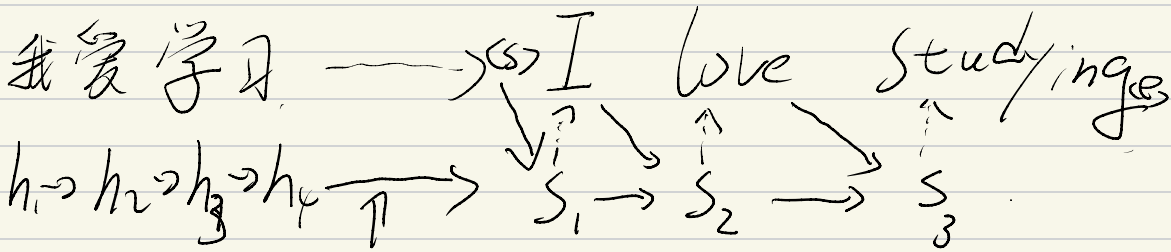
25.02.11

seq2seq

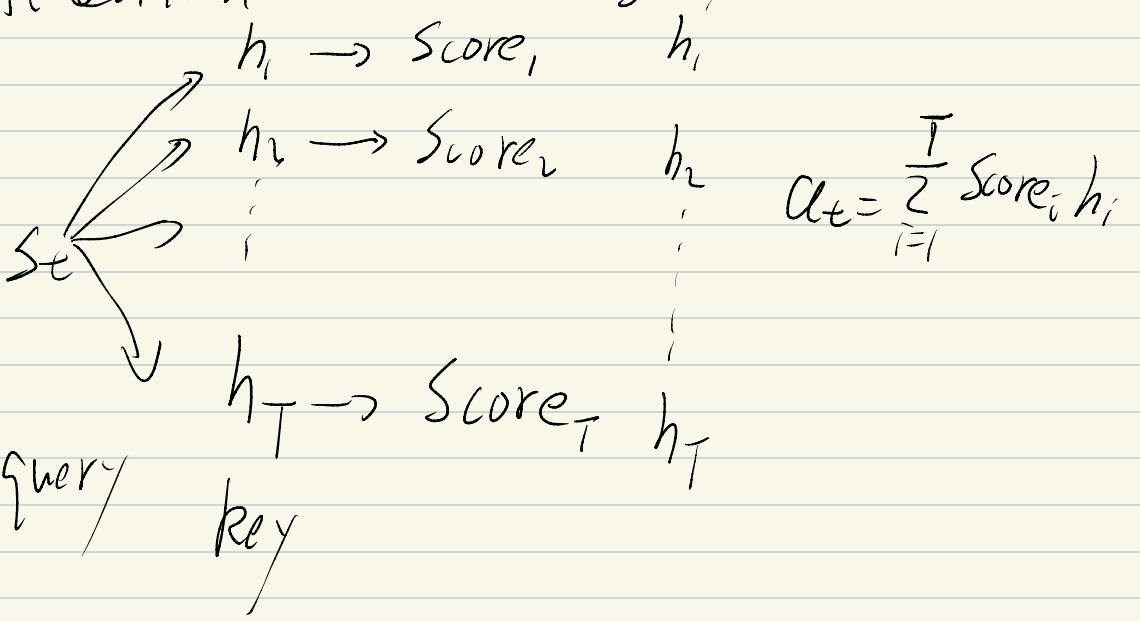


① Information Bottle Neck

② Recurrent structure is not parallelizable

③ O(Seq length)

Attention Mechanism



$$a_t = \sum_{i=1}^{T} \text{Score}_i \, h_i$$

Attention is All U need

我爱学习

$x_1 \quad x_2 \quad x_3 \quad x_4 \in \mathbb{R}^d$

$q_i = W_q \cdot x_i \;; \quad k_i = \underset{key}{W_k} \cdot x_i, \quad v_i = \underset{value}{W_v} \cdot x_i$

query $\quad W_q, W_k, W_v \in \mathbb{R}^{d \times d}$

$$\tilde{w}_{ij} = \frac{q_i^T k_j}{\sqrt{d}}, \quad i, j \in \{1, 2, \cdots n\}$$

↓ Softmax

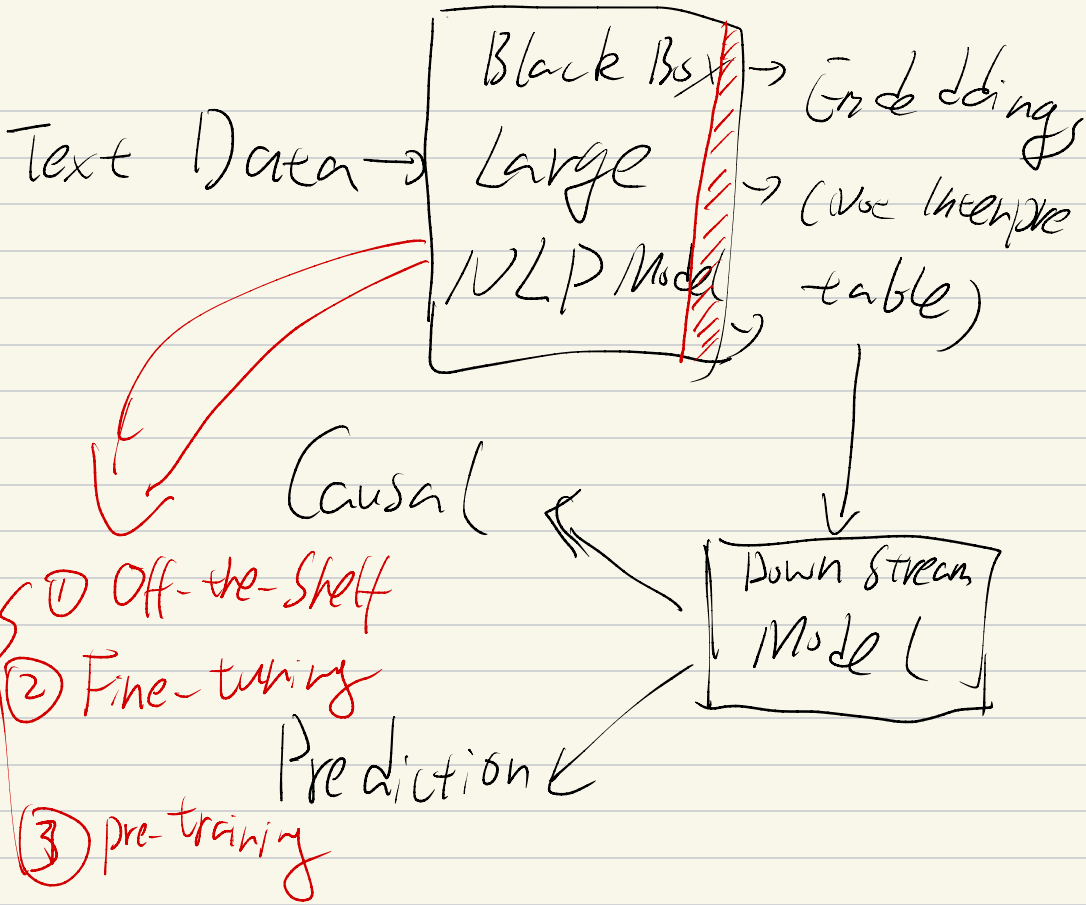
$$w_{ij} = \frac{\exp(\tilde{w}_{ij})}{\sum_{l=1}^{n} \exp(\tilde{w}_{il})}$$

$$y_i = \sum_{j=1}^{n} w_{ij} v_j$$

① Parallelization ✓

② Information Bottleneck

③ O(seq. length)


Text Data
Black Box
Large
NLP Model
Embedding
(Use Inference
table)
Causal
Down Stream
Model
Prediction
① Off-the-Shelf
② Fine-tuning
③ pre-training