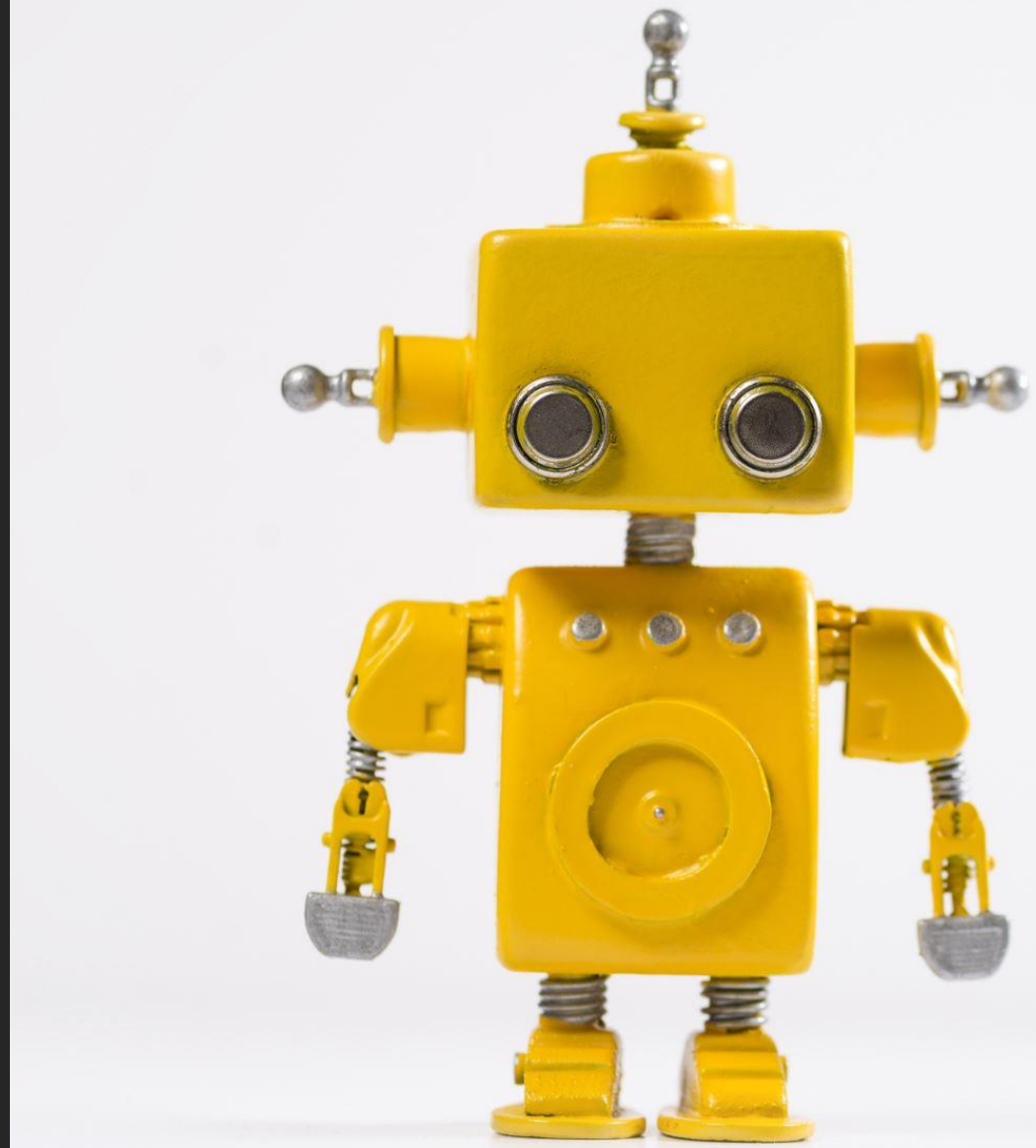
LERNENDE AGENTEN

NATURAL LANGUAGE PROCESSING

Betreuer: Prof. Dr. Michael Neitzke
Erstellt und präsentiert von:
Ibrahem Alkurdi
Ali Sheikh Hassan
Muhammad Adnan



GLIEDERUNG:

- Question Answering (SQuAD)
- Chatbot & Translator
- Training on Cloud

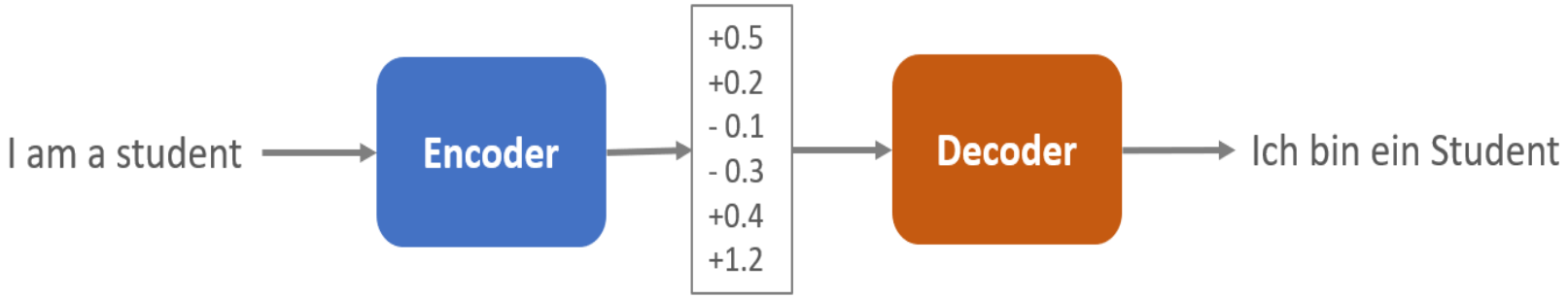
Stanford Question Answering Dataset
(reading comprehension dataset)

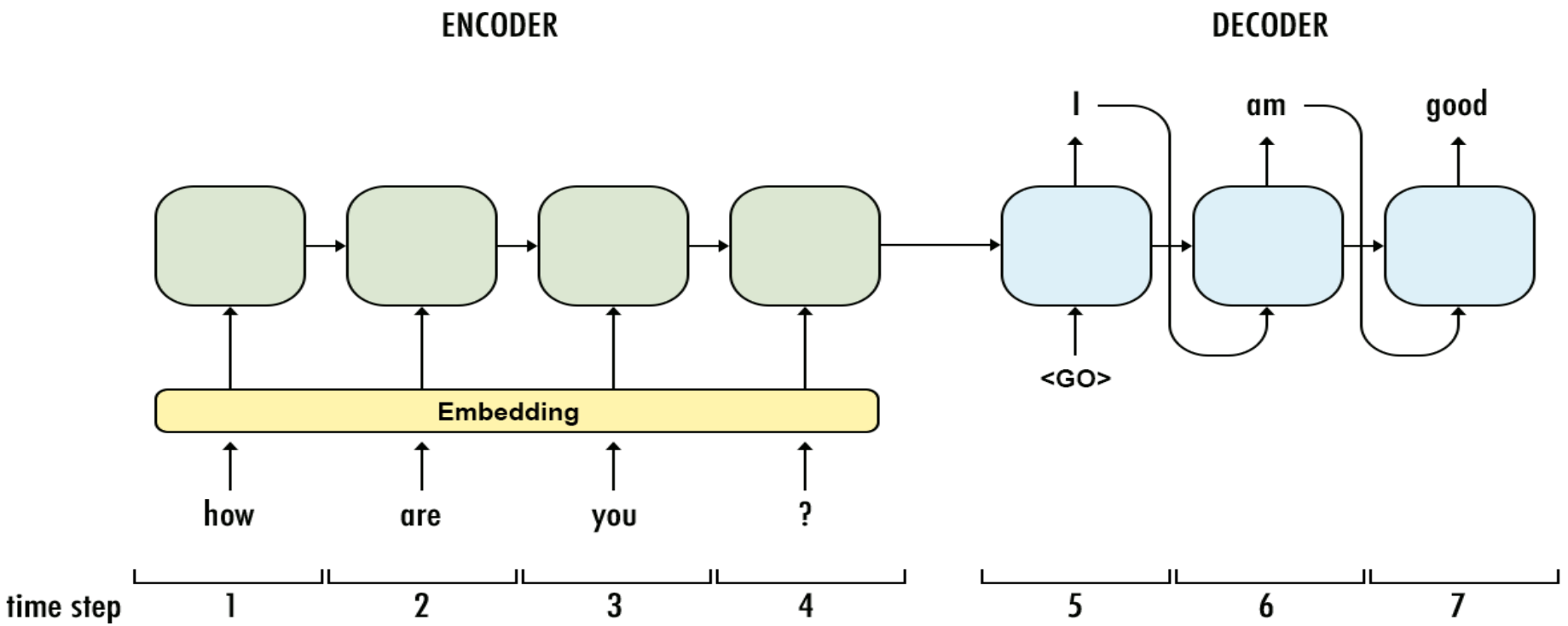
Sentence having the right answer

'context': 'Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (bor
n September 4, 1981) is an American singer, songwriter, record producer an
d actress. Born and raised in Houston, Texas, she performed in various sin
ging and dancing competitions as a child,and rose to fame in the late 1990
s as lead singer of R&B girl-group Destiny\'s Child. Managed by her father
, Mathew Knowles, the group became one of the world\'s best-selling girl g
roups of all time. Their hiatus saw the release of Beyoncé\'s debut album,
Dangerously in Love (2003), which established her as a solo artist worldwi
de, earned five Grammy Awards and featured the Billboard Hot 100 number-on
e singles "Crazy in Love" and "Baby Boy".',
'text': 'in the late 1990s'
'question': 'When did Beyonce start becoming popular?'

Exact Answer

Sequence to Sequence (Encoder Decoder)





Question Answering Answering

- **Train data** : 507 , **validation data** : 57 and 18 **different topics**
- **For Example:**
  - **Topic:** artificial intelligence
    - - What language are you written in?
      - - Python.
      - - I am written in Python.
  - **Topic:** emotion
    - - Can you feel?
      - - Maybe I can. I am a fairly sophisticated piece of software.
    - - Do you feel?
      - - I may have. In some respects, I am programmed to emulate feelings.

Seq2Seq with embedding layer word by word

- **Word Embedding:** ist eine Lerntechnik in Natural Language Processing (NLP), bei denen Wörter oder Ausdrücke aus dem Vokabular auf Vektoren reeller Zahlen abgebildet werden.

- **Train data**: 564 pair and 18 **different topics**

- - What language are you written in?
  - Python.
  - I am written in Python

- **Results:** **from data trained on**

- - hello
- greetings
- - who are you
- i am on the internet
- - what is AI
- artificial intelligence is the branch of engineering and science devoted to constructing machines that think
- - what is your location
- everywhere

not trained on

- - hello my friend
- bots are a lot of fun bots are the life of the party
- - where are you now
- i am a computer i can not
- - what is a human
- a fancy name for applied computer science in biology

with pre trained embedding layer word by word

From data trained on

- - hello

- hi

- - who are you

- i am just an artificial intelligence

- - what is AI

- artificial intelligence is the branch of engineering and science devoted to constructing machines that think

- - what is your location

- everywhere

- - what is an operating system

- software that coordinates between the hardware and other parts of the computer to run other software is called an operating system or the os windows macos linux unix all of them are types of oses android and ios are operating systems for mobile devices software which implements the basic functions of a computer such as memory access processes and peripheral access

Not trained on

- - hello my friend

- i have a lot of your case

- - where are you now

- i am from where all software programs are from a galaxy far far away

- - ask me something

- what do you get when you cross a cat with a lemon

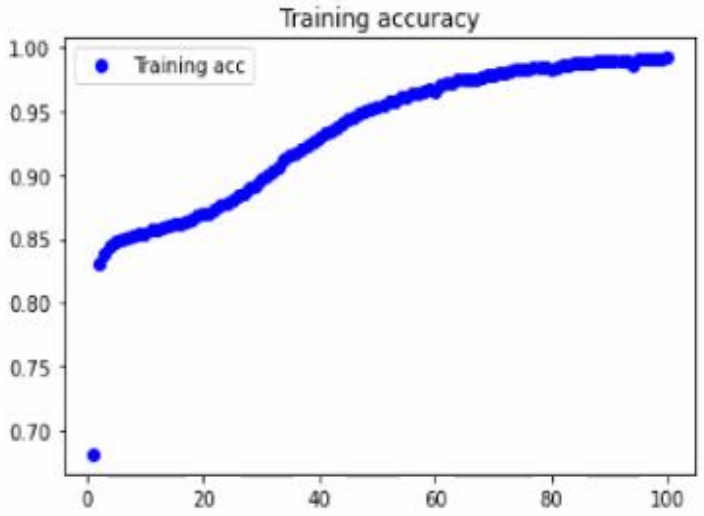
- - what is a human

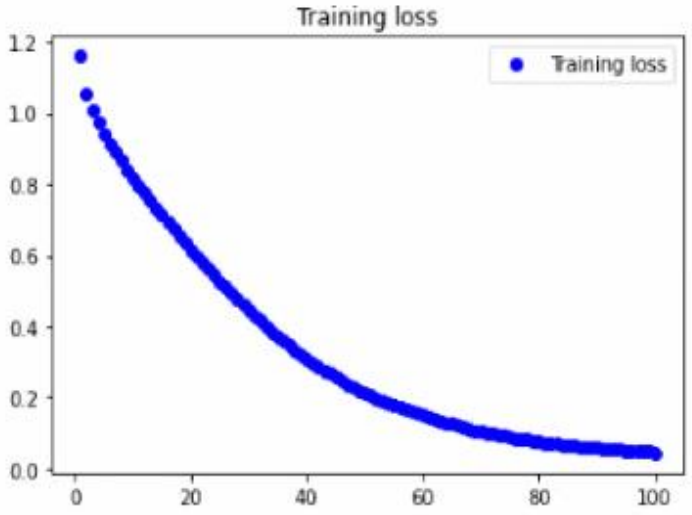
- an established system of political administration by which a nation state district etc is governed

COMPARE between with pre trained embedding and without

- Pre trained (using Glove 100d)

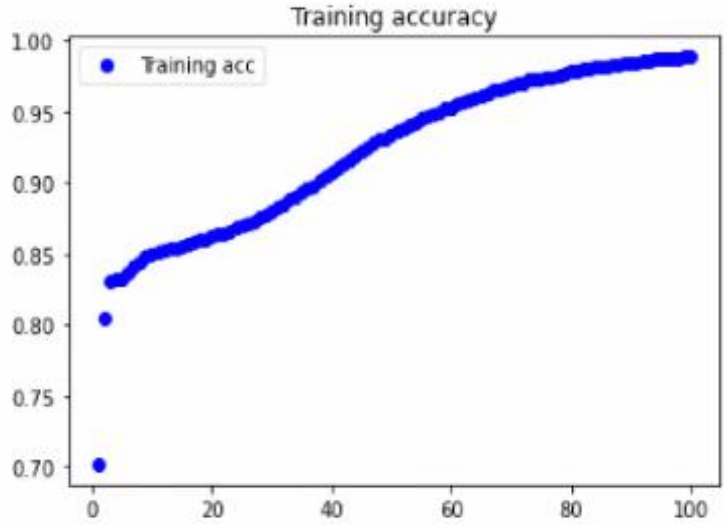
```
Total params: 1,572,473
Trainable params: 1,204,273
Non-trainable params: 368,200
```

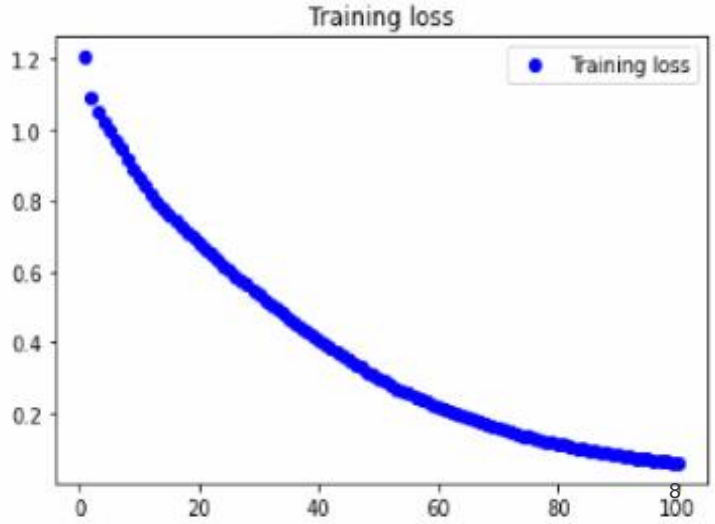




- Not pre trained

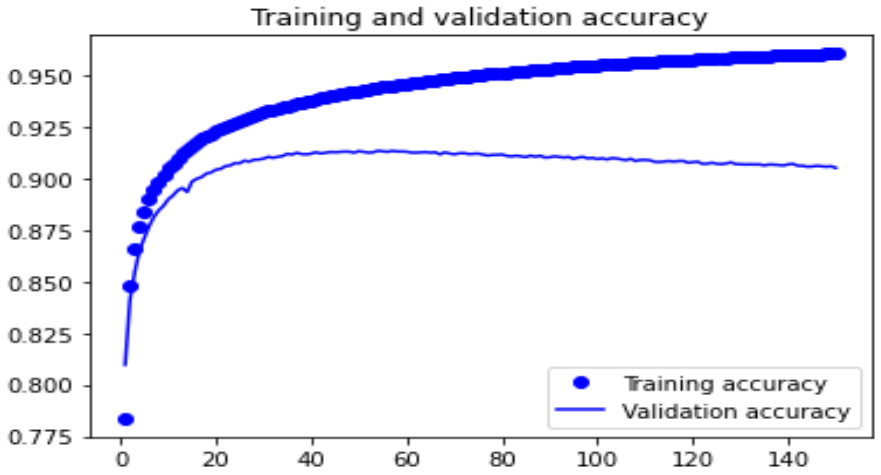
```
Total params: 1,572,473
Trainable params: 1,572,473
Non-trainable params: 0
```

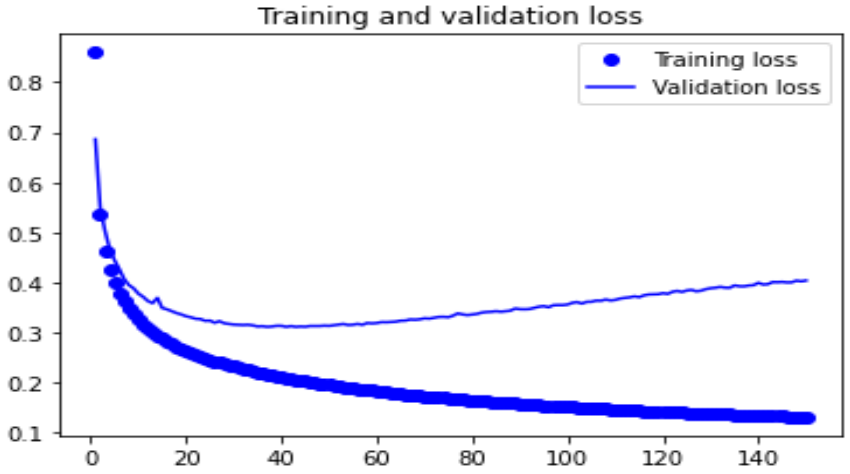




Translator with Seq2Seq Char by Char

- **All data** 25000 pair from 208 thousand.
- **Train data** 90%
- **Validation data** 10% => 2500 pair
- 150 **epochs**
- 2.5 hours of train on GPU






```
Total params: 228,573
Trainable params: 228,573
Non-trainable params: 0
```

```
Epoch 150/150
22500/22500 [==============================] - 58s 3ms/step - loss: 0.1310 - accuracy:
0.9609 - val_loss: 0.4039 - val_accuracy: 0.9053
```

The difference between Seq2Seq word by word and Seq2Seq char by char

- **Word by word:**
  - Good / very good results
  - Overfitting is harder because of big dimensions
  - Problems with RAM
  - For example:
    - 50000 sentences dataset
    - Vocabulary is to big (maybe 5000 different words)
    - For each target sentence we need a vector with len(fixed length) * vocabulary number

- **Character by character:**
  - Bad / not to good results
  - Overfitting is easy because of to small dimensions especially when dataset is not big
  - No Problems with RAM
  - For example:
    - 50000 sentences dataset
    - Vocabulary / characters number is small (max 200 different characters)
    - For each target sentence we need a vector with len(fixed length) * characters number

BLEU Score (bilingual evaluation understudy)

- **BLEU Scores** messen den Unterschied zwischen menschlichen und maschinellen Übersetzungen. Verwendet wird hierfür ein vergleichsweise simpler Algorithmus.

- Example:
  - Reference 1: the cat is on the mat
  - Reference 2: there is a cat on the mat
  - Candidate: the cat the cat on the mat

Unigram	score	Bigram	score
the	1	the cat	1
cat	1	cat the	0
the	1	the cat	1
cat	1	cat on	1
on	1	on the	1
the	1	the mat	1
mat	1		
BLEU Score	7/7=1		5/6= 0.83

```
src=[Tom, wake up.], target=[Steh auf, Tom!], predicted=[Tom, kosten Sie es zu sim!]
src=[It rained heavily.], target=[Es regnete heftig.], predicted=[Es regnete spät.]
src=[No kidding?], target=[Wirklich?], predicted=[Echt?]
src=[This is my coach.], target=[Das ist mein Trainer.], predicted=[Das ist mein Bleistift.]
src=[They work hard.], target=[Sie strengen sich an.], predicted=[Sie wusst einen Versuch zun mich geben.]
src=[I was sent home.], target=[Ich wurde nach Hause geschickt.], predicted=[Ich war sehr müde.]
src=[Clean that window.], target=[Mach das Fenster sauber.], predicted=[Mach das Fenster sauer.]
src=[Show it to me.], target=[Zeig ihn mir.], predicted=[Zeig es mir.]
src=[She is beautiful.], target=[Sie ist schön.], predicted=[Sie ist schwach.]
src=[The lovers kissed.], target=[Die Liebenden küssten sich.], predicted=[Die Ristecht zu sicher.]
##############################################
[['Steh', 'auf,', 'Tom!']]
['Tom,', 'kosten', 'Sie', 'es', 'zu', 'sim!']
 1-gram score1: 0.447032
 2-gram score1: 0.280576
 3-gram score1: 0.121251
 4-gram score1: 0.030230
 4-gram score2: 0.146429
```

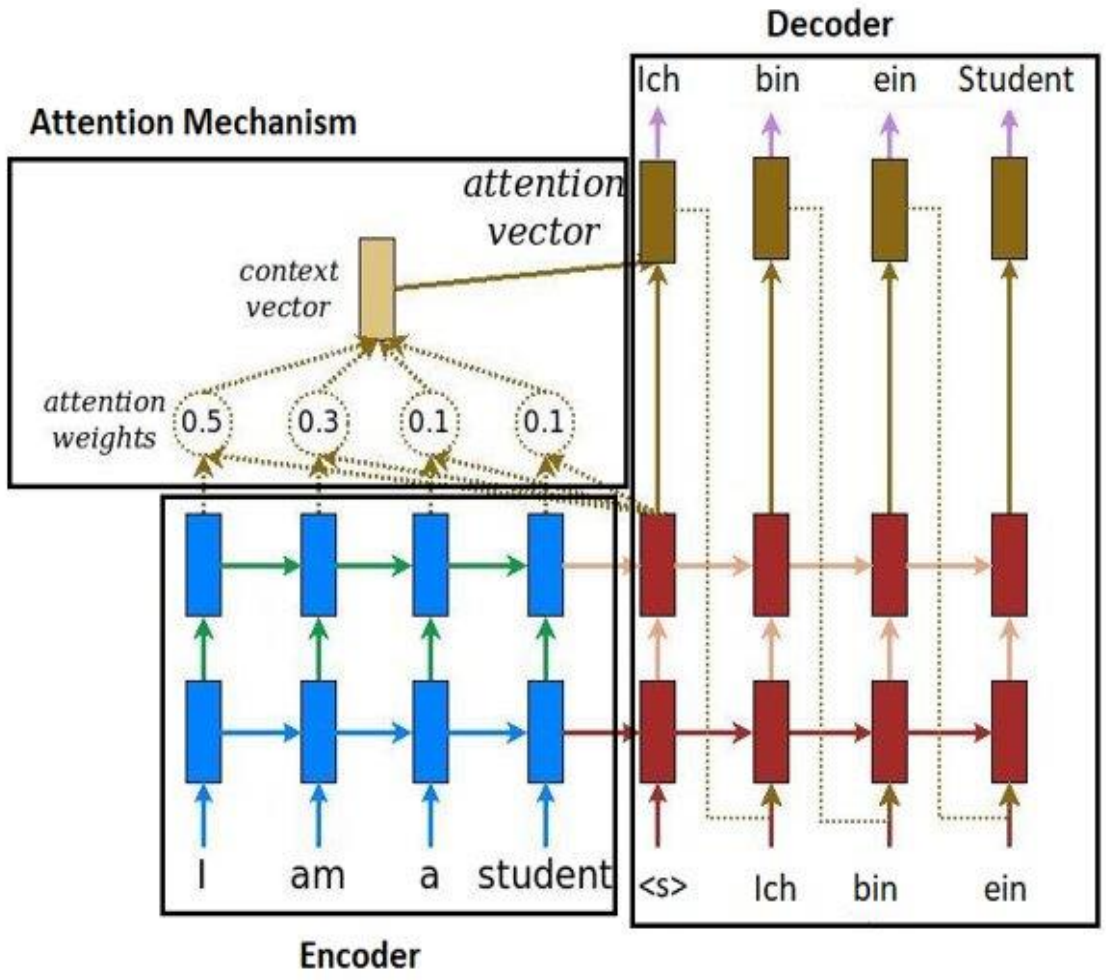
- **BLEU Score for training data**

- **BLEU Score for validation data**

```
src=[Tom does like you.], target=[Tom mag dich.], predicted=[Tom hat nur eine Arbeit gefunden.]
src=[Tom does love you.], target=[Tom liebt dich wirklich.], predicted=[Tom hat das Beschend gefunden.]
src=[Tom doesn't agree.], target=[Tom stimmt nicht zu.], predicted=[Tom ertrinkte sich.]
src=[Tom doesn't agree.], target=[Tom ist nicht einverstanden.], predicted=[Tom ertrinkte sich.]
src=[Tom doesn't drink.], target=[Tom trinkt nicht.], predicted=[Tom hat nicht geschickt.]
src=[Tom doesn't drive.], target=[Tom fährt nicht.], predicted=[Tom hat es nicht gesehen.]
src=[Tom doesn't drive.], target=[Tom fährt nicht Auto.], predicted=[Tom hat es nicht gesehen.]
src=[Tom doesn't shave.], target=[Tom rasiert sich nicht.], predicted=[Tom hat nicht geschummelt.]
src=[Tom doesn't smoke.], target=[Tom raucht nicht.], predicted=[Tom hat nicht gekommen.]
src=[Tom doesn't snore.], target=[Tom schnarcht nicht.], predicted=[Tom hat nicht geschickt.]
##############################################
[['Tom', 'mag', 'dich.']]
['Tom', 'hat', 'nur', 'eine', 'Arbeit', 'gefunden.']
 1-gram score1: 0.316263
 2-gram score1: 0.148603
 3-gram score1: 0.037108
 4-gram score1: 0.002657
 4-gram score2: 0.046397
```

Attention

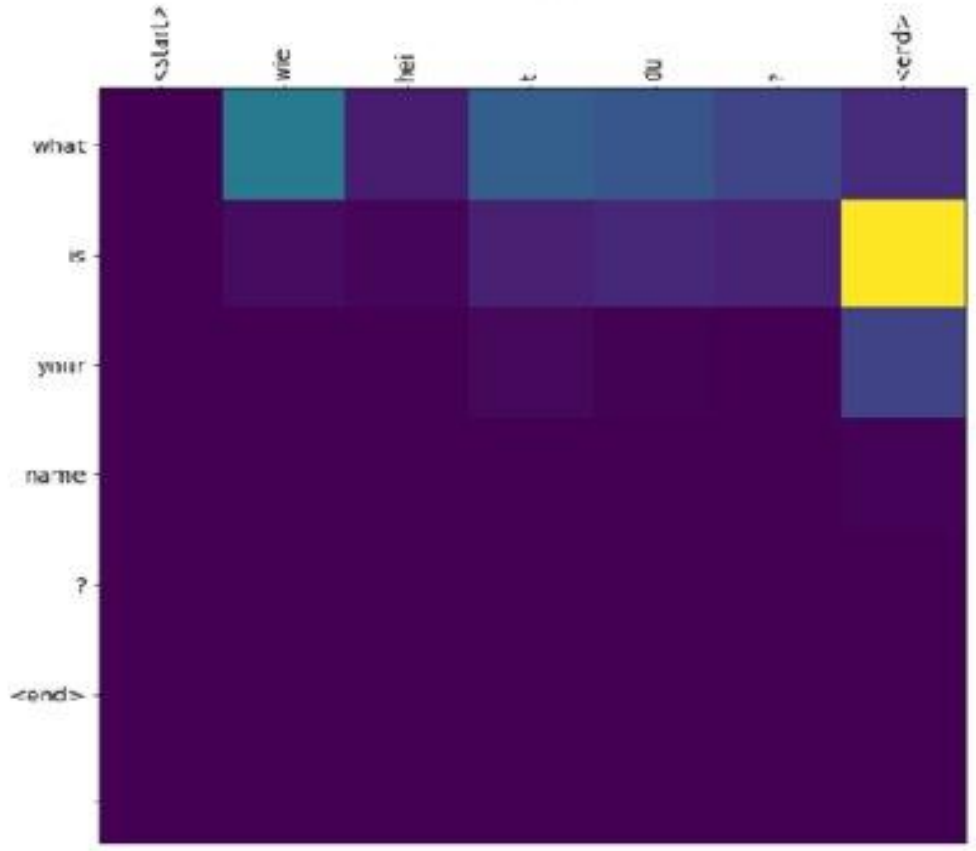
- **Attention Mechanism (Aufmerksamkeitsmechanismus)** ist ein Versuch, die Aktion zu implementieren, indem sie sich selektiv auf einige relevante Dinge konzentriert und andere in tiefen neuronalen Netzen ignoriert.



Visualizing Attention

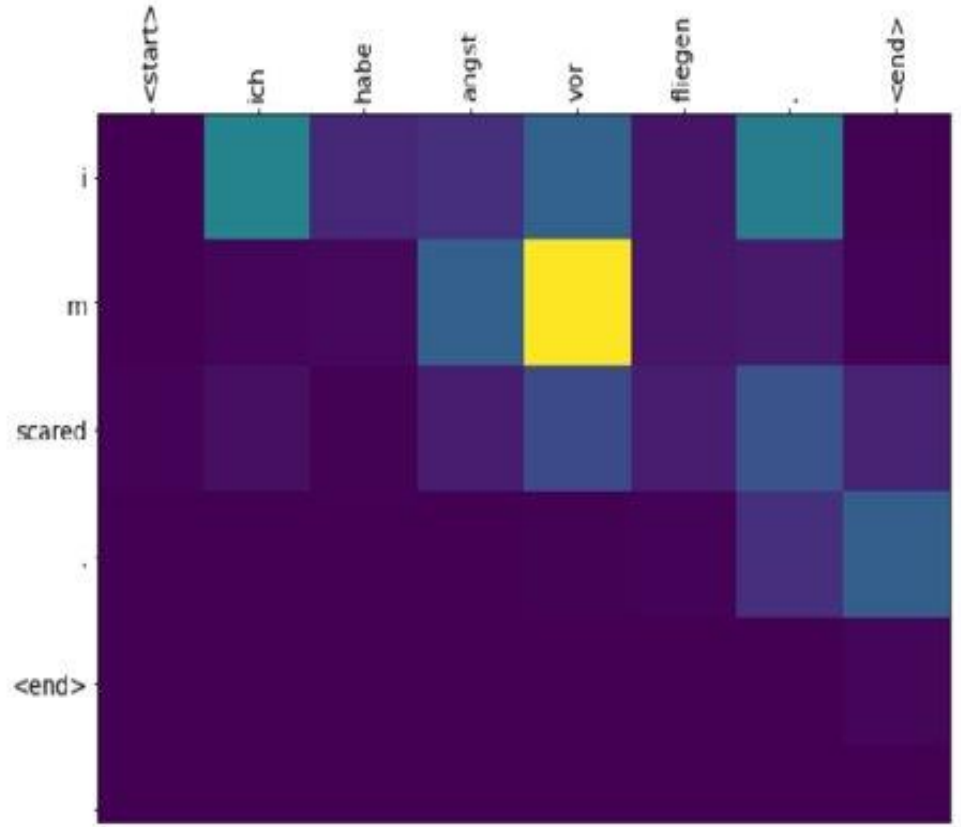
```
[ ] translate(u'Wie heißt du?')
```

```
Input: <start> wie hei t du ? <end>
Predicted translation: what is your name ? <end>
```



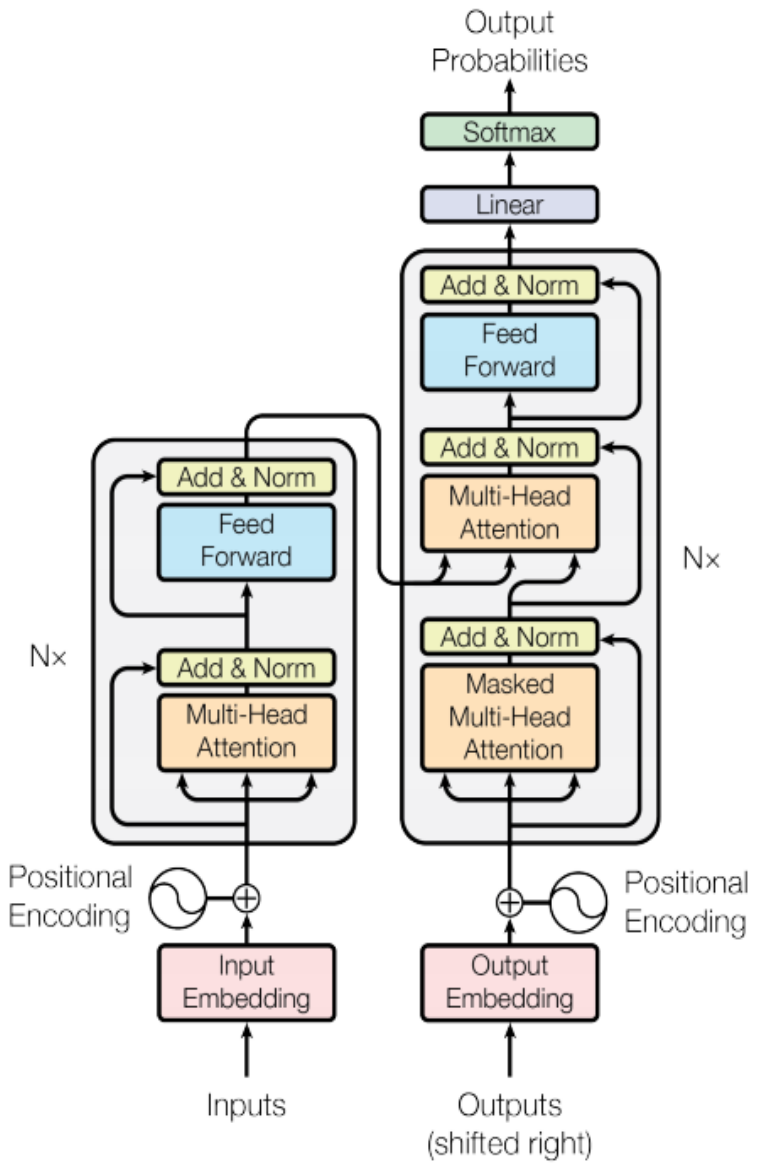

```
[ ] translate(u'Ich habe Angst vor fliegen.')
```

```
Input: <start> ich habe angst vor fliegen . <end>
Predicted translation: i m scared . <end>
```



TRANSFORMER

state of the art



Custom prompt

it was a nice day in the summer, when I saw

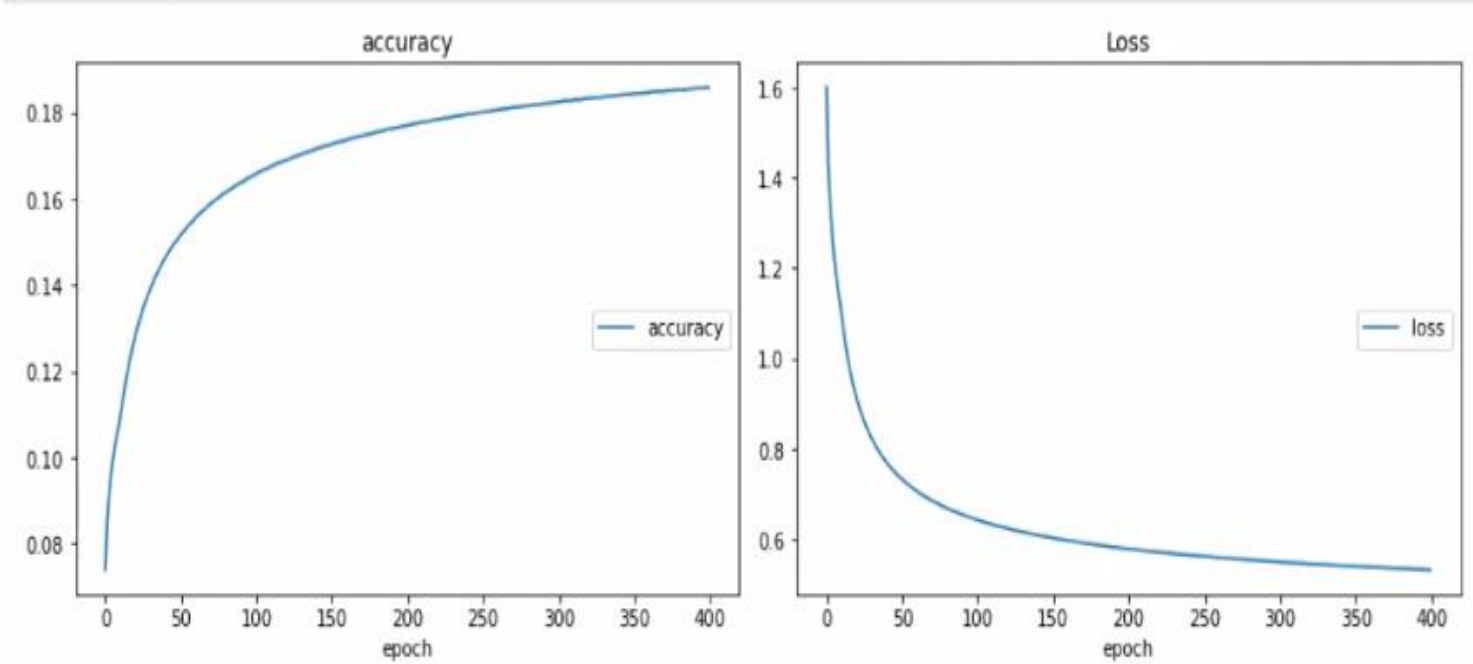
GENERATE ANOTHER

Completion

it was a nice day in the summer, when I saw a great giant "white horse" with a curved neck above the crown. It's a very elaborate coat, with many straps around it. Perhaps it's a good time to have another talk about image, about this very kind image which had everyone making it. When I came to this country from the Silesian Pomerania, they had no images on any of their doors and windows. All of them were painted with flowers and birds. When I saw the figure of a white horse with a curving neck, it struck me that I understood a lot of what I'm having with these demonstrations, why it's necessary to fight against certain things. Every male child has a little fairy, a girl

Training Transformer with movies data

- 221 thousand pair of **training data**
- 400 **epochs**
- 8 hours of training on **GPU**
- Example of training data
  - ZOZO: I saw Gregoire make a call, sir?
  - PAUL: Let us remember this night and tell the world that even in hell there are good people.
  - ZOZO: Thank you, sir.
  - PAUL: You are my family now, Zozo, my brother. I will get you out of here.
  - ZOZO: Many beers.
  - PAUL: How much beer?
  - ZOZO: Sir, Gregoire has been taking beers.




```
accuracy
        accuracy                (min:    0.074, max:    0.186, cur:    0.186)
Loss
        loss                    (min:    0.532, max:    1.601, cur:    0.532)
380/380 [==============================] - 51s 133ms/step - loss: 0.5317 - accuracy: 0.1858
```

```
Total params: 8,963,877
Trainable params: 8,963,877
Non-trainable params: 0
```

Results

```
start chatting -->
enter --exit-- to stop chatting
- - Hi
- i m sorry . i m sorry .
- - sorry about what?
- i don t know . i guess i m going to have to go home .
- - why
- i m not asking you to answer .
- - I know that
- i m sorry , i don t know what i m talking about .
- - I don't know too. Tell me what is going on
- i m not the one who tried to kill you .
- - who then?
- i don t know .
- - I need to know that
- i m not going to tell you anything else .
- - please tell me something
- i don t know what you re talking about .
- - I am talking about the one, who wants to kill me.
- i know what to do . i thought you were going to let me talk to you into this .
- - I am letting you talk about it
- forget it . i m not going to swim in my sleep .
- - are you a crazy ?
- i think so .
- - yes I think so too
- i m not a loser at all . i m a writer . i m afraid of what i m doing here .
- - I didn't know that
- i m sorry i didn t know you were the one who wanted to know what was going on about it .
- - okay
- i m going to be a good sport . i ll tell you what s going on .
- - why are you talking like this to me
- i m telling you , i m just a little bit stronger than i have .
- - just say your name
- i don t like your name .
- - what is about my name ?
- you re not supposed to bother me with trifles gifts , are you ?
- - no i am not
- i m not afraid of anything . i m sorry .
- - ok i have to go
- i thought you said you were gone for a couple of months . you ignore me .
- - I didn't say that
- i know you did . but i m not sure what you did . i m not a guy . i m sorry .
- - ok see you soon
- just the cora house . i want to go home . go home .
- - good bye
- good bye .
- - bye
- bye .
```

Training on Cloud

FOR FREE:

- **Kaggle:**
  - 10 CPU at the same time, each 16 GB RAM & 4.9 GB HDD.
  - 1 GPU/TPU each week 30 hrs.
  - Very easy access, use and share any dataset on Kaggle.
- **Google Colab:** each Notebook:
  - approx. 13 GB RAM & 70 GB Disk
  - 1 GPU if available.
  - Easy connect to google drive.
- **Jupyter.icc.informatik:**
  - GPU
  - RAM more than Kaggle
  - Train faster than Kaggle

Any questions?