03_TransformerModel_101

December 15, 2024

[1]:
```
from IPython.display import Image, display
```

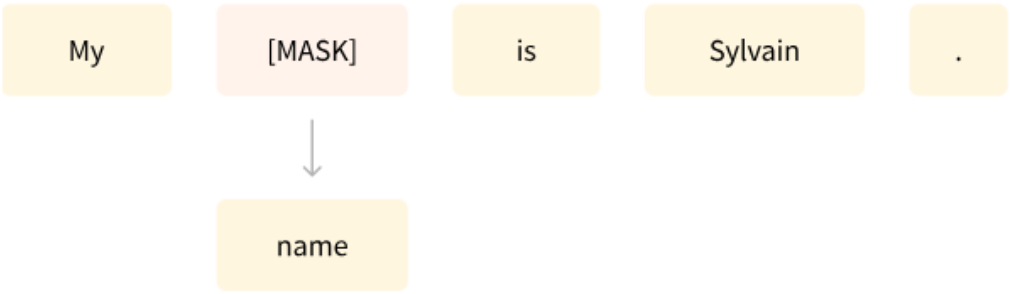
0.1 Transformers are language models

All the Transformer models, have been trained as **language models**. They have been trained on large amounts of raw text in a self-supervised fashion. By default **language models** are **causal language models** as they try to model:

$$P(x_t|x_{t-1}, x_{t-2}, ...x_0)$$

However, there is another type of **language models** called **masked language modeling**.

[2]:
```
display(Image(filename='./Pics_04_MLM.png', width=300, height=100))
```



0.2 General Architecture

The model is primarily composed of two blocks: - **Encoder**: The encoder receives an input and builds a representation of it (its features). This means that the Encoder block is optimized to **acquire understanding from the text**. - **Decoder**: The decoder uses the encoder's representation (features) along with other inputs to generate a target sequence. This means that the Decoder block is optimized for **generating outputs**.

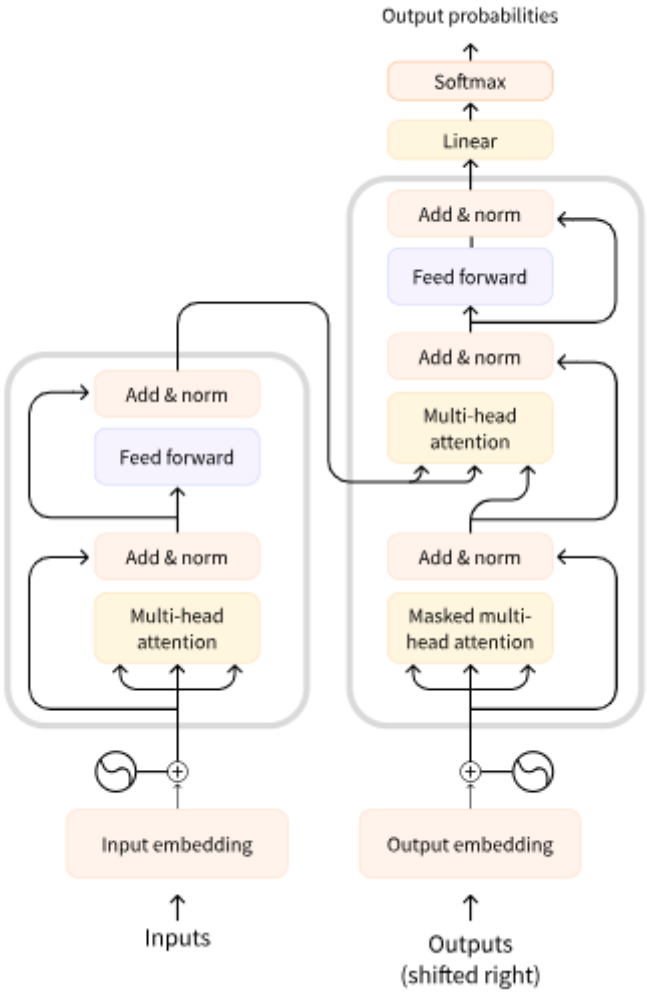
Each of these parts can be used independently, depending on the task:

- **Encoder-only models**: Good for tasks that require understanding of the input, such as sentence classification and named entity recognition. These models are often characterized

as having **"bi-directional" attention**, and are often called ***auto-encoding models***. The pretraining of these models usually revolves around somehow *corrupting a given sentence* (for instance, **masking tokens** as in *Masked Language Modelling.*) and tasking the model with finding or reconstructing the initial sentence. For ex: ALBERT, BERT, DistilBERT, ELECTRA, RoBERTa.

- **Decoder-only models**: Good for generative tasks such as text generation. At each stage, for a given word the attention layers can only access the words positioned before it in the sentence. These models are often called ***auto-regressive models***. The pre-training task here is *Causal Language Modelling.* For ex: GPT Based Models (GPT-2, GPT-3, Transformer XL, CTRL)
- **Encoder-decoder models or sequence-to-sequence models**: Good for generative tasks that require an input, such as translation or summarization. The pretraining of these models can be done using the objectives of encoder or decoder models, but usually involves something a bit more complex. For instance, **T5** is pretrained by *replacing random spans of text (that can contain several words) with a* ***single mask special word****, and the objective is then to predict the text that this mask word replaces.* For Ex: BART, mBART, Marian, T5.

[3]:
```
display(Image(filename='./Pics_05_TransformerArchitecture.png', width=400,␣
↪height=100))
```


Output probabilities
Softmax
Linear
Add & norm
Feed forward
Add & norm
Multi-head attention
Add & norm
Masked multi-head attention
Add & norm
Feed forward
Add & norm
Multi-head attention
Input embedding
Output embedding
Inputs
Outputs (shifted right)

[ ]: