

神经网络语言模型的性能优化研究

On Optimization Perspective of Neural Language Model

姜楠 (*nanjiang@buaa.edu.cn*)

北京航空航天大学计算机学院研究生开题答辩

2016 年 12 月 20 日



概览

1 论文选题的背景与意义

2 国内外研究现状及发展动态

■ 隐写分析

3 论文的研究内容及拟采取的技术方案

■ 拟采取的技术方案

■ 论文的研究内容

4 论文研究计划



题目来源

论文题目《神经网络语言模型的性能优化研究》为自拟课题。

题目来源

论文题目《神经网络语言模型的性能优化研究》为自拟课题。

语言模型 (Language Model)

- 语言模型可以对一段文本的概率进行估计，对信息检索、机器翻译、语音识别等任务有着重要的作用
- 形式化讲，统计语言模型的作用是为一个长度为 m 的字符串确定一个概率分布 $P(w_1; w_2; \dots; w_m)$ ，表示其存在的可能性



题目来源

论文题目《神经网络语言模型的性能优化研究》为自拟课题。

语言模型 (Language Model)

- 语言模型可以对一段文本的概率进行估计，对信息检索、机器翻译、语音识别等任务有着重要的作用
- 形式化讲，统计语言模型的作用是为一个长度为 m 的字符串确定一个概率分布 $P(w_1; w_2; \dots; w_m)$ ，表示其存在的可能性

神经网络

- 人工神经网络是一个能够学习，能够总结归纳的系统，也就是说它能够通过已知数据的实验运用来学习和归纳总结。
- 与之不同的基于符号系统下的学习方法，它们也具有推理功能，只是它们是建立在逻辑演算法的基础上，也就是说它们之所以能够推理，基础是需要有一个推理演算法则的集合。



语言模型

语言模型可以对一段文本的概率进行估计，对信息检索、机器翻译、语音识别等任务有着重要的作用。形式化讲，统计语言模型的作用是为一个长度为 m 的字符串确定一个概率分布 $P(w_1; w_2; \dots; w_m)$ ，表示其存在的可能性，其中 w_1 到 w_m 依次表示这段文本中的各个词。一般在实际求解过程中，通常采用下式计算其概率值：

$$\begin{aligned} P(w_1; w_2; \dots; w_m) &= P(w_1)P(w_2|w_1)P(w_3|w_1; w_2)\cdots P(w_i|w_1; w_2; \dots; w_{i-1}) \\ &\quad \cdots P(w_m|w_1; w_2; \dots; w_{m-1}) \end{aligned} \tag{1}$$

在实践中，如果文本的长度较长，公式1右部
 $\cdots P(w_m|w_1; w_2; \dots; w_{m-1})$ 的估算会非常困难。因此，研究者们提出使用一个简化模型： n 元模型 (n -gram model)。在 n 元模型中估算条件概率时，距离大于等于 n 的上文词会被忽略，也就是对上述条件概率做了以下近似：

$$P(w_i|w_1; w_2; \dots; w_{i-1}) \approx P(w_i|w_{i-(n-1)}; \dots; w_{i-1}) \tag{2}$$

隐写分析

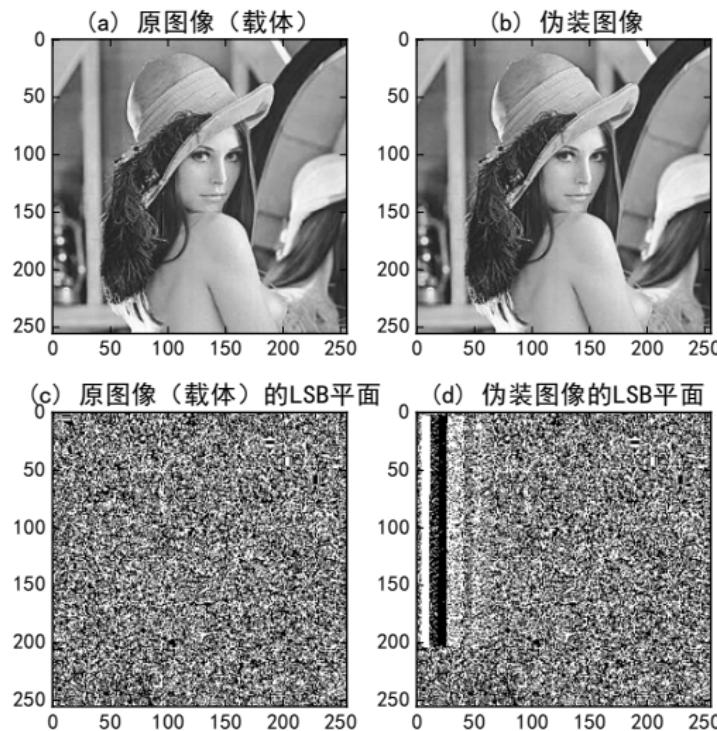
隐写分析

- 视觉隐写分析
- 结构隐写分析
- 统计隐写分析
- 学习隐写分析

隐写分析

隐写分析

- 视觉隐写分析
- 结构隐写分析
- 统计隐写分析
- 学习隐写分析





循环神经网络

样本集

容量为 N 的训练样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

循环神经网络

样本集

容量为 N 的训练样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

- 特征向量 \mathbf{x}_i 为图像块的特征



循环神经网络

样本集

容量为 N 的训练样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

- 特征向量 \mathbf{x}_i 为图像块的特征
- 标签 $y_i \in \{-1, 1\}$ 为安全评估结果，在训练集中由隐写方法评估得到，在使用隐写系统时预测结果作为选择位置的参考指标



循环神经网络

样本集

容量为 N 的训练样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

- 特征向量 \mathbf{x}_i 为图像块的特征
- 标签 $y_i \in \{-1, 1\}$ 为安全评估结果，在训练集中由隐写方法评估得到，在使用隐写系统时预测结果作为选择位置的参考指标

SVM 分类器

追求最大“间隔”的分类

$$\max_{\mathbf{w}, b} \frac{2}{\|\mathbf{w}\|}$$

$$s.t. \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, \dots, N$$



循环神经网络

样本集

容量为 N 的训练样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

- 特征向量 \mathbf{x}_i 为图像块的特征
- 标签 $y_i \in \{-1, 1\}$ 为安全评估结果，在训练集中由隐写方法评估得到，在使用隐写系统时预测结果作为选择位置的参考指标

SVM 分类器

追求最大“间隔”的分类

过度拟合 & 线性不可分

$$\max_{\mathbf{w}, b} \frac{2}{\|\mathbf{w}\|}$$

$$s.t. \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, \dots, N$$



循环神经网络

样本集

容量为 N 的训练样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

- 特征向量 \mathbf{x}_i 为图像块的特征
- 标签 $y_i \in \{-1, 1\}$ 为安全评估结果，在训练集中由隐写方法评估得到，在使用隐写系统时预测结果作为选择位置的参考指标

SVM 分类器

追求最大“间隔”的分类

过度拟合 & 线性不可分

$$\max_{\mathbf{w}, b} \frac{2}{\|\mathbf{w}\|}$$

■ 核函数：变换特征空间至高维

$$s.t. \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, \dots, N$$



循环神经网络

样本集

容量为 N 的训练样本集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

- 特征向量 \mathbf{x}_i 为图像块的特征
- 标签 $y_i \in \{-1, 1\}$ 为安全评估结果，在训练集中由隐写方法评估得到，在使用隐写系统时预测结果作为选择位置的参考指标

SVM 分类器

追求最大“间隔”的分类

$$\max_{\mathbf{w}, b} \frac{2}{\|\mathbf{w}\|}$$

$$s.t. \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, i = 1, 2, \dots, N$$

过度拟合 & 线性不可分

- 核函数：变换特征空间至高维
- 软间隔：以权重 C 容忍分类错误



实验平台和设置

Linux 操作系统

R 主要用于数据统计和图表处理

Python2.7 使用的开发语言和开发环境

Theano 主要的建模语言

实验平台和设置

Linux 操作系统

R 主要用于数据统计和图表处理

Python2.7 使用的开发语言和开发环境

Theano 主要的建模语言

- 同时还依赖于其他的处理脚本，需要对 bash script 和 C/C++ 有足够的了解和掌握；
- GPU 的设备是使用 Titan X，并且对应的 CUDA 版本为 8.0(需要 CUDNN/CUSPARSE 等库的支持)。

SVM 的训练

使用 80 组不同参数进行训练，得到的 SVM 在错误率方面的表现

时间安排

- 2016 年 12 月 ~ 2017 年 1 月 : 整理资料 , 学习研究语言模型的领域知识 ;
- 2017 年 2 月 ~ 2017 年 4 月 : 研究学习深度学习模型的知识 , 特别是循环神经网络的建模过程 ;
- 2017 年 5 月 ~2017 年 7 月 : 调研并实现解决大词表问题的主要手段 , 并实现基本代码框架 ;
- 2015 年 8 月 ~2015 年 10 月 : 实验验证与完善 ;
- 2015 年 11 月 ~2015 年 12 月 : 资料整理和论文撰写 .

Thanks

谢谢各位老师和同学！请大家批评指正；
论文中用到的全部源代码（包括本幻灯片），数据，图像，文档见：
👉 https://github.com/jiangnanHugo/Graduate_Design