17.11.21

OPENCAMPUS.sh

# Natural Language Processing with Transformers

# USING TRANSFORMERS

- Quiz
- Breakout Diskussion
- Definition der Projekte
- Preprocessing und Postprocessing für Transformer

# QUIZ

https://forms.office.com/r/F6z11321Jn

# BREAKOUT DISKUSSION

- What is the advantage of character-based tokenization in comparison to word-based tokenization?

- How does the tokenizer used by a model influence its capability?
  → For which tasks might a character-based Tokenization be beneficial?

# PROJEKTIDEEN

- Vorhersage der Beantwortungsschwierigkeit von Aufgaben (Karo, Sina)
- Klassifikation von Antwort-Mails hinsichtlich Höflichkeit und ggf. hinsichtlich von fachlichen Kriterien (Chris und Sabrina)
- Die richtigen Worte für das perfekte Produkt (Leon)
- Sentiment-Analyse & Themen-Tagging von Nachrichtenartikeln (Leon)
- Sentiment-Analyse zur Vorhersage der Volatilität von Aktienkursen (Jule)
- SHU-T: Generierung von Antworten auf Hass-Artikel (Martin)
- Paraphrasing Texts (Peyman)
- Automatische Erkennung/Ergänzung fehlender Produktattribute (Laura)
- Vorhersage von Produktkategorien anhand des Produkttextes (Laura)

# FIRST PROJECT TASKS

(1) Complete the Project Proposal form.

(2) Setup a project channel in the Chat.

(2) Define a common repository or GoogleDrive to exchange the program code.

(3) Define first steps in the project.

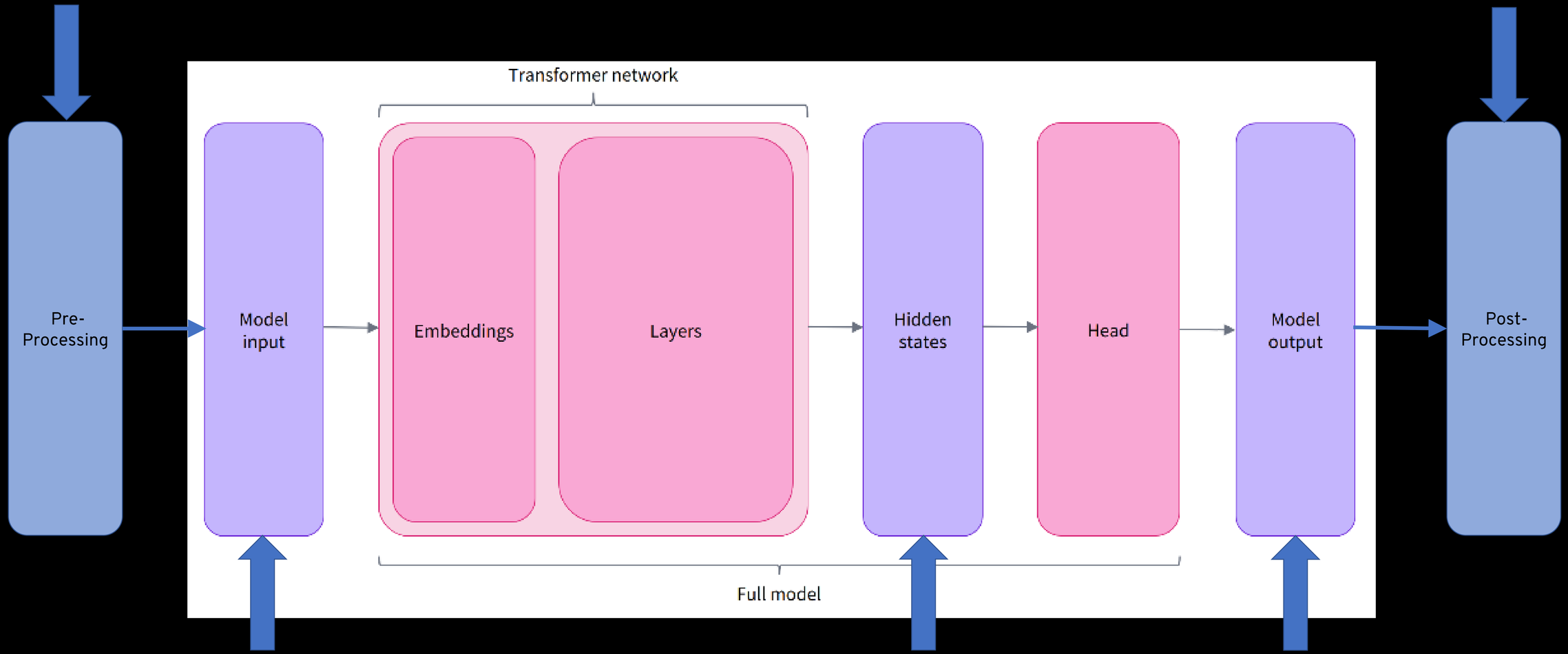
(4) Decide on times for regular project meetings.

# TOKENIZATION

- Byte-Pair Encoding (BPE; e.g., GPT-2)
- WordPiece (e.g., BERT)
- Unigram (e.g., T5, XLNet)

- Splitting
- Mapping to integers
- Adding model dependent tokens/integers

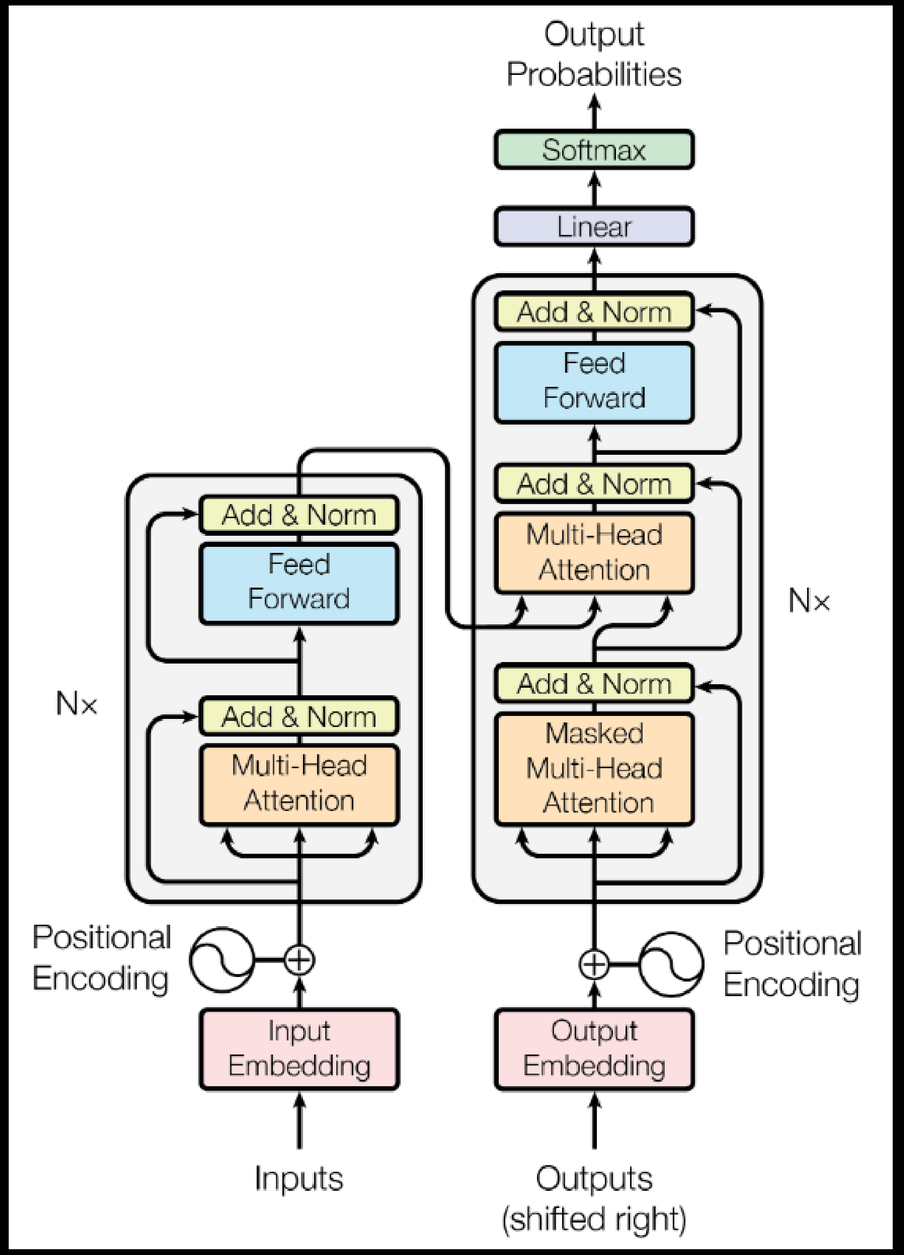
- Logits to probs
- Probs to classes
- (Classes to tokens/text)




Tokenized subwords (integers)

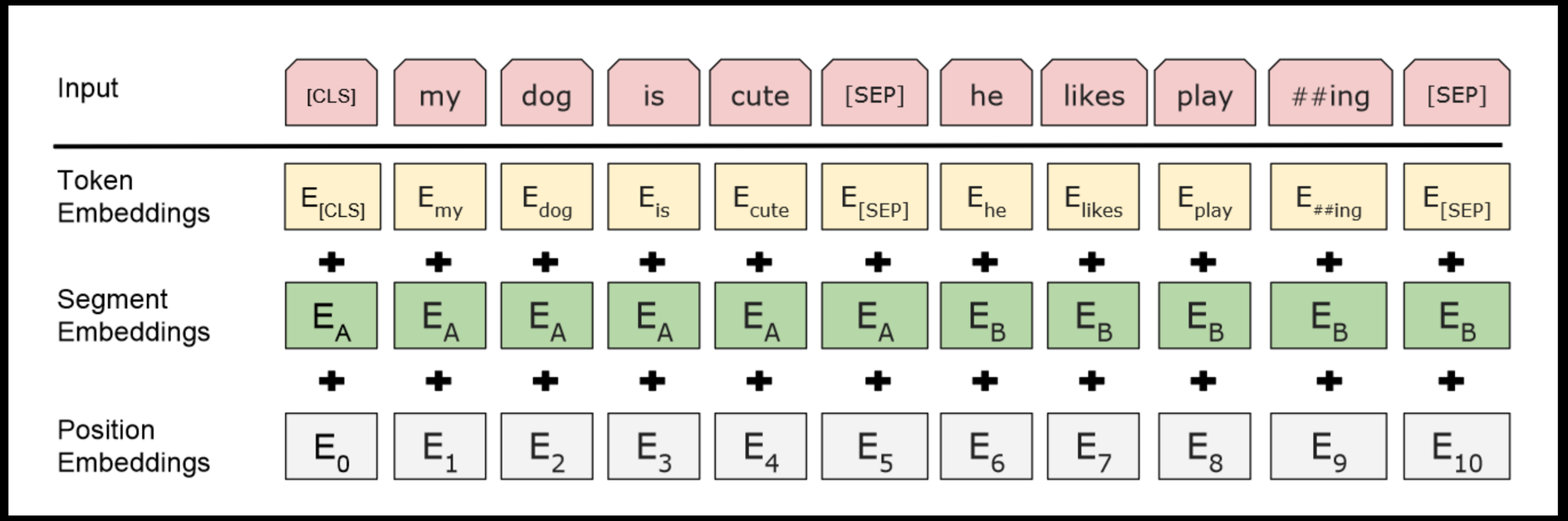
Contextualized embeddings for each token (high dimensional vectors of floats)

Logits (one float for each trained class)




Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. Retrieved from http://arxiv.org/abs/1706.03762

# BERT EMBEDDINGS




Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. Retrieved from http://arxiv.org/abs/1810.04805

```
from transformers import TFAutoModel

checkpoint = "distilbert-base-uncased-finetuned-sst-2-english"
model = TFAutoModel.from_pretrained(checkpoint)
outputs = model(inputs)
outputs.last_hidden_state.shape
```

```
(2, 16, 768)
```

# TODOS BIS ZUR NÄCHSTEN WOCHE

- Chapter 3 des Hugging Face Kurses absolvieren: https://huggingface.co/course/chapter3

- Erstmalig im Team treffen und die zuvor beschrieben Aufgaben angehen

# MASKING LEVELS

| Sentence | Harry | Potter | is | a | series | of | fantasy | novels | written | by | British | author | J. | K. | Rowling |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Basic-level Masking | [mask] | Potter | is | a | series | [mask] | fantasy | novels | [mask] | by | British | author | J. | [mask] | Rowling |
| Entity-level Masking | Harry | Potter | is | a | series | [mask] | fantasy | novels | [mask] | by | British | author | [mask] | [mask] | [mask] |
| Phrase-level Masking | Harry | Potter | is | [mask] | [mask] | [mask] | fantasy | novels | [mask] | by | British | author | [mask] | [mask] | [mask] |

Liu, B. (n.d.). NLP Pretraining–From BERT to XLNet. Retrieved December 15, 2020, from Title website: https://bangliu.github.io/survey/2019/07/01/NLP-Pretraining/