

09.11.22

Application of Transformer Models

INTRODUCTION TO TRANSFORMER MODELS

- **Quiz**
- **Breakout Discussion**
- **From Pre- to Post-Processing
in Transformers**
- **Projects**

QUIZ

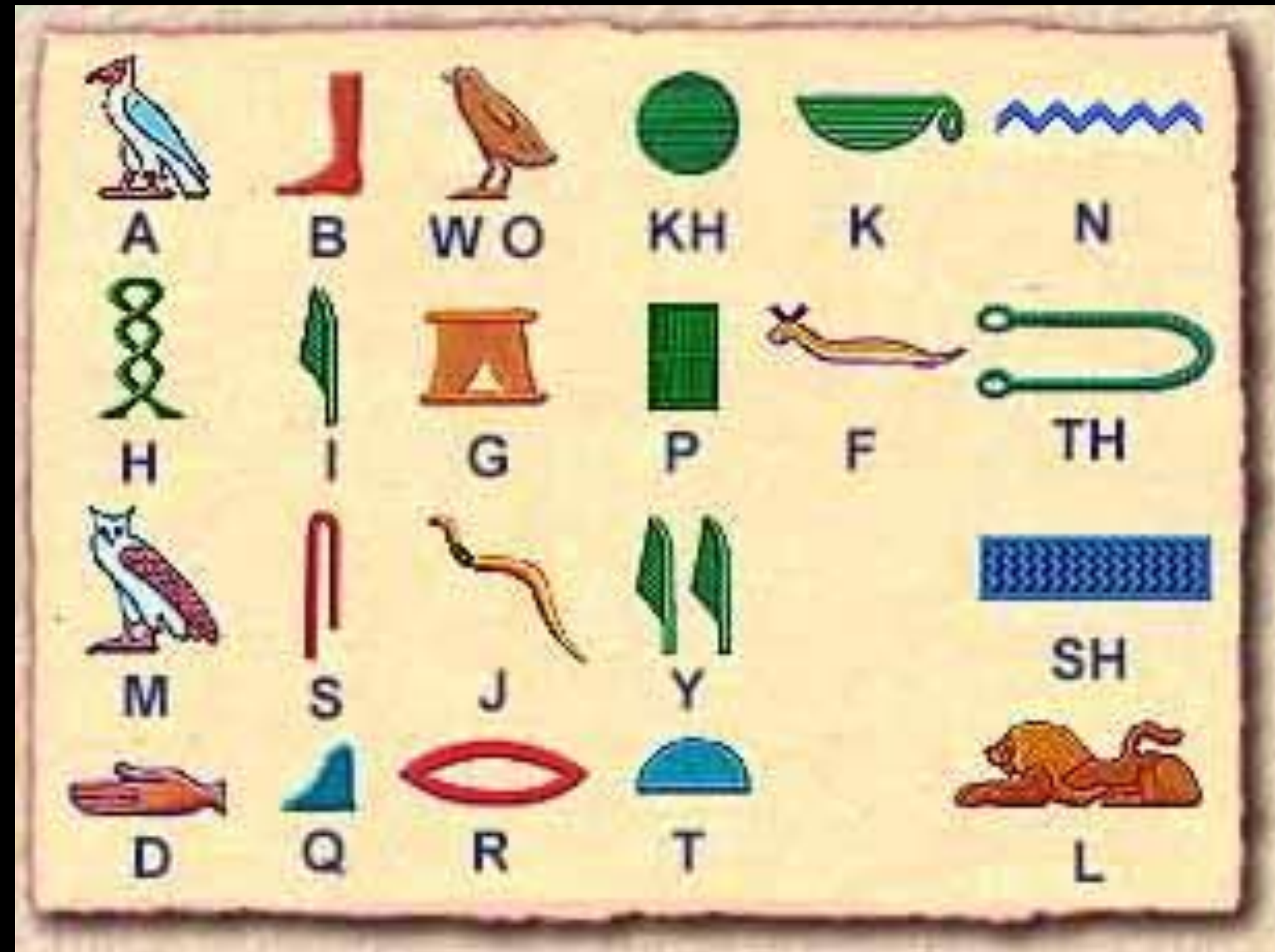


<https://forms.office.com/r/z75QWkMFm6>



BREAKOUT DISCUSSION

- **How does the tokenizer used by a model influence its capability?**
- **What might be an approach to tokenize ancient hieroglyphics?**



Input IDs



Special tokens



Tokens



Raw text

[101, 2292, 1005, 3046, 2000, 19204, 4697, 999, 102]



[[CLS], let, ', s, try, to, token, ##ize, !, [SEP]]



[let, ', s, try, to, token, ##ize, !]



Let's try to tokenize!

TOKENIZATION

- **Byte-Pair Encoding (BPE; e.g., GPT-2)**

Chetna Khanna - Medium. (n.d.). Retrieved May 2, 2022, from <https://chetnakhanna.medium.com/>

- **Unigram (e.g., T5, XLNet)**
- **WordPiece (e.g., BERT)**

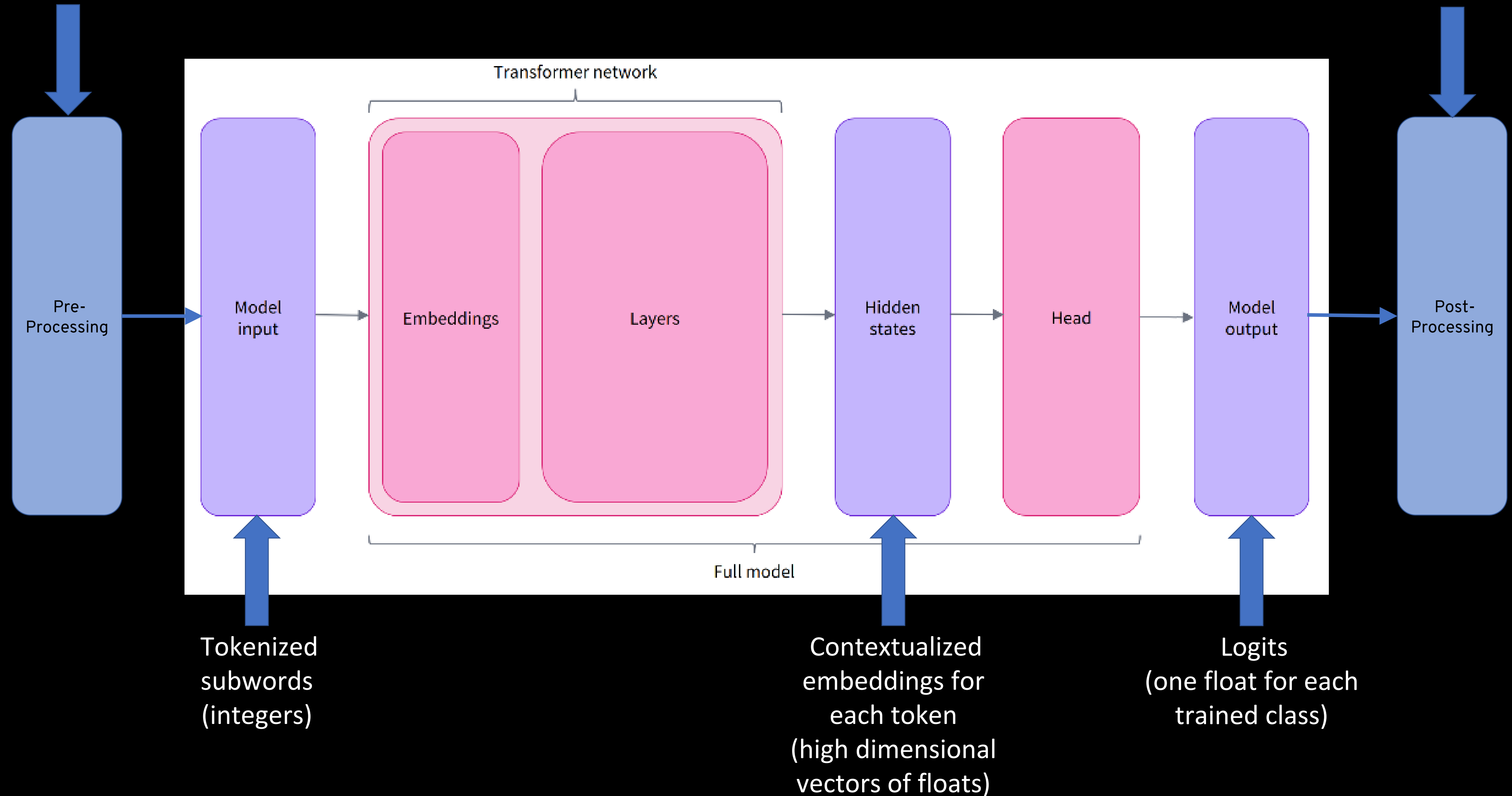
Tokenizers: How machines read. (2020, January 28). FloydHub Blog.
<https://blog.floydhub.com/tokenization-nlp/>

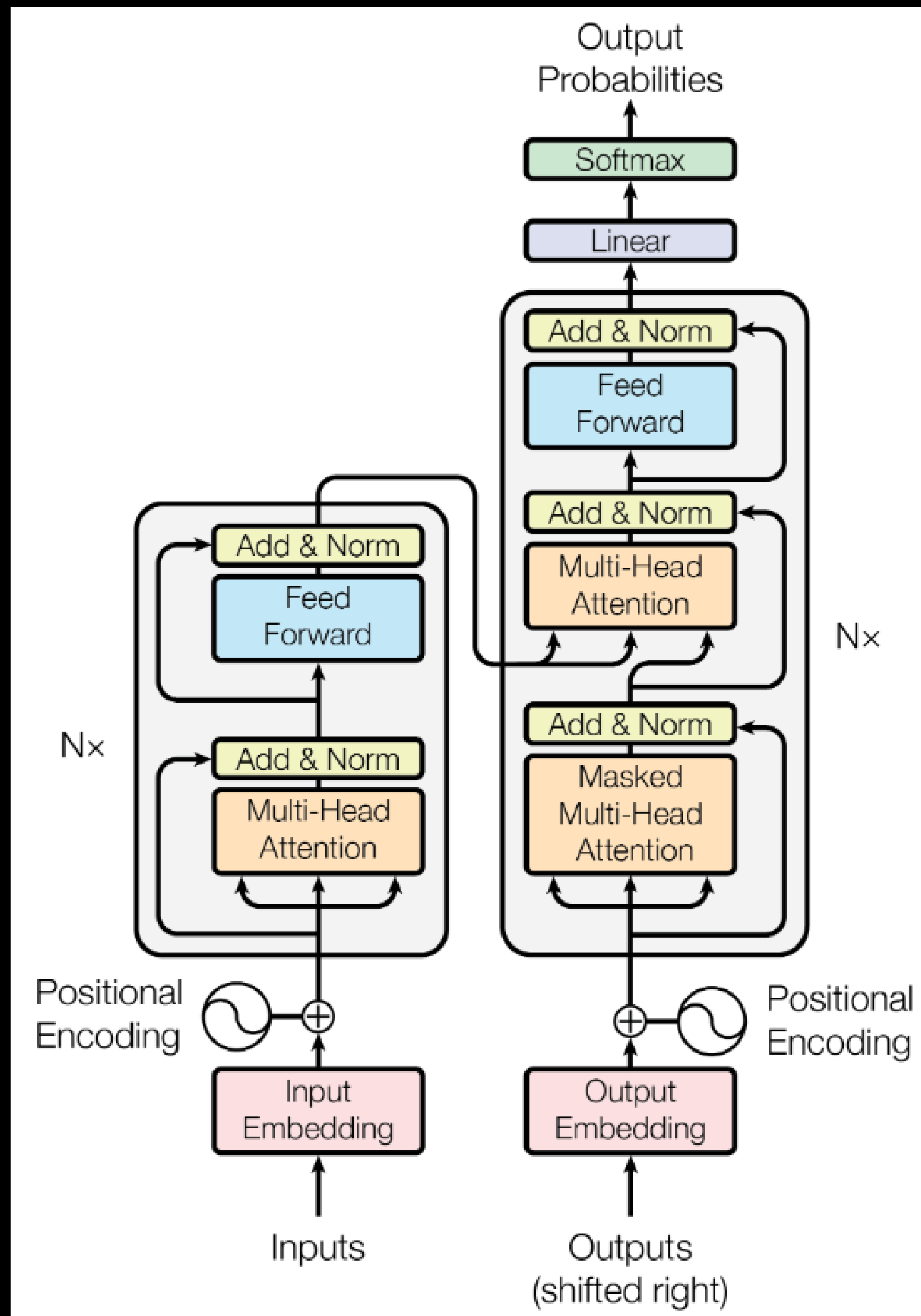
ATTENTION MASK

```
{  
  'input_ids': <tf.Tensor: shape=(2, 16), dtype=int32, numpy=  
    array([[ 101,  1045,  1005,  2310,  2042,  3403,  2005,  1037, 17662, 12172,  2607,  2026,  2878,  
           [ 101,  1045,  5223,  2023,  2061,  2172,   999,   102,    0,    0,    0,    0,    0,  
          ], dtype=int32)>,  
  'attention_mask': <tf.Tensor: shape=(2, 16), dtype=int32, numpy=  
    array([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],  
           [1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]  
          ], dtype=int32)>  
}
```

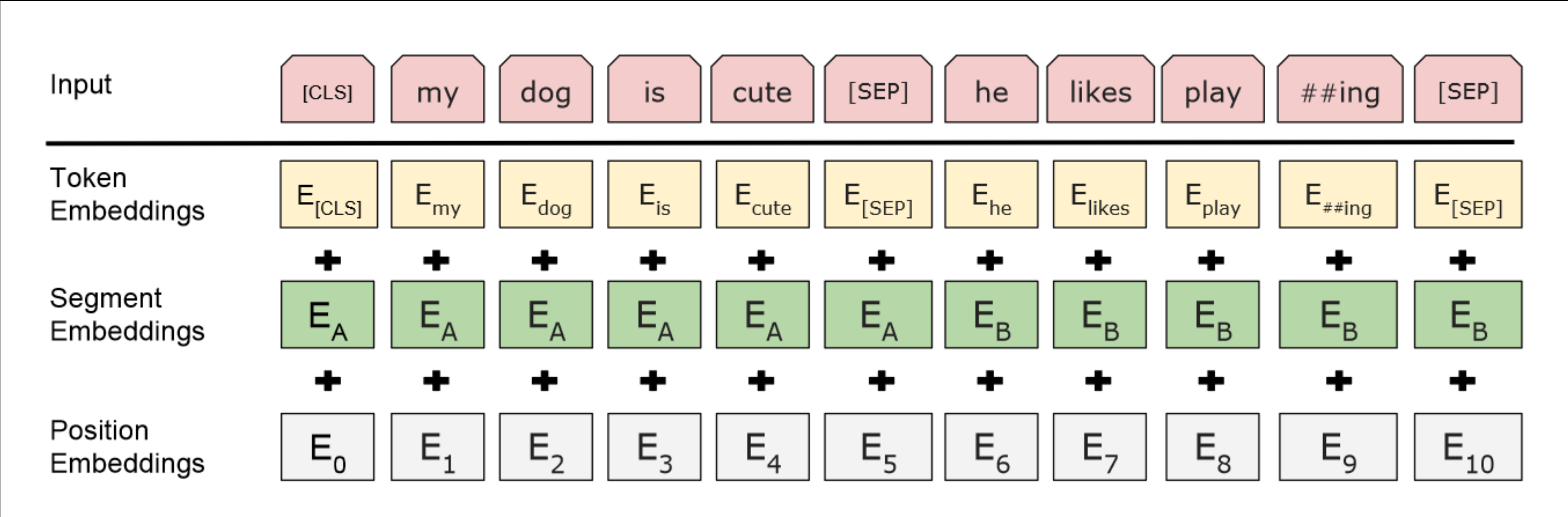

- Splitting
- Mapping to integers
- Adding model dependent tokens/integers

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





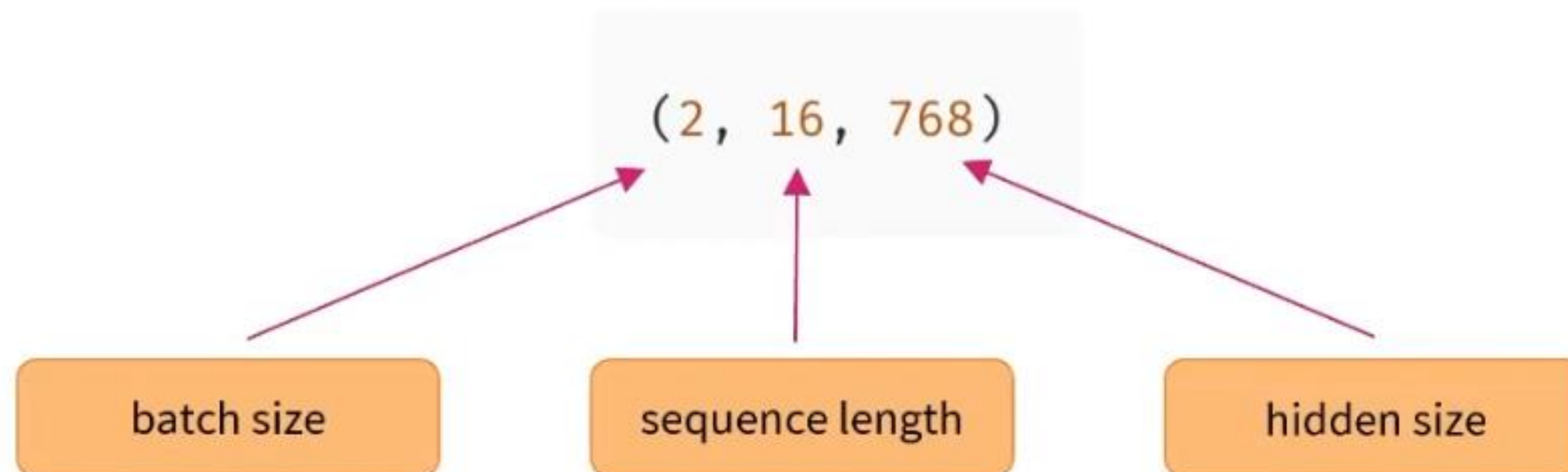
BERT EMBEDDINGS



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
```



PROJECTS

- **Saif/ Emmanuel/ Kristian/ Atul: Time Series Prediction
Financial/Climate**
- **Jonathan/Julian: Arguments Mining / NER Task on data already
collected**
- **Malte/ T.-Niklas: Speech to speech models including translation**
- **Jeremy/ Veit/ Christian: Transcribing and summarizing Podcasts**
- **Laura/ Janosch/ Valentin : Training a model to produce text written
in different author's style**
- **Khan: Classification of activity descriptions according to keywords**
- **Manpreet: Unsupervised training of log data from a simulation
software to predict user behavior**
- **Max: Q&A model**

PROJECT MILESTONES

- 16.11. Form project groups**
- 23.11. Literature review**
- 30.11. Dataset characteristics**
- 04.01. Baseline model**
- 11.01. Project presentations**

TODOS UNTIL NEXT WEEK

- Complete [chapter 3](#) (Fine-Tuning a Pretrained Model) of the Hugging Face course
- Meet with the other members of the project you are interested in to get a more concrete idea of the project goal.