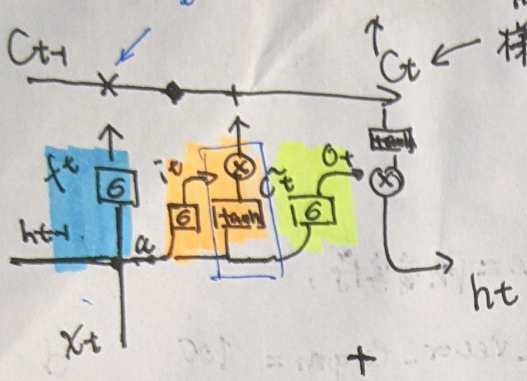
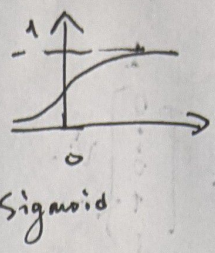
LSTM.

element-wise multiplication



解決梯度消失



~~ML (DL)~~
~~DB~~

Deep Learning & NLP

$$\tanh\left[\sigma(W_f\begin{bmatrix}h_{t-1}\\x_t\end{bmatrix})\ \ \sigma(W_i\begin{bmatrix}h_{t-1}\\x_t\end{bmatrix}) * \tanh(W_c\begin{bmatrix}h_{t-1}\\x_t\end{bmatrix})\right] * \sigma\left[W_o\begin{bmatrix}h_{t-1}\\x_t\end{bmatrix}\right]$$

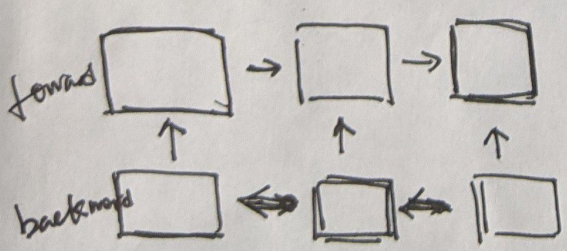
① Forget Gate ② Input Gate ③ new value ④ Output gate

$= h(t)$

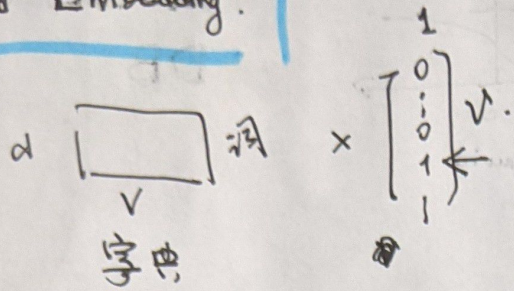
W_f, W_i, W_c, W_o → $h_{t-1}+x_t$ → h_t [ ]

params: $4 \times \text{Shape}(h) \times (\text{Shape}(h) + \text{Shape}(x))$

Bidirectional Lstm.



Word Embedding.

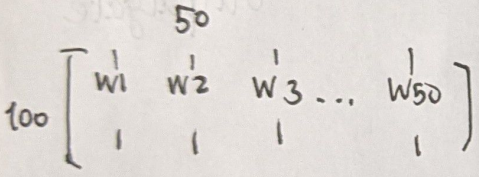


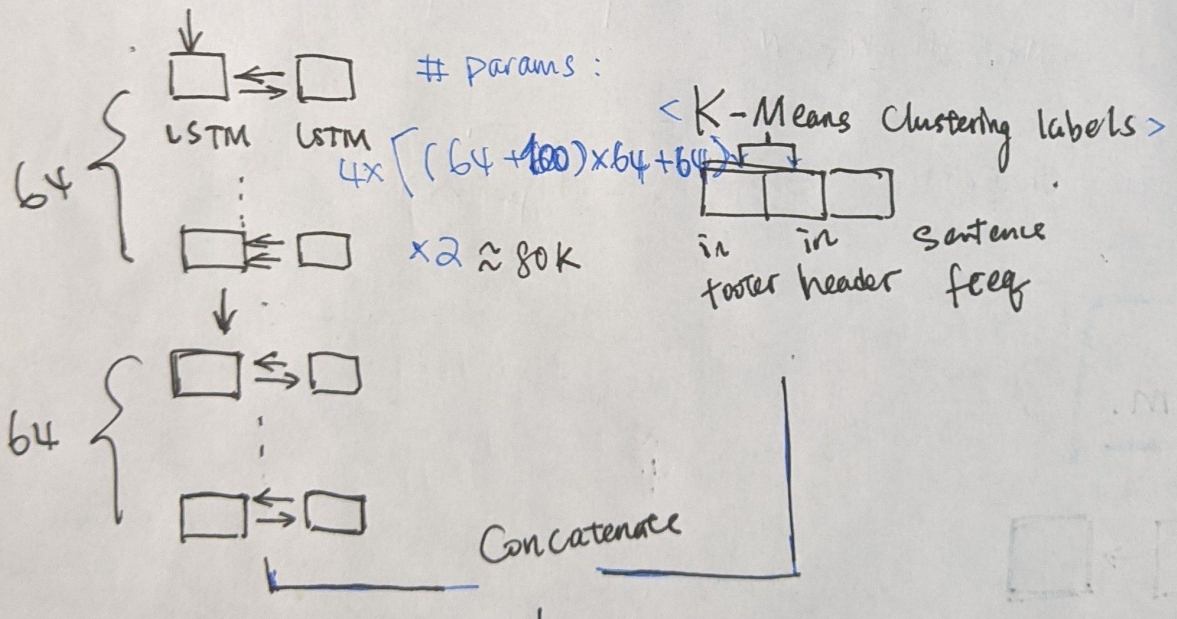
(vocab_size = 所有字符, V

embedding_vector_length = 100 d

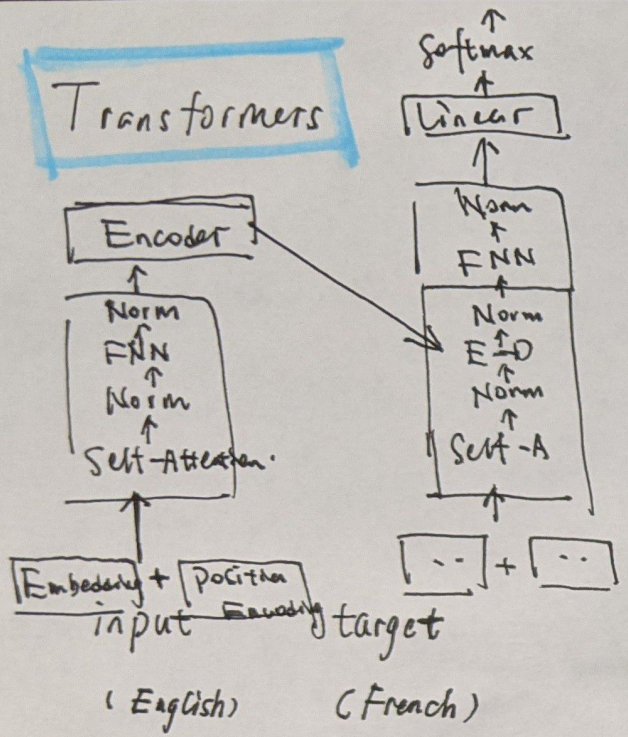
max_length = 50 ,

Our model (DP Intern)





Transformers



Tensorflow : ✓ Production (Industry)

vs

Pytorch : ✓ (Academy)
✓ more complex model.

Self-Attention model: $\begin{bmatrix} w_1' & w_2' & \cdots \\ | & | & \end{bmatrix} \begin{bmatrix} - w_1 - \\ - w_2 \\ \vdots \end{bmatrix}$

high score if inner prod. is high.

Transformer 模型,

Tensorflow vs. Pytorch.

BERT

predict ① masked words ② next sentence

① 15% words randomly

Softmax + Cross Entropy Loss

判断是否是masked word

② 50% true consequtive 2 sentences

50% False (faked)

C

↓

Binary Classifier, Cross Entropy Loss.

- 结构为 transformer 的 Encoder

被作 classification, 非Regression.

Bert-KPE