VietAI - NLP

Week 1 : NLP and word embeddings

- Goals:

language → [Computer] → language

← Understanding (NLU) →

← Generation (NLG) →

- Deep Learning for NLP:
  - ☐ Recurrent neural networks
  - ☐ Long short-term memory
  - ☐ Convolutional neural networks
  - ☐ Autoencoders
  - ☐ transformer

- Vector Representation of Words

  word → số / vector → +, -, ×, /

  ↳ one-hot vector, Bag-of-words or BOW (1)

  ↳ Contextual representation (2)

(1) Nhược điểm: chiều giảm? Sửa sau

→ sparse matrix (Giống với trên 1)

Khuyết: tính độ tương đồng

(2) Contextual Representation

- word vector
- using dot product

Word2Vec:

↳ Skip gram: $\mathcal{L}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log P(w_{t+j} | w_t; \theta)$

↳ CBOW: $\mathcal{L}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \log P(w_t | w_{t-m}, \dots, w_{t+m})$

- ☐ ~~2.7~~ + DL Prob
- ☐ Machine Learning + Bắt tập VietAI
- ☐ Ielts Reading + Listening
- ☐ Ielts Writting
- ☐ Coursera
- ☐ Personal Projects

RNN Variants

Dữ liệu dạng chuỗi, khác với feature

- Tính thứ tự

Sequence data Processing

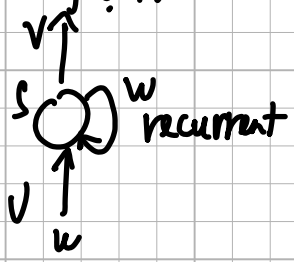
↳ Recurrent Network

→ ảnh khó → hiện tại → tương lai

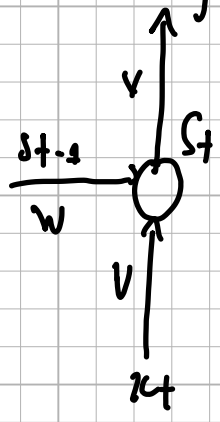
Tham số RNN chỉ có 3 ma trận

U, V, W

chia sẻ giữa các time step



unfolded form → compressed form



thường tính bằng

kết hợp → phi tuyến hóa

$$S_t = \sigma(U x_t + W s_{t-1})$$

U x_t: với tại

W s_{t-1}: với cũ

$$\tilde{y}_t = \text{softmax}(V S_t)$$

ŷ_t: dự đoán

decode/giải mã

Step 1 Model Design

sigmoid: $\frac{1}{1+e^{-x}}$

softmax: $\frac{e^x}{1+e^x}$

Step 2 loss function Design

$$\mathcal{L}(\theta) = -\sum_{j=1}^{K} y_{t,j} \log \tilde{y}_{t,j}$$

with t steps

$$\mathcal{L}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(\theta)$$

Kiến thức: Cấu trúc LSTM, GRU

Some Use cases:

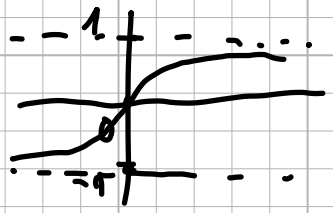
↳ one to one, one to many, many to one

many to many (1), many to many (2)

- Training process: loss function

→ Relu ⇒ đổi sigmoid/tanh.
→ khởi tạo W = I
→ LSTM

⇒ tanh function



Ablation Study: nghiên cứu loại bỏ

- Problem of Parallel computing

HCMUS
- ☐ Prob and Stats
- ☐ toán logic tổ hợp

- ☐ ~~3.7~~ + D2L Prob + D2L 2.4.6
- ☐ ~~Machine Cơ bản~~ + Bài tập VietAI
- ☐ Ielts Reading + Writing + Listening + Speaking
- ☐ Ielts Writting
- ☐ Coursera
- ☐ Personal Projects
- ☐ 100 DS

Certi
- ☐ C8229
- ☐ .NL
- ☐ NLP
- ☐ RL

Attention(Q, K, T)

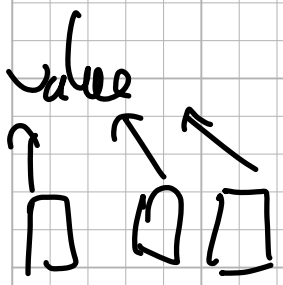
$$= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right).V$$

Attention is all you need

- Hidden stages (key & query)
- Self Attention
- Transformer Architecture

they self Attention → RNN

Key, query, value



h...

weighting

key

query

- Multihead Attention

↳ FFNN Concate and shorten ?

Ex 8.

- D2L Prob + D2L C3, 1 3 5 7

Estimation of the mean

→ Normal IID samples

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\bar{X}_n) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \cdot n\mu = \mu$$

$$Var[\bar{X}_n] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$p = 1$

Ex 3: Prob we see a head after drawing N examps

① $P_N(p^n) = C_N^n \frac{1}{2^n} \cdot \frac{1}{2^{N-n}} = C_N^n \frac{1}{2^N}$ $(n = 1 \dots N)$

$$E_N[\tilde{p}_n] = \sum_{n=1}^{N} \frac{n}{N} C_N^n \cdot \frac{1}{2^n}$$

$\left(\tilde{p} = \frac{n}{N}\right)$

$$= \frac{1}{2^N}\left(\sum_{n=1}^{N} \frac{n}{N} \frac{N!}{(N-n)!\,n!}\right) = \frac{1}{2}$$

$$V_N[n] = N \cdot p \cdot (1-p) = \frac{N}{4}$$

$$Var\left[\frac{n}{N}\right] = \frac{1}{N^2} \cdot \frac{N}{4} = \frac{1}{4N}$$

② Chebychev Inequality (Deviation $\delta > 0$)

$$P_N\left(\left|\tilde{p} - \frac{1}{2}\right| \geq \delta\right) \leq \frac{1}{4N\delta^2}$$

③ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. $\tilde{p} \sim N\left(\frac{1}{2}, \frac{1}{4N}\right)$

Ex 4: $x \sim N(0,1)$

Compute $z_m \overset{\text{def}}{=} m^{-1} \sum_{i=1}^{m} x_i$

- $E(z_m) = E\left(\frac{1}{m} \cdot \sum_{i=1}^{m} x_i\right) = \frac{1}{m} \cdot E\left(\sum_{i=1}^{m} x_i\right) = \frac{1}{m} \cdot 0 = 0$

$\text{Var}(z_m) = \frac{1}{m^2} \cdot \sum_{i=1}^{m} \text{Var}(x_i) = \frac{1}{m^2} \cdot m \cdot 1 = \frac{1}{m}$

- Cannot apply Chebyshev's inequality for every z_m independently.
→ Not independent events, Dependence on sample size, collective behavior

Ex 5: $P(A)$, $P(B)$

$0 \leq P(A \cup B) \leq P(A) + P(B)$

$0 \leq P(A \cap B) \leq \max(P(A), P(B))$

Ex 6: Markov Chain, B only depends on A, C only depends on B

$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|B)$

Ex 7: Assume two test are not independent

$P(D=1 \mid H=0) = 0,1$ and $P(D=0 \mid H=1) = 0,01$

$H=1$: $P(D_1, D_2 \mid H=1) = P(D_1 \mid H=1) \cdot P(D_2 \mid H=1)$

and $P(D_1 = D_2 = 1 \mid H=0) = 0,02$

	$H=1$	$H=0$		$H=1$	$H=0$
$P(D_1=1 \mid H)$	0,99	0,1	$P(D_2=1 \mid H)$	0,99	0,1
$P(D_1=0 \mid H)$	0,01	0,9	$P(D_2=0 \mid H)$	0,01	0,9

$P(D_1, D_2 \mid H=0) \cdot P(H=0)$ $(= P(D_1, D_2, H=0))$

$= P(D_2 \mid D_1, H=0) \cdot P(D_1 \mid H=0) \cdot P(H=0)$

(1)

	$H=1$	$H=0$
$P(D_1=D_2=1 \mid H)$	0,5801	0,102
$P(D_1=1, D_2=0 \mid H)$	0,0095	0,08
$P(D_1=0, D_2=1 \mid H)$	0,0095	0,08
$P(D_1=D_2=0 \mid H)$	0,0001	0,82

$P(D_2=1 \mid D_1=0, H=0) = 0,08$ YES!

$P(D_2=0 \mid D_1=0, H=0) \rightarrow 0,82 \mid 0,92 = 0,891$

(2) $P(H=1) = 0,6015$

$$P(H=1 \mid D_1=1) = \frac{P(D_1=1 \mid H=1) \cdot P(H=1)}{P(D_1=1)} = 0,0146 = P(H=1 \mid D_2=1)$$

$P(D_1=1)$

$P(D_1=1, H=1) + P(D_1=1, H=0)$

$= P(D_1=1 \mid H=1) \cdot P(H=1) + P(D_1=1 \mid H=0) \cdot P(H=0)$

(3) $P(H=1 \mid D_1=1, D_2=1)$

$$= \frac{P(D_1=1, D_2=1 \mid H=1) \cdot P(H=1)}{P(D_1=1, D_2=1)} = 0,4240$$

$P(D_1=1, D_2=1)$

$P(D_1=1, D_2=1 \mid H=1) \cdot P(H=1) + P(D_1=1, D_2=1 \mid H=0) \cdot P(H=0)$

$(*) = P(D_1=1 \mid D_2=0, H=0) \cdot P(D_2=0 \mid H=0)$

Tại $D_1 = D_2 = 1$

$\Rightarrow P(D_1=1 \mid D_2=0, H=0) = 0,08$

$0,02 = P(D_2=1 \mid D_1=1, H=0)$

$P(D_1=0 \mid D_2=0, H=0) = 0,92$

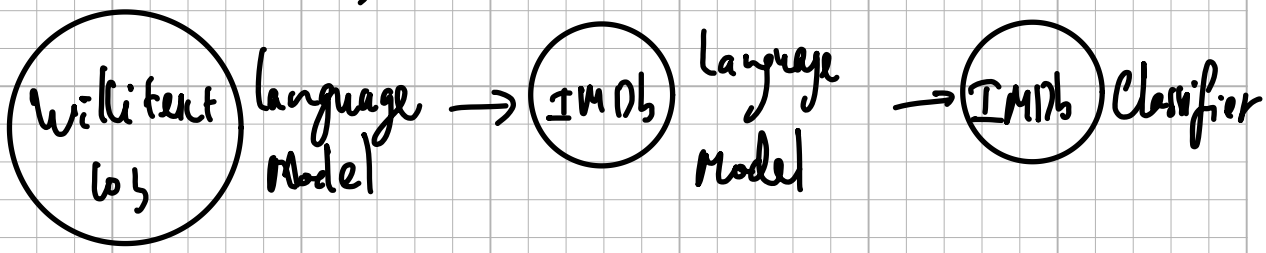
$0,1$

$\Rightarrow P(D_2=1 \mid D_1=1, H=0) = 0,2$

$\Rightarrow P(D_2=0 \mid D_1=1, H=0) = 0,8 \Rightarrow P(D_1=1, D_2=0 \mid H=0)$ $(*)$

[Coding 2.1] Hugging face Datasets and Text Summarization

Transfer learning → Mehr Kontext als Trainig



Bert: Bidirectional Encoder.

Input Representation

- WordPiece: Tokenizer / [CLS]

  FFNN + Softmax

→ Masked Language Model

  why IS? Too little: too expensive

  Too much: Not enough content

→ Next sentence prediction (Also commonly used)

- Byte-Pair Encoding

  Unknown tokens

- Long texts

+) Compact Pretrained Model

Downstream tasks

- Sequence Classification
- Token classification

+) token classification

Last Bert

+) Fine-tuning Strategy

Adapter LORA

+) Word tokenization

+ <UNK>

+) Sub word tokenization (units such as morphemes or phonemes)

1994 pair encoding

→ byte-pair encoding

3.1.C. Exercises

1. Data $x_1 \dots x_n \in \mathbb{R}$. find constant b

$$\text{Argmin}_b \sum_i (x_i - b)^2$$

(1) $f(b) = \sum_i (x_i - b)^2 \Rightarrow f' = -\sum_i 2(x_i - b) = 0$

$$\Rightarrow \sum_i x_i = nb \Rightarrow b = \frac{\sum_i x_i}{n}$$

$f'' > 0 \Rightarrow$ convex $\Rightarrow$

$$\min (X - b)^T (X - b)$$

$$\left((x_i - b)^2\right)' = \sum_i 2 \cdot (x_i - b) = 0$$

$$\Sigma_i x_i = nb.$$