

# P大树洞生成器

匡宇轩 (2100013089), 陈红韵 (2100013130), 王天源 (1700017703)

联系邮箱: 2100013089@stu.pku.edu.cn,  
2100013130@stu.pku.edu.cn, 1700017703@pku.edu.cn

指导教师: 邓志鸿

小组编号:

北京大学, 人工智能引论课程  
2021-2022, 春季学期

## 摘要

本项目基于因果语言建模 (CLM) 的原理开发, 使用 Python 和 PyTorch 实现。我们在树洞文本数据集上训练 LSTM 模型, 并根据用户的输入, 使用训练好的模型产生回复。我们将模型命名为 HoleAI。

**关键词:** 自然语言处理; 长短期记忆网络; 因果语言建模

## 引言

本项目为北京大学信息科学技术学院国家精品课程《人工智能引论》的课程项目。本文为该项目的总结报告。

本项目受到北京大学树洞的启发。北京大学树洞作为校内学生匿名交流平台, 是一个完美的语料库, 本项目通过收集大量树洞文本数据, 用神经网络进行拟合, 旨在创造出一个具有北大学生气质的 AI。

## 方法

### 1. 数据集收集与预处理

收集: 由于树洞的反爬虫机制, 我们手动收集约 3000 条树洞均为, 约一百万字, 包括日常话题、经典话题、神洞

清洗: 去除洞号、日期、收藏信息; 去除低频字符; 去除过多重复字符

### 2. 神经网络构建与训练

Word Embedding 词嵌入

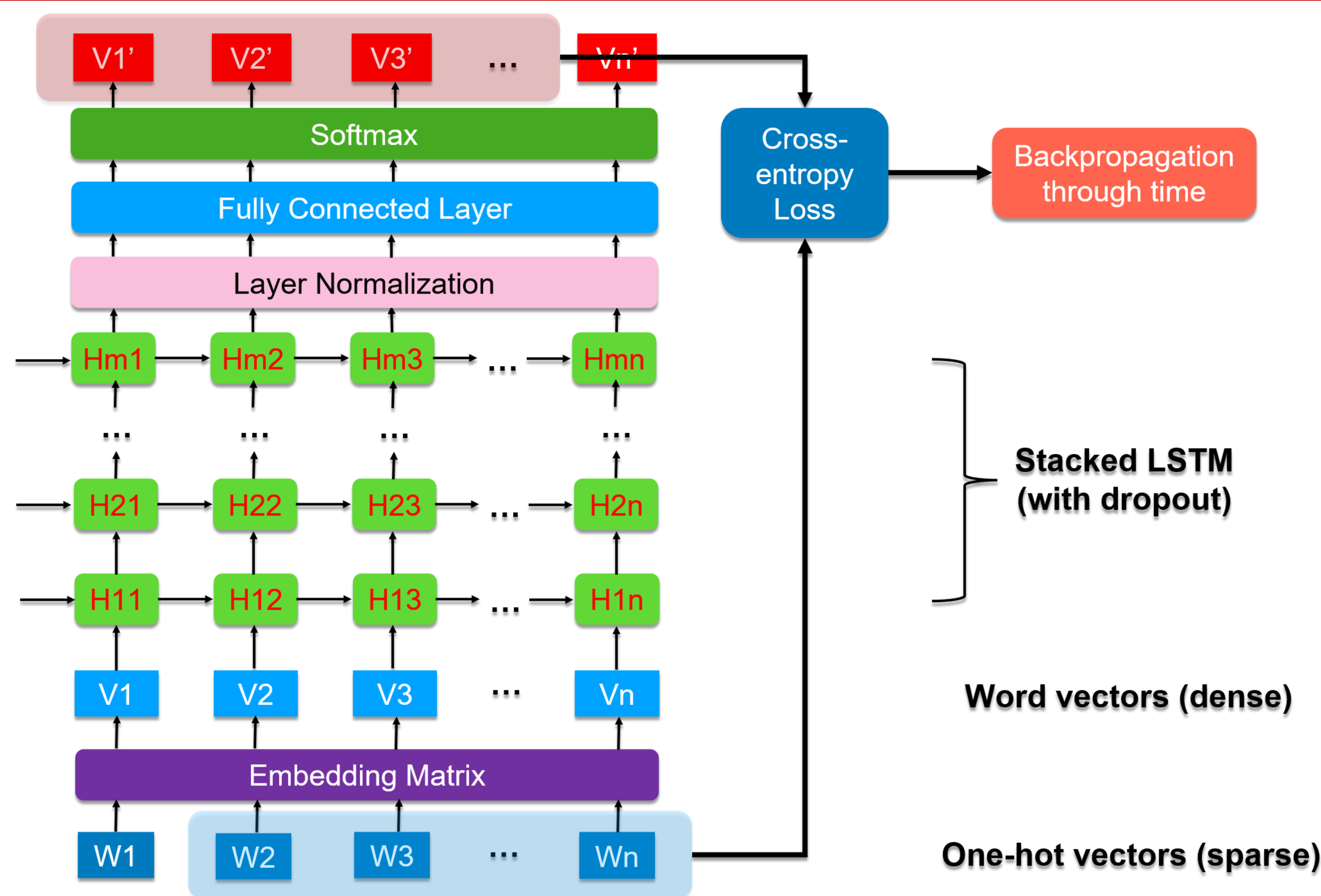
Multi-layer LSTM 多层 LSTM

Dropout 丢弃层

Layer Normalization 层归一化

Fully Connected Layer 全连接层

### 3. 前端可视化、人机互动



## 实验

在对模型的优化过程中, 我们得出以下结论:

1.

(1) 训练结果较好, 损失函数值单调下降, 且收敛到了一个比较低的水平

(2) 随机分割数据集可以起到一定正则化的作用, 这使得训练最后 validation loss 低于 train loss

2.

(1) Dropout rate 不宜过大, num\_layers 不宜过多, 不利于参数更新 (优化梯度回传)

(2) 随机分割数据集, 也因此丢失了部分语义信息

(3) 隐层大小至关重要。隐空间越大, 模型的表现力越强, 对数据的拟合能力越强

## 成果



## 参考文献

- [1] Efficient Estimation of Word Representations in Vector Space, Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, 2013, <https://arxiv.org/abs/1301.3781>
- [2] Long short-term memory, S Hochreiter, J Schmidhuber, 1997, <https://pubmed.ncbi.nlm.nih.gov/9377276/>
- [3] Understanding LSTM Networks, colah's blog, 2015, <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

## 总结

1. 这个项目十分新颖有趣, 且具有一定社会科学方面的研究价值。同时, 我们也可通过该模型体验千万个“我”的生活。
2. 数据集的缺乏和数据集噪声过多给我们带来的很大挑战。
3. 模型大小和数据集大小的匹配是另一个挑战。
4. 通过本项目, 我们加深了对神经网络的认识, 拓宽了对前沿任务的了解, 也学习了许多人工智能以外的知识。

