基于词表分解优化的循环神经网络语言模型

Vocabulary Factorized Optimization for Recurrent Neural Network Language Model

是黎彬

(*olivier.shi@buaa.edu.cn*)

北京航空航天大学中法工程师学院研究生开题答辩

2017 年 12 月 20 日

概览

1. 论文选题的背景与意义
2. 论文工作计划
   - 论文研究目标
   - 论文主要研究内容
3. 已经完成的工作
   - 词表过大带来的问题
   - 三大类模型变种
   - 基于词表分解的语言模型
4. 关键技术和难点
   - 循环神经网络及其变种
   - 词表分解优化
5. 下一阶段工作计划

题目来源

论文题目《基于词表分解优化的循环神经网络语言模型》为实验室研究课题。

题目来源

论文题目《基于词表分解优化的循环神经网络语言模型》为实验室研究课题。

语言模型（Language Model）

1. WHAT？语言模型是学习人类语言的模式，并对一段文本的概率进行估计
2. WHY？对信息检索、机器翻译、语音识别等自然语言处理领域中的人物有着重要的作用
3. HOW？统计语言模型的作用是为一个长度为 T 的词序列确定一个联合概率分布 $P(w_1; w_2; \cdots; w_T)$，表示其存在的可能性

题目来源

论文题目《基于词表分解优化的循环神经网络语言模型》为实验室研究课题。

语言模型（Language Model）

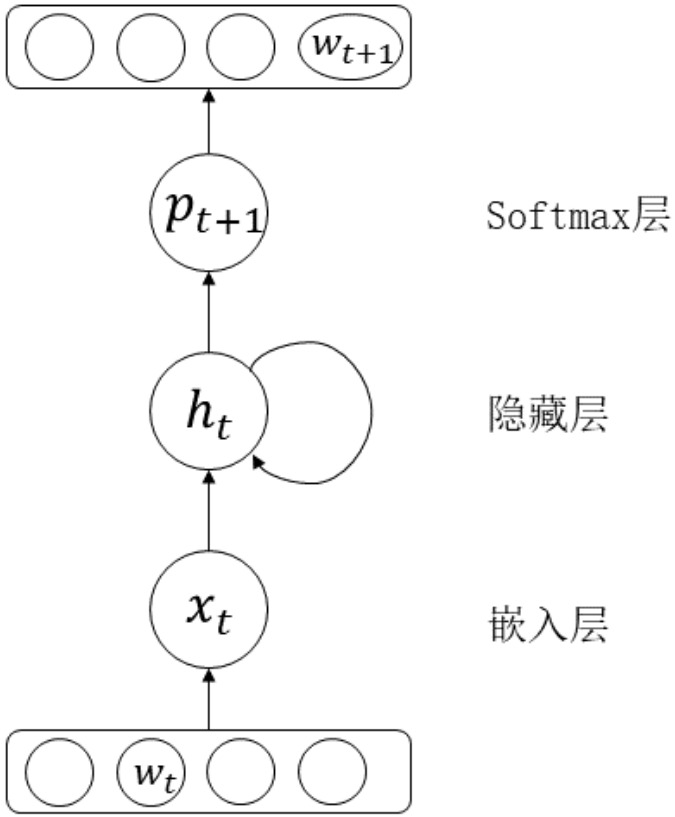
1. WHAT？语言模型是学习人类语言的模式，并对一段文本的概率进行估计
2. WHY？对信息检索、机器翻译、语音识别等自然语言处理领域中的人物有着重要的作用
3. HOW？统计语言模型的作用是为一个长度为 T 的词序列确定一个联合概率分布 $P(w_1; w_2; \cdots; w_T)$，表示其存在的可能性

深度学习与自然语言处理（Deep learning for NLP）

1. 由于人类语言结构的复杂性和数据多样性，传统的方法建模能力有限
2. 深度学习的提出与发展大大提高了计算机在复杂数据上建模能力，为从结构性较差的数据上学习一定的知识提供模型支撑

论文研究目标

1. 探究利用深度学习对自然语言建模的方法——循环神经网络语言模型及其变种
2. 探究目前循环神经网络语言模型遇到的一个最大的瓶颈——“词表过大”问题



论文主要研究内容

1. 调研语言模型的相关背景与基础知识，了解在基于统计的语言模型方法过时之后，目前主流的基于深度学习的语言建模方法——循环神经网络语言模型。
2. 深入探究循环神经网络语言模型，包括其理论知识以及代码实现，并从初步的实验中发掘“大词表”问题的根源所在。
3. 考察目前存在的前沿的方法是如何解决语言模型中“大词表”问题的，包括基于采样技术，基于字符级别编码和基于词表分解三大类方法，本课题将重点探讨基于词表分解这一类方法。
4. 深入探究基于词表分解的循环神经网络语言模型，包括基于预测层词表分解的方法和输入层词表分解的方法这两种，总结目前这一类方法存在的问题，并探索对其提升的性能的方法。
5. 针对基于词表分解的循环神经网络语言模型，如何将词表进行分解是一个值得讨论与研究的问题，包括对词表的初始化分解方法，以及在建模过程中动态调整词表分布的方法。

词表过大带来的问题

高计算复杂度（High computational complexity）

RNNLM 中的计算复杂度主要受 softmax 的影响，其中归一化项 $\sum_{w' \in \mathcal{V}} \exp(z_{w'})$ 需要遍历整个词表上对所有词的预测分值进行求和，会带来非常高的计算量并使得模型低效。对于一个大小为 $|\mathcal{V}|$ 的词表，Softmax 的计算复杂度等于 $|\mathcal{V}|$，那么当词表过大的时候，该模块的计算量就会巨大。

庞大的模型大小（Huge model size）

RNNLM 中的参数规模受词表大小影响的有嵌入层（Embedding layer）和 Softmax 层，因为嵌入层要将所有的词都映射到向量空间以及 Softmax 层要在整个词表空间内去预测。它们分别拥有 $\mathcal{O}(|\mathcal{V}| \times |D|)$ 和 $\mathcal{O}(|\mathcal{V}| \times |H|)$ 的参数量，其中 $|D|$ 和 $|H|$ 分别表示嵌入层和隐藏层的维度。

三大类模型变种

有非常多研究者在 RNNLM 基础上做了很多改进，可以大致分成以下 3 大类：

三大类模型变种

有非常多研究者在 RNNLM 基础上做了很多改进，可以大致分成以下 3 大类：

基于采样近似方法

通过采样技术从词表中选取一部分的词来近似计算 RNNLM 中的 softmax。这一类的方法有重要性采样（Importance Sampling, IS）、噪声对比估计（Noise Contrastive Estimation）以及负采样（Negative Sampling），它们的采样技术不同，但核心的思想都是一样的。

三大类模型变种

有非常多研究者在 RNNLM 基础上做了很多改进，可以大致分成以下 3 大类：

基于采样近似方法

通过采样技术从词表中选取一部分的词来近似计算 RNNLM 中的 softmax。这一类的方法有重要性采样（Importance Sampling, IS）、噪声对比估计（Noise Contrastive Estimation）以及负采样（Negative Sampling），它们的采样技术不同，但核心的思想都是一样的。

基于字符级别建模方法

所有英语单纯单词都由 26 个字母组合构成，Kim 基于此发明 CharCNN，用字符级别的输入代替原始的词级别的输入，那么这能将词嵌入层的参数从原始的 $\mathcal{O}(|\mathcal{V}| \times |\mathcal{D}|)$ 缩减至 $\mathcal{O}(|Char| \times |\mathcal{D}|)$。

三大类模型变种

有非常多研究者在 RNNLM 基础上做了很多改进，可以大致分成以下 3 大类：

基于采样近似方法

通过采样技术从词表中选取一部分的词来近似计算 RNNLM 中的 softmax。这一类的方法有重要性采样（Importance Sampling, IS）、噪声对比估计（Noise Contrastive Estimation）以及负采样（Negative Sampling），它们的采样技术不同，但核心的思想都是一样的。

基于字符级别建模方法

所有英语单纯单词都由 26 个字母组合构成，Kim 基于此发明 CharCNN，用字符级别的输入代替原始的词级别的输入，那么这能将词嵌入层的参数从原始的 $\mathcal{O}(|\mathcal{V}| \times |\mathcal{D}|)$ 缩减至 $\mathcal{O}(|Char| \times |\mathcal{D}|)$。

基于词表分解预测方法

利用词表的结构性和层级性以及条件概率把原来的词的一步预测分解成多步预测问题。这一类的方法有很多，包括基于双层结构 Softmax（Class-based Hierarchical Softmax, cHSM）的方法和基于树层级结构 Softmax（Tree-based Hierarchical Softmax, tHSM）的方法。

基于词表分解的语言模型

主要包括基于双层结构 Softmax（Class-based Hierarchical Softmax, cHSM）的方法和基于树层级结构 Softmax（Tree-based Hierarchical Softmax, tHSM）的方法

基于词表分解的语言模型

主要包括基于双层结构 Softmax（Class-based Hierarchical Softmax, cHSM）的方法和基于树层级结构 Softmax（Tree-based Hierarchical Softmax, tHSM）的方法

cHSM

通过条件概率把传统的 Softmax 预测方法分裂成两个步骤。原来的 Softmax 是直接把一个词从词表中预测出来，而 cHSM 是先预测这个词属于的类别，然后再在这个类别中预测出具体哪个词，

$$p(w|h) = p(c|h)p(w|c,h) \tag{1}$$

基于词表分解的语言模型

主要包括基于双层结构 Softmax（Class-based Hierarchical Softmax, cHSM）的方法和基于树层级结构 Softmax（Tree-based Hierarchical Softmax, tHSM）的方法

cHSM

通过条件概率把传统的 Softmax 预测方法分裂成两个步骤。原来的 Softmax 是直接把一个词从词表中预测出来，而 cHSM 是先预测这个词属于的类别，然后再在这个类别中预测出具体哪个词，

$$p(w|h) = p(c|h)p(w|c,h) \tag{1}$$

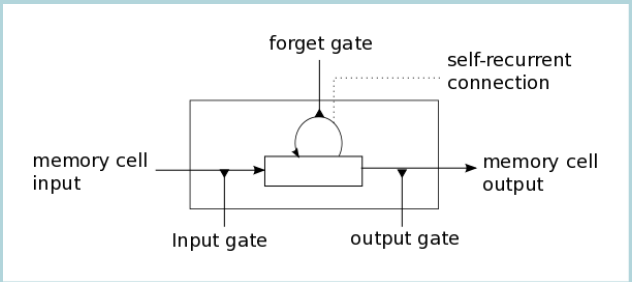
tHSM

它可以被看作是 cHSM 的延伸。tHSM 扩展了 cHSM 的思想，把 Softmax 层继续分裂，直到把原始的 Softmax 分裂成一个树结构（Tree）。所有的词都列在在这棵树的叶子节点（Leaf Node）上，而这棵树的中间节点表示词表的深层次结构。

循环神经网络及其变种

循环神经网络及其变种

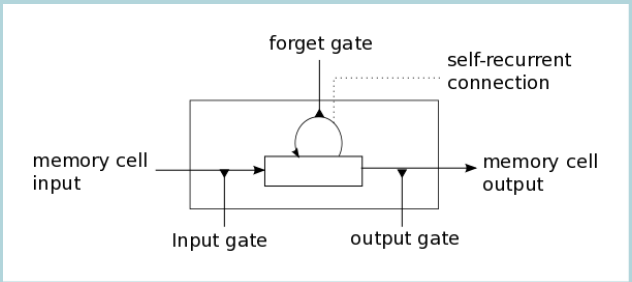
Long Short-Term Memery (LSTM)



$$
\begin{aligned}
i_t, f_t, o_t &= \sigma(Wx_t + Uh_{t-1} + b) \\
\tilde{c}_t &= \psi(W'x_t + U'h_{t-1} + b') \\
c_t &= c_t \odot i_t + c_{t-1} \odot f_t \\
h_t &= c_t \odot \psi(o_t)
\end{aligned}
$$

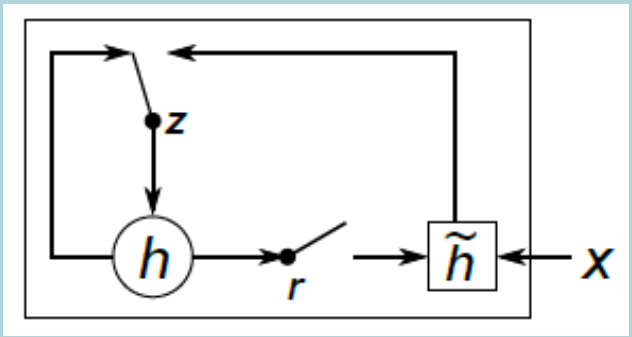
循环神经网络及其变种

Long Short-Term Memery (LSTM)



$$
\begin{aligned}
i_t, f_t, o_t &= \sigma(Wx_t + Uh_{t-1} + b) \\
\tilde{c}_t &= \psi(W'x_t + U'h_{t-1} + b') \\
c_t &= c_t \odot i_t + c_{t-1} \odot f_t \\
h_t &= c_t \odot \psi(o_t)
\end{aligned}
$$

Gated Recurrent Units (GRU)



$$
\begin{aligned}
z_t, r_t &= \sigma(Wx_t + Uh_{t-1} + b) \\
\tilde{h}_t &= \tanh(W'x_t + U'(h_{t-1} \odot r_t) + b') \\
h_t &= (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}
\end{aligned}
$$

拟采取的技术方案

词表分解管理

词表具有一定的结构性

1. 调研聚类方法：将一定意义上相似的词聚类，初始化词的二级表示
2. 调研交换算法：初始化的词类关系并不一定符合期望并适合模型的构建，于是要在模型训练过程中动态调整交换词在类中的位置

模型优化

现有的基于词表分解的语言模型存在一定的问题，比如在预测一个词的类标和类内位置的时候，它们都是依赖同一个隐藏层状态，即根据同一个信息源去预测两个不同空间的量。这在一定程度上可能会减小预测准确率。

时间安排

1. 2018 年 5 月 ∼ 2018 年 7 月：调研目前前沿的基于词表分解的语言模型
2. 2018 年 7 月 ∼ 2018 年 9 月：调研并优化基于词表分解循环神经网络
3. 2018 年 9 月 ∼ 2018 年 10 月：调研并优化词表分解方案
4. 2018 年 10 月 ∼ 2018 年 11 月：代码实现我们设计的方法
5. 2018 年 11 月 ∼ 2018 年 12 月：补充完整实验验证，统计分析实验结果
6. 2018 年 12 月 ∼ 2019 年 1 月：整理资料和论文撰写

存在的问题

1. 基于 GPU 的编程增加了 coding 难度
2. 计算资源的匮乏使得实验进行缓慢
3. 数据量极大，建模实验周期长
4. 一些前沿的方法还没有公开其代码实现
5. 基础模型采用 RNN，当神经模型变得复杂时，建模过程变的不可解释，这对寻找提升模型方法增大了难度

Thank You !

谢谢各位老师和同学！请大家批评指正；

本次开题报告，文献综述，PPT 已经上传至 Github：

`https://github.com/OlivierShi/ecpkn`