Sentence representation via Recursive Nerual Networks

Zhe Han

iampkuhz@gmail.com

2015 年 1 月 9 日

Outline

- 3 个表示句子向量的模型
  - (展开的) 递归自编码机解决复述问题 (2011NIPS)
  - 递归自编码机 + 解决情感分析问题 (2011EMNLP)
  - 基于语义依存树构建的递归神经网络解决图像描述问题 (2014TACL)
- 不同句子向量模型的分析比较
  - 穿插在模型介绍中

socher's paper: www.socher.org

- Dynamic pooling and unfolding recursive autoencoders for paraphrase detection
- Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions
- Grounded Compositional Semantics for Finding and Describing Images with Sentences

socher

- stanford 毕业的博士
  - Chris Manning 和 Andrew Ng 的学生, 创业公司 MetaMind 的 CTO
- 此人非常喜欢用递归神经网络 (Recursive Nerual Network)
  - 罗炳峰: 2014NIPS Global Belief Recursive Neural Networks

Motivation

- 为什么要表示句子的语义向量
  - 单词的语义向量以及有比较好的表示了 (word2vec)
  - 中文维基百科谓词归一

```
Word: 出生  Position in vocabulary: 345

                                   Word       Cosine distance
------------------------------------------------------------------------
                                   生于              0.773387
                                 出生地        0.622590
                                   出身              0.605404
                                   现居              0.595559
                                   移居              0.585317
                                   旅居              0.570838
                                   长大              0.570293
                               他的父亲        0.560490
```

```
Word: 坐标  Position in vocabulary: 1986

                                   Word       Cosine distance
------------------------------------------------------------------------
                                   座标              0.692477
                                   向量              0.664867
                                   原点              0.659200
                                   矢量              0.653525
                                 法向量        0.646993
```

Motivation

- 为什么要表示句子的语义向量
  - 单词的语义向量以及有比较好的表示了 (word2vec)
  - 中文维基百科谓词归一
    - 1500/16000 谓词含有 word2vec 向量，其余为低频词或组合词
  - 利用客体的语义信息提取特征
    - 客体一般是短语 or 句子，需要从词向量提取短语向量

复述检测

socher 2011NIPS

- Dynamic pooling and unfolding recursive autoencoders for paraphrase detection

Paraphrase identification

- definition
  - 给定一组句子, 判断其是否是复述
    - binary classification
- Microsoft Research Paraphrase Corpus (MSRP)
  - train: 4,076 sentence pairs (2,753 positive: 67.5 %)
  - test: 1,725 sentence pairs (1,147 positive: 66.5 %)
  - 2 个标注者, 83% 的一致性, 第三个人更正

Sample data

- Sentence 1: Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
- Sentence 2: Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
- Class: 1 (true paraphrase)

Paraphrase identification

- 常用方法
  - 提取词汇特征, 语义特征
    - n-gram features, skip-gram fatures; POS tag, wordnet similarity, dependency tree relation, ...
  - SVM 分类
    - 或是投票分类
- Challenge
  - 没有提取句子的全局信息 (dependency features 利用不足)
  - 对句子涵义的特征提取不足 (没有真正理解句子)

Paraphrase identification

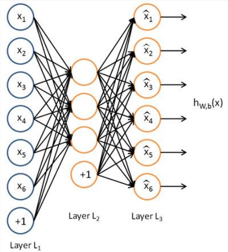
socher 的方法

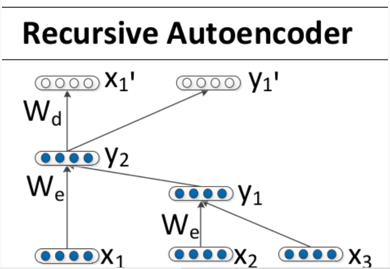
- 利用 NYT 新闻训练每个单词的向量 (100 维)
- 对于每个句子 (多个单词向量) 采用训练一个递归的自动编码机, 得到一个句子级别的语义向量.
- 通过判断两个句子的语义向量的相似性得到语义相似性特征

- 递归的自动编码机 (Unfolding Recursive Autoencoder)
  - 抽取句子的语义向量, 得到语法数上每个节点 (单词, 短语) 的向量
- Dynamic Pooling
  - 对于长度变化的两个句子, 抽取固定维数的特征

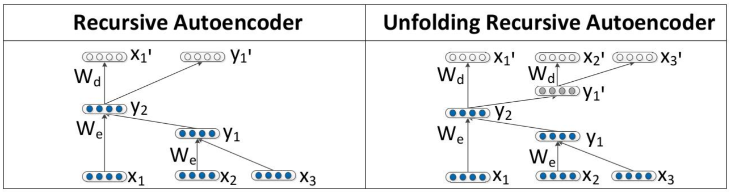
Unfolding Recursive Autoencoder

- Autoencoder
  - 希望压缩的特征 (L2 层) 能表示原数据 (L1 层)
    - 能表示等价于可以还原 (L3 层向量约等于 L1 层向量)
- Recursive Autoencoder
  - Autoencoder in recursive structure
    - Pollack 提出 (1990)
    - 词向量没有压缩: $(0, ..., 0, 1, 0, ..., 0)$
  - 进一步的, 对于深层的网络 (语法树), 递归使用同一个简单的 Autoencoder



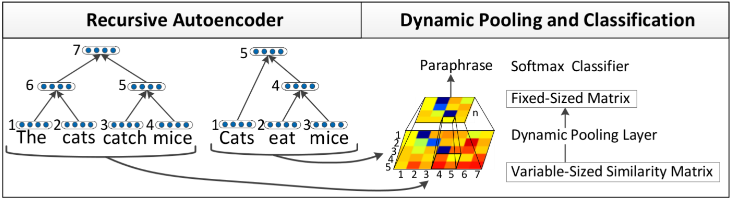


Unfolding Recursive Autoencoder

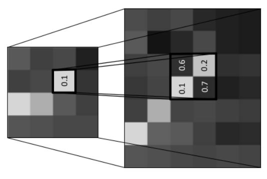


- Recursive Autoencoder
- Comparison (on $y_2 - > x_1 y_1$)
  - Neural network
    - minimum $\| y_2' - y_2 \|$
  - Recursive Autoencoder
    - minimum $\| [x_1'; y_1'] - [x_1; y_1] \|$
  - Unfolding Recursive Autoencoder
    - minimum $\| [x_1'; x_2'; ...; x_j'] - [x_1; x_2; ...; x_j] \|$

Dynamic Pooling(简要了解)



- motivation
  - 如何对两个长度变化的句子抽取固定维数的特征?
    - 长度为 n 的句子，cTree 有（2n-1）个节点
  - 把不同长度的句子压缩（扩张）到相同的维数



Paraphrase identification

- QA
  - 为何使用 uRAE 而不是 RAE 或者两个子节点的向量平均?
    - 多个单词组成的句子/短语（高层节点），需要更多的单词信息，RAE 只关心最近的 2 个儿子节点
    - 向量平均：两个儿子向量的平均忽视了结构关系
    - 实验证明，向量平均找不出来；RAE 对 2 个单词组成的短语，识别其近义词效果很好；uRAE 对于 2-3 个单词组成的短语的效果很好，甚至 5 个单词组成的短语有些也可以正确找到。

Center Phrase	Recursive Average	RAE	Unfolding RAE
the U.S.	the U.S. and German	the Swiss	the former U.S.
suffering low morale	suffering a 1.9 billion baht UNK 76 million	suffering due to no fault of my own	suffering heavy casualties
to watch hockey	to watch one Jordanian border policeman stamp the Israeli passports	to watch television	to watch a video
advance to the next round	advance to final qualifying round in Argentina	advance to the final of the UNK 1.1 million Kremlin Cup	advance to the semis
a prominent political figure	such a high-profile figure	the second high-profile opposition figure	a powerful business figure
Seventeen people were killed	"Seventeen people were killed, including a prominent politician "	Fourteen people were killed	Fourteen people were killed
conditions of his release	"conditions of peace, social stability and political harmony "	conditions of peace, social stability and political harmony	negotiations for their release

情感分析

socher 2011EMNLP

- Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions

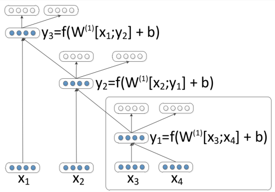
Predicting Sentiment

- identification
  - 给定一段话, 判断其情感极性 (积极/消极, 正面评价/负面评价)
  - 给定一段话, 判断其评分分布 (1-5 颗星)
- Experience project dataset(Potts, 2010)
  - 31,676 个段子, 74,859 条评分
  - 选取点评次数大于等于 4 次的段子

Sample data

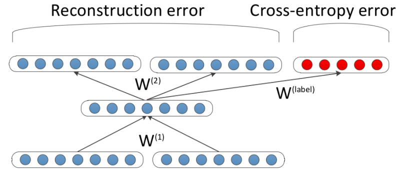
KL	Predicted&Gold	V.	Entry (Shortened if it ends with ...)
.03	 .16 .16 .16 .33 .16	6	I reguarly shoplift. I got caught once and went to jail, but I've found that this was not a deterrent. I don't buy groceries, I don't buy school supplies for my kids, I don't buy gifts for my kids, we don't pay for movies, and I dont buy most incidentals for the house (cleaning supplies, toothpaste, etc.)...
.03	 .38 .04 .06 .35 .14	165	i am a very succesfull buissnes man.i make good money but i have been addicted to crack for 13 years.i moved 1 hour away from my dealers 10 years ago to stop using now i dont use daily but once a week usally friday nights. i used to use 1 or 2 hundred a day now i use 4 or 5 hundred on a friday.my problem is i am a funcational addict...
.05	 .14 .28 .14 .28 .14	7	Hi there, Im a guy that loves a girl, the same old bloody story... I met her a while ago, while studying, she Is so perfect, so mature and yet so lonely, I get to know her and she get ahold of me, by opening her life to me and so did I with her, she has been the first person, male or female that has ever made that bond with me,...

Predicting Sentiment



- Unsupervised Recursive Autoencoder for Structure
  - 贪心的构造二叉树
    - 每次计算当前状态任意一对相邻节点的合并代价，取代价最小的一对合并，直到结束
    - $p = f(W^{(1)}[c1 : c2] + b^{(1)})$, $[c_1' : c_2'] = W^{(2)}p + b^{(2)}$
  - 考虑两个子节点的块大小
    - 块越大越重要．体现在重构误差中
    - $E_{rec}([c_1 : c_2]; \theta) = \frac{n_1}{n_1+n_2} \parallel c_1 - c_1' \parallel^2 + \frac{n_2}{n_1+n_2} \parallel c_2 - c_2' \parallel^2$

Predicting Sentiment



- Semi-supervised Recursive Autoencoder for Structure
  - 扩展向量，加入情感分布向量 d(维数为分类的个数)
  - 预测分布：$d(p;\theta) = softmax(W^{label}p)$
  - 真实分布：t
  - 采用交叉熵估计损失：$E_{cE}(p,t;\theta) = -\sum_{k=1}^{K} t_k \log d_k(p;\theta)$
  - 总体的损失函数为：

    $$E([c_1 : c_2]_s, p_s, t, \theta) = \alpha E_{rec}([c_1 : c_2]; \theta) + (1-\alpha) E_{cE}(p,t;\theta)$$

Predicting Sentiment

- 对比之前的 RNN 模型
  - 加入了情感分布特征
    - 单纯的语言模型是不带情感极性的: good 和 bad 词向量很像
  - 没有使用句法树作为递归结构
    - 采用贪心的方法逐次向上递归
    - 句子的情感极性和句法结构并没有必然联系 (情感性一般蕴含于修饰词)

理解图像描述问题

socher 2014TACL

- Grounded Compositional Semantics for Finding and Describing Images with Sentences

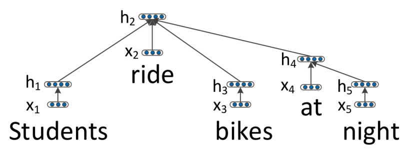
Finding and Describing Images with Sentences

- definition
  - 给定一段描述 (一个句子), 找出其描述的图片
  - 给定一张图片, 找出描述他的句子
- Rashtchian et al., 2010 dataset
  - 1000 images, each with 5 sentences

Sample data

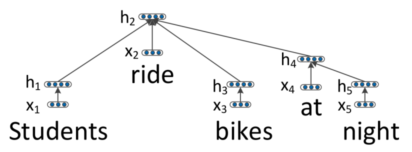
1. A woman and her dog watch the cameraman in their living with wooden floors.
2. A woman sitting on the couch while a black faced dog runs across the floor.
3. A woman wearing a backpack sits on a couch while a small dog runs on the hardwood floor next to her.
4. A women sitting on a sofa while a small Jack Russell walks towards the camera.
5. White and black small dog walks toward the camera while woman sits on couch, desk and computer seen in the background as well as a pillow, teddy bear and moggie toy on the wood floor.

DT-RNN



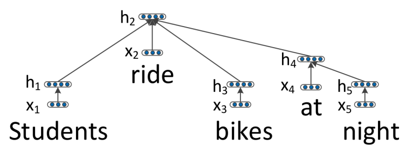
- 语义表征: 自底向上求解
  - $h_c = g_\theta(x_c) = f(W_v x_c)$
  - $h_2 = g_\theta(x_2, h_1, h_3, h_4) = f(W_v x_2 + W_{l1} h1 + W_{r1} h_3 + W_{r2} h_4)$
    - 训练集 $W_{r\cdot} = (W_{r1}, \dots, W_{rk_r}), W_{l\cdot} = (W_{l1}, \dots, W_{lk_l})$
    - 测试集如果 W_{rk_t} 有 $k_t > k_r \to W_{rk_t} = I$
  - 加权: 越大的子块越重要
    - $h_i = f(\frac{1}{l(i)}(W_v x_i + \sum_{j\in C(i)} l(i) W_{pos(i,j)} h_j))$
  - SDT-RNN: Semantic Dependency Tree RNNs
    - 递归矩阵和节点在语法树上的关系类型有关
    - 和在语法树的左右或位置无关: $W_r, W_l \to W_{suj}$

Describing Images



- 模型缺点
  - 中心动词缺失导致结果差别大
    - A blue and yellow airplane flying straight down while emitting white smoke
    - Airplane in dive position

Describing Images



- 对比之前的模型
  - 与传统 RNN 模型区别
    - 叶节点和中间节点不要求维数相同 (通过 W_v 转换)
  - 可以接受多元子节点 (dependency tree vs constituency tree)
  - CTree 的上层节点的重要性明显高, 不平均
  - CTree 更适合情感分析, DTree 更适合提取句子的语义表征
    - 非实词 ("but") 在 CTree 位于较高节点
    - CTree 更能把握句子的中心语义 (中心动词, 主体, 客体)