# Doc2vec 模型结构和训练

**方敏**

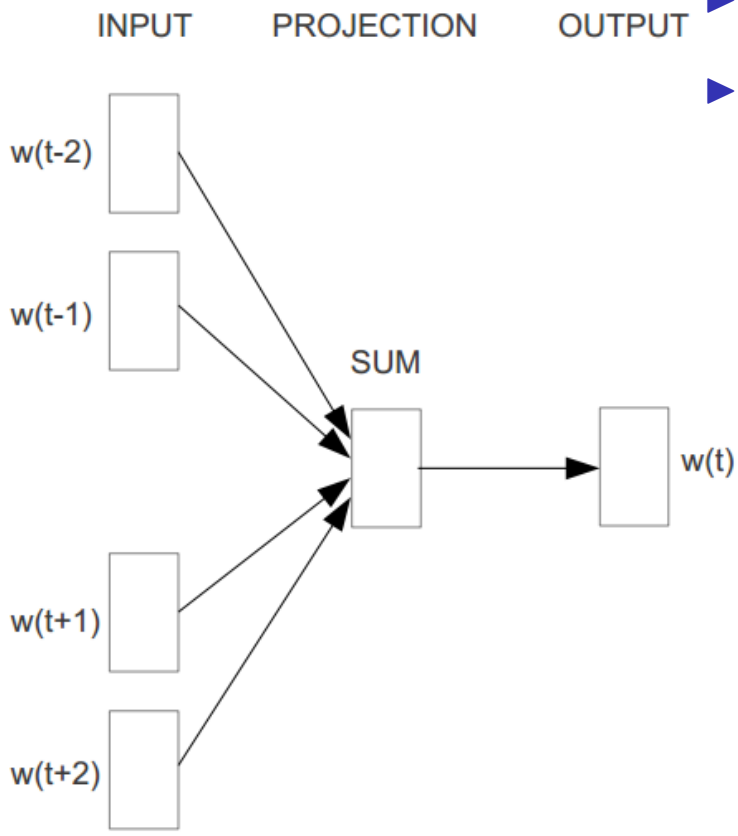
**驰声研发**

2016 年 10 月

# overview

- 背景知识：word2vec 模型
- Doc2vec 模型结构介绍
- 模型训练与 doc 向量获取
- Optimizing Computational Efficiency
- 其它模型

# 背景知识：word2vec 模型




图：CBOW(Continuous Bag-of-Words Model)

- 预测 $p(w_t|w_{t-k},...,w_{t+k})$
- 目标函数如下

$$\frac{1}{T}\sum_{t=k}^{T-k} logp(w_t|w_{t-k},...,w_{t+k})$$

$$p(w_t|w_{t-k},...,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

$$y_{w_t} = b + Uh(w_{t-k},...w_{t+k}) \quad (1)$$

$h$ 表示 INPUT 到 PROJECTION 的转化关系，如求和

# Doc2vec 模型结构介绍

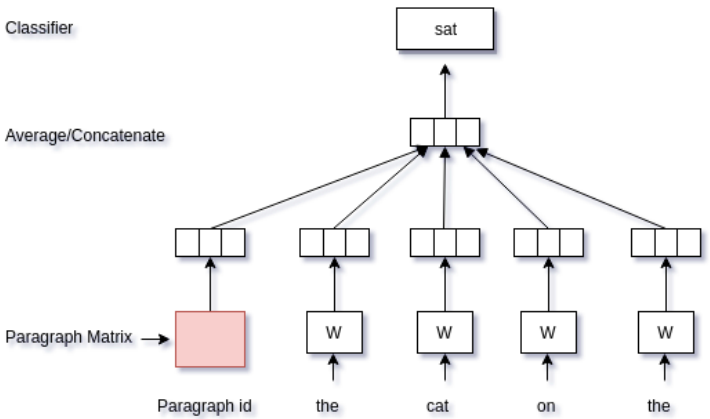



图: Doc2vec PV-DM

- **预测** $p(w_{t,j}|w_{t-k,j}, \ldots, w_{t+k,j}, d^j)$
  $w_{t,j}$ **表示第** $j$ **篇** doc **的第** $t$ **个词，** $d^j$ **表示第** $j$ **篇** doc,$k$ **为** $1/2$ **窗长**
- **分类器**

$$p(w_{t,j}|w_{t-k,j}, \ldots, w_{t+k,j}) = \frac{e^{y_{w_{t,j}}}}{\sum_i e^{y_i}}$$

$$y_{w_{t,j}} = b + Uh(w_{t-k,j}, \ldots w_{t+k,j}, d_j) \quad (2)$$

# Doc2vec 模型结构介绍

目标函数：

$$\frac{1}{J}\sum_{j=1}^{J}\frac{1}{N_j-2k}\sum_{t=k}^{N_j-k}logp(w_{t,j}|w_{t-k,j},...,w_{t+k,j},d_j) \tag{3}$$

其中，$N_j$ 表示第 $j$ 篇 doc 的字符长度，$J$ 表示语料中包含的 doc 数目

# 模型训练与 doc 向量获取

若不考虑使用输出反馈：

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1}) \tag{4}$$

$$y_t = g(W_{hy}h_t) \tag{5}$$

# Optimizing Computational Efficiency

- Negative Sampling
- Hierarchical Softmax

$$\frac{1}{T}\sum_{t=k}^{T-k} logp(w_t|w_{t-k}, ..., w_{t+k})$$

$$p(w_t|w_{t-k}, ..., w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{6}$$

$$y_{w_t} = b + Uh(w_{t-k}, ...w_{t+k})$$

# 其它模型

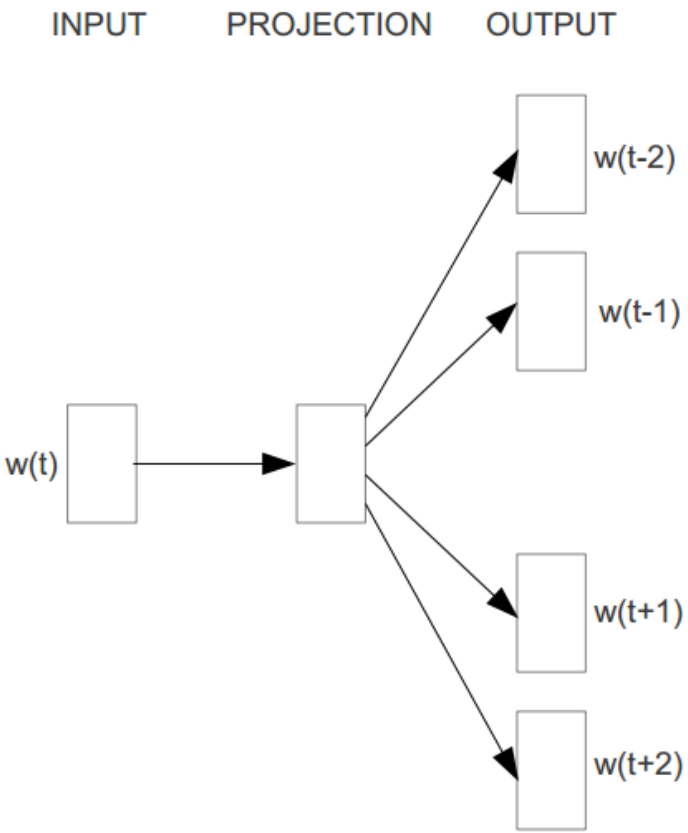



图: CBOW(Continuous Bag-of-Words Model)

- 预测 $p(w_t|w_{t-k},...,w_{t+k})$
- 目标函数如下

$$\frac{1}{T}\sum_{t=k}^{T-k}\sum_{-k<=l<=k,l!=0} logp(w_{t+l}|w_t)$$

$$p(w|w_t) = \frac{e^{y_w}}{\sum_i e^{y_i}}$$

$$y_w = b + Uh(w_t) \tag{7}$$

# 其它模型

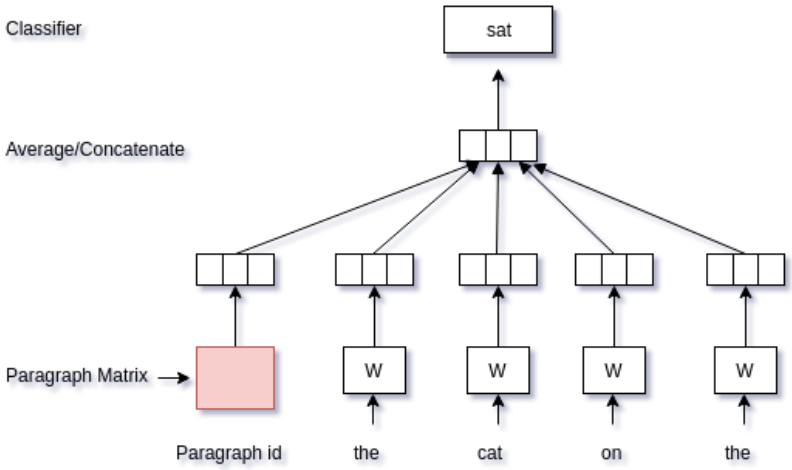



图: Doc2vec PV-DM

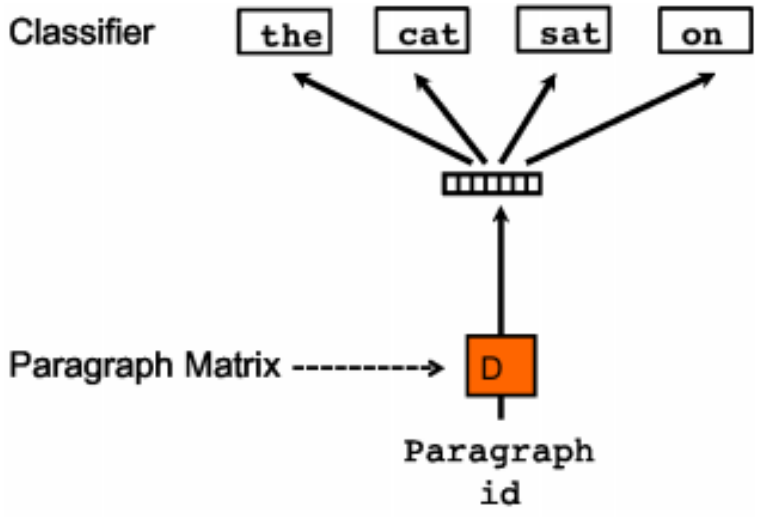



图: Doc2vec PV-DBOW