

## 本周进度

- 学习 CS224n 课程
- 学习 pytorch / transformers 框架

## 学习笔记

### CS224n - MLP

#### Lecture 01.

- Directory:
- ① Human language and word meaning.
  - ② Word2vec objective function gradients. 目标函数梯度
  - ③ Optimization basics.

#### Course:

- ① How do we represent the meaning of a word? 符号意义  
Signifier (symbol)  $\leftrightarrow$  signified (idea or thing) = denotational semantics
- ② How do we have usable meaning in a computer? WordNet  $\rightarrow$  一个包含同义词集 (synonym sets) 和超义词 (hypernyms) 的词典
- ③ Representing words by their context.

Distributional semantics: 根据附近词来推断词的意义 (分布语义学) “上下文”

- ④ 词向量: Word vectors — “以所有相似文本上下文中的相似词”

- ⑤ word2vec:
  1. 大量文本;
  2. 每个词由一个向量表示;
  3. 每个词的位置有一个中心词  $C$  和上下文词  $O$  center word (contexts / co-occurrences) words
  4. 使用词向量的相似性来计算与  $C$  的相似性。
  5. 调节词向量最大化可能。

#### \* Word2vec: object function — 目标函数

对于每一个位置  $t=1, \dots, T$ , 在一个固定大小  $m$  的窗口内, 给定中心词  $w_t$  来预测上下文词。

$$\text{可能性 (likelihood)} = L(\theta) = \prod_{t=1}^T \prod_{\substack{j=1 \\ j \neq t}}^m P(w_{tj} | w_t; \theta)$$

$\theta \rightarrow$  所有要优化的变量。

objective function  $J(\theta) \Rightarrow$  (平均) 负对数似然函数  $\Leftrightarrow$  the (average) negative log likelihood.

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{\substack{j=1 \\ j \neq t}}^m \log P(w_{tj} | w_t; \theta)$$

Minimizing objective function  $\Leftrightarrow$  Maximizing predictive accuracy  
最小化目标函数  $\Leftrightarrow$  最大化预测精度。

Question: How to calculate  $P(w_{tj} | w_t; \theta)$ ?

Answer: use two vectors per word  $w$ :  
 $v_w$  when  $w$  is a center word  
 $u_w$  when  $w$  is a context word

$\hookrightarrow$  for a center word  $C$  and a context word  $O$ .

$$P(O|C) = \frac{\exp(u_O^T v_C)}{\sum_{w \in V} \exp(u_w^T v_C)}$$

Word2Vec: 预测词  $\vec{v}$  prediction function.

$$p(c|c) = \frac{\exp(u^T v_c)}{\sum_{w \in V} \exp(u^T v_w)}$$

指数化可以任何数变为

对整个词汇归一化 (Normalized), 生成概率分布.

点积表示 D.C 的相似度

$$u^T v = u \cdot v = \sum_{i=1}^n u_i v_i$$

更大的点积 = 更大的概率

证:  $\partial x^T a / \partial x = \partial a^T x / \partial x = a$

## 下周任务

- 继续学习课程
- 开始写 **Demo**
- 阅读CS课程第一节所涉及到的论文:
  - **A Neural Probabilistic Language Model**
  - **Efficient Estimation of Word Representations in Vector Space**
  - **Natural Language Processing (Almost) from Scratch**
  - **word2vec Parameter Learning Explained**