Recap

- Motivation: Sequence to sequence tasks (like machine translation)

Encoder-Decoder architecture

- Encoder reads in the input, generates internal representation
- Decoder produces output, consuming internal representation

Attention

- Developed for image classification, then transfered to machine translation
- Let the model learn the relevant input tokens for each output token

BERT

- Breakthrough in natural language processing
- Pre-training vs. fine-tuning
- Huggingface: Platform to make such models easy to use
  - Good documentation on transformers: huggingface.co/docs/transformers

Machine Learning 7: Exercises and BERT in Action

VL Sprachliche Informationsverarbeitung

Nils Reiter
nils.reiter@uni-koeln.de

January 19, 2023
Winter term 2022/23

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

CRETA
CENTER FOR REFLECTED TEXT ANALYTICS

Exercises

- Type-Token-Ratio
- Annotation
- Ziffernerkennung mit log. Regression

Introduction

- Libraries: `transformers`, `datasets` (huggingface), `keras`
- Models: `bert-base-cased` other options
  - Some numbers: Context width: 512 tokens, embedding: 768 dimensions

Introduction

- Libraries: `transformers`, `datasets` (huggingface), `keras`
- Models: `bert-base-cased` other options
  - Some numbers: Context width: 512 tokens, embedding: 768 dimensions
- Important preprocessing step: tokenization
  - Needs to match training tokenization

Introduction

- Libraries: `transformers`, `datasets` (huggingface), `keras`
- Models: `bert-base-cased` other options
  - Some numbers: Context width: 512 tokens, embedding: 768 dimensions
- Important preprocessing step: tokenization
  - Needs to match training tokenization
- Use cases
  - Fill mask
  - Get BERT embeddings (to use in a non-BERT neural network)
  - Fine-tune for our own task