

基于深度学习与词嵌入的情感分析系统研究与实现

- 答辩人：聂国庆
- 班级：计算机 2015-1
- 指导老师：赵中英

目录

CONTENTS

01 课题选题

02 算法设计

03 模型调优与测试

04 总结回顾

01

课题选题

01 课题选题

01

选题来源

情感分析研究的课题是在指导老师帮助下的选择的，来自生产实际（社会实践）课题。

对电商网站商品的评价、以及社交媒体的评论等大量信息的有效情感分析，能够帮助我们进行更准确的、更及时的做出正确的响应。

02

研究内容

针对文本进行句子和段落级的情感倾向性分析，利用算法来判断句子的情感色彩。

研究的目的在于提高情感分析算法的准确性，不断学习，不断提高和优化算法。在实际数据集上的进行模型训练与调优，并对模型进行简单的封装和部署。

02

算法设计

02 算法设计

深度学习模型

随着计算机硬件的发展，GPU的高效利用，深度学习在一些图像和自然语言处理领域展现很多出优于传统机器学习的特性。

机器学习模型

传统机器学习又称为统计学习，相对于深度学习来说计算代价小，在一些问题上传统机器学习就可以达到比较好的效果。

模型集成融合

深度学习模型与传统的树形机器学习模型的集成融合效果不错。

传统的集成融合往往只是简单的投票加权，而我使用基于机器学习的方法进行两层模型集成融合，效果不错。

02 算法设计



数据预处理

英文本身就是单词，按空格分开即可。

数据预处理部分需要尽可能的降低数据的杂乱噪声，清洗掉无用的字符、乱码，删除“停用词”，尽可能清洗掉无用信息。



词向量

利用fit分词器(tokenizer)的方法建立word index 词典，之后就可以按照词典序（词频）的值将文本处理成数字型的词向量。

只有数学的向量才能进一步被学习和预测。

02 算法设计



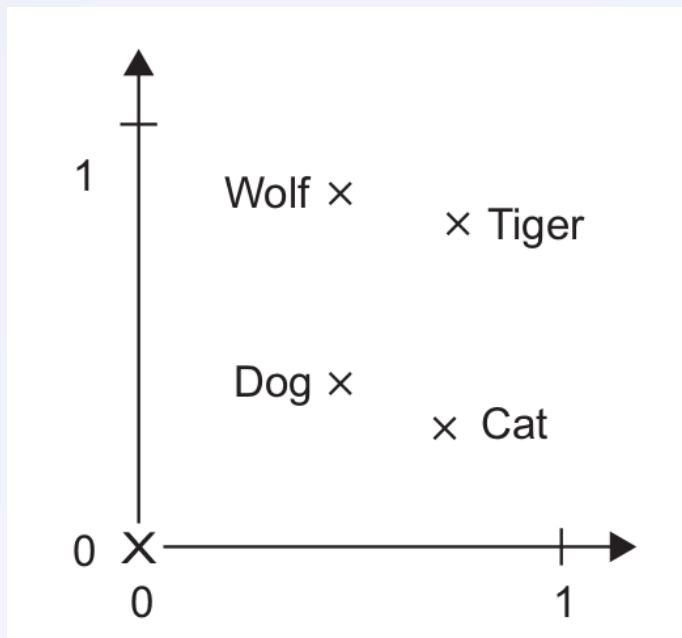
Embedding 词嵌入

One-Hot编码太过稀疏，运算压力大，所以使用词嵌入的方式，既降低了运算矩阵的维度，同时又一定程度上赋予了词与词之间关联关系。

使用了两种词嵌入的方式，

一种是在数据集相对较大的时候，自己利用word2vector模型训练词嵌入。

另一种是在数据集并不大时使用的，预训练的词嵌入，是其他人在较大且泛化性较强的数据集上训练得到的词嵌入。

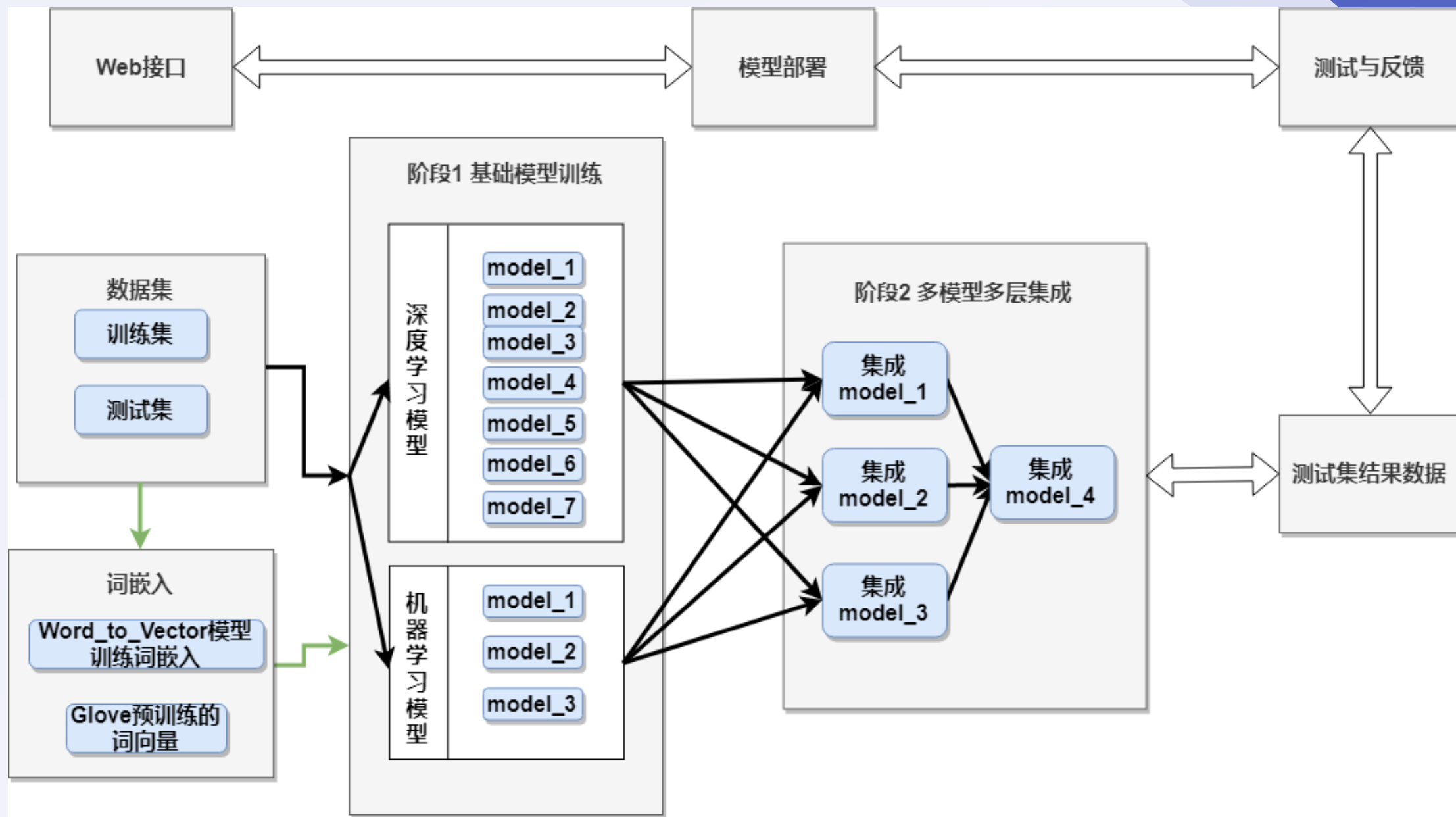


“从宠物到野生动物” 向量

“从猫科到犬科” 向量

“国王” - “男性” + “女性” = “皇后”
King male female Queen

02 算法设计



02 算法设计



RNN，循环神经网络；

LSTM，GRU 和 双向LSTM/GRU

RNN 一般用于处理时间序列问题

LSTM 长短时记忆（Long short-term memory）三个门控单元

GRU 是LSTM的简化，只有两个门控单元，表示能力不如LSTM 但是 运算负载小

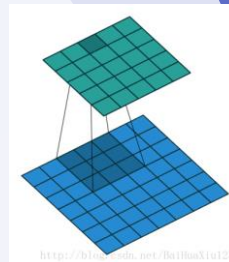
双向RNN，从两个顺序上进行运算和观察，效果好于或等于单向RNN



CNN，卷积神经网络

二维卷积神经网络一般用于处理图像问题。卷积层和池化层，通过一定步长的窗口扫描，与卷积核进行运算，生成带有参数信息（深度方向）的新数据，利用池化层进行关键提取。

一维的CNN 在文本时序信息处理上 也可以 发挥这些特性，有利于关键信息的采集处理。



Attention 注意力

注意力机制可以用于模型的最后一层之前，对于即将输出的结果进行重要性的选择加权，使结果更加合理



Inception 和 残差连接

Inception 模块通过不同步长不同卷积池化的结构，可以提取到不同的有效信息，避免有效信息的遗漏。

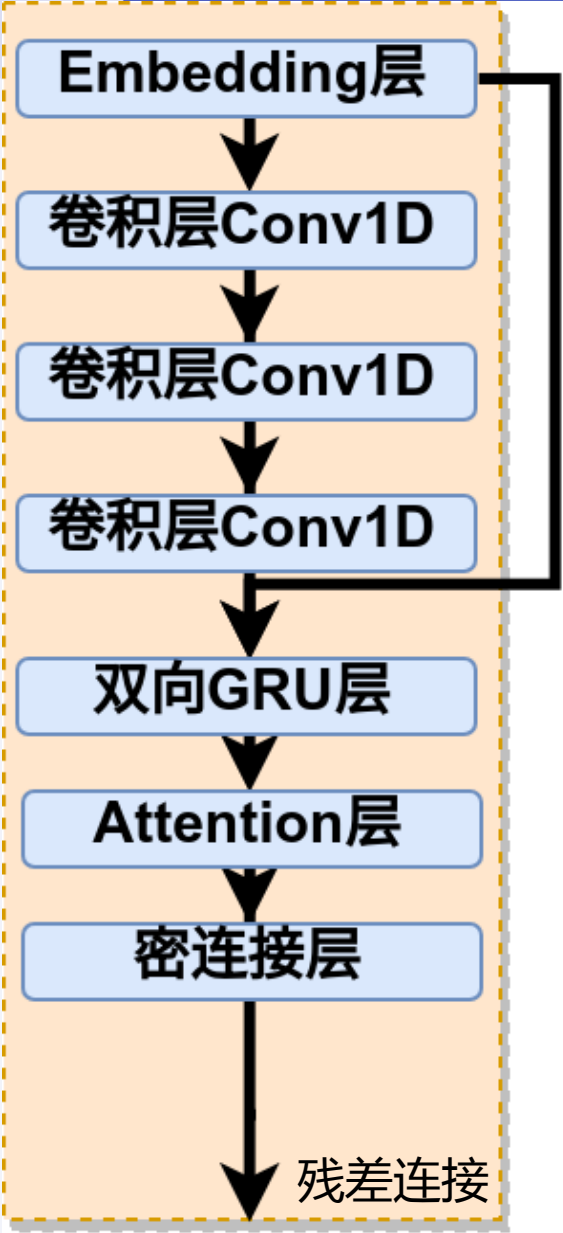
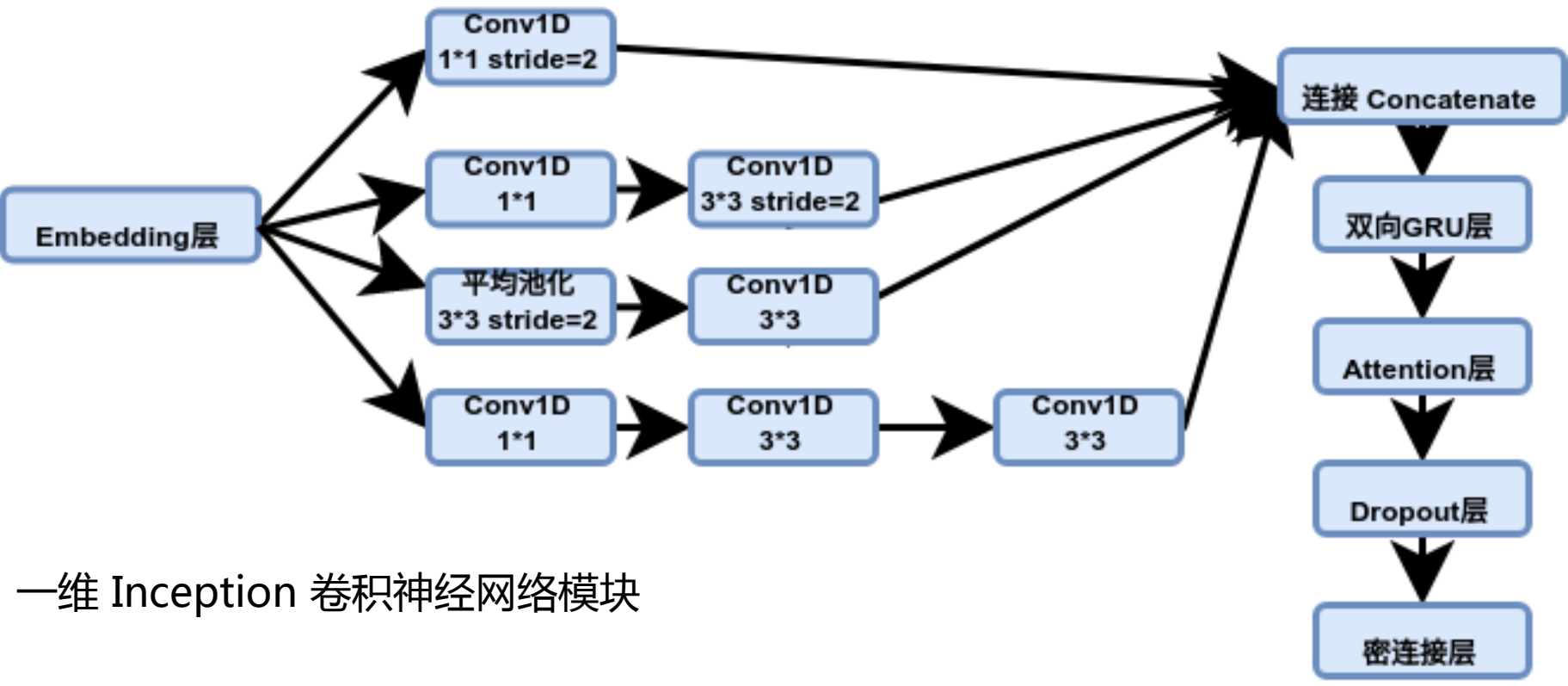
残差连接 也是避免信息遗漏的方法，通过某一层神经网络去和本层之前的某一层进行连接来实现。



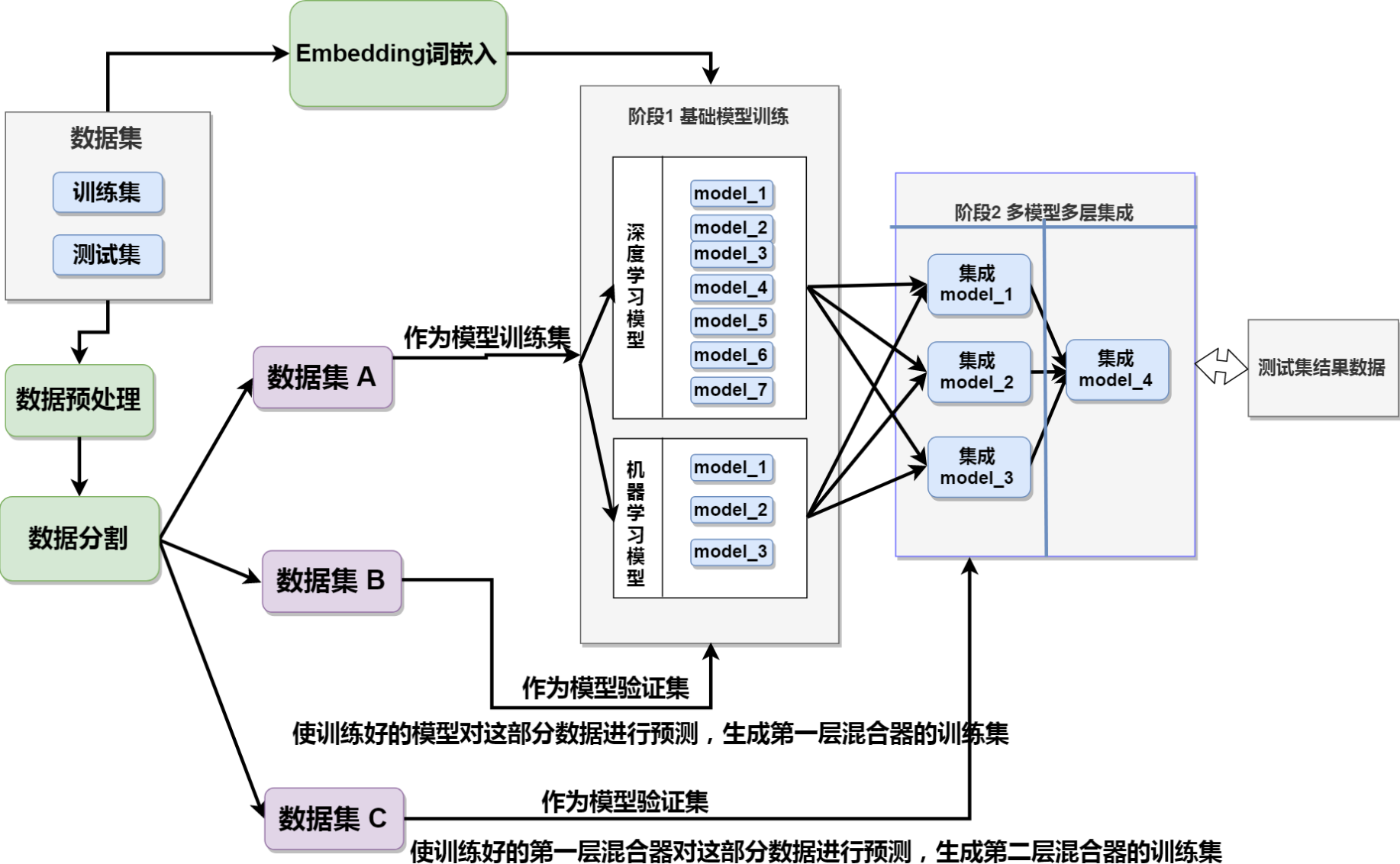
Dropout 正则化

可以制定数量的使神经元随机失活，是避免模型训练过早产生过拟合的重要方法，同样的方法还有 Batch Normalization

02 算法设计

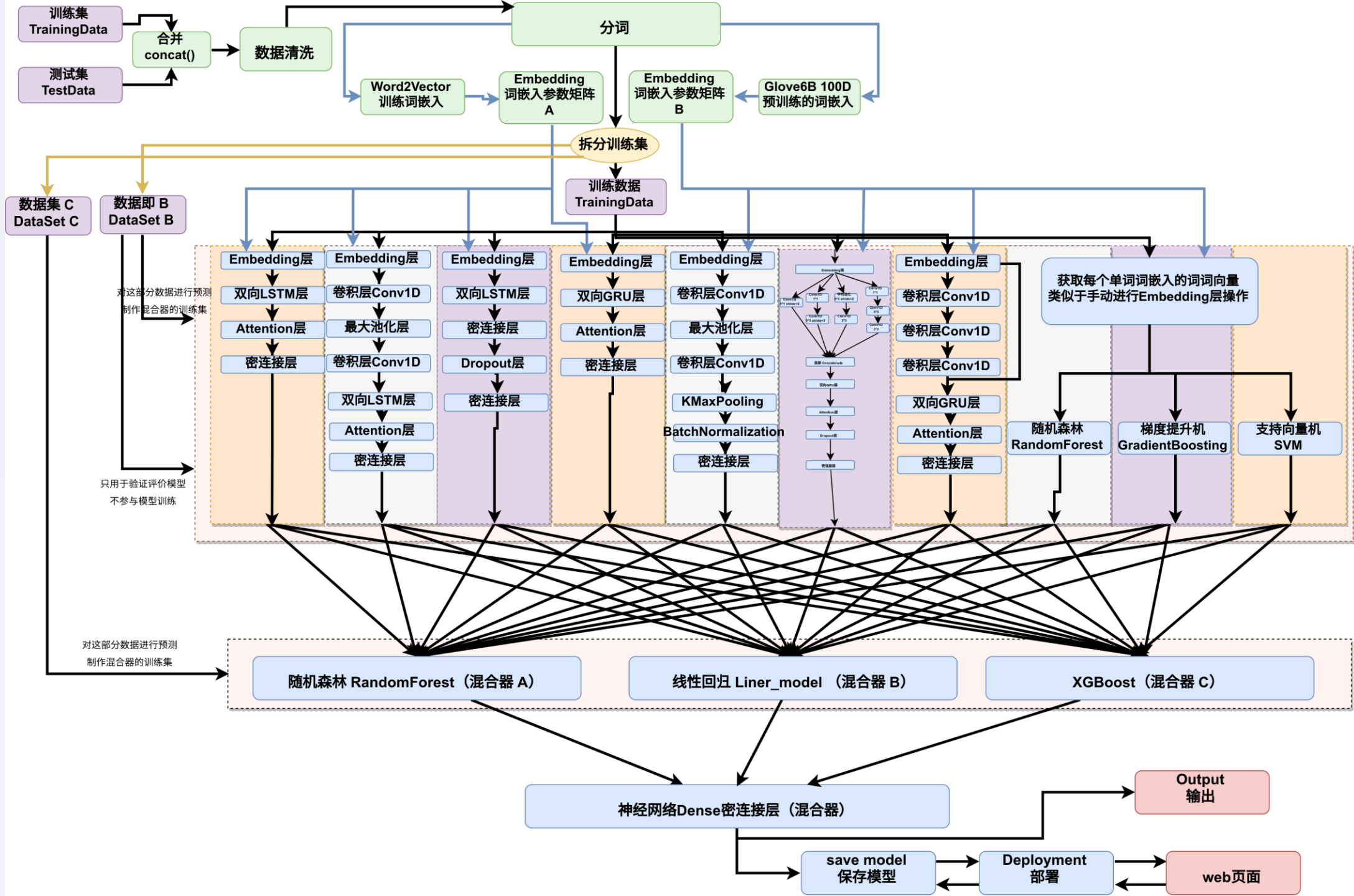


02 算法设计



模型集成融合

- ~ 三层结构
- ~ 数据集拆分 (clean)
- ~ 构造下一层的训练集



03

模型调优与测试

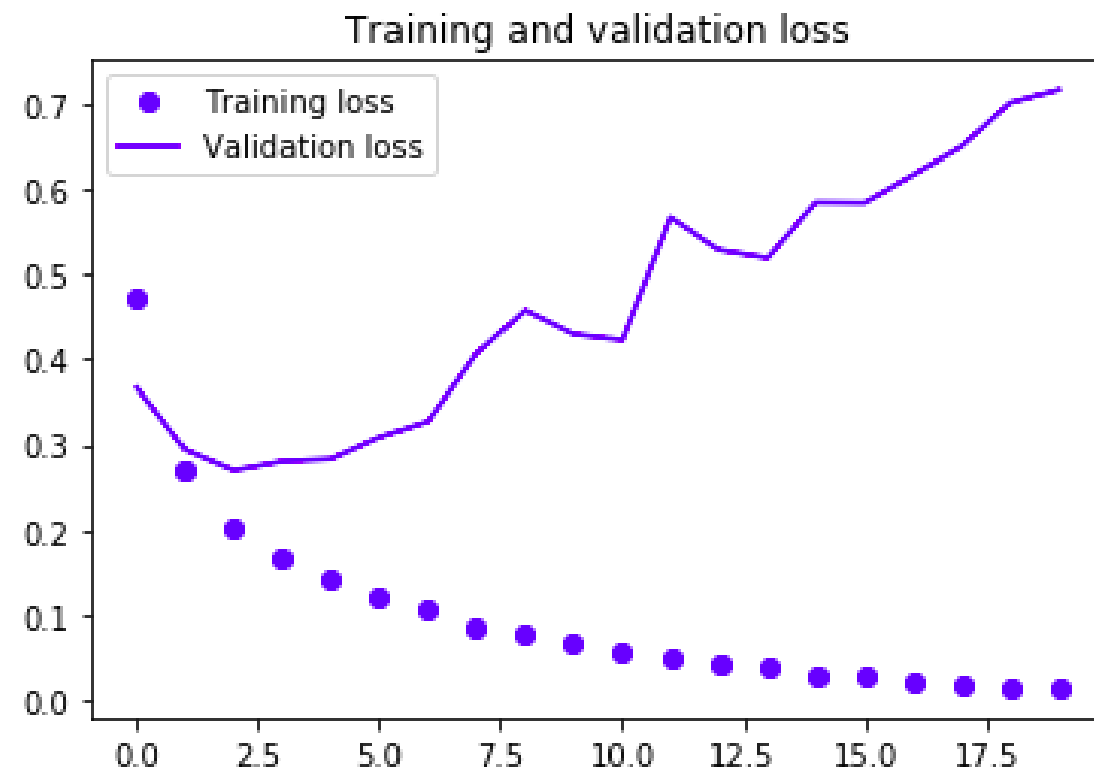
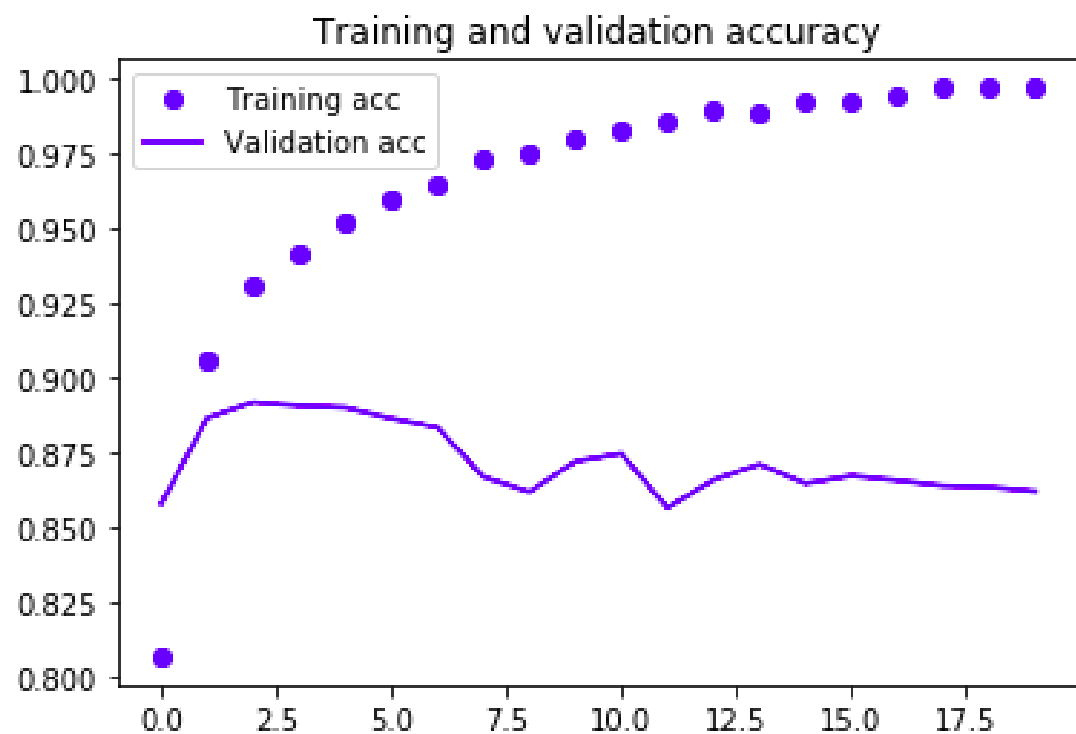
03 模型调优与测试

id	情感	评论
5814_8	1	With all this stuff going down at the moment with MJ i've started listening to his music, watching the odd documentary here and there,...
2381_9	1	\The Classic War of the Worlds\" by Timothy Hines is a very entertaining film that obviously goes to great effort and lengths to faithfully recreate H. G. Wells' classic
7759_3	0	The film starts with a manager (Nicholas Bell) giving welcome investors (Robert Carradine) to Primal Park. A secret project mutating a primal animal using fossilized
3630_4	0	It must be assumed that those who praised this film (\the greatest filmed opera ever,\" didn't I read somewhere?) either don't care for opera, don't care for Wagner,

03 模型调优与测试

id	评论
8348_2	This movie is a disaster within a disaster film. It is full of great action scenes, which are only meaningful if you throw away all sense of reality. Let's see, ...
5828_4	All in all, this is a movie for kids. We saw it tonight and my child loved it. At one point my kid's excitement was so great that sitting was impossible. However, I am a
7186_2	Afraid of the Dark left me with the impression that several different screenplays were written, all too short for a feature length film, then spliced together clumsily into

03 模型调优与测试



03模型调优与测试












在第一个IMDB数据集上经过AUC评分，计算重合的面积，可以达到95.97%分，排名能达到前15%。

Name	Submitted	Wait time	Execution time	Score
fffff_01.csv	4 days ago	0 seconds	0 seconds	0.9597

Complete
























[Jump to your position on the leaderboard](#)

49	eugeneyan
50	LaoGanMa
51-578	+ Load 528 More

78	Jason Ament		0.95986	5	4y
79	ziflit		0.95977	45	4y
80	Random Typing Monkeys	  	0.95977	24	4y
81	hecatombe		0.95972	10	4y
82	Quique Wolff		0.95927	6	4y
83	Soldados de Fontela	   	0.95916	64	4y

03模型调优与测试

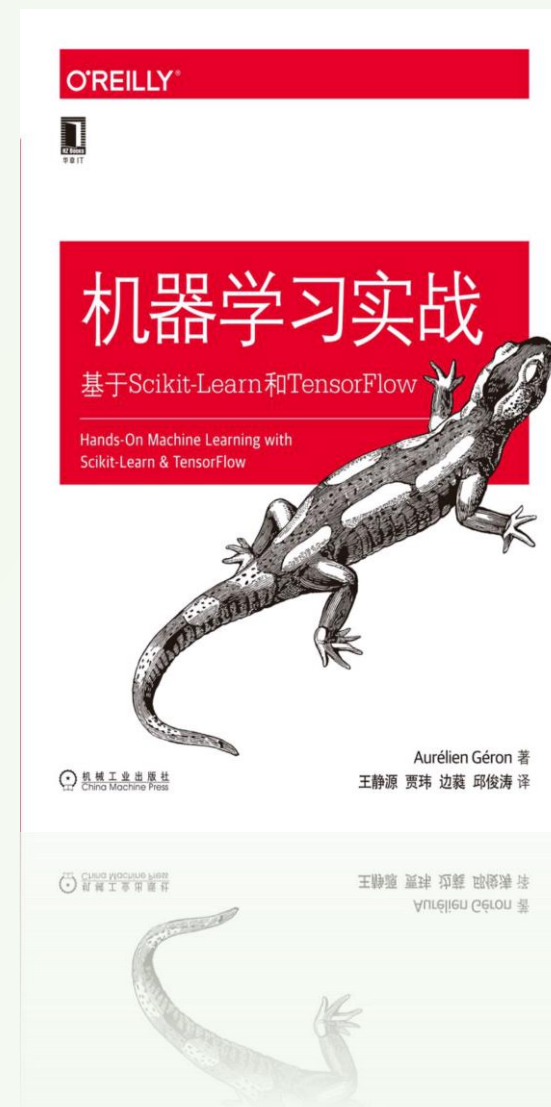
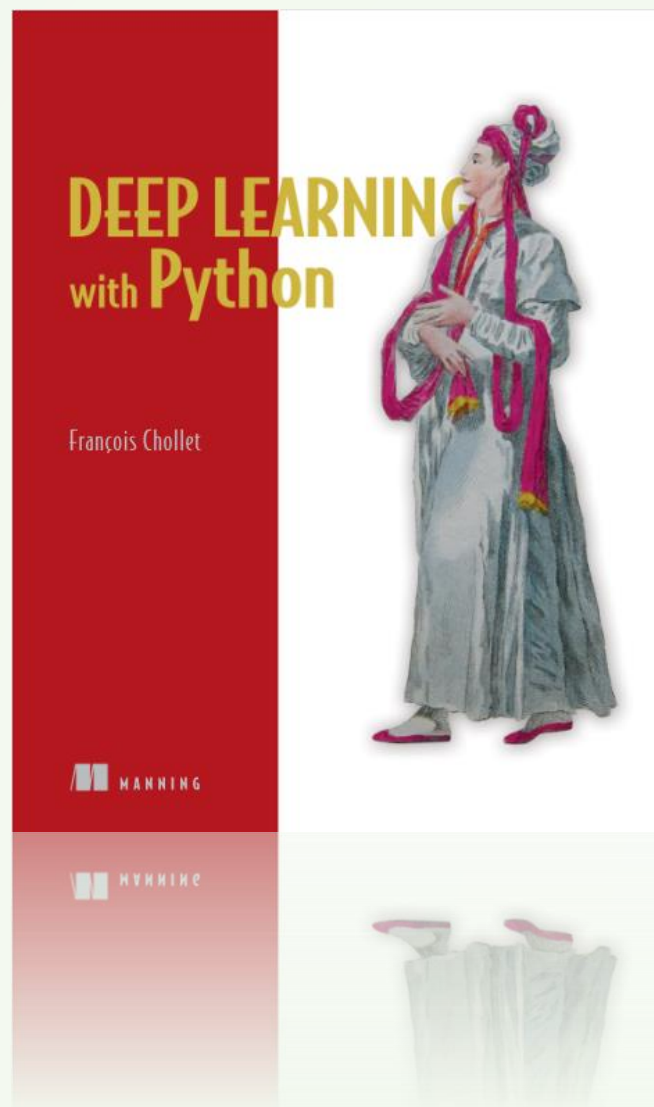
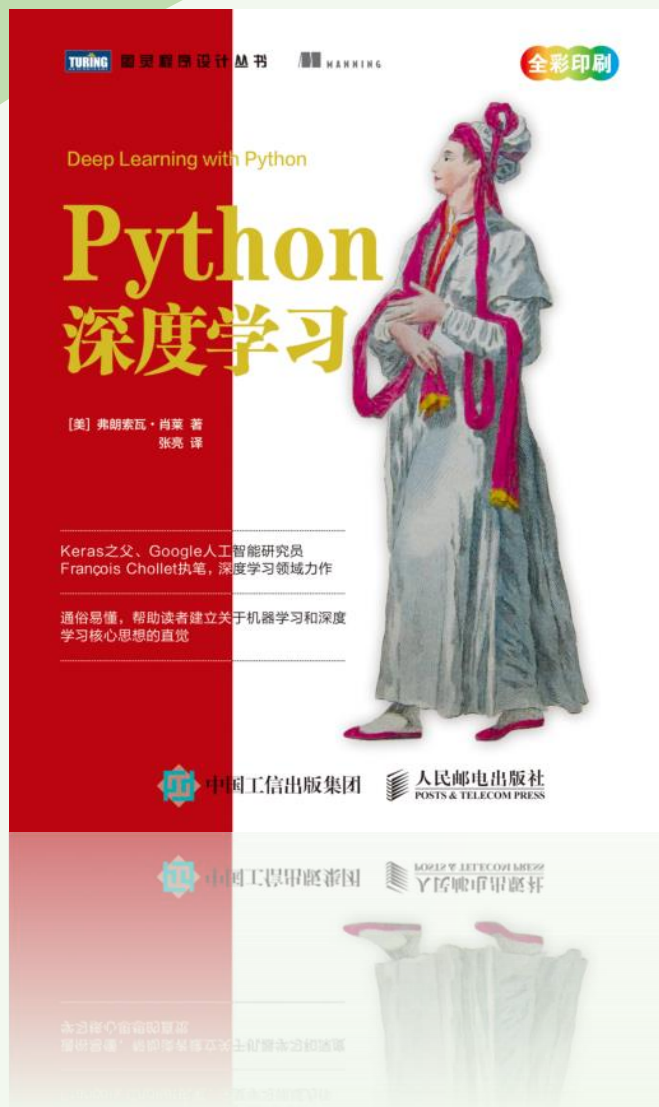
在第二个twitter数据集上经过F1 Score的评分方法，得到了0.7131280389的分数，排名196/614，30%左右。

186	 ArjunDutta	0.7155425220	
188	 Sr_Harsh	0.7155025554	
189	 puna	0.7152542373	
190	 moumita2212	0.7149460709	
191	 devanshkhs	0.7147766323	
192	 sdinesh718	0.7144906743	
193	 saswata10	0.7142857143	
194	 zmlao	0.7134502924	
195	 Siddaram	0.7133105802	
196	 woaikangyanyan	0.7131280389	 Add approach
197	 Georgios_Sarantis	0.7131011609	
198	 thaoho	0.7123655914	

04

总结回顾

04 总结回顾



感谢各位老师批评指正！

- 答辩人：聂国庆
- 班级：计算机 2015-1
- 指导老师：赵中英