Report of Deep Learning for Natural Langauge Processing

Yuxiang Cheng　ZY2303301
cyx1313@126.com

Abstract

给利用给定语料库，利用 1～2 种神经语言模型（如：基于 Word2Vec，LSTM， GloVe 等模型）来训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联、或者其他方法来验证词向量的有效性。

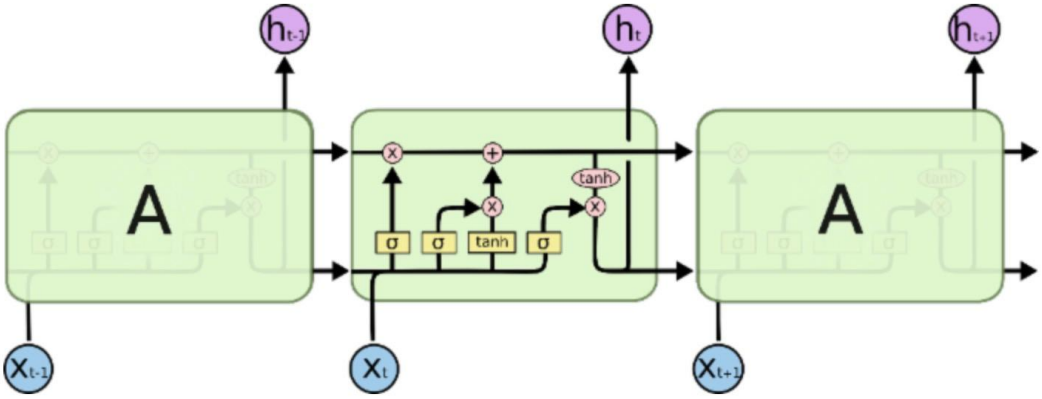
Introduction

LSTM 模型

Long Short Term 网络—— 一般就叫做 LSTM ——是一种 RNN 特殊的类型，可以学习长期依赖信息。LSTM 由 Hochreiter & Schmidhuber (1997) 提出，并在近期被 Alex Graves 进行了改良和推广。在很多问题，LSTM 都取得相当巨大的成功，并得到了广泛的使用。长短时记忆网络（LSTM）是一种特殊类型的 RNN，由 Hochreiter 和 Schmidhuber 于 1997 年提出，目的是解决传统 RNN 的问题。

- 解决梯度消失问题: 通过引入“记忆单元”，LSTM 能够在长序列中保持信息的流动。
- 捕捉长依赖性: LSTM 结构允许网络捕捉和理解长序列中的复杂依赖关系。
- 广泛应用: 由于其强大的性能和灵活性，LSTM 已经被广泛应用于许多序列学习任务，如语音识别、机器翻译和时间序列分析等。

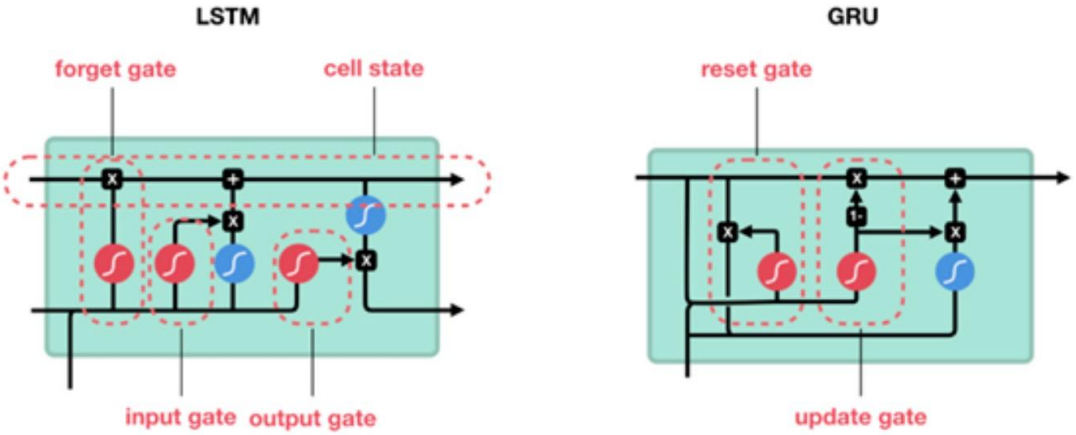
LSTM 的提出不仅解决了 RNN 的核心问题，还开启了许多先前无法解决的复杂序列学习任务的新篇章。标准 RNN 中的重复模块包含单一的层。LSTM 同样是这样的结构，但是重复的模块拥有一个不同的结构。不同于 单一神经网络层，这里是有四个，以一种非常特殊的方式进行交互。



门的相互作用

- 遗忘门：负责控制哪些信息从单元状态中遗忘。
- 输入门：确定哪些新信息被存储。
- 输出门：控制从单元状态到隐藏状态的哪些信息流动。

这些门的交互允许 LSTM 以选择性的方式在不同时间步长的间隔中保持或丢弃信息。



Methodology& Experimental Studies

选择《倚天屠龙记》、《鹿鼎记》、《射雕英雄传》、《神雕侠侣》、《笑傲江湖》、《碧血剑》作为样本，使用了 LSTM 对其中的角色进行了聚类，选取关联度最高的五个名词，其中 LSTM 模型的训练 epoch 为 15。

书名	角色	关联角色	关联度
倚天屠龙记	周芷若	张无忌	0.808
倚天屠龙记	殷素素	张无忌	0.805
倚天屠龙记	张翠山	张无忌	0.780
倚天屠龙记	赵敏	张无忌	0.762
倚天屠龙记	都大锦	张无忌	0.741
鹿鼎记	康熙	韦小宝	0.883
鹿鼎记	吴之荣	韦小宝	0.810
鹿鼎记	陈近南	韦小宝	0.804
鹿鼎记	老者	韦小宝	0.798
鹿鼎记	女尼	韦小宝	0.780
射雕英雄传	黄蓉	郭靖	0.782
射雕英雄传	众人	郭靖	0.728
射雕英雄传	六子	郭靖	0.714
射雕英雄传	欧阳克	郭靖	0.690
射雕英雄传	黄药师	郭靖	0.658
神雕侠侣	小龙女	杨过	0.833
神雕侠侣	李莫愁	杨过	0.803
神雕侠侣	陆无双	杨过	0.798
神雕侠侣	法王	杨过	0.784

神雕侠侣	周伯通	杨过	0.774
笑傲江湖	岳不群	令狐冲	0.800
笑傲江湖	桃花仙	令狐冲	0.792
笑傲江湖	盈盈	令狐冲	0.791
笑傲江湖	任我行	令狐冲	0.786
笑傲江湖	向问天	令狐冲	0.781
碧血剑	崔秋山	袁承志	0.858
碧血剑	青青	袁承志	0.777
碧血剑	焦公礼	袁承志	0.774
碧血剑	温青	袁承志	0.770
碧血剑	张朝唐	袁承志	0.765

、由图表可以看出 LSTM 能够捕捉大部分的人物关系，但也会有个别例子，就像‘六子’和‘众人’这些非人名称但是在文章中频率很高也会被识别出来，可能是 epoch 的轮数或者是 embedding 不充分导致的。

References

[1] https://blog.csdn.net/lyc_yongcai/article/details/73201446

[2] https://blog.csdn.net/hust_tsb/article/details/79485268

[3] https://developer.aliyun.com/article/1333079

[4] https://easyai.tech/ai-definition/lstm/