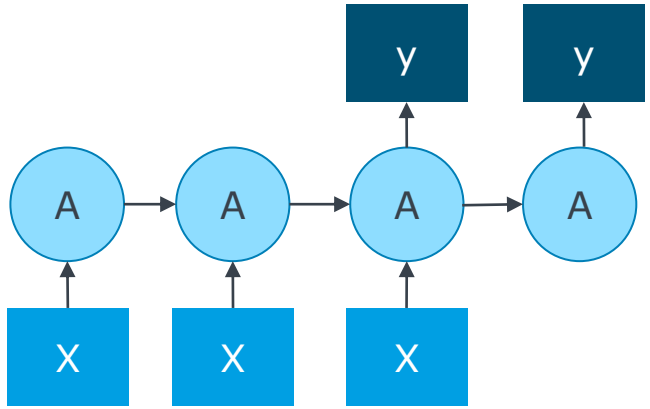
Sequence Output:
Bspw. automatisierte **Bilderkennung und -beschreibung**
(1 Bild mit n Wörten beschreiben)

Viele Inputs, ein Output



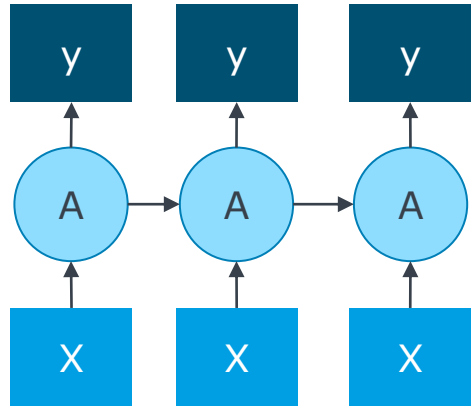
Sequence Input:
Bspw. automatisierte **Sentiment Analysis/ Gefühlsanalyse**
(1 Satz wird analysiert ob positive/ negative Aussagen)

Viele Inputs, viele Outputs



Sequence input and output:
Bspw. Automatisiertes **Übersetzen Texte** (Deutsch – Englisch), aber Input und Output können verschiedene Längen haben.

Viele Inputs, viele Outputs



Gleich lange Sequence input and output:
Bspw. Videoklassifikation, jeder Frame wird gelabelt

Frage: welche Struktur wird eingesetzt für
a. Bewerten ob positives/ negatives Kundenfeedback
b. Automatisiertes Übersetzen eines Satzes
c. Labeln Videoframes (je Frame)
d. Beschreiben eines Bildes mit Wörtern

Quelle: Link

ÜBERSICHT BEKANNTE VERFAHREN.

- **LSTM (1997)[1]:**
  - Entwickelt von Sepp Hochreiter und Jürgen Schmidhuber an der TU München.
  - aktuell (noch?) das häufigst eingesetzte Verfahren für sequentielle Daten, bspw. Spracherkennung in Handys (Link).
  - Kann viel längere Zeitsequenzen verarbeiten als normale RNN (keine Probleme mit Vanishing Gradient[2])
  - Ressourcenaufwendig: hoher Speicherbedarf und (sehr) aufwendige Trainingszeit.
  - Und neigt zum Overfitting (das man durch Tunen Hyperparameter und Struktur LSTM aber reduzieren kann).

Wiederholungsfrage: was ist Overfitting?

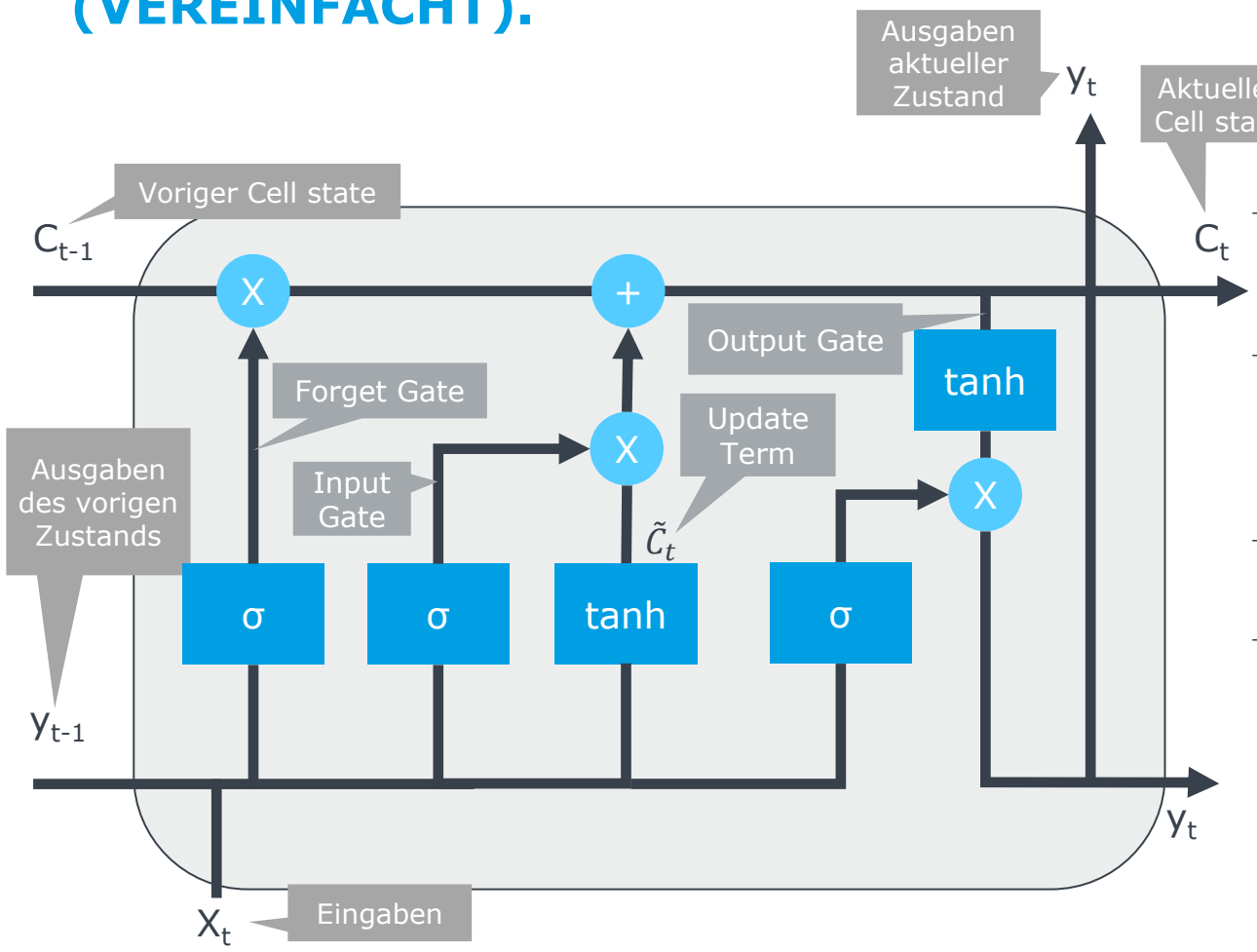
- **Attention (2017)[3]:**
  - Entwickelt von Google.
  - Hat LSTM in vielen Gebieten abgelöst, v.a. für maschinelles Übersetzen aufgrund höherer Flexibilität.
  - Schnelleres Training als LSTM aufgrund Aufteilen Lernen auf mehrere Rechner (Parallelisieren).

~~- **Gated Recurrent Units:** ähnlich LSTM (aber ohne Outputgate) und mit schlechterer Laufzeit-Performance~~

~~- Bidirectional Encoder Representations from Transformers (2018)[4]: Quasi Transfer Learning für Transformer.~~

Quellen:
1 Hochreiter, Schmidhuber: "Long Short-Term Memory ", 1997. Link
2 Vanishing Gradient: beim Trainieren von Netzen mit vielen Schichten und Nutzung Expontentialfunktionen als Aktivierungsfunktion kann bei der Backpropagation der Fall entstehen, daß die Gradienten sehr, sehr klein werden und damit die Gewichte und Bias des Modells, v.a. der ersten Schichten, nicht mehr geändert/ trainiert werden. Vgl. Hochreiter, S.: „Untersuchungen zu dynamischen neuronalen Netzen", 1991.
3 Vaswani et al.: "Attention is all you need ", 2017. Link
4 Devlin et al.: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019. Link

LONG SHORT-TERM MEMORY: STRUKTUR UND ABLAUF (VEREINFACHT).



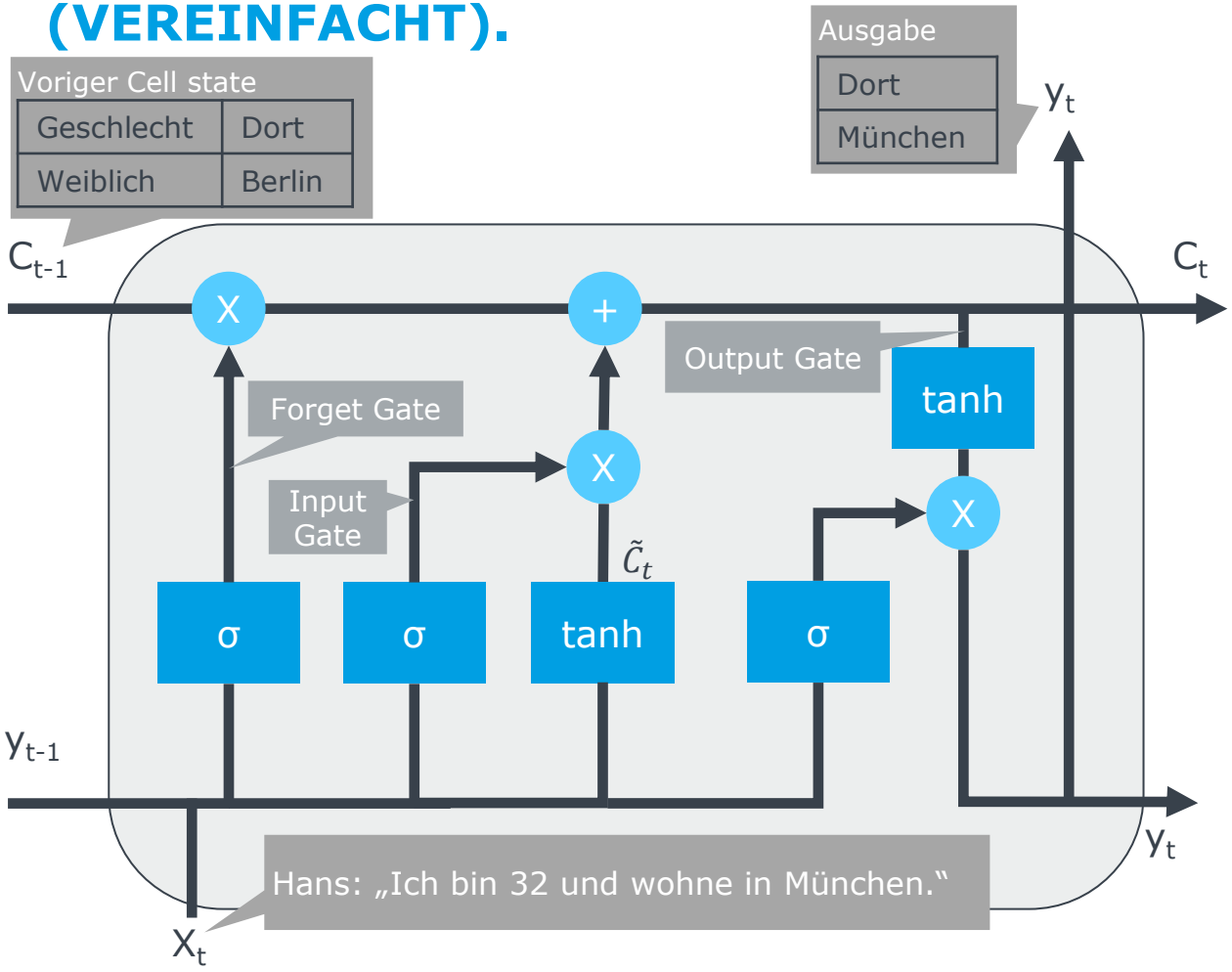
Wir schauen uns eine von mehreren LSTM-Zelle an, die miteinander verknüpft sind, d.h. die abgebildete Zelle hat „Nachbarzellen".

- **Forget Gate:** entscheidet, welche Infos des Cell State C_{t-1} zu vergessen oder weiter zu erinnern sind. Geschieht per Parameter mit Wert 0 für Vergessen, 1 für Erinnern oder Wert dazwischen, falls Gate unsicher ist.
- **Input Gate:** hat zwei Schritte.
  Schritt 1: Entscheidet, welche Infos für Cell State C_t zu aktualisieren sind. Erfolgt per Parameter mit Werten von 0 (irrelevant) bis 1 (relevant).
  Schritt 2: Bestimmt Update-Terme, d.h. die neuen Werte für den Cell State.
- **Cell State:** Hier wird die Zelle C_t aktualisiert, basierend auf Inputs des Forget Gate und Input Gates.
- **Output Gate:** entscheidet was aus der Zelle ausgegeben wird.

LSTM vermeidet Vanishing Gradient durch:
- Steuerung Verhalten Gradienten durch Werte Gates: Wert 0 verhindert bspw. Änderung Gradient.
- Werte des Gates für Vermeiden Explosion werden gelernt.
- Formel für Ausgabe C_t enthält Addition. Dadurch hat die den Gradienten erzeugende Ableitung ein „besseres Verhalten" als bei bspw. Multiplikation

Quellen:
Bild: Link.
LSTM: Hochreiter, Schmidhuber: „Long Short-Term Memory", 1997 sowie Link, Link.
Erklärung Vermeidung Vanishing Gradient: Link

LONG SHORT-TERM MEMORY: FALLBEISPIEL (VEREINFACHT).



Ziel: „Verstehen" einer Konversation durch Algorithmus
Anna: „Ich bin 30 Jahre alt und wohne in Berlin"
Hans: „Ich bin 32 und wohne in München."
Anna: „Wie viele Menschen leben dort?"

Frage: worauf bezieht sich dort?

Start mit Hans' Satz: Cell state C_{t-1}:

Geschlecht	Dort
Weiblich	Berlin

- **Forget Gate:** vergiss Werte für Geschlecht und Wohnort (Forget gate = 0), da mit Hans andere Person mit anderem Geschlecht und Wohnort spricht.
- **Input Gate:**
  - Schritt 1: bestimme zu aktualisierende Werte
  - Schritt 2: bestimme Änderungswerte $\tilde{C}_t$

Geschlecht	Dort
Männlich	München

- **Cell State:** schreibe die Werte in die Zelle C_t

Geschlecht	Dort
Männlich	München

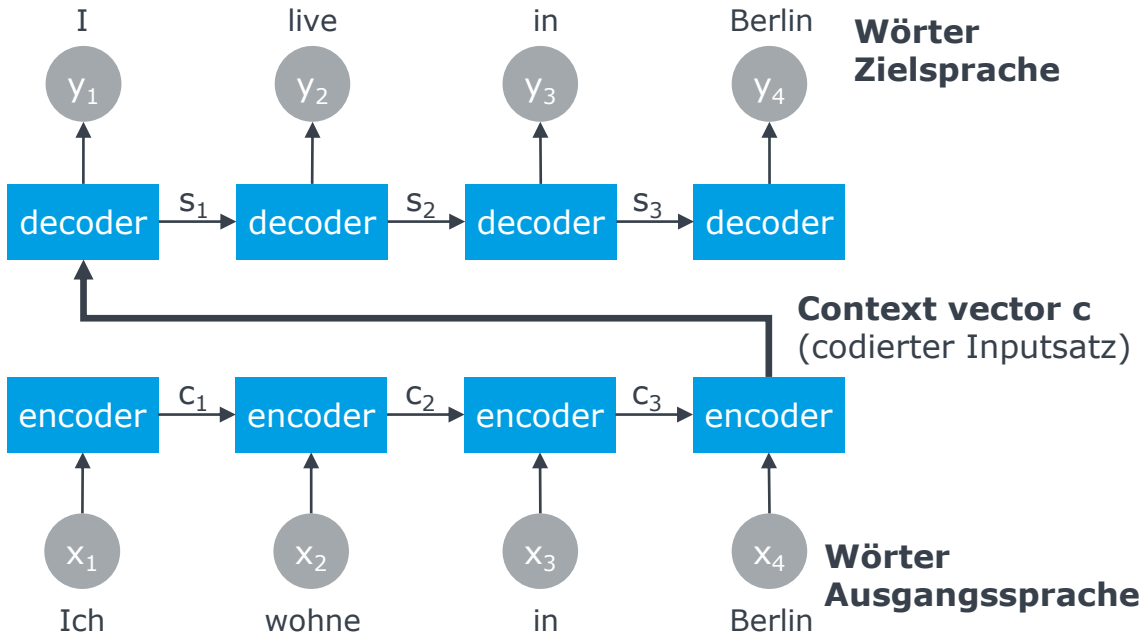
- **Output Gate:** schreibe in Ausgabe y_t Wert für dort, da dieser Begriff für Annas nächsten Satz relevant ist. LSTM lernt also, daß Annas Frage sich auf München bezieht!

Dort
München

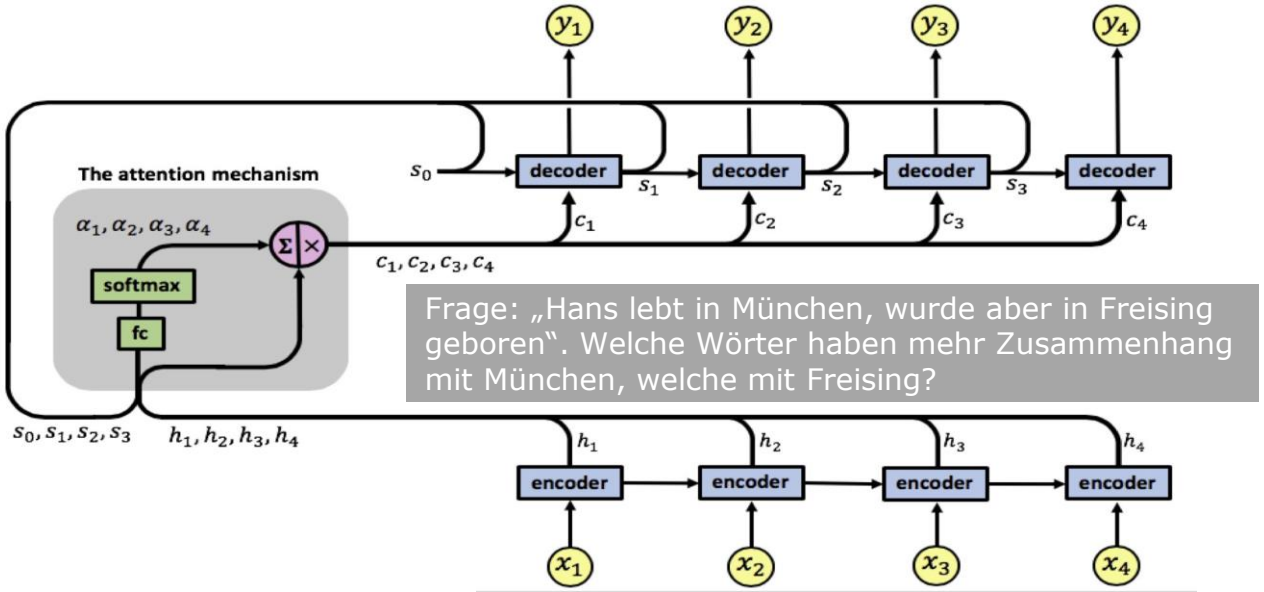
Quellen:
Bild: Link.
LSTM: Hochreiter, Schmidhuber: „Long Short-Term Memory", 1997 sowie Link, Link.
Erklärung Vermeidung Vanishing Gradient: Link

(VEREINFACHTES) ATTENTION – MOTIVATION.

Übersetzen Ausgangs- in Zielsprache



Attention Mechanismus



Herausforderungen/ Probleme anderer Verfahren:

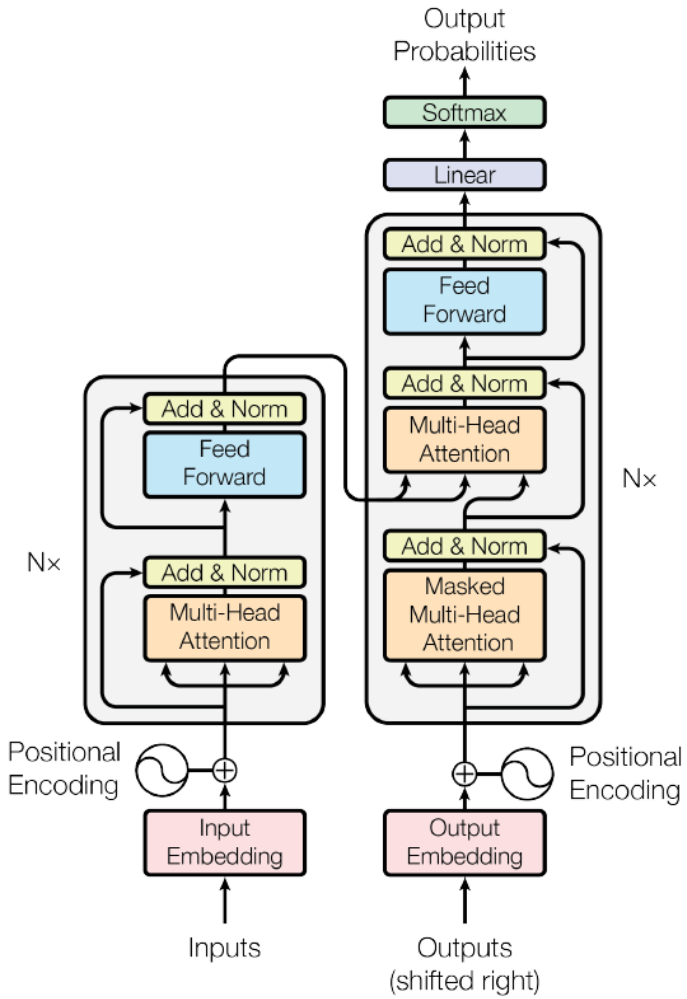
- Basis RNN: Länge Decoder (Zielsprache) festgelegt, was zu Problemen beim Lernen von langen Sätzen führt (umgangssprachlich: Modell „vergißt“ Kontext).
- LSTM: Input erstes Wort für Zielsprache ist letztes Wort Ausgangssprache; dabei sind oft erste Wörter ähnlich.

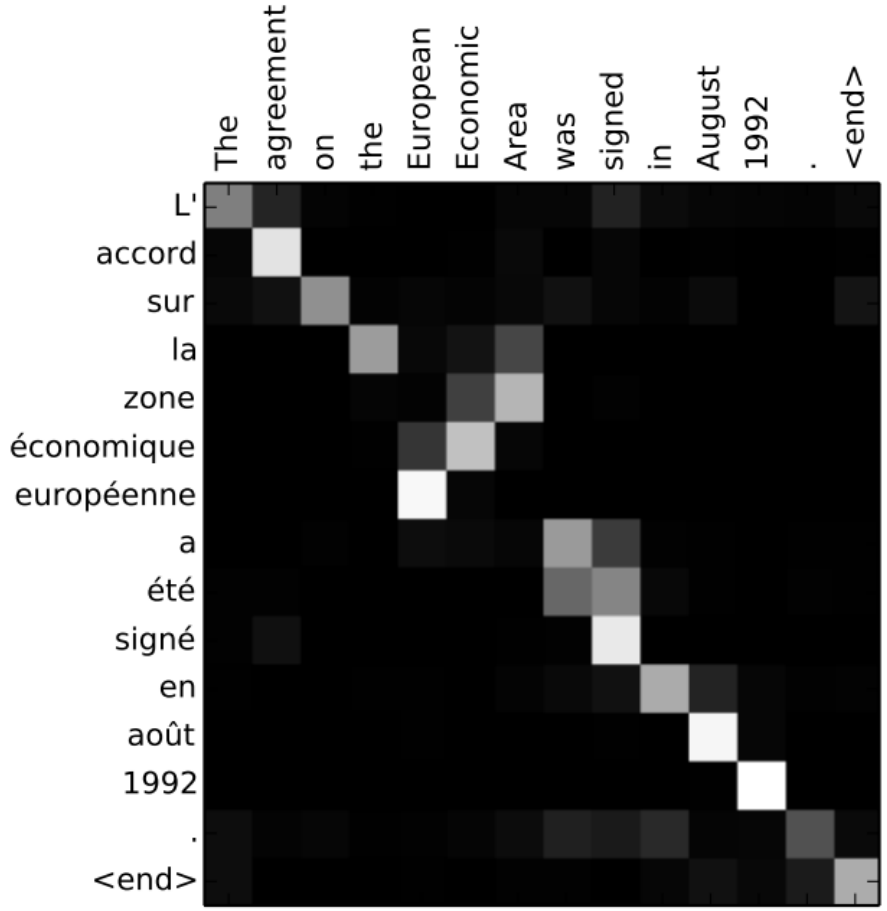
Attention löst die erwähnten Herausforderungen

- Attention ermöglicht Decoder, bestimmte Teile des Inputs stärker zu gewichten bei Vorhersage Output (über c_1 – c_4).
  Lernen dieser Gewichte Attention (a_1 – a_4) per Neuronales Netz.
- Problem Satzlänge wird durch individuelle Gewichte für jeden Output gelöst. Von dieser unterschiedlichen Fokussierung stammt der Name Attention.

Quelle Bild: Link

ATTENTION – VERANSCHAULICHUNG.





Schön zu sehen: Modell berücksichtigt französische Grammatik bei der Übersetzung European Economic Area (im französischen gedreht).

Quelle Bilder: linkes Bild: Vaswani et al.:"Attention is all you need", (2017). Link
rechtes Bild: Link

4. NATURAL LANGUAGE PROCESSING

NATURAL LANGUAGE PROCESSING (NLP) BIETET ENORME POTENTIALE.

Ausgewählte Anwendungsfälle:

- Schrifterkennung (Lesen von Rechnungen, Adressen, …).
- Spracherkennung (Siri, Alexa, …).
- Sprache-zu-Text (automatisierte Untertitel).
- Sentiment Analysis (automatisierte Auswertung Kunden-Feedback).
- Natural Language Understanding (Chatbots, Email-Überwachung, …).
- Natural Language Generation (Automatisierte Berichtserstellung, Schreiben Artikel, …).
- Maschinelles Übersetzen.
- ….

NATURAL LANGUAGE PROCESSING IST EIN SCHWERES THEMA MIT VIELEN HERAUSFORDERUNGEN

Sprachen unterscheiden sich deutlich:

- Satzstellung: Stellung Verb im Satz, Adjektiv zum Nomen, …
- Zeiten (viele Zeiten vs. keine Zeiten).
- Inflektionen wegen Deklinationen/ Kasus (Englisch vs. Russisch vs. Finnisch).
- Akzente und Dialekte.
- Redewendungen („Equal goes it los“ ☺).
- …

Texte unterscheiden sich deutlich:

- Schriften (lateinisches Alphabet, kyrillisch, griechisches Alphabet, Chinesisch/ Japanisch/ Koreanisch [Kanji], …).
- Wortabstand (oder keinen).
- Schriftrichung: links/ rechts (bspw. Hebräisch)
- …

Erkennen Kontext:

- Zusammenhänge zwischen Wörtern (wie Redewendungen).
- Manche Sprachen haben keine Zeiten, sondern nur aus Kontext ersichtlich (Chinesisch, Finnisch).
- Ironie….

BEISPIELHAFTE NLP-TECHNIKEN:

- Text Segmentierung: Aufteilung eines Texts in verschiedene Einheiten (z.B. Wörter, Sätze oder Themen).
- Sentence breaking: Aufspalten Text in Sätze, bspw. anhand Punkt.
- Lemmatization: Vereinheitlichung eines Wortes und seiner Varianten mithilfe Grammatiken (Wörterbuch).
- Stemming: Vereinheitlichung Wort und seiner Varianten durch Abschneiden Endungen (play für played/playing/plays/ …).
- Text/Word Embedding/ vectorization: Umwandlung einzelner Buchstaben/ ganzer Wörter in eindeutige Zahlen.

5. CASE STUDY

WHEN TAYLOR MEETS HELENE…

Use Case 1: Taylor Generator: basierend auf bestehenden Taylor Swift Lyrics automatisiert möglichst ähnliche Texte erstellt (Natural Language Generation)

Use Case 2: Helenes neue Zielgruppe Amerika: Automatisiertes Übersetzen Helene Fischer Lieder in Englisch (**maschinelles Übersetzen)**

When Taylor meets Helene: wir geben übersetzte Helene Texte in den Taylor-Generator ein.

WIEDERHOLUNG GENERISCHER ABLAUF SUPERVISED LEARNING. ÜBERSICHT.

1. Daten organisieren und hochladen
2. Daten aufbereiten/ Data cleaning
3. Daten aufteilen in Test- und Trainingsmenge (sowie ggf. Validierungsmenge)
4. Vorbereitungen: Machine Learning Verfahren wählen, Gewichte initialisieren, Kostenfunktion wählen
5. Training: Schrittweise Optimierung Modellparameter bis Modell möglichst gute Performance für die Trainingsmenge hat
6. Modell(-güte) validieren anhand der Testmenge
7. Deployment: Modell einsetzen im „Live"-Betrieb inkl. kont. Überprüfen Güte Modell und Aktualisierung

WHEN TAYLOR MEETS HELENE
SCHRITT 1: DATEN ORGANISIEREN.

Textgenerierung

"Shake it off"

"I knew you were trouble"

"Tim McGraw"
He said the way my blue eyes shined
Put those Georgia stars to shame that night
I said, "That's a lie"
Just a boy in a Chevy truck
That had a tendency of gettin' stuck
On back roads at night
And I was right there beside him all summer long
And then the time we woke up to find that summer gone

Automatisiertes Abgreifen der Texte aller Lieder und Speichern als Textdatei

Maschinelles Übersetzen

Tab-delimited Bilingual Sentence Pairs

These are selected sentence pairs from the Tatoeba Project.

Updated: 2020-08-23

- Afrikaans - English afr-eng.zip (807)
- Albanian - English sqi-eng.zip (451)
- Algerian Arabic - English arq-eng.zip (155)
- Arabic - English ara-eng.zip (11446)
- Assamese - English asm-eng.zip (1655)
- Azerbaijani - English aze-eng.zip (2208)
- Basque - English eus-eng.zip (677)
- Belarusian - English bel-eng.zip (3825)
- Bengali - English ben-eng.zip (4342)
- Berber - English ber-eng.zip (120079)
- Bosnian - English bos-eng.zip (102)
- Bulgarian - English bul-eng.zip (14979)
- Burmese - English mya-eng.zip (347)
- Cantonese - English yue-eng.zip (3213)
- Catalan - English cat-eng.zip (665)
- Cebuano - English ceb-eng.zip (205)
- Central Dusun - English dtp-eng.zip (1521)
- Chavacano - English cbk-eng.zip (1905)

Introducing Anki

- If you don't already use Anki, vist the website at http://ankisrs.net/ to download this free application for Macintosh, Windows or Linux.

About These Files

- Any flashcard program that can import tab-delimited text files, such as Anki (free) can use these files.
- **Warning!** There are errors in the Tatoeba Corpus. (Detailed Warning)
- **In order to minimize the number of errors**, I only used sentences that were owned by identified native speakers working on the Tatoeba Project and English sentences that I've personally checked and did not reject.
- **Warning!** Please remember that even doing this may not have eliminated all errors.

How the Data Looks

English + TAB + The Other Language + TAB + Attribution

```
This work isn't easy.    この仕事は簡単じゃない。    CC-BY 2.0 (France) Attributi
Those are sunflowers.    それはひまわりです。        CC-BY 2.0 (France) Attributi
Tom bought a new car.    トムは新車を買った。        CC-BY 2.0 (France) Attributi
This watch is broken.    この時計は壊れている。      CC-BY 2.0 (France) Attributi
```

The attribution gets imported into Anki as a tag, by default This attribution contains the

Herunterladen Deutsch – Englisch „Wörterbuch" mit ~221'000 Deutschen Sätzen.

WHEN TAYLOR MEETS HELENE
SCHRITT 2: DATEN SÄUBERN.

Textgenerierung

Embedding auf Basis Buchstaben

He said the way my blue eyes shined
Put those Georgia stars to shame that night
I said, "That's a lie"

he said the way my blue eyes shined
put those georgia stars to shame that night
i said thats a lie

{0: ' ', 1: 'a', 2: 'b', 3: 'c', 4: 'd', 5: 'e', 6: 'f', 7: 'g', 8: 'h', 9: 'i', 10: 'j', 11: 'k', 12: 'l', 13: 'm', 14: 'n', 15: 'o', 16: 'p', 17: 'q', 18: 'r', 19: 's', 20: 't', 21: 'u', 22: 'v', 23: 'w', 24: 'x', 25: 'y', 26: 'z'}

Maschinelles Übersetzen

Embedding auf Basis Wörter

Tom kommt heute.

tom kommt heute

Input Language embedding
1 ----> <start>
5 ----> tom
144 ----> kommt
114 ----> heute
3 ----> .
2 ----> <end>

Tom is coming today.

Tom is coming today

Target Language embedding
1 ----> <start>
5 ----> tom
8 ----> is
185 ----> coming
133 ----> today
3 ----> .
2 ----> <end>

- Alle Wörter in Kleinschreibung umwandeln
- Sonderzeichen entfernen
- Embedding für Wörter/ Buchstaben: zuweisen eindeutige Nummer

AUTOMATISIERTES ERSTELLEN TAYLOR SWIFT SONGS. SCHRITT 3: DATEN AUFTEILEN IN TRAININGS- UND TESTMENGE.

Textgenerierung

```
seq_length = 100

Data_X_Taylor = []
Data_y_Taylor = []

for i in range(0, n_chars_Taylor - seq_length, 1):
    # Input sequence(will be used as samples)
    seq_in  = Cleaned_Taylor_lyrics[i:i+seq_length]

    # Output sequence (will be used as target)
    seq_out = Cleaned_Taylor_lyrics[i + seq_length]

    # Store samples in data_X
    Data_X_Taylor.append([chars2int_Taylor[char] for char in seq_in])

    # Store targets in data_y
    Data_y_Taylor.append(chars2int_Taylor[seq_out])

n_patternsTaylor = len(Data_X_Taylor)

print( 'Total Patterns : ', n_patternsTaylor)
```

```
Total Patterns :  444330
```

```
# Vektor anpassen damit er in LSTM RNN eingespeist werden kann
X_Taylor = np.reshape(Data_X_Taylor, (n_patternsTaylor, seq_length, 1))

# Normalizing input dat
X_Taylor = X_Taylor/ float(n_vocab_Taylor)

# One hot encode the output targets :
y_Taylor = np_utils.to_categorical(Data_y_Taylor)
```

```
print(X_Taylor.shape[1], X_Taylor.shape[2])
```

Verhältnis 70% - 30%

Maschinelles Übersetzen

```
def load_dataset(path, num_examples=None):
  # creating cleaned input, output pairs
  #targ_lang, inp_lang = create_dataset(path, num_examples)
  eng, deu = create_dataset(path, num_examples)

  input_tensor, inp_lang_tokenizer = tokenize(deu)
  target_tensor, targ_lang_tokenizer = tokenize(eng)

  return input_tensor, target_tensor, inp_lang_tokenizer, targ_lang_tokenizer
```

```
num_examples = 50000
input_tensor, target_tensor, inp_lang, targ_lang = load_dataset(path2file, num_examples)

# Calculate max_length of the target tensors
max_length_targ, max_length_inp = target_tensor.shape[1], input_tensor.shape[1]
```

```
# Creating training and validation sets using an 80-20 split
input_tensor_train, input_tensor_val, target_tensor_train, target_tensor_val = train_test_split(input_tensor, target_tensor, test_size=0.2)

# Show length
print(len(input_tensor_train), len(target_tensor_train), len(input_tensor_val), len(target_tensor_val))
```

```
40000 40000 10000 10000
```

Verhältnis 80% - 20%

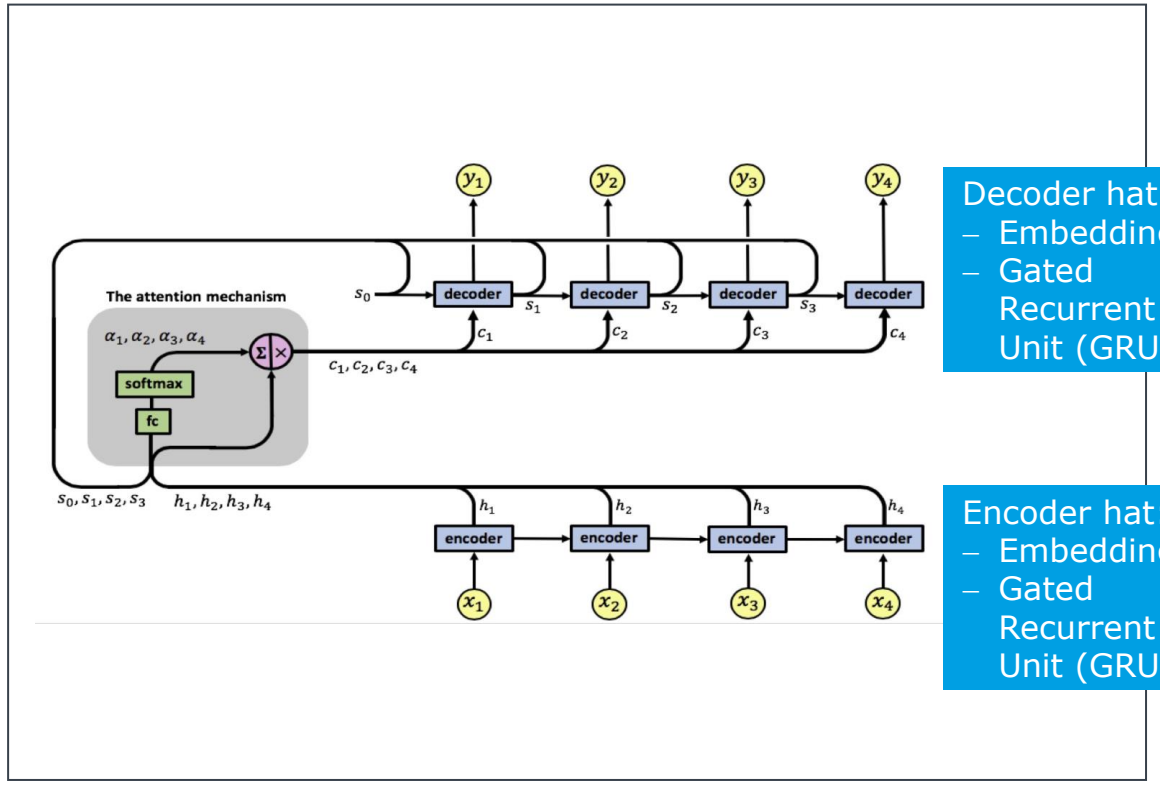
AUTOMATISIERTES ERSTELLEN TAYLOR SWIFT SONGS. SCHRITT 4: MACHINE LEARNING VERFAHREN WÄHLEN.

Textgenerierung

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_4 (LSTM)                (None, 100, 256)          264192
_________________________________________________________________
dropout_4 (Dropout)          (None, 100, 256)          0
_________________________________________________________________
lstm_5 (LSTM)                (None, 100, 256)          525312
_________________________________________________________________
dropout_5 (Dropout)          (None, 100, 256)          0
_________________________________________________________________
lstm_6 (LSTM)                (None, 100, 256)          525312
_________________________________________________________________
dropout_6 (Dropout)          (None, 100, 256)          0
_________________________________________________________________
lstm_7 (LSTM)                (None, 100, 256)          525312
_________________________________________________________________
dropout_7 (Dropout)          (None, 100, 256)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 25600)             0
_________________________________________________________________
dense_1 (Dense)              (None, 29)                742429
_________________________________________________________________
activation_1 (Activation)    (None, 29)                0
=================================================================
Total params: 2,582,557
Trainable params: 2,582,557
Non-trainable params: 0
_________________________________________________________________
```

4 gekoppelte LSTM mit jeweils einem Dropout zur Reduzierung Overfit

Maschinelles Übersetzen



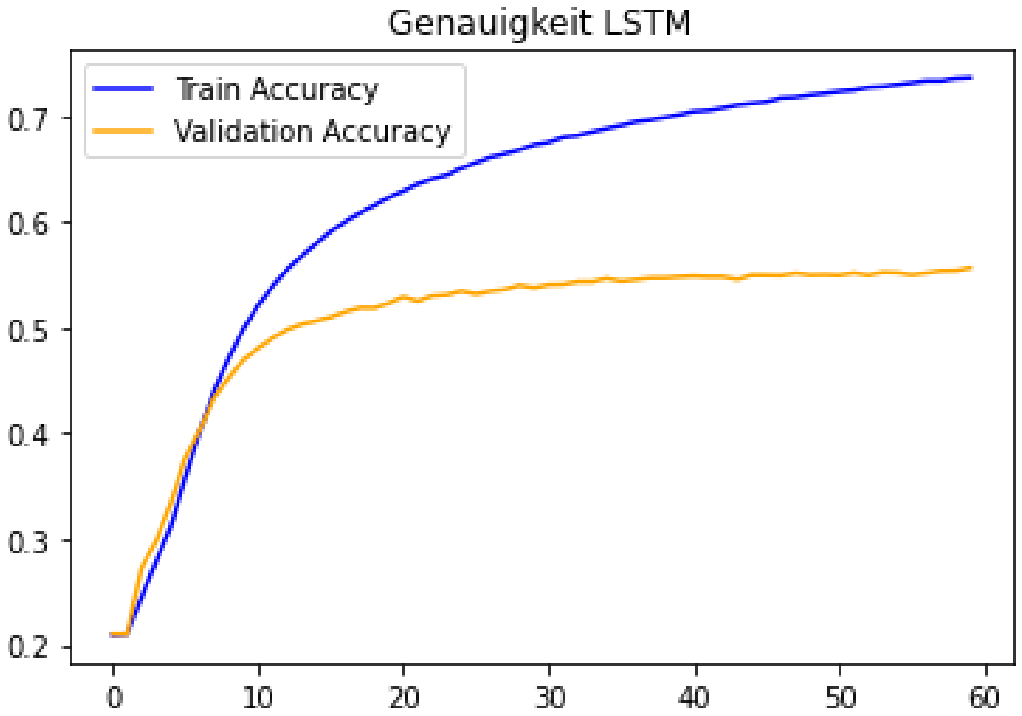
Decoder hat:
- Embedding
- Gated Recurrent Unit (GRU)

Encoder hat:
- Embedding
- Gated Recurrent Unit (GRU)

Vereinfachtes Beispiel: 1 GRU für En/Decoder und Attention mit 10x Fully Connected Layer

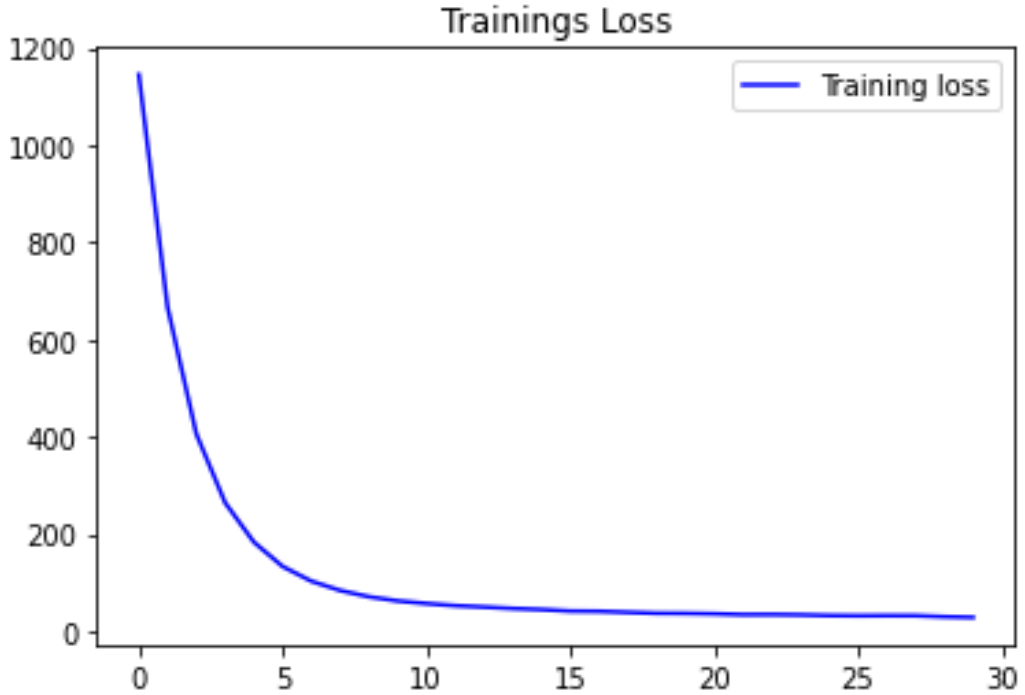
AUTOMATISIERTES ERSTELLEN TAYLOR SWIFT SONGS. SCHRITT 5: TRAINING

Textgenerierung



60 Epochen

Maschinelles Übersetzen



30 Epochen

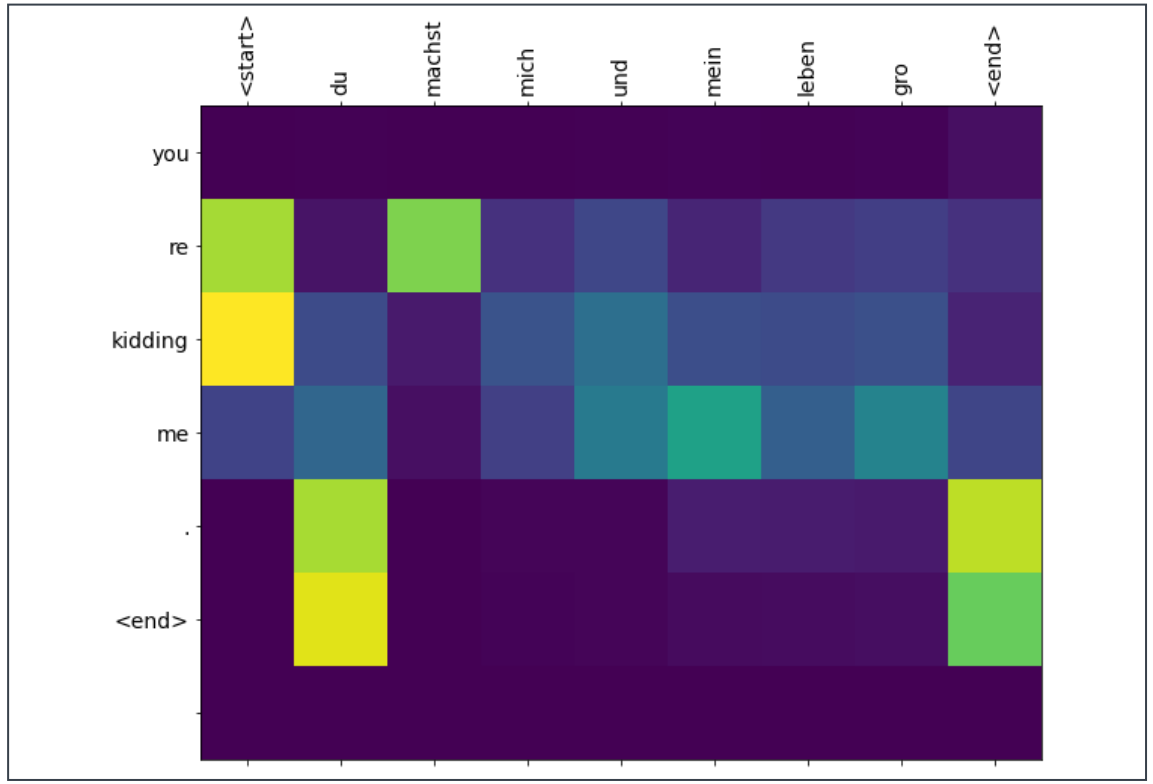
AUTOMATISIERTES ERSTELLEN TAYLOR SWIFT SONGS. SCHRITT 6: VALIDIERUNG MODELLGÜTE PER TESTDATEN.

Textgenerierung

e because i mean the crowd was going so loud that if he would have said no they would have probably " boy i had the time i was an american gor i lifht be the things they did to say that i dont know i want to be alone i need you and i can see you staring in the world was riding in your all and you dont leave me like the way you want me with me and you were the way you want to walk away the way at wo Done

hate hate hate hate hate baby im just gonna shake shake shake shake shake i shake it off i shake it off heartbreakers gonna break break break break break and the fakers gonna fake fake fake fake fake baby im just gonna shake shake shake shake shake i shake it off i shake it off heartbreakers gonna break break break break break and the fakers gonna fake fake fake fake fake baby im just gonna shake shake shake shake shake i shake it off i shake it off heartbreakers gonna break break break break br Done

Maschinelles Übersetzen



Qualität bei Textgenerierung und Übersetzen nicht ausreichend; deutlich längeres Trainieren als 60 Epochen notwendig

4. ZUSAMMENFASSUNG

QUELLEN:

Künstliche Intelligenz:

- Gröner, Heinecke: Kollege KI
- Burkov: The Hundred-Page Machine Learning Book, online verfügbar unter Link
- Nielsen: Neural Networks and Deep Learning, online verfügbar unter Link
- Russel, Norvig: Artificial Intelligence – a modern approach

Web-Links:

- https://karpathy.github.io/2015/05/21/rnn-effectiveness/
- https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- https://distill.pub/2018/building-blocks/
- Attention: http://mlexplained.com/2017/12/29/attention-is-all-you-need-explained/
- https://weberna.github.io/blog/2017/11/15/LSTM-Vanishing-Gradients.html
- https://r2rt.com/written-memories-understanding-deriving-and-extending-the-lstm.html

Notebooks:

- Sprache zu Text: https://colab.research.google.com/drive/1qFt8qxKtM05hRuRxsA1Lq4JtP7tstcgc