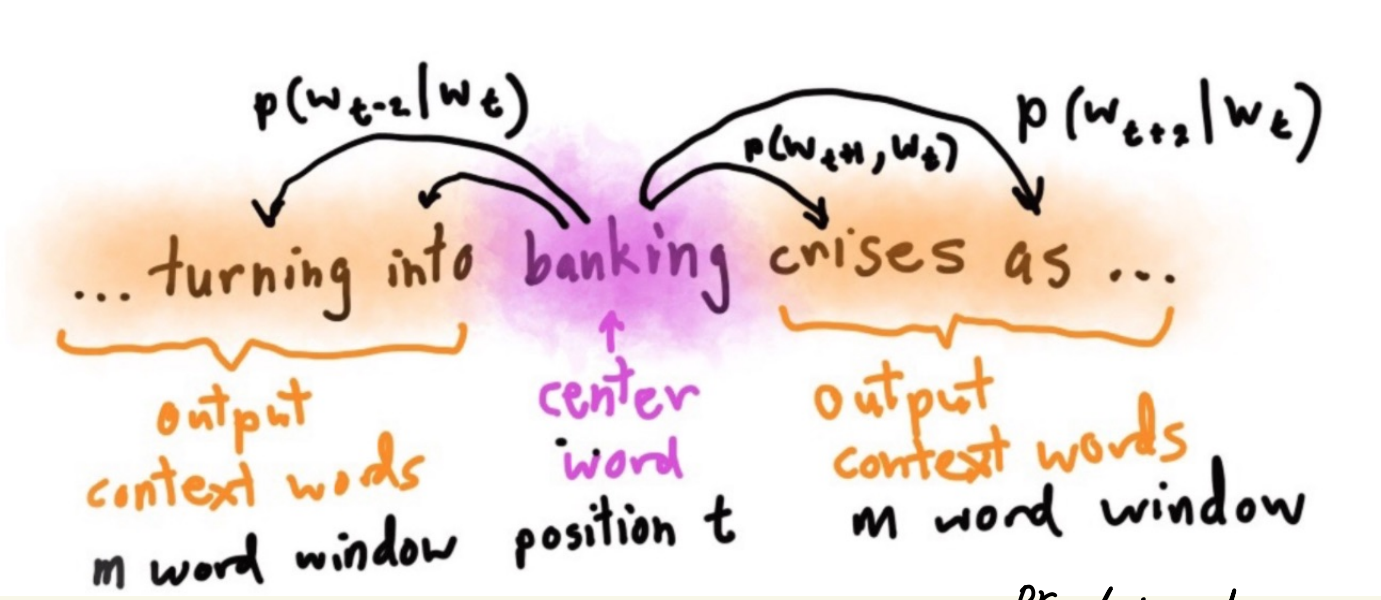
HXH 2022.10.25

Word2Vec: 文本转化为向量，one-hot不可行 { 维度高 / 失去语义关系

那如何通过向量表达句子中词语(整句话)的意思？

通过调整一个单词和该单词的上下文单词的向量，然后根据两个向量来计算相似度 / 或者根据向量推测上下文

Word2Vec { skip-gram 中心词预测上下文 / CBOW 上下文预测中心词



Pr. christopher manning

首先，定义一个目标函数，定义为预测结果的乘积。

$$\max J(\theta) = \prod_{t=1}^{T} \prod_{-m \le j \le m} p(w_{t+j} | w_t ; \theta)$$

m：上下文窗口大小
T：总的词数量，
每一个词前后m个词的概率都要计算

⇓ 方便计算，用log

$$\min J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \le j \le m} \log p(w_{t+j} | w_t)$$

那么 $p(w_{t+j} | w_t)$ 怎么计算呢？ ⇒ 通过softmax

将向量代入softmax中：

↓

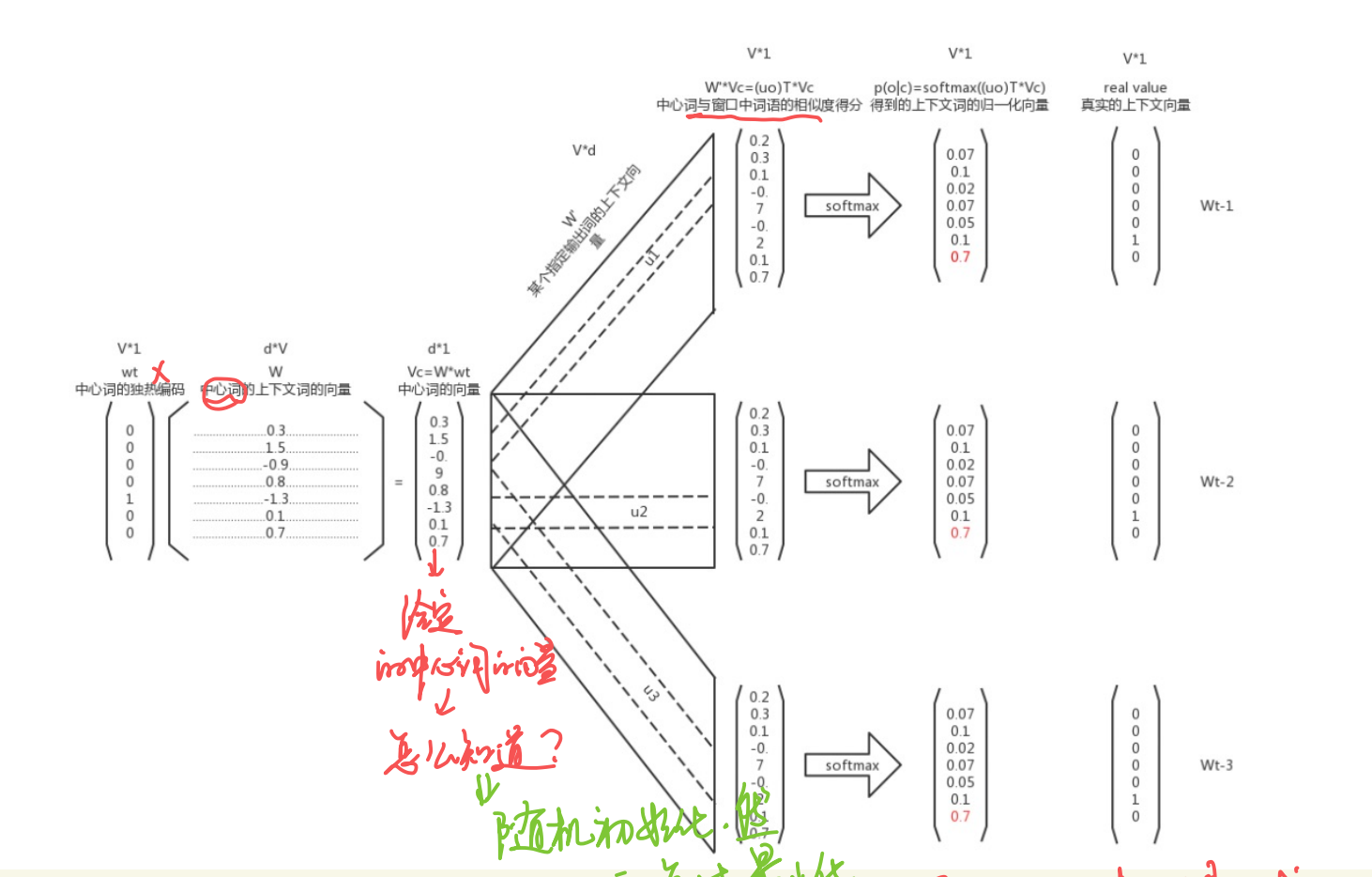
转化为0-1的概率分布

$$p(o|c) = \frac{\exp(u_o^T \cdot v_c)}{\sum_{w=1} \exp(u_w^T \cdot v_c)}$$

o：上下文中某一个词
c：中心词
u：对应in上下文词向量
v：指词向量

所以 u_o 表示中心词 c in 上下文词向量中 m 一个
v_c 表示中心词 in 词向量

重点看 skip-gram in 算法流程：



除去 in 中心词 in 词向量

怎么知道？

随机初始化，通过最小化损失函数最大化输出概率来更新得到最终 in 词向量

最终损失函数就是比较计算出 in 词向量与真实词向量 in 差距

目标 {
- 最小化损失函数
- 最大化输出概率

梯度下降法

将常数去掉，损失函数就可以表示为：

$$J(\theta) = \log \frac{\exp(u_o^T \cdot v_c)}{\sum_{w=1}^{V} \exp(u_w^T \cdot v_c)}$$

对 v_c 求偏导：

$$\frac{\partial J(\theta)}{\partial v_c} = \frac{\partial}{\partial v_c} \log \frac{\exp(u_o^T \cdot v_c)}{\sum_{w=1}^{V} \exp(u_w^T \cdot v_c)}$$

$$= \frac{\partial}{\partial v_c} \log \exp(u_o^T \cdot v_c) - \frac{\partial}{\partial v_c} \log \sum_{w=1}^{V} \exp(u_w^T \cdot v_c)$$

第一部分：u_o　　第二部分

第二部分：$\frac{\partial}{\partial v_c} \log \sum_{w=1}^{V} \exp(u_w^T \cdot v_c)$

$$= \frac{1}{\sum_{w=1}^{V} \exp(u_w^T \cdot v_c)} \cdot \frac{\partial}{\partial v_c} \sum_{x=1}^{V} \exp(u_x^T \cdot v_c)$$

$$= \frac{1}{\sum_{w=1}^{V} \exp(u_w^T \cdot v_c)} \cdot \sum_{x=1}^{V} \exp(u_x^T \cdot v_c) \cdot u_x$$

$$= \sum_{x=1}^{V} \frac{\exp(u_x^T \cdot v_c)}{\sum_{w=1}^{V} \exp(u_w^T \cdot v_c)} \cdot u_x$$

$$= \sum_{x=1}^{V} p(x|c) \cdot u_x$$

$p(x|c)$：表示 x 词在中心词 c 为condition情况下为观测的概率

整体来看 x 词上下文的向量，这样就是上下文向量的期望了

所以，第2次就是所有上下文词向量的期望

那么，整个损失函数可表示为：

$$J(\theta) = u_o - \sum_{x=1}^{V} p(x|c) \cdot u_x = observed - expected$$

observed：观测值（真实）　expected：期望值

所以，要最小化损失函数，让期望值和真实值尽可能的接近

最小时对应的词向量即为我们所需要的

ps：参考 CSDN 博主：

深圳湾以北.

<<Word2vec 算法理解和基础实现>>