



西安交通大学
XI'AN JIAOTONG UNIVERSITY

实习课题展示

——文本分类

小组成员：
汪宇豪



CONTENS



一、数据预处理

二、传统机器学习算法

三、神经网络





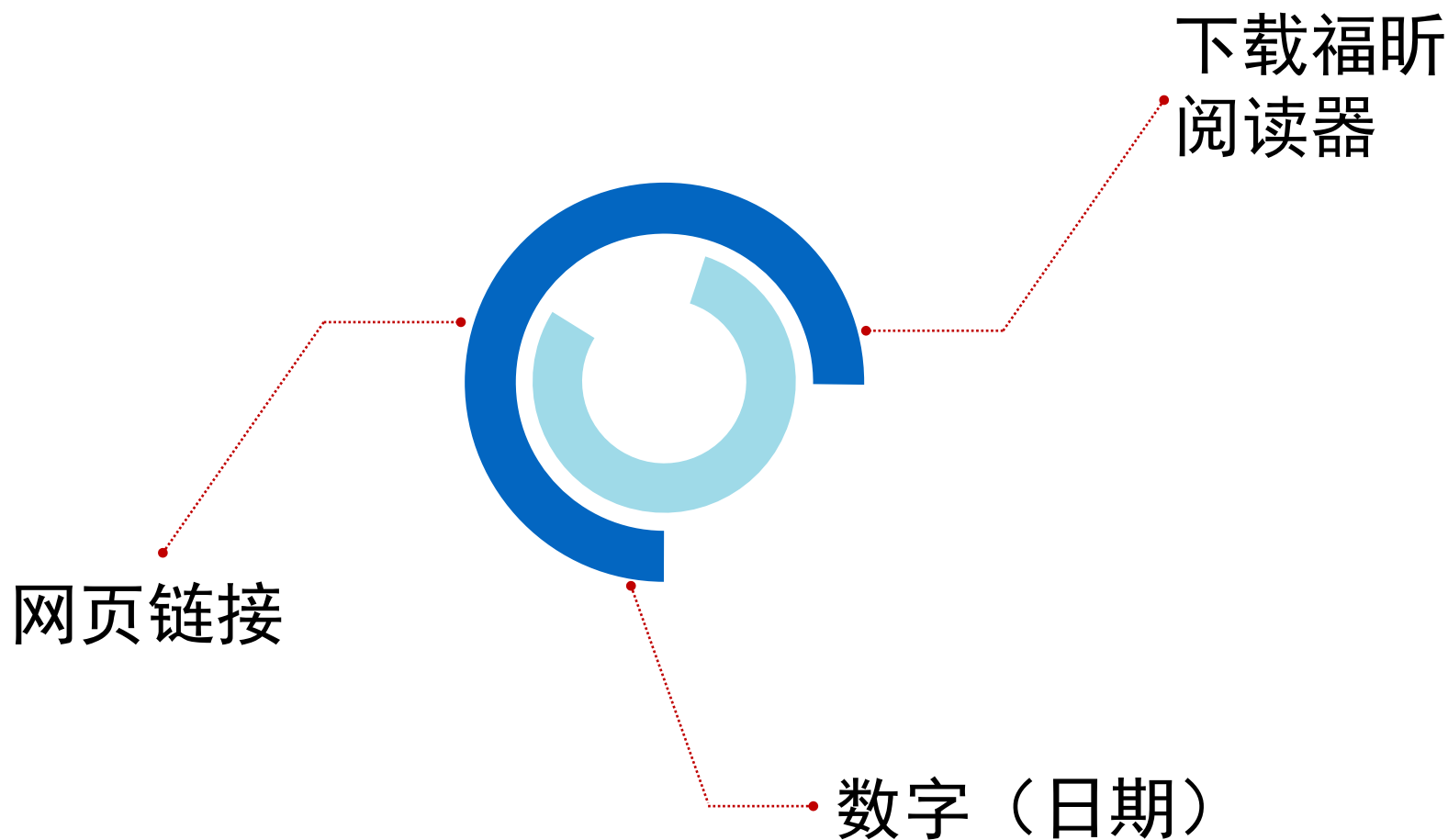
西安交通大学
XI'AN JIAOTONG UNIVERSITY

数据预处理


01



无关信息



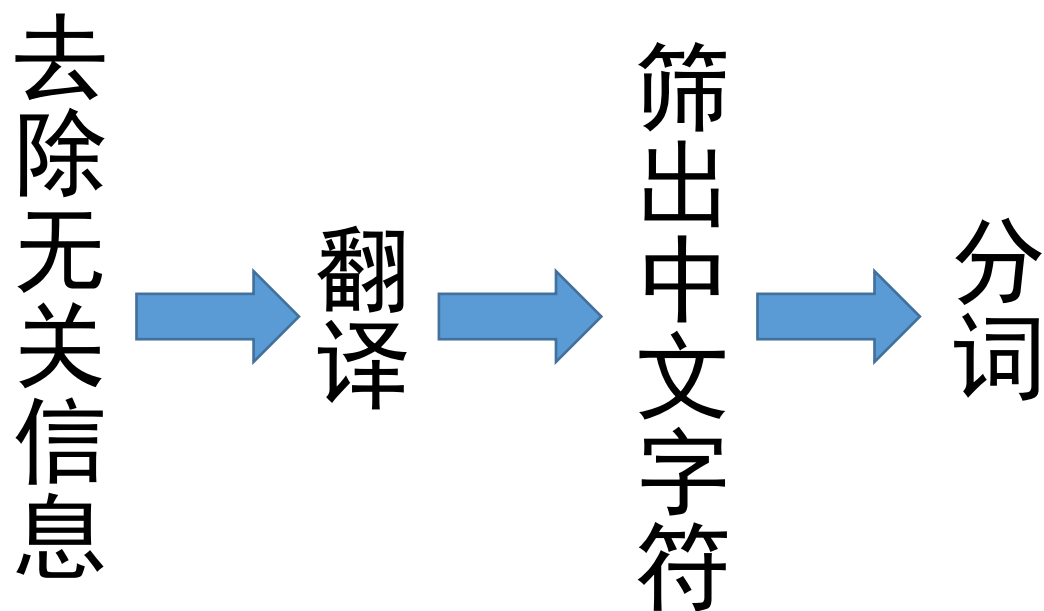
英文/中英文混合文本

jie  分词

She has 100 refereed publications
including 44 papers published at
leading.....



Shehas100refereedpublicationsincluding44paperspublishedatleading:



爬取谷歌翻译网站

爬取百度翻译网站

调用百度翻译API

问题

≡ Google 翻译

📄 文字

📄 文档

检测到中文

中文

英语

德语



交流 吕荣文教授于2018年8月17-20日参加了在加拿大大瀑布市举办的第十屆全球華人化工學者研討會 吕荣文教授于2018年8月17-20日参加了在加拿大大瀑布市举办的第十屆全球華人化工學者研討會 (10th Global Chinese Chemical Engineers Symposium, GCCESC-10) 。全球華人化工學者研討會 (Global Chinese Chemical Engineers Symposium) 起源于2009年8月，由加拿大西安大略大學杰出教授祝



点击图标下载 App

 Android  iOS

↔ 英语 中文(简体) 日语 ▼

交流 吕荣文教授于2018年8月17-20日参加了在加拿大大瀑布市举办的第十屆全球華人化工學者研討會 吕荣文教授于2018年8月17-20日参加了在加拿大大瀑布市举办的第十屆全球華人化工學者研討會 (10th Global Chinese Chemical Engineers Symposium, GCCESC-10)。全球華人化工學者研討會 (Global Chinese Chemical Engineers Symposium) 起源于2009年8月，由加拿大西安大略大学



pkuseg中文分词工具

1. 多领域分词。支持了新闻领域，网络领域，医药领域，旅游领域，以及混合领域的分词预训练模型。
2. 更高的分词准确率
3. 支持用户自训练模型
4. 支持词性标注

对比结果

预处理方法	在测试集的准确率
j i e b a中文分词	86. 6%
谷歌翻译+j i e b a中文分词	89. 2%
谷歌翻译+pkuseg中文分词	87. 0%
pkuseg中文分词+空格英文分词	87. 1%
pkuseg中文分词+n l t k英文分词	87. 9%
pkuseg中文分词+英文谷歌翻译	86. 5%



西安交通大学
XI'AN JIAOTONG UNIVERSITY

传统机器学习算法

02



对比结果

	算法	测试集的准确率
	朴素贝叶斯	85.5%
→	随机梯度下降	93.2%
	logistic回归	90.8%
	支持向量机	36.3%
→	线性支持向量机	93.5%
→	MLP	93.4%
	K近邻	82.9%
	随机森林	88.6%
	梯度boosting	90.1%
	adaboost	58.6%
	决策树	85.6%
	rocchio	74.8%



西安交通大学
XI'AN JIAOTONG UNIVERSITY

神经网络

03



第一步：调参

五个参数

`max_features` 作为特征的单词个数

`maxlen` 每个样本取特征词个数

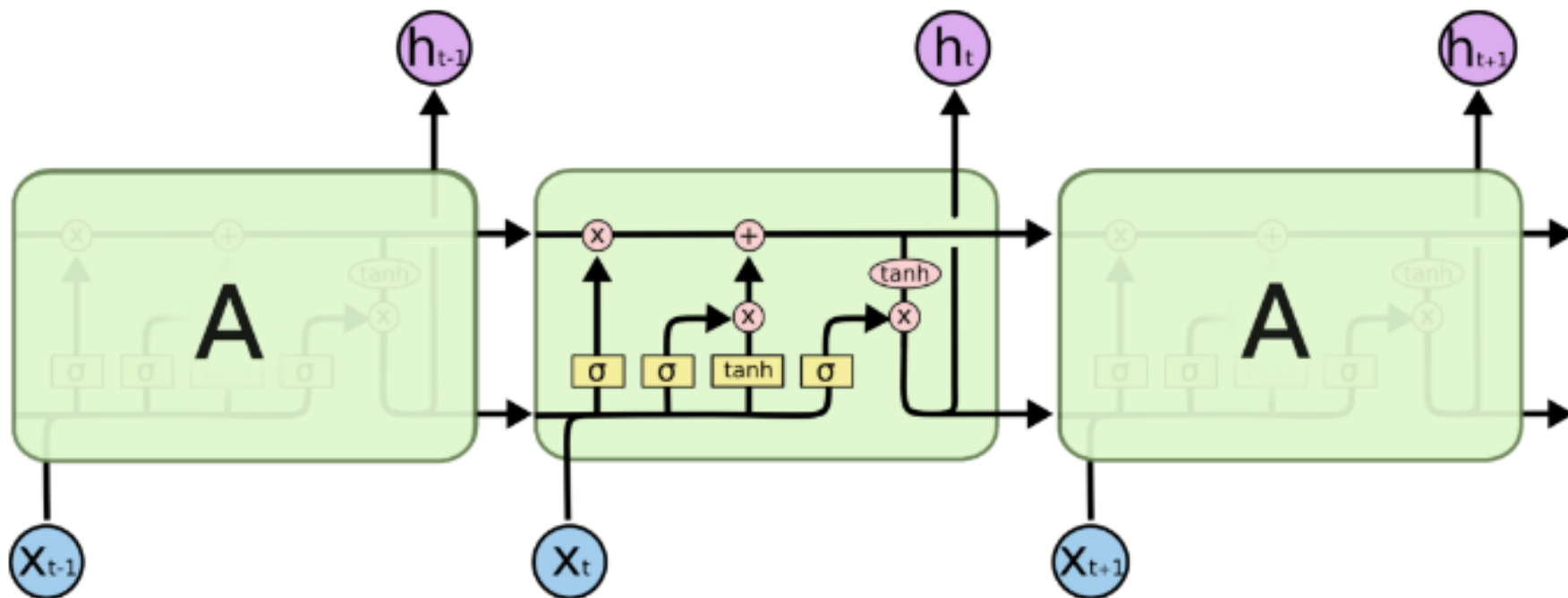
`embedding_dim` 嵌入层维度之一

`epochs` 训练轮数

`batch_size` 每次训练的batch大小

第二步：循环神经网络

1. LSTM



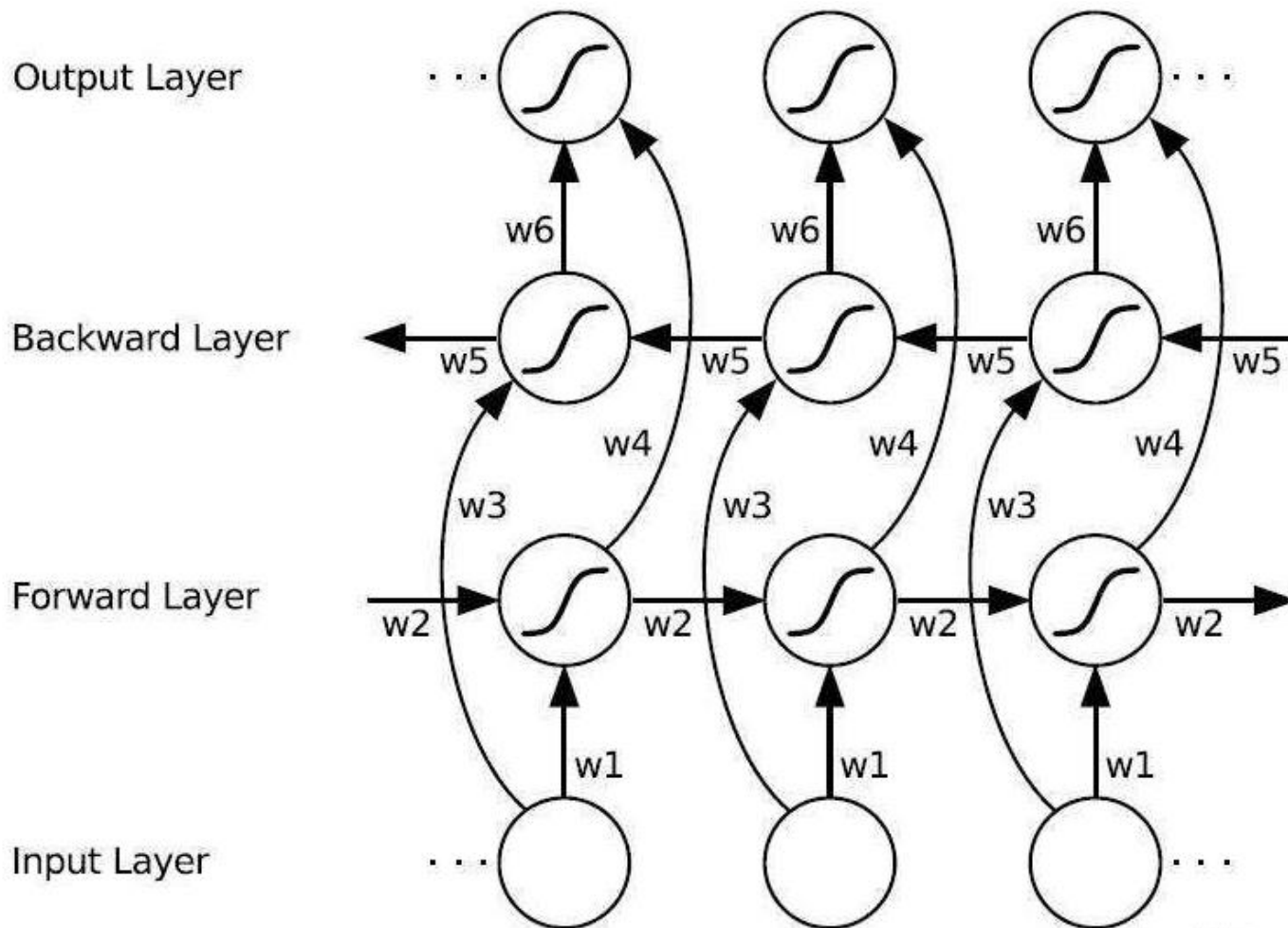
第二步：循环神经网络

嵌入层+LSTM+全连接层

88.7%

第二步：循环神经网络

2. Bi-LSTM



第二步：循环神经网络

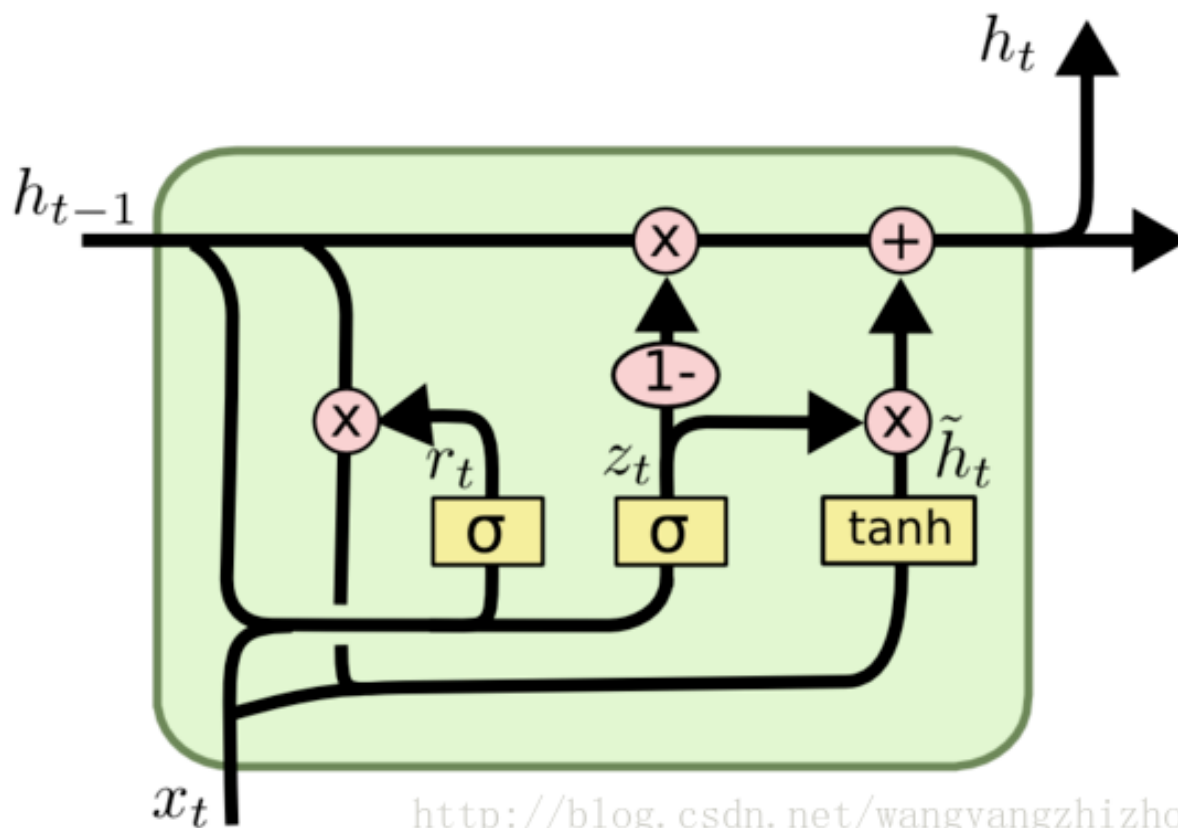
嵌入层+Bi-LSTM+全连接层

89.1%

训练时间略长

第二步：循环神经网络

3. GRU



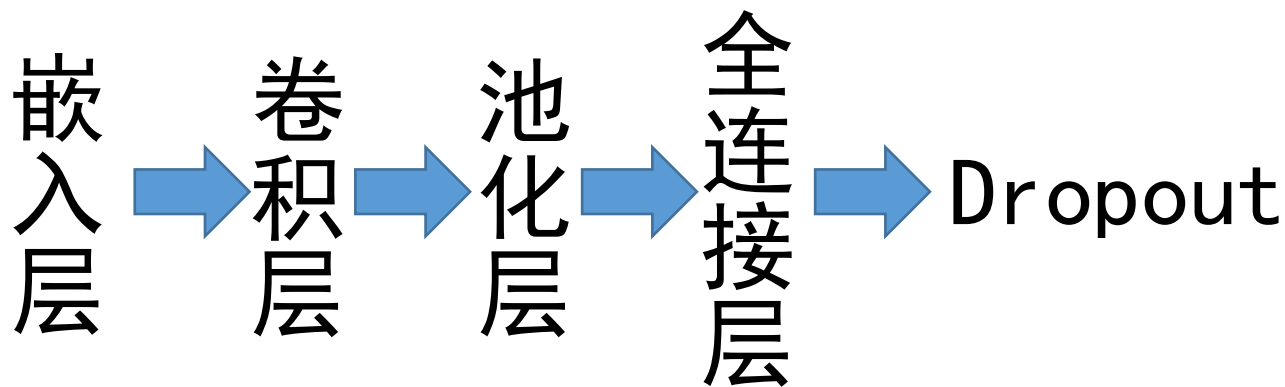
第二步：循环神经网络

嵌入层+GRU+全连接层

86.4%

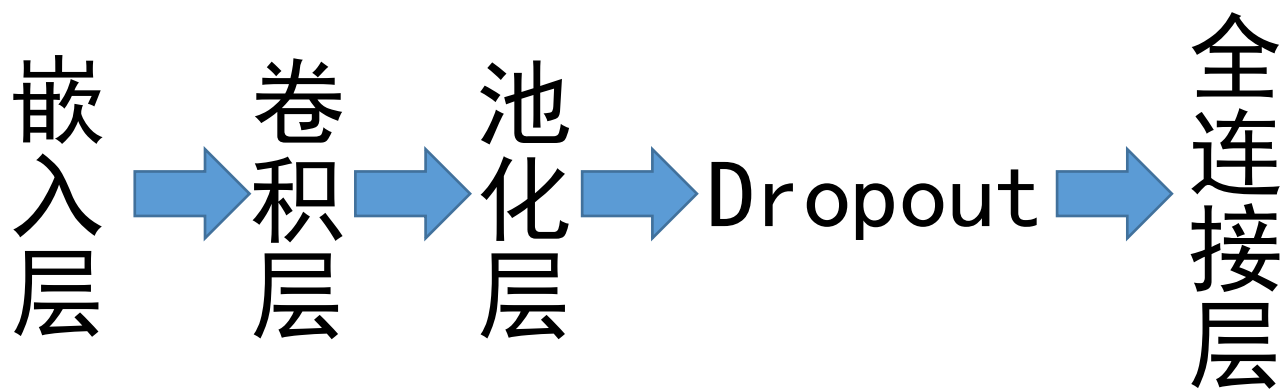
训练时间短

第三步：卷积神经网络（CNN）



92.3%

第三步：卷积神经网络（CNN）

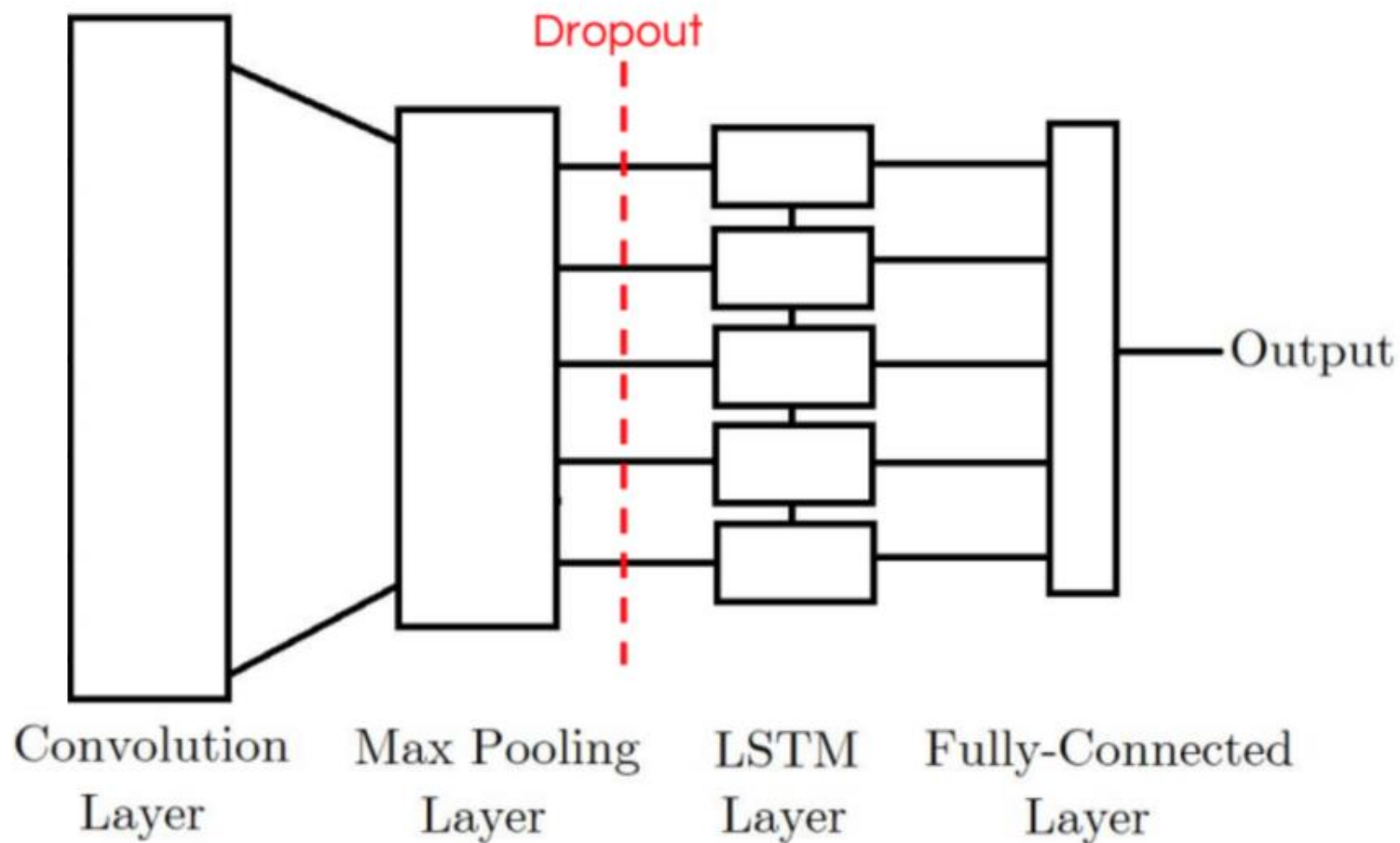


92.2%

第四步：CNN和LSTM联和

1.1 CNN-LSTM

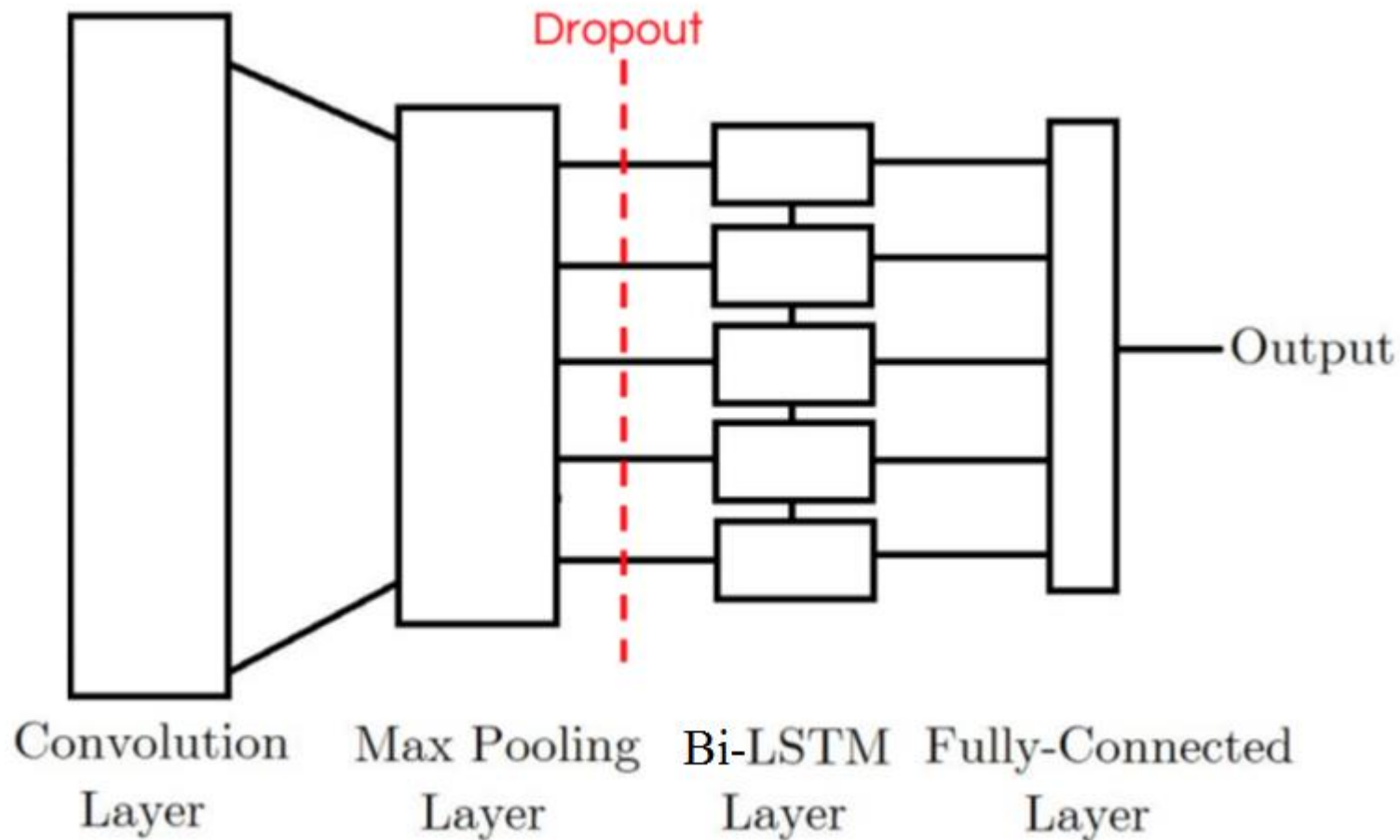
88.9%



第四步：CNN和LSTM联和

1.2 CNN-BiLSTM

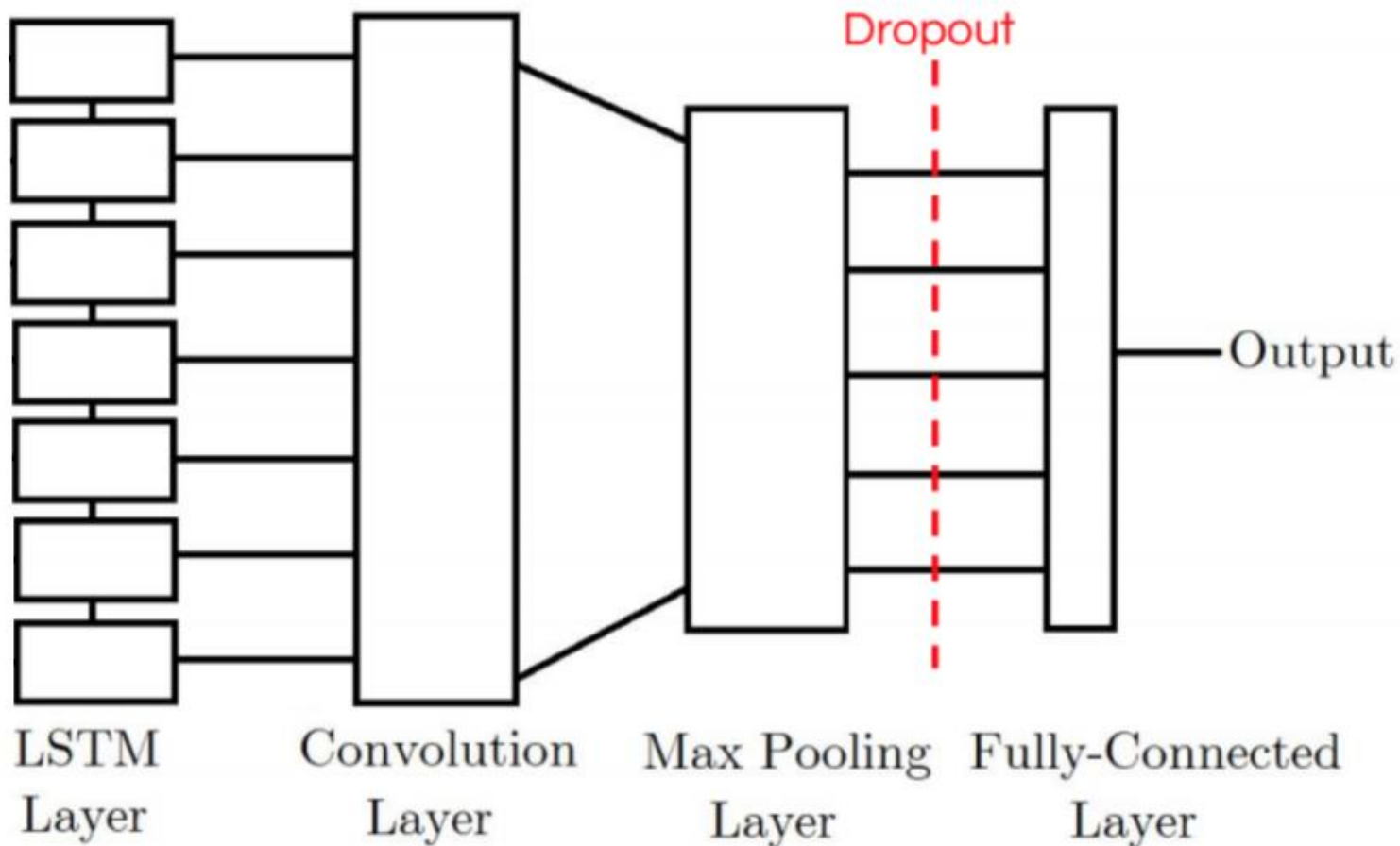
89.6%



第四步：CNN和LSTM联和

2.1 LSTM-CNN

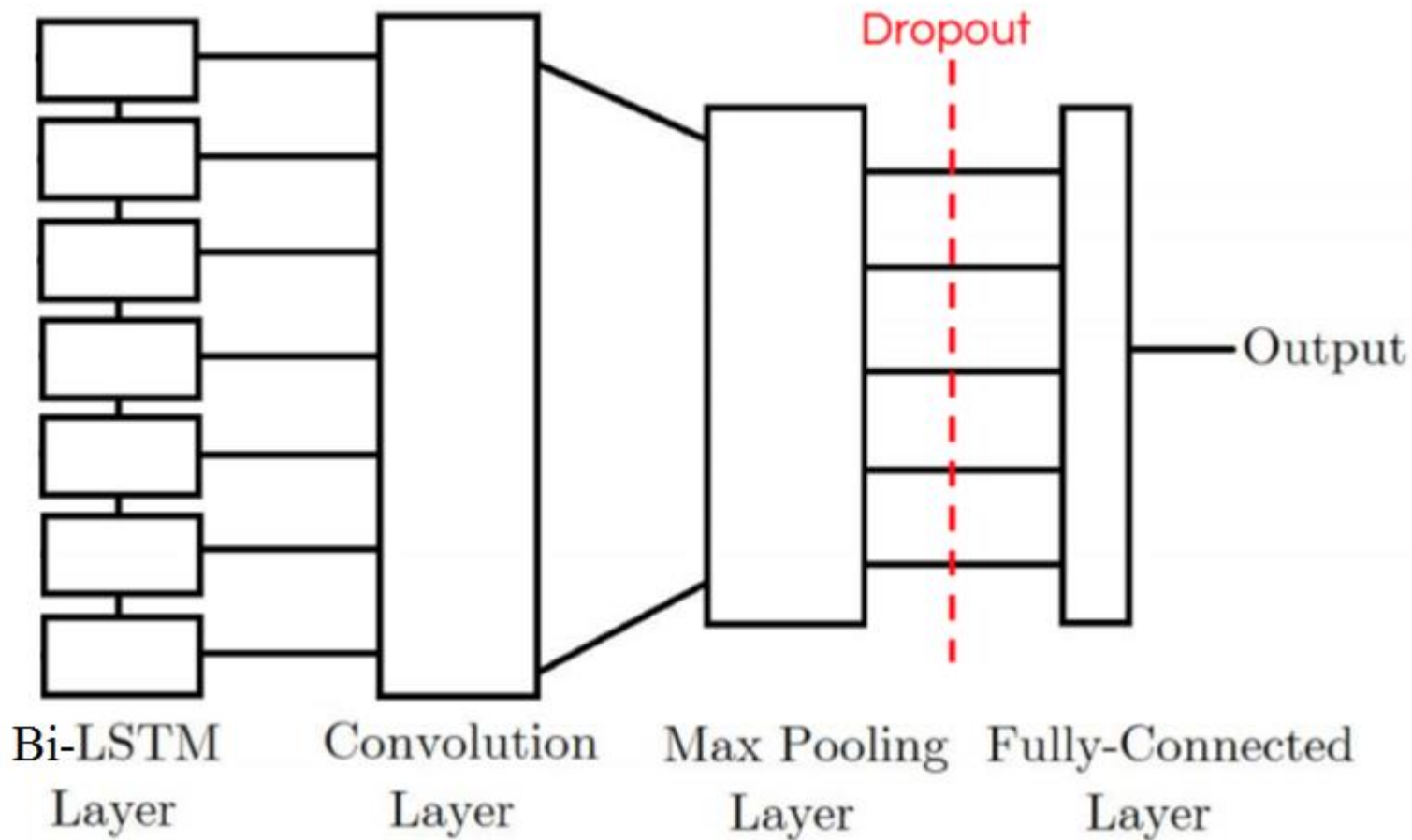
89.7%



第四步：CNN和LSTM联和

2.2 BiLSTM-CNN

90.2%



第四步：CNN和LSTM联和

LSTM-CNN-BiLSTM?



西安交通大学
XI'AN JIAOTONG UNIVERSITY

结 论





西安交通大学
XI'AN JIAOTONG UNIVERSITY

问题与解决





西安交通大学
XI'AN JIAOTONG UNIVERSITY

谢 谢！

