MIT NLP科研预习情况

上海大学 王泓杰

在本次预习的过程中，我结合了以下资料和内容在NLP领域进行了自主学习和探索。

[”【斯坦福CS224N】(2021|中英) 深度自然语言处理”](https://www.bilibili.com/video/BV18Y411p79k/?p=2&spm_id_from=333.1007.top_right_bar_window_history.content.click&vd_source=3cf6a25e20fce619b2d35930f06672e8)

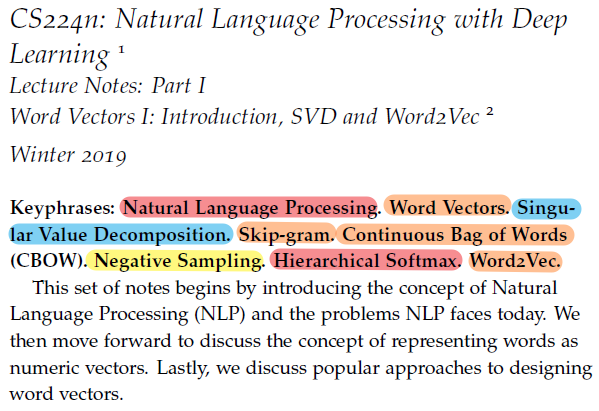
[CSDN上博主的笔记](https://caoyang.blog.csdn.net/article/details/125020572)

[ShowMeAI社区](https://www.showmeai.tech/tutorials/36?articleId=231)

* **预习完成度 (以百分比估计，图片中记号笔颜色由暖到冷表示由熟练到陌生)**

由于在这段时间我需要花大部分时间在备考GRE上，预习可能不够充分。

cs224n-2019-notes01-wordvecs1 80%

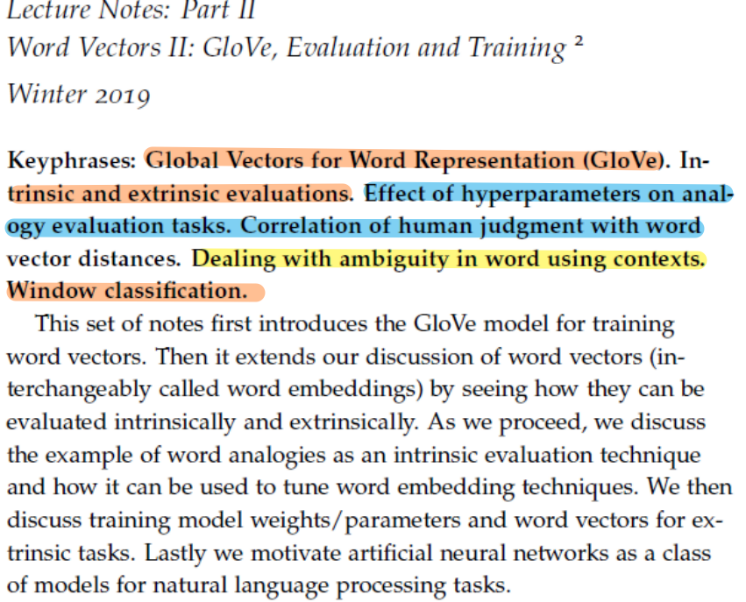


# 对Word2Vec中机器学习的一些理念有了了解。中心词和周围词的关系，word vector中的多维度；两个词向量相互点乘得到的值可以描述两个词的相似度；

# 暂时放弃了一些细节概念的学习

Skip-gram/Continuous BagOfWords

cs224n-2019-notes02-wordvecs2 70%



因为corpus中太多词，用GD去update会很久

Stochastic Gradient Descent (SGD)

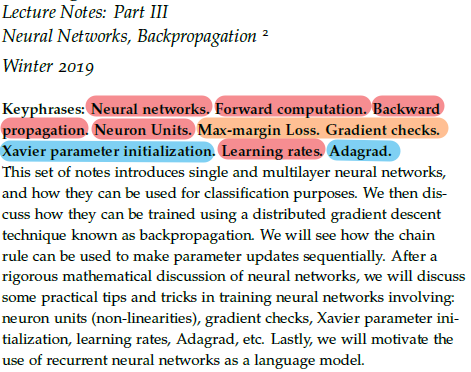
GloVe

Co-occruence matrix

Window classfications

Intrinsic / Extrinsic evaluation

cs224n-2019-notes03-neuralnets 85%



神经网络的概念引入

正反向传播的目的，各自训练的是什么 是模型参数or what

梯度检查

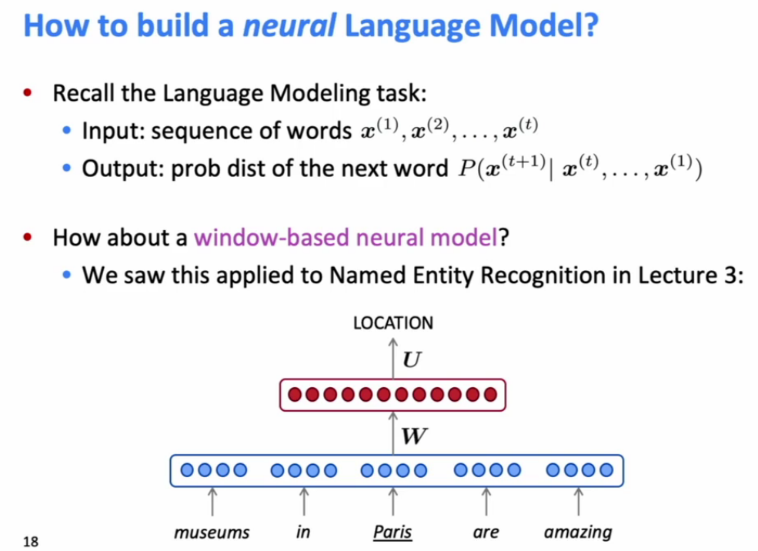
Single , multiple layer

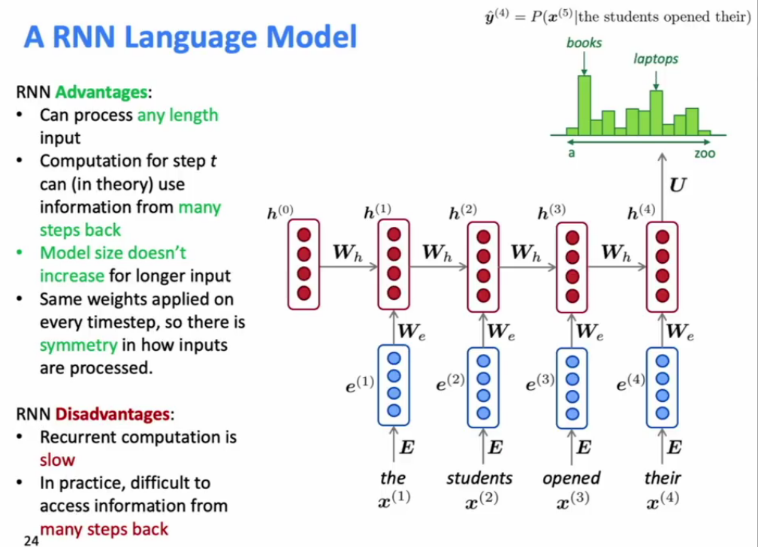
Learning rates /

Dropout 用于正则化

Neuron units: sigmoid/ReLU..

cs224n-2019-notes05-LM\_RNN 70%

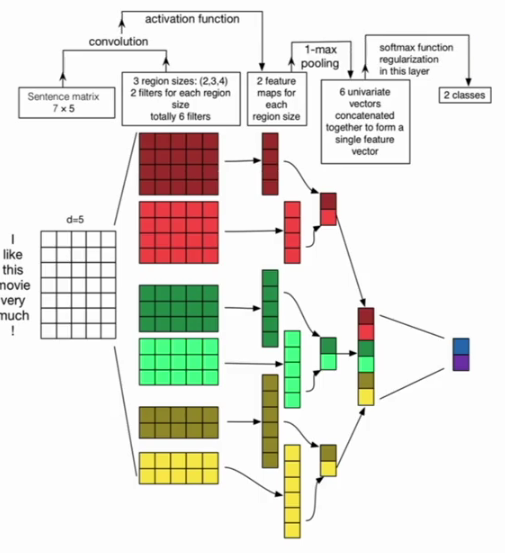




RNN 中的weight是一样的

RNN Language Model 的评估

cs224n-2019-notes08-CNN 75%





* **难易度**

由于我之前选修过机器学习课程，对一些prerequiste knowledge有概念和了解。E.g. Gradient descent, loss funtion, neural networks, activated funtion, softmax/sogmoid function,

总体来说，不少概念目前只是停留在表层，大概只是知道这些concepts的目的和以此解决的问题，然后一些数学上的推到和计算就不太愿意去啃。

* **难点**

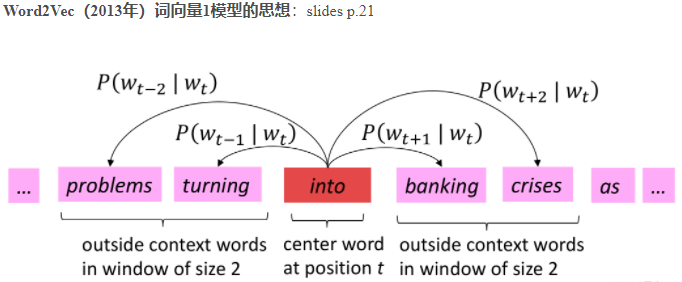
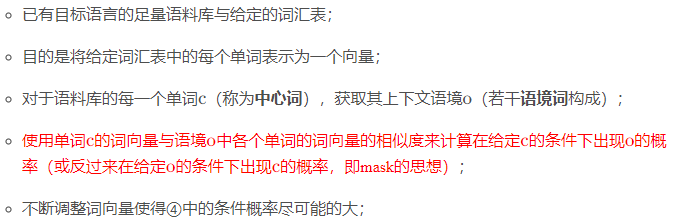


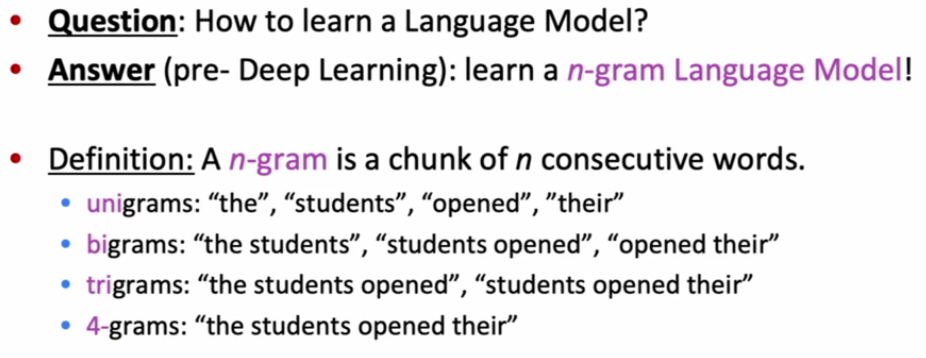
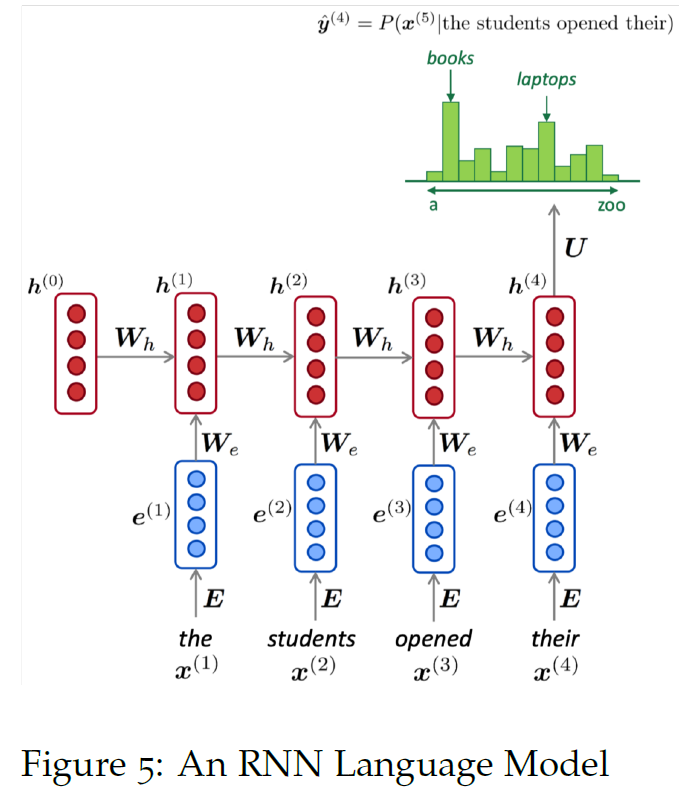
Image1. Idea of word vector



002

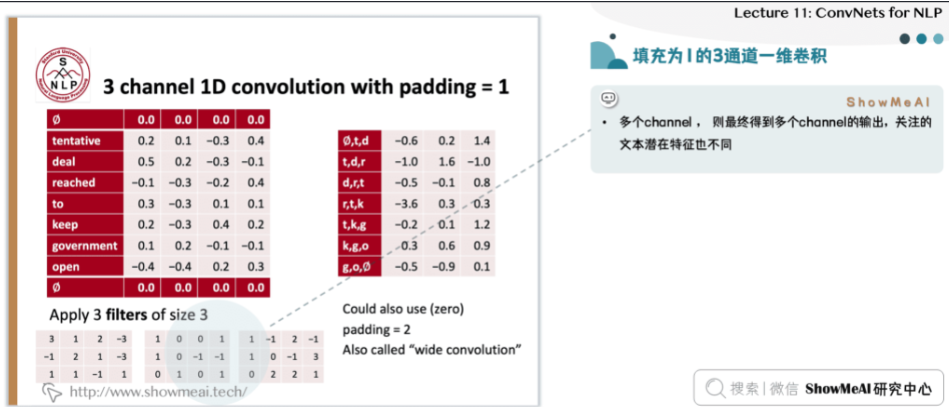
# 什么是GloVe？

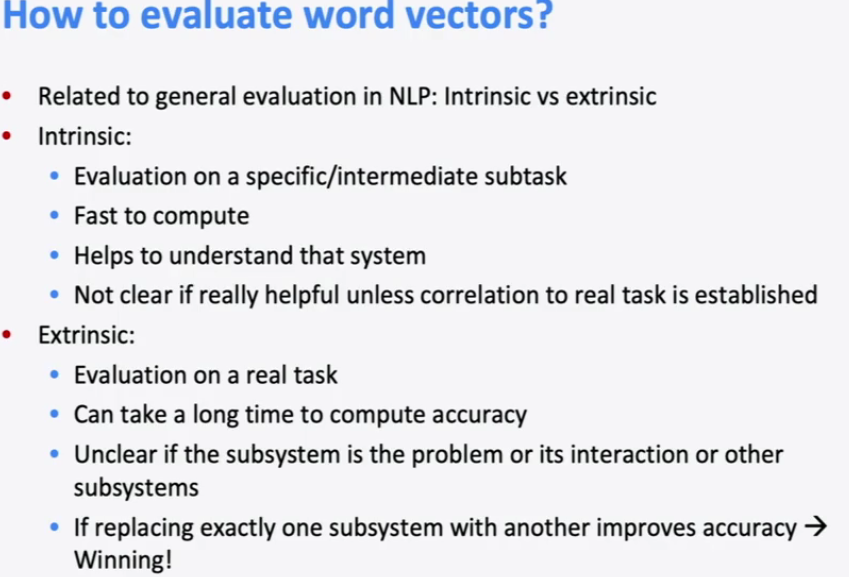
GloVe的全称叫**Glo**bal **Ve**ctors for Word Representation，它是一个基于全局词频统计（count-based & overall statistics）的词表征（word representation）工具，它可以把一个单词表达成一个由实数组成的向量，这些向量捕捉到了单词之间一些语义特性，比如相似性（similarity）、类比性（analogy）等。我们通过对向量的运算，比如欧几里得距离或者cosine相似度，可以计算出两个单词之间的语义相似性



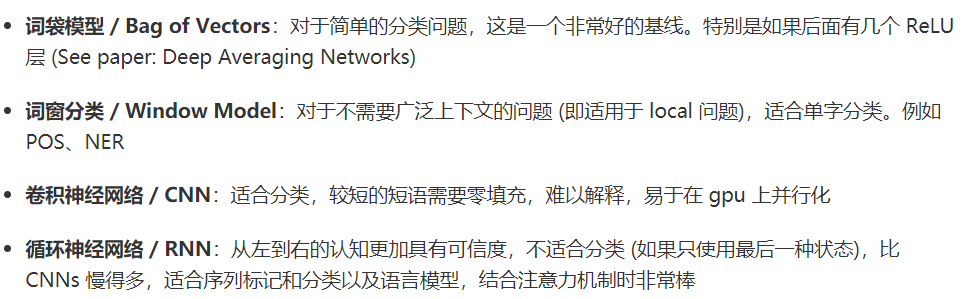
GRU 和 LSTM

文本形式的一维卷积



* **困惑点**

**CNN 和 RNN 的对比**



因为没有动手实践 所以一些应用层面和理论相结合的地方会感受到不太清楚