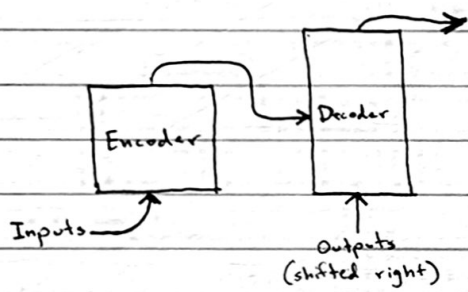
Transformer Architecture

* Training/Implementation overview on slides

* Slides on "seq-trans 1 - transformer_architecture.pdf"

- **Encoder** (left): Receives input and builds representation of it (its features). Model is optimized to acquire understanding from input
- **Decoder** (right): Uses encoders features along w/ other inputs to generate target sequence.



* Alternate multi-head self attention w/ feedforward layers that operate over each word individually

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

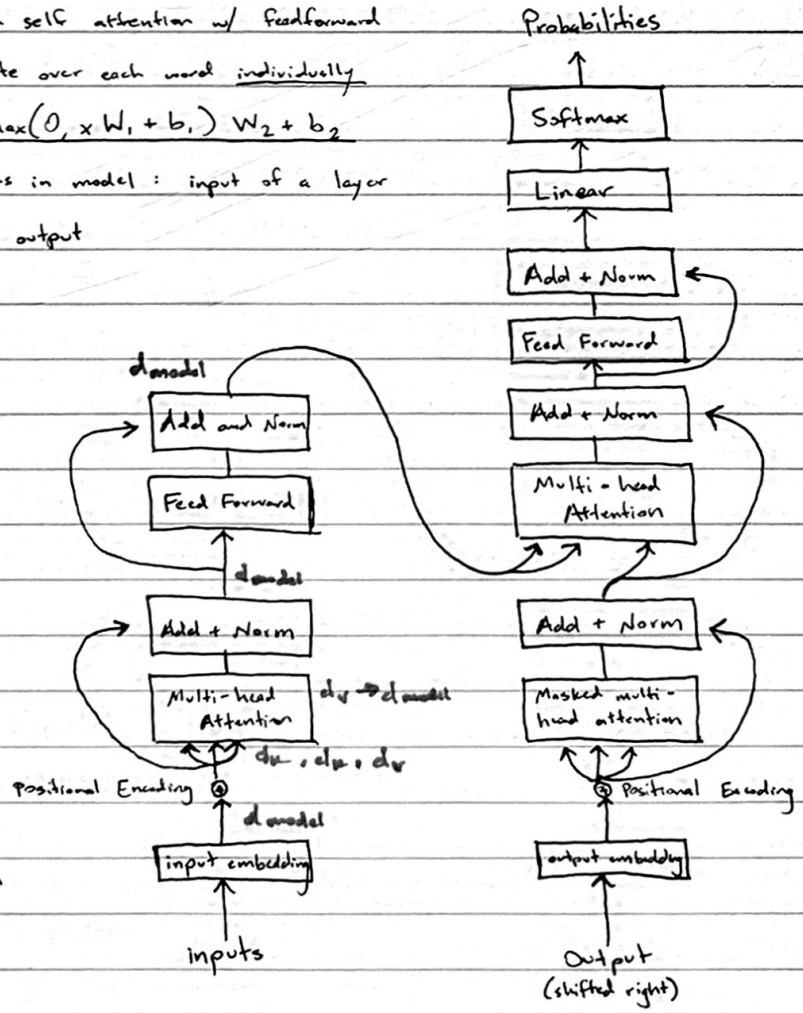
* Residual Connections in model: input of a layer is added to its output

* Vectors: d_{model}

* Queries/keys: d_k
(Always smaller than d_{model})

* Values: Separate dimension d_v, output is multiplied by W^o, which is $d_v \times d_{model}$ so we can get back to d_{model} before the residual

* FFN can explode the dimension with W_1 and collapse it back with W_2



* Encoder - Decoder Model