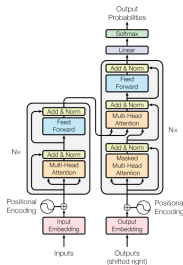# Transformer

BytePair encoding (BPE)

Learning goals

- Understand BPE
- Understand the Transformer Encoder + Decoder
- Understand how they are connected
- Understand the limitations for long sequences

BYTEPAIR ENCODING (BPE)

Data compression algorithm ▸ Gage (1994)

- Considering data on a *byte*-level
- Looking at pairs of bytes:
  1. Count the occurrences of all byte pairs
  2. Find the most frequent byte pair
  3. Replace it with an unused byte
- Repeat this process until no further compression is possible

BYTEPAIR ENCODING (BPE)

Open-vocabulary neural machine translation ▶ Sennrich et al. (2016)

- Translation as an open-vocabulary problem
- Word-level NMT models:
    - Handling out-of-vocabulary word by using back-off dictionaries
    - Unable to translate or generate previously unseen words
- Subword-level models alleviate this problem

BYTEPAIR ENCODING (BPE)

Adapt BPE for word segmentation ▸ Sennrich et al. (2016)

- *Goal:* Represent an open vocabulary by a vocabulary of fixed size
  $\rightarrow$ Use variable-length character sequences
- Looking at pairs of characters:

  1 Initialize the the vocabulary with all characters plus
  end-of-word token

  2 Count occurrences and find the most frequent character pair,
  e.g. "A" and "B" (⚠ Word boundaries are **not** crossed)

  3 Replace it with the new token "AB"

- Only one hyperparameter: Vocabulary size
  (Initial vocabulary + Specified no. of merge operations)
  $\rightarrow$ Repeat this process until given $|V|$ is reached

BYTEPAIR ENCODING (BPE)

test

BYTEPAIR ENCODING (BPE)