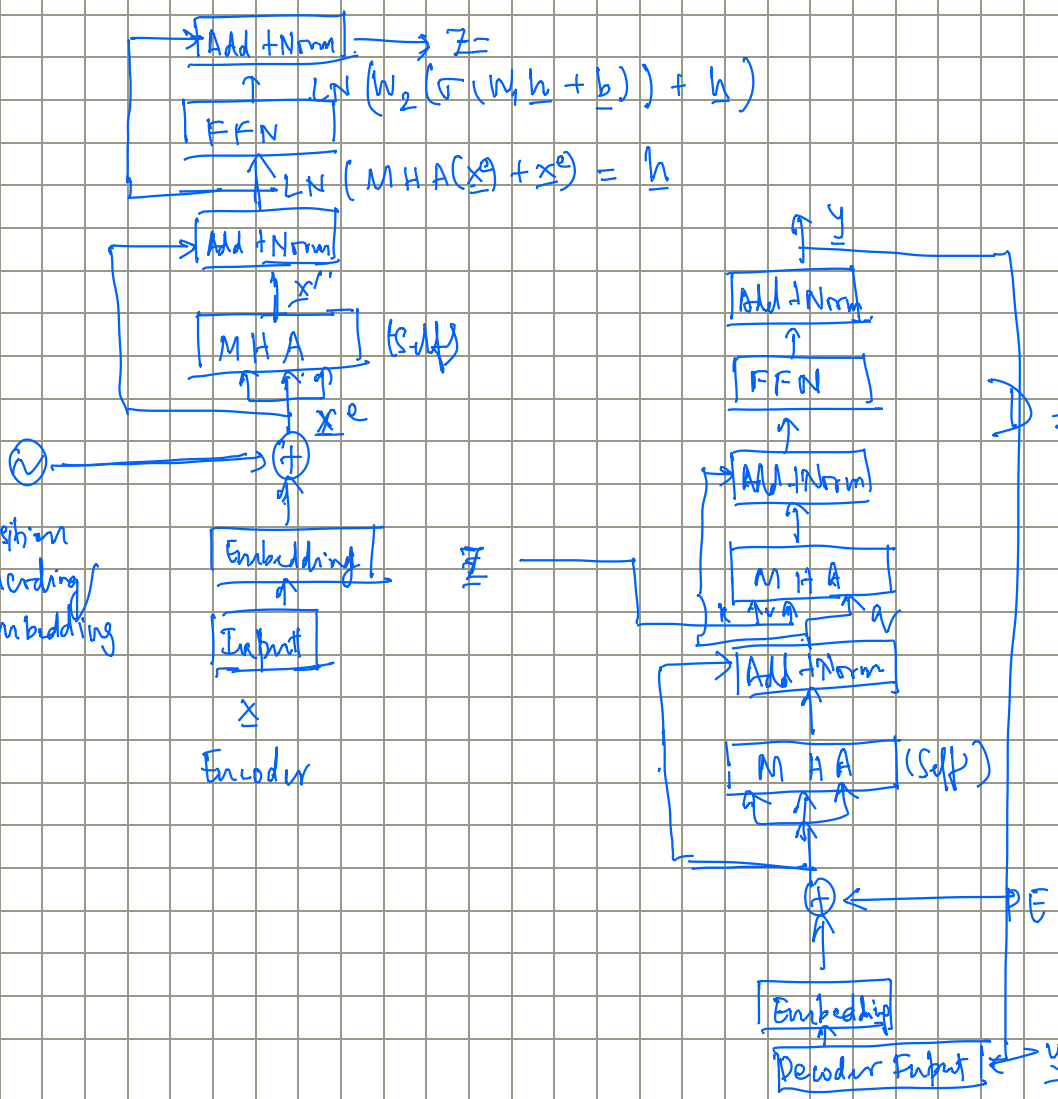
17/10/23

Deep learning

- Recap: Attention mechanism
- Transformer: Architecture, attention, I/O relation
- Attention: $J(i,j) = \sum_{a,b \in N_k(i,j)} \sum \text{softmax}_{ab}\left(\langle q_f(i,j), k(a,b)\rangle\right) \cdot v(a,b)$.
  - $q_f(i,j) = W_q\, x(i,j)$ query
  - $k(i,j) = W_k\, x(i,j)$ key
  - $v(i,j) = W_v\, x(i,j)$ value

  } Self-attention
  - $x(i,j) \in \mathbb{R}^c$, $q_f(i,j), k(i,j), v(i,j) \in \mathbb{R}^d$
  - $W_q, W_k, W_v \in \mathbb{R}^{d \times c}$
  - $N_k(i,j)$ is a neighborhood around (i,j) of size $k \times k$
- Transformer architecture (Vaswani et al. 2017)



* Lookup GELU activation

$GELU(x) = x \cdot \Phi(x)$

$\Phi(x) = CDF(N(0,1))(x)$

$= \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots (x^{(T)}, y^{(T)})\}$ (time)