

# GloVe: Global Vectors for Word Representation

*EMNLP2014)*

**Jeffrey Pennington**

**Richard Socher**

**Christopher D. Manning**

Stanford

2015 年 6 月 16 日

- 词向量模型
  - global matrix factorization (LSA)
  - local context window (skip-gram)
- word2vec 的实现方式
- GloVe
  - 模型（目标函数）
  - 对比 skip-gram
  - 复杂度分析
- 实验
  - Word analogies ( $A-B=C-?$ )
  - Word Similarity
  - Named entity recognition

global matrix factorization

- 可以利用预料中出现频率信息，但是处理  $A-B=C-?$  问题效果不好

local context window

- 处理  $A-B=C-?$  效果很好，但是利用局部单词的左右窗口内词作为上下文信息，不能利用全局层面的共现信息

GloVe

- 综合
- 建立满足  $A-B=C-D$  的模型，利用共现信息（共现频率）作为权值标准，最优化模型误差

## vector models

### global matrix factorization

- 可以利用预料中出现频率信息，但是处理  $A-B=C?$  问题效果不好

### local context window

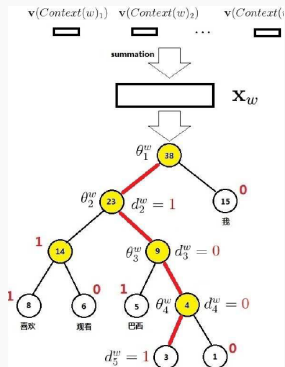
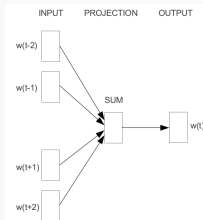
- 处理  $A-B=C?$  效果很好，但是利用局部单词的左右窗口内词作为上下文信息，不能利用全局层面的共现信息

### GloVe

- (综合) 建立满足  $A-B=C-D$  的模型，利用共现信息（共现频率）作为权值标准，最优化模型误差

## word2vec 的实现方式

# word2vec 实现: CBOW

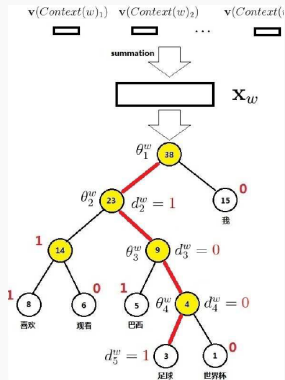
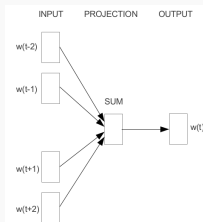


- 前后窗口（宽度 1-4 随机）内所有单词的词向量平均作为隐层词向量
- 所有单词按照词频构建 huffman 树，每个中间结点对应一个判别向量，和隐层词向量、目标词向量维数相同
- 每个中间结点的词向量和隐层词向量的点积做 sigmoid 变换后的值，作为选择左分支（0）的概率，否则为走右支（1）的概率
- 每个单词是对应一个 huffman 编码  
足球 *fb* (1001), 观看 (110)

$$p(fb|context(fb)) = \prod_{j=2..5} p(d_j^w | X_w, \theta_{j-1}^w)$$

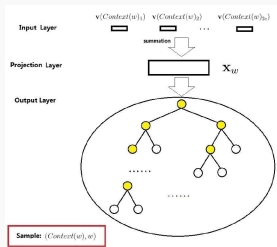
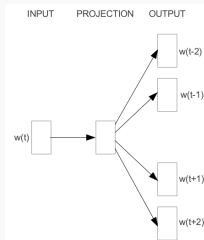
$$p(d_j^w | X_w, \theta_{j-1}^w) = \begin{cases} \sigma(x_w^T \theta_{j-1}^w) & d_j^w = 0 \\ 1 - \sigma(x_w^T \theta_{j-1}^w) & d_j^w = 1 \end{cases} \quad (1)$$

# word2vec 实现: CBOW



- 前后窗口（宽度 1-4 随机）内所有单词的词向量平均作为隐层词向量
- 所有单词按照词频构建 huffman 树，每个中间结点对应一个判别向量，和隐层词向量、目标词向量维数相同
- 每个中间结点的词向量和隐层词向量的点积做 sigmoid 变换后的值，作为选择左分支（0）的概率，否则为走右支（1）的概率
- 每个单词是对应一个 huffman 编码
- 最优化  $\sum_{w \in C} \log p(c|context(c))$

## word2vec 实现:Skip-gram



- 类似 CBOW
- 隐层就是输入的当前单词向量
- 不同的是每个隐层向量需要生成多个单词（窗口内的所有单词都预测，不分顺序）
  - 多一个  $\prod$  循环
- 最优化  $\sum_{w \in C} \prod_{w_j \in \text{window}_c} \log p(w_j | c)$



# GloVe Model

# GloVe Model

- $X_{ij}$ :  $word_j$  作为  $word_i$  上下文出现次数
- $X_i = \sum_k X_{ik}$
- $P_{ij} = P(j|i) = X_{ij}/X_i$
- $F(w_i, w_j, \tilde{w}_k) = P_{ik}/P_{jk}$

- ① 改用向量差别 
$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$
- ② 保持线性结构 
$$F((w_i - w_j)^T \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$
- ③ 我们采用  $F = \exp$  或者  $w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$ , 使得下面的式子成立 
$$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i} \quad F((w_i - w_j)^T \tilde{w}_k) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$$
- ④ 如果是采用  $w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$ , 通过增加偏置项使得模拟  $w_i^T \tilde{w}_k$  与  $X_{ij}$  之间的关系: 
$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

# GloVe Model

- $X_{ij}$ :  $word_j$  作为  $word_i$  上下文出现次数
- $X_i = \sum_k X_{ik}$
- $P_{ij} = P(j|i) = X_{ij}/X_i$
- $F(w_i, w_j, \tilde{w}_k) = P_{ik}/P_{jk}$

- ① 改用向量差别 
$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$
- ② 保持线性结构 
$$F((w_i - w_j)^T \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$
- ③ 我们采用  $F = \exp$  或者  $w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$ , 使得下面的式子成立 
$$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i} \quad F((w_i - w_j)^T \tilde{w}_k) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$$
- ④ 如果是采用  $w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$ , 通过增加偏置项使得模拟  $w_i^T \tilde{w}_k$  与  $X_{ij}$  之间的关系: 
$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

# GloVe Model

- $X_{ij}$ :  $word_j$  作为  $word_i$  上下文出现次数
- $X_i = \sum_k X_{ik}$
- $P_{ij} = P(j|i) = X_{ij}/X_i$
- $F(w_i, w_j, \tilde{w}_k) = P_{ik}/P_{jk}$

① 改用向量差别

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

② 保持线性结构

$$F((w_i - w_j)^T \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

③

$$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i} \quad F((w_i - w_j)^T \tilde{w}_k) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$$

④

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

⑤

目标函数

$$J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

# GloVe Model

- $X_{ij}$ :  $word_j$  作为  $word_i$  上下文出现次数
- $X_i = \sum_k X_{ik}$
- $P_{ij} = P(j|i) = X_{ij}/X_i$
- $F(w_i, w_j, \tilde{w}_k) = P_{ik}/P_{jk}$

① 改用向量差别

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

② 保持线性结构

$$F((w_i - w_j)^T \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

③

$$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i} \quad F((w_i - w_j)^T \tilde{w}_k) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$$

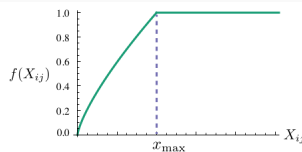
④

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

⑤

目标函数  $J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$



# GloVe Model

GloVe 的目标函数  $J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$  对比 window-based methods (skip-gram and ivLBL) 的目标函数:

- 当前词  $word_i$  预测其上下文单词  $word_j$ :

$$Q_{ij} = \frac{\exp(w_i^T \tilde{w}_j)}{\sum_{k=1}^V \exp(w_i^T \tilde{w}_k)}$$

- 目标函数为  $J = - \sum_{\substack{i \in \text{corpus} \\ j \in \text{context}(i)}} \log Q_{ij}$

- 合并目标函数中的相同项:

$$J = - \sum_{i=1}^V \sum_{j=1}^V X_{ij} \log Q_{ij}$$

- 根据  $X_i = \sum_k X_{ik}$ ,  $P_{ij} = P(j|i) = X_{ij}/X_i$ , 转化目标函数

$$J = - \sum_{i=1}^V X_i \sum_{j=1}^V P_{ij} \log Q_{ij} = \sum_{i=1}^V X_i H(P_i, Q_i)$$

# GloVe Model

GloVe 的目标函数  $J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$  对比 window-based methods (skip-gram and ivLBL) 的目标函数:

- 转化目标函数  $J = - \sum_{i=1}^V X_i \sum_{j=1}^V P_{ij} \log Q_{ij} = \sum_{i=1}^V X_i H(P_i, Q_i)$
- 归一化的  $P_{ij}, Q_{ij}$  计算复杂, 不使用交叉熵, 改用压缩过的最小均方

$$\begin{aligned} \hat{J} &= \sum_{i,j} X_i (\log \hat{P}_{ij} - \log \hat{Q}_{ij})^2 \\ \hat{J} &= \sum_{i,j} X_i (\hat{P}_{ij} - \hat{Q}_{ij})^2 = \sum_{i,j} X_i (w_i^T \tilde{w}_j - \log X_{ij})^2. \end{aligned}$$

误差衡量  $P, Q$  差别大小

- Mikolov 提出限制权重来消除高频词的过度影响

$$\hat{J} = \sum_{i,j} f(X_{ij}) (w_i^T \tilde{w}_j - \log X_{ij})^2$$

- GloVe 的目标函数是合理的 (?)

## 复杂度分析





# Experiments

# Experiments

## Word analogies

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	<u>67.5</u>	<u>54.3</u>	<u>60.3</u>
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	<u>64.8</u>	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	<u>80.8</u>	61.5	<u>70.3</u>
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW <sup>†</sup>	300	6B	63.6	<u>67.4</u>	65.7
SG <sup>†</sup>	300	6B	73.0	66.0	69.1
GloVe	300	6B	<u>77.4</u>	67.0	<u>71.7</u>
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	<b><u>81.9</u></b>	<b><u>69.3</u></b>	<b><u>75.0</u></b>

- 19544 组问题: *a is to b as c is to \_\_\_*?
  - SVD-S: 矩阵元素值压缩  $\sqrt{X_{trunc}}$
  - SVD-L: 压缩  $\log(1 + X_{trunc})$
  - SVD- $f(X)$  ?
- CBOW 效果提升不明显

# Experiments

Word Similarity

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	<u>72.7</u>	75.1	56.5	37.0
CBOW <sup>†</sup>	6B	57.2	65.6	68.2	57.0	32.5
SG <sup>†</sup>	6B	62.8	65.2	69.7	<u>58.1</u>	37.2
GloVe	6B	<u>65.8</u>	<u>72.7</u>	<u>77.8</u>	53.9	<u>38.1</u>
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	<b><u>75.9</u></b>	<b><u>83.6</u></b>	<b><u>82.9</u></b>	<b><u>59.6</u></b>	<b><u>47.8</u></b>
CBOW <sup>*</sup>	100B	68.4	79.6	75.4	59.4	45.5

## Experiments

### NER

Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	<b>88.7</b>	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	<b>93.2</b>	88.3	<b>82.9</b>	<b>82.2</b>

- Discrete: 直接使用 Stanford NER 的输出结果
- 20 维，窗口大小为 5