北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

文本分类-研究综述

小组成员：郑昊、张才华、张颢耀、白云、孙华山（报告顺序）

指导老师：张华平

时间：2022/3/29

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目 录

CONTENTS

1 背景意义与发展历史

2 TC基本流程和特征工程

3 基于传统机器学习的分类器

4 基于深度学习的分类器

5 文本分类前沿进展

6 总结与展望

1 背景意义与发展历史

人工智能


人工智能
自然语言处理
机器学习
深度学习
神经网络
自然语言
知识图谱
计算机系统
识别方法
用户界面
词向量
数据库
数据处理
上下文
数据集
文本数据
计算机存储
关键词
预处理
检测方法
语料库
传感器
客户端
特征向量
预测模型
多模态
显示器
自动生成
生成模型
接收用户
存储系统
数据分析
信息处理设备
操作方法
文本处理
实体识别
增强现实
训练数据
计算系统
麦克风
预训练
用户设备
机器学习模型
神经网络模型
输入数据
相似度
命名实体
语言模型
社交网络
元数据
准确性
提取方法
关键字
情感分析
深度神经网络
预测方法
机器人
语音识别
数据源
卷积神经网络
相关性
图像处理
信息处理方法
交互式
自适应
结构化
终端设备
大数据
推荐方法
训练方法
注意力
计算机视觉
知识库
分布式
存储器
对话系统
语音输入
数字助理
构建方法
目标文本
模型训练
分类器
数据处理系统
命名实体识别
学习方法
分析方法
学习模型
分类模型
音频数据
匹配方法
语音交互
图像数据
无线通信系统
语音数据
分类方法
标识符
区块链
特征提取
待识别
文本信息

文本挖掘

扒一扒Angelababy黄晓明背后的故事

贺公子：来源丨吃瓜有料 原文丨扒一扒angelababy黄晓明背后的故事 前不…

9756 赞同 · 645 评论 · 2019-07-15

娱乐

2022俄乌战争，如何看待俄罗斯乌克兰双方军队作战水平？

琳琳：俄军战报里第一天就摧毁了一切技术装备的乌克兰草莓兵今天整…

424 赞同 · 119 评论 · 03-16

军事

习近平：《中国共产党领导是中国特色社会主义最本质的特征》

新华社：[图片]《求是》杂志发表习近平总书记重要文章《中国共产党…

708 赞同 · 2020-07-15

政治

不吹不黑，梅西和 C 罗未来在世界足坛的历史地位孰高孰低？

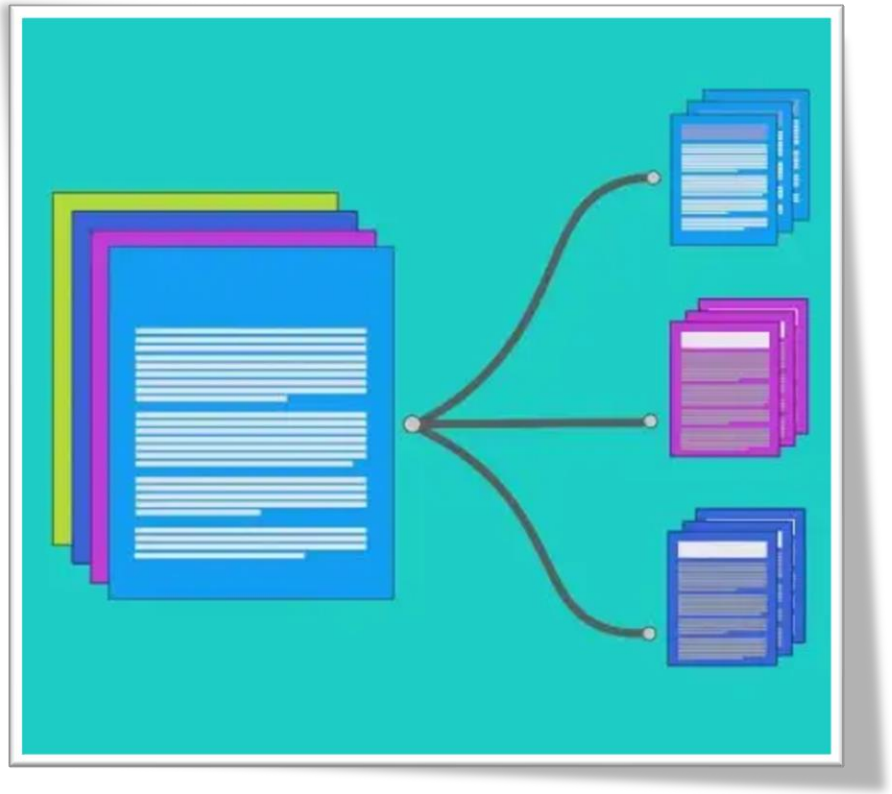
麦香包：两人的差距只会越来越大。不过我也要为梅西说几句话，拿梅…

3171 赞同 · 221 评论 · 2021-09-20

体育

背景

V2.5



- **文本分类**指的是计算机通过算法对输入的文本按照一定的类目体系进行自动化归类的过程
- **文本分类**实现的目标就是从大量的文本中分类发现有价值的信息
- **文本分类**成为了有效组织和管理文本数据重要方式

应用

标签

根据标题为图文视频打标签
政治、体育、娱乐等

邮件

垃圾邮件的判定
邮件检测和短信过滤

医疗领域

智能分诊技术节约大量医疗资源
提升服务质量和效率

网络安全

根据用户历史访问记录进行分类
机构从而决定是否允许其访问

标签


12:48
存草稿
发布
上传完成
修改封面
分区 * 美食-美食制作
标题 * 蛋糕的制作工艺
7/80
类型 * 自制 转载
转载权限 未经作者授权 禁止转载
标签 * 请选择合适的标签
简介 可填写简介
0/250
添加水印


12:48
添加标签
完成
输入自定义标签，最多20个字符
0/6
推荐标签
料理制作 美食vlog 家常菜 蛋糕
烘焙 甜点 美食 美食圈 制作教程
甜品 生活 吃货 美食VLOG 自制
美食制作 厨艺 美食视频 日常
料理 面包
参与活动
迷惑动物图鉴
沙雕动物欢乐多
B站小剧场
等你来拍最好看的短剧~~
热 金牌美食制作人
制作美食，赢14万元投稿激励!
热 这个真好吃
拍摄生活领万元大奖
跟着UP主创作吧（第三期）
投稿瓜分23万，新UP必得5元


烘焙
甜点
美食
美食圈
制作教程
甜品
生活
吃货
美食VLOG
自制
美食制作
厨艺
美食视频
日常
料理
面包

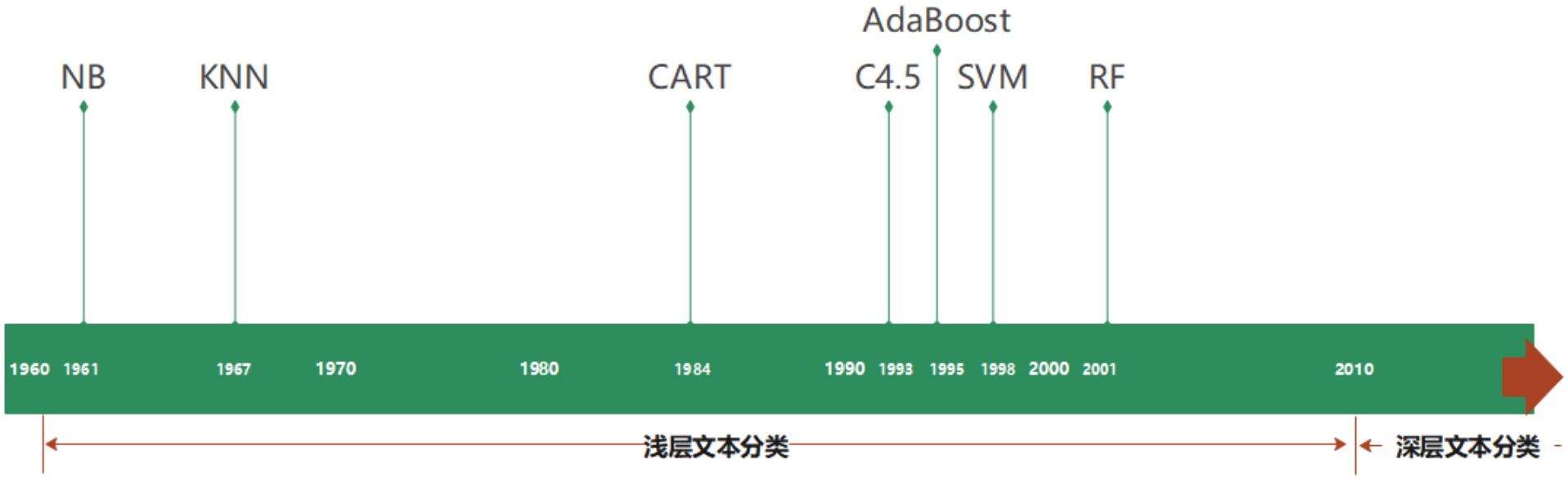
邮件

医疗


推荐医生
您好，请问有什么可以帮助您的吗?
我小腹很痛
您还有其他症状吗?
身体发热
本结果仅用于辅助挂号诊断，不代表任何诊断及治疗意见
推荐您去以下科室挂号:
外科
挂号>

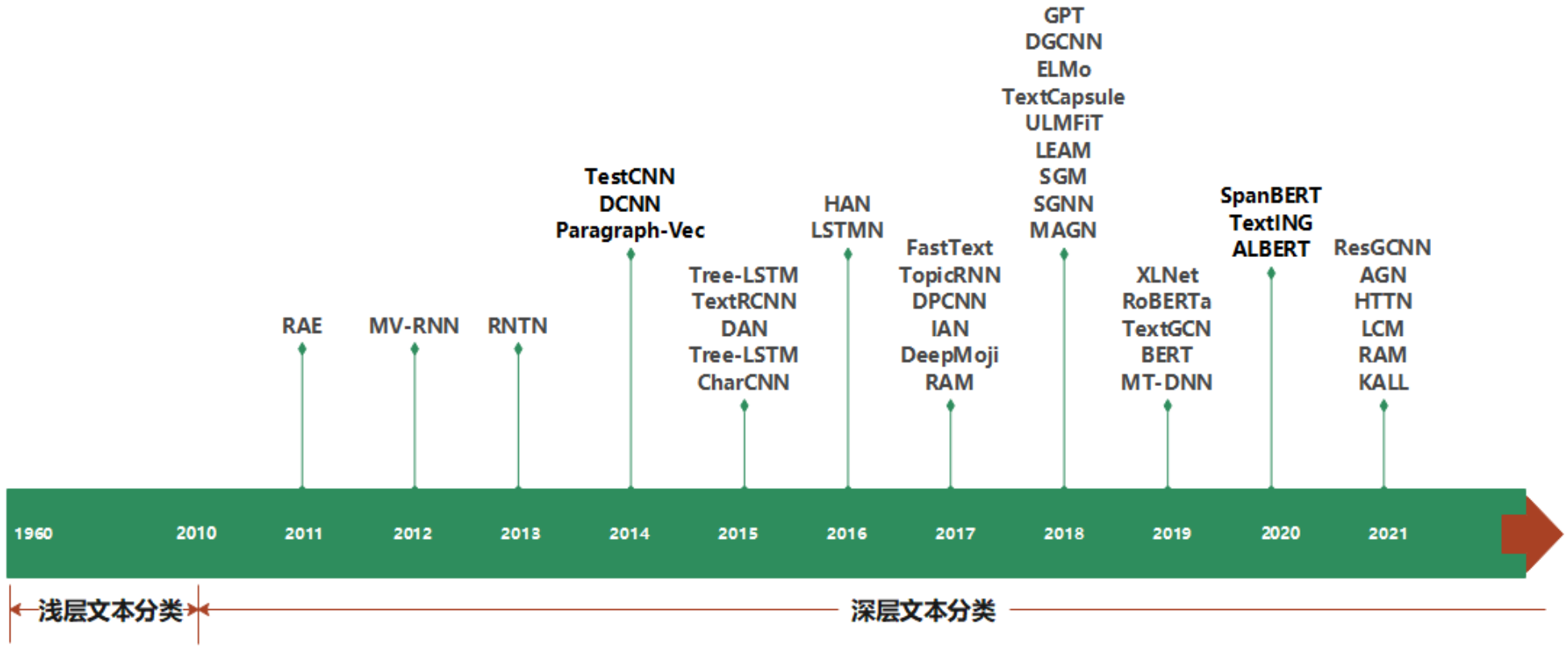
社交网络异常账号检测

浅层文本分类


AdaBoost
NB
KNN
CART
C4.5
SVM
RF
1960
1961
1967
1970
1980
1984
1990
1993
1995
1998
2000
2001
2010
浅层文本分类
深层文本分类

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

深层文本分类



对比

文本分类

机器学习

- 准确性和稳定性较好
- 不具语义学习能力
- 特征工程费时费力
- 模型表征能力有限
- 结果有效性受限
- 分类效果较差

深度学习

- 自动获取特征表达能力
- 模型提取特征能力好
- 具备语义学习能力
- 分类效果好
- 分类效率高

对比

特 离散→连续 高维→连续 征

学 浅层→深层 习

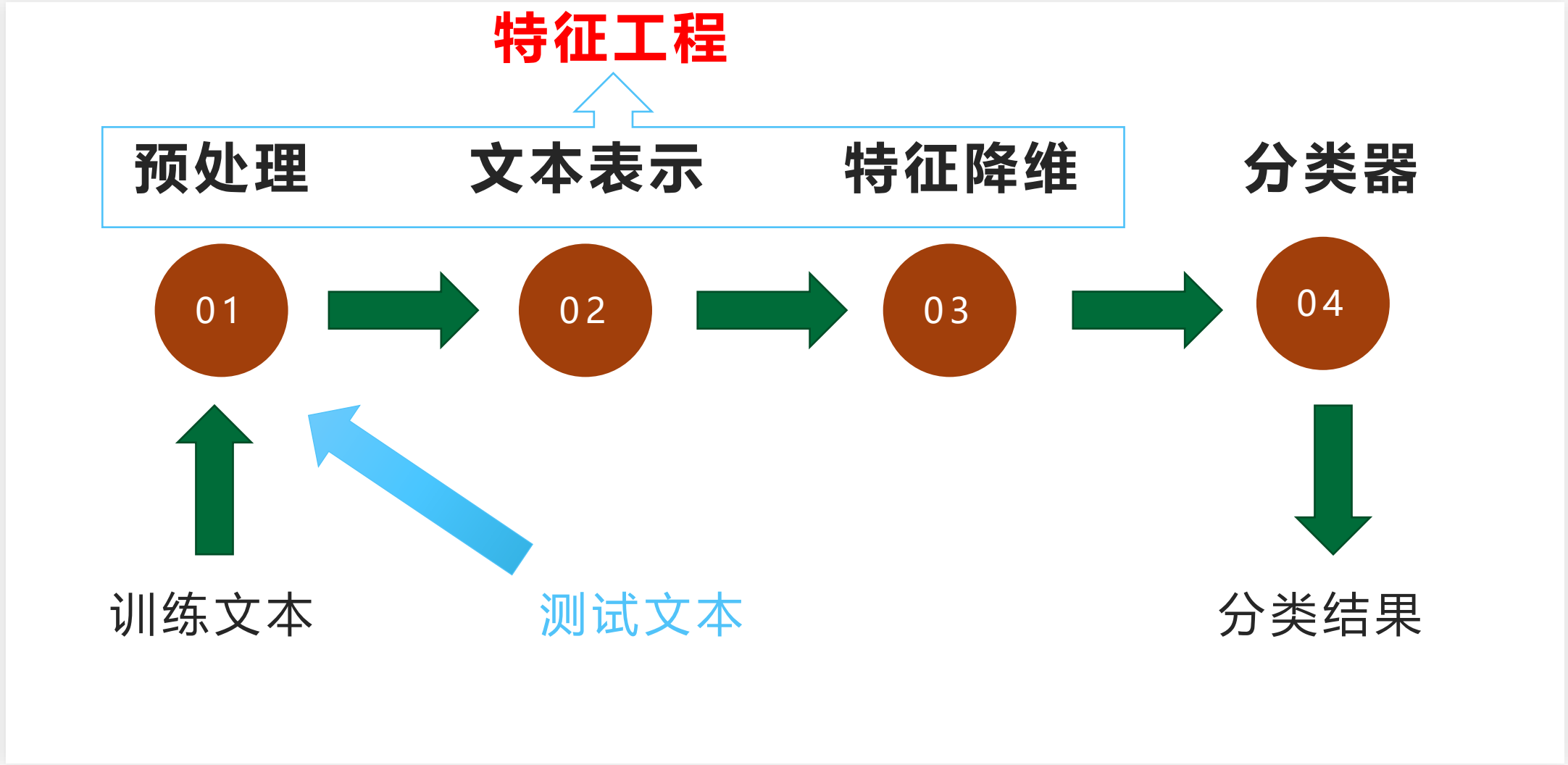
文本 步骤式→整体学习 处理

文本 浅层→深层 理解

文本 单一→集成 分类

2 基本流程及特征工程

基本流程



部分常用数据集

任务类型	名称	类型	数据量	任务类型	名称	类型	数据量
情感分析	Yelp	商户点评	Yelp-2：59.8w Yelp-5：70w	主题分类	DPpedia	Wikipedia常用信息	63w 14类
	IMDB	影评	5w		Ohsumed	医学摘要	7400文档 23类心血管病
	MPQA	新闻意见	10606		EUR-Lex	欧盟法律	2w 文档 4k标签
	Amazon	产品评论	400w	QA问答	SQuAD	Wikipedia文章	SQuAD 1.1 107,785个问题-答案对
新闻分类	AGNews	学术新闻	12.76w 4个类别		MS MARCO	Bing用户查询	10w 问题 20w文档
	Reuters	短新闻	11228 46个类别	NLI	SNLI		52w 句子 3标签
	搜狗新闻集	CA及CS新闻语料库	51w 5个类别		SICK		1w 3标签

特征工程

文本预处理

处理文本标记、分词、词干提取、去停用词

文本表示

离散表示

分布式表示

特征表示

特征选择

特征提取

文本表示——离散表示方法

1.One-Hot 独热表示法

语料：
I like deep learning.
I like NLP.
I enjoy flying.

词表：
{ I， like ,deep, learning, NLP ， enjoy, flying }

One-Hot表示:
I [1,0,0,0,0,0,0]
like [0,1,0,0,0,0,0]
deep [0,0,1,0,0,0,0]

—简单、速度快

—词之间的相似性无法衡量、词之间的重要性差异无法体现

文本表示——离散表示方法

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

2.Bag of Word(BoW) 词袋模型

https://pic4.zhimg.com/80/v2-1924d71f0025cb6ae5bcc89bb3ecd74f_1440w.jpg

0/1:

I like NLP. I enjoy flying. [1,1,0,0,1,1,1]

TF:

I like NLP. I enjoy flying. [2,1,0,0,1,1,1]

TF*IDF:

I like deep learning. [0.33,0.5,1,1,0,0,0]
I like NLP. [0.33,0.5,1,0,1,0,0]
I enjoy flying. [0.33,0,0,0,0,1,1]

TF—只能表达词在当前文本中的重要程度；很多停用词由于频次高，权重大

TF*IDF—词之间是独立的，无法提供词序信息和上下文信息；数据稀疏

3.N-gram N元组表示法

重构词表：N=2, 长度 7→13

{ I， like ,deep, learning, NLP , enjoy, flying， I like, deep learning, like deep, like NLP, I enjoy, enjoy flying}

TF*IDF:

I like deep learning. [0.33,0.5,1,1,0,0,0， 0.5, 1,1,0,0,0]

I like NLP. [0.33,0.5,1,0,1,0,0， 0.5， 0,0,1,0,0]

I enjoy flying. [0.33,0,0,0,0,1,1， 0,0,0,0,1,1]

—增加前后文信息，可以获取局部的上下文信息

—词表维度增大、数据稀疏

文本表示——分布式表示法（词向量、词嵌入）

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

1.Co-Occurrence 词向量

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

共现矩阵
+
降维

—考虑到句子中词的顺序

—词表的长度很大，导致词的向量长度也很大；共现矩阵也是稀疏矩阵

2.Word2Vec

CBOW—利用上下文的词预测中心词

SKIP-GRAM—利用中心词预测上下文的词

—学到语法和语义信息、词向量维度小、通用性比较强

—无法解决多义词的问题；静态模型；无法针对特定任务动态优化

3.GloVe

—基于全局词频统计，结合全局矩阵分解和局部上下文窗口的优点

—考虑词语上下文，全局语料库信息；词向量维度小

—无法解决多义词的问题、静态模型，无法针对特定任务动态优化

4.ELMO

语言模型训练神经网络，使用词嵌入时，词已具备上下文信息，神经网络可以根据上下文信息对词嵌入进行调整，调整后的词嵌入更能表达在上下文中的具体含义，解决了静态词向量无法表示多义词的问题

—可以表示多义词

特征降维——降低特征向量维度

特征提取

通过属性间的关系，改变原特征空间，如组合不同属性得到新的属性。

主要方法：PCA（主成分分析）、LSI（潜在语义索引）、NMF（非负矩阵分解）...

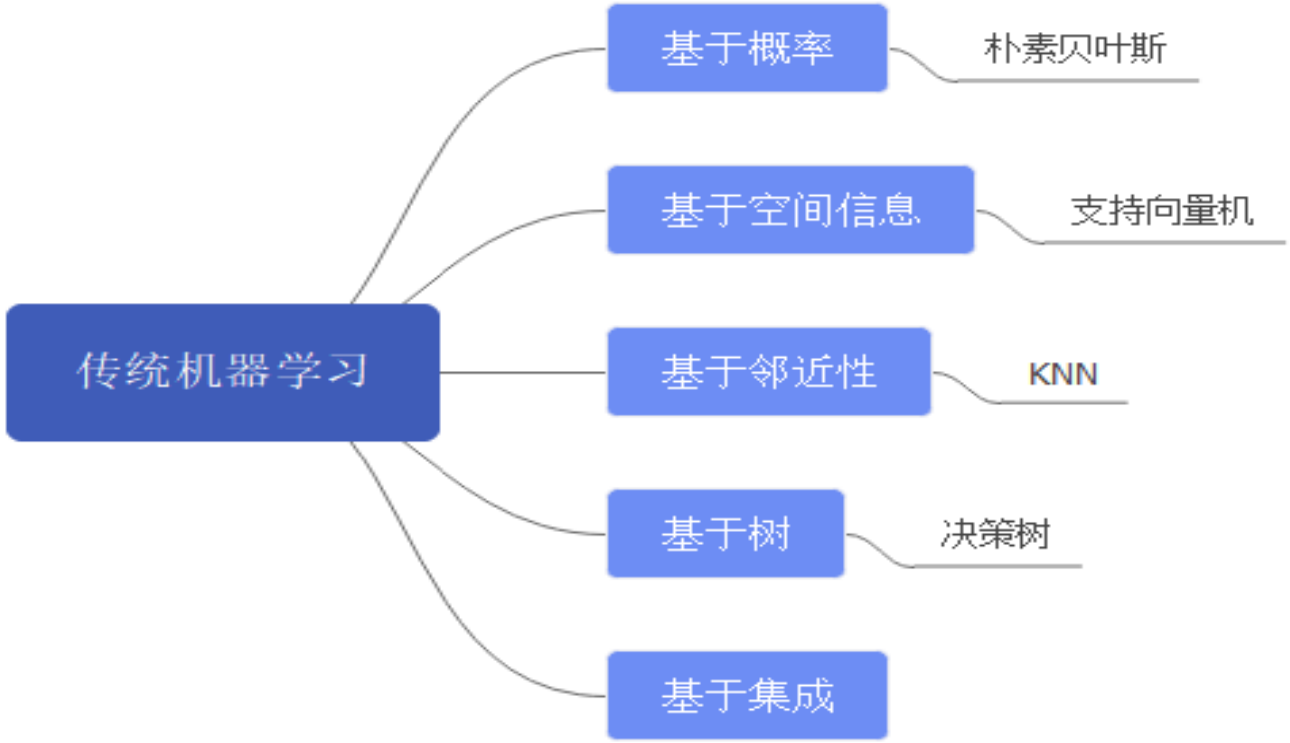
特征选择

对原特征空间中的特征进行筛选，不改变其原有属性。

统计量：特征频度、互信息、信息增益、X^2统计量、期待交叉熵...

2 基于传统机器学习算法文本分类

传统机器学习



朴素贝叶斯

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} = \frac{P(Y)P(x1,x2,\ldots,xn|Y)}{P(x1,x2,\ldots,xn)}$$

贝叶斯公式

$$P(x1,x2,\ldots,xn|Y) = P(x1|Y)P(x2|Y)\ldots P(xn|Y)$$

条件独立性假设

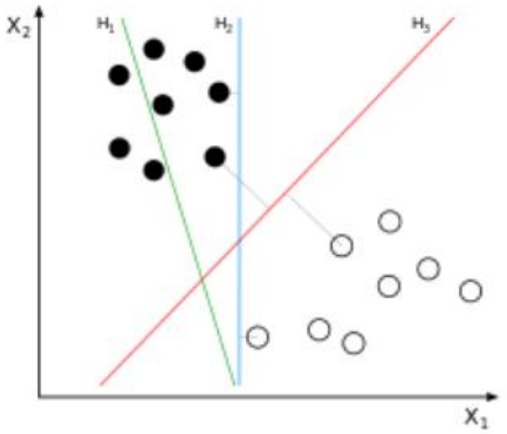
分类方法：比较P(Y)P(X | Y)

文本分类中：P(x1 | Y)是在语料库中Y类文本中x1出现的频率；
P(Y) 是语料库中Y类的频率

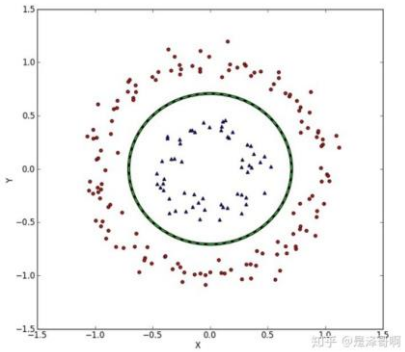
支持向量机

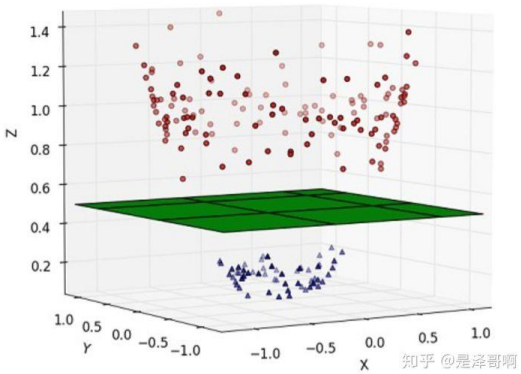
SVM

$$\frac{|w^T x + b|}{||w||}$$



核函数　　映射到高维空间



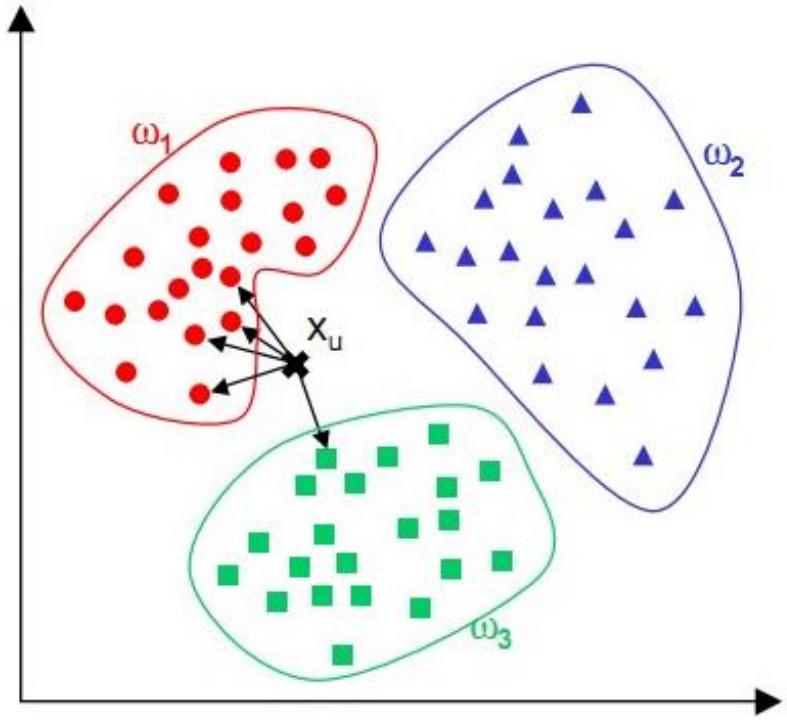


多分类：　　一对多
一对一

图片来自https://zhuanlan.zhihu.com/p/77750026

KNN

核心思想：如果一个样本在特征空间中的K个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性

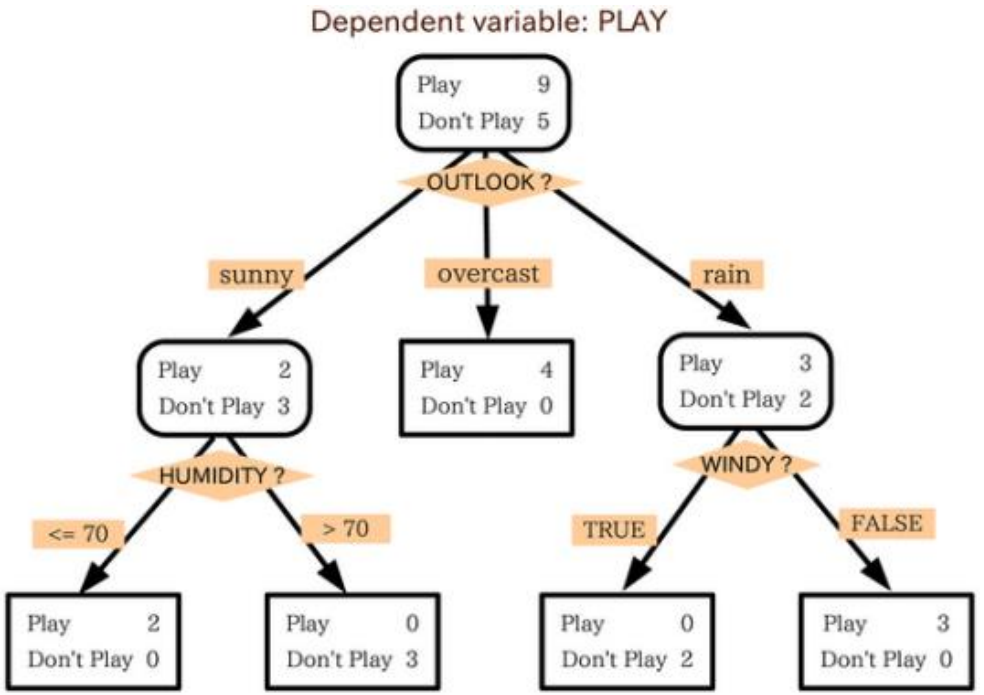


距离度量：

欧式距离
曼哈顿距离
……

决策树

决策树	对于复杂的问题，通过建立树模型产生分支节点，被划分成两个或多个较为简单的子集，从结构上划分为不同的子问题。
构建步骤	特征选择、决策树生成、决策树的修剪
生成算法	ID3, C4.5和C5.0等



集成的方法

- 假设我们有25个基分类器
- 每个分类器的分类错误率 ε=0.35
- 假设各个分类器之间是相互独立的
- 集成方法得到的模型预测错误的可能性就为：

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

主要的集成方法

- 传统集成——随机森林
- 基于增强——AdaBoost
- 基于堆叠——Stacking

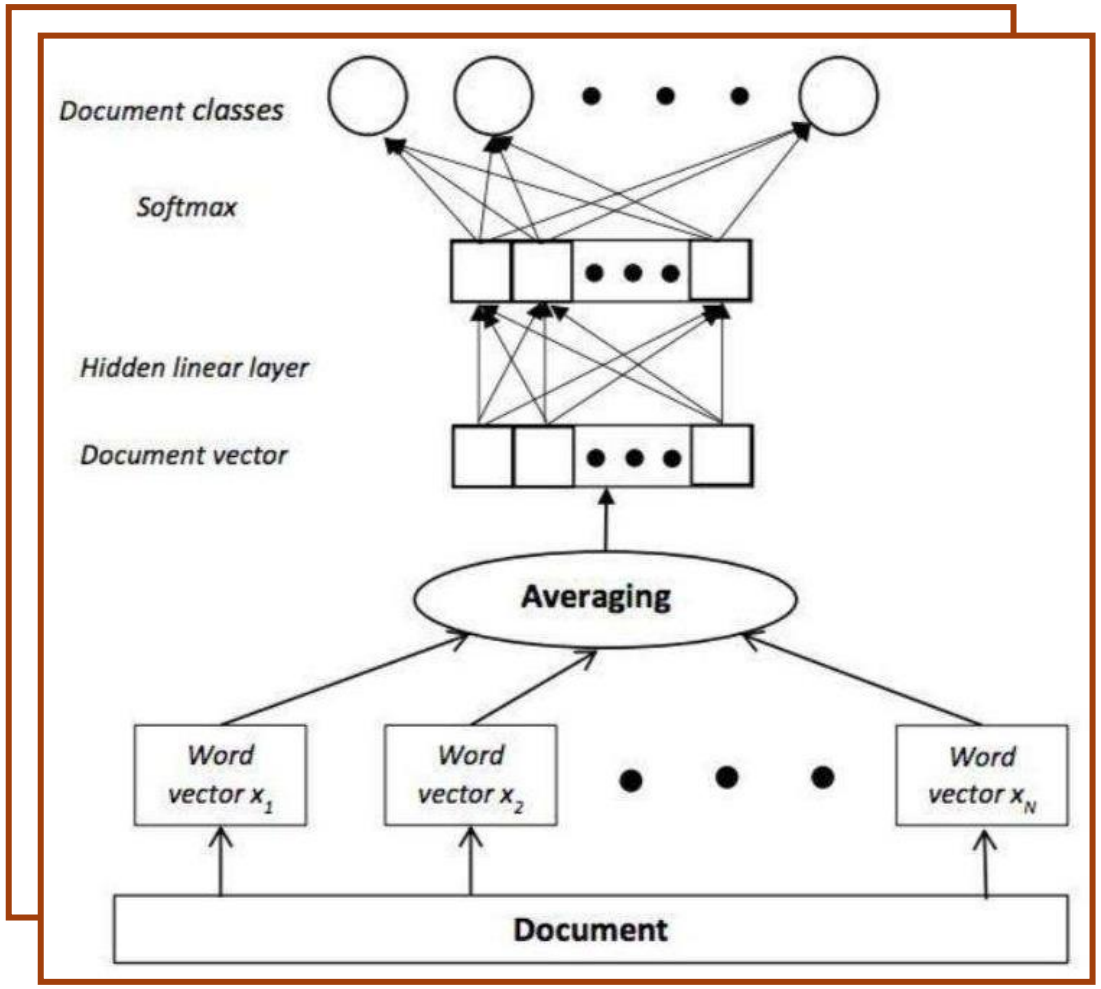
各种方法的优缺点

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

	优点	缺点
朴素贝叶斯	结构简单，分类有效，对缺失数据不敏感	条件独立假设在实际中不成立
支持向量机	高维稀疏、小样本数据集处理效果好，可解决非线性问题	对于多分类问题处理较为复杂，大规模训练样本难以实施
KNN	训练时间复杂度比SVM低，适用于样本容量较大本文集合	计算量大，样本不平衡时，对稀有类别的预测准确率低，K值的选择影响效果
决策树	易于理解和解释	处理连续型、缺失数据较为困难。
集成的方法	提高预测性能，直接级联不同模型，容易实现，参数较少	多个模型混合在一起使得预测结果难以理解，黑盒系统

4 基于深度学习的模型

1 深度神经网络（Deep Neural Networks）



- 深度神经网络可以自动从数据中学习高级特征，在语音识别、图像处理、文本理解等方面比浅层学习模型取得了更好的效果。
- 近年来，深度学习在文本分类等自然语言处理任务中的研究和应用得到了学术界的广泛关注，并且取得了一些重大的进展。

BASIC INFORMATION BASED ON DIFFERENT MODELS. TRANS: TRANSFORMER. TIME: TRAINING TIME.

Model	Year	Method	Venue	Applications	Code Link	Metrics	Datasets
ReNN	2011	RAE [45]	EMNLP	SA, QA	[46]	Accuracy	MPQA, MR, EP
	2012	MV-RNN [47]	EMNLP	SA	[48]	Accuracy, F1	MR
	2013	RNTN [49]	EMNLP	SA	[50]	Accuracy	SST
	2014	DeepRNN [51]	NIPS	SA;QA	-	Accuracy	SST-1;SST-2
MLP	2014	Paragraph-Vec [52]	ICML	SA, QA	[53]	Error Rate	SST, IMDB
	2015	DAN [54]	ACL	SA, QA	[55]	Accuracy, Time	RT, SST, IMDB
RNN	2015	Tree-LSTM [2]	ACL	SA	[56]	Accuracy	SST-1, SST-2
	2015	S-LSTM [3]	ICML	SA	-	Accuracy	SST
	2015	TextRCNN [57]	AAAI	SA, TL	[58]	$Macro$-$F1$, etc.	20NG, Fudan, ACL, SST-2
	2015	MT-LSTM [8]	EMNLP	SA,QA	[59]	Accuracy	SST-1, SST-2, QC, IMDB
	2016	oh-2LSTMp [60]	ICML	SA, TL	[61]	Error Rate	IMDB, Elec, RCV1, 20NG
	2016	BLSTM-2DCNN [62]	COLING	SA, QA, TL	[63]	Accuracy	SST-1, Subj, TREC, etc.
	2016	Multi-Task [64]	IJCAI	SA	[65]	Accuracy	SST-1, SST-2, Subj, IMDB
	2017	DeepMoji [66]	EMNLP	SA	[67]	Accuracy	SS-Twitter, SE1604, etc.
	2017	TopicRNN [68]	ICML	SA	[69]	Error Rate	IMDB
	2017	Miyato et al. [70]	ICLR	SA	[71]	Error Rate	IMDB, DBpedia, etc.
	2018	RNN-Capsule [72]	TheWebConf	SA	[73]	Accuracy	MR, SST-1, etc.
CNN	2014	TextCNN [18]	EMNLP	SA, QA	[74]	Accuracy	MR, SST-2, Subj, etc.
	2014	DCNN [7]	ACL	SA, QA	[75]	Accuracy	MR, TREC, Twitter
	2015	CharCNN [5]	NeurIPS	SA, QA, TL	[76]	Error Rate	AG, Yelp P, DBPedia, etc.
	2016	SeqTextRCNN [9]	NAACL	Dialog act	[77]	Accuracy	DSTC 4, MRDA, SwDA
	2017	XML-CNN [78]	SIGIR	NC, TL, SA	[79]	DCG@k, etc.	EUR-Lex, Wiki-30K, etc.
	2017	DPCNN [80]	ACL	SA, TL	[81]	Error Rate	AG, DBPedia, Yelp.P, etc.
	2017	KPCNN [82]	IJCAI	SA,QA,TL	-	Accuracy	Twitter, AG, Bing, etc.
	2018	TextCapsule [83]	EMNLP	SA, QA, TL	[84]	Accuracy	Subj, TREC, Reuters, etc.
	2018	HFT-CNN [85]	EMNLP	TL	[86]	$Micro$-$F1$, etc.	RCV1, Amazon670K
	2020	Bao et al. [87]	ICLR	TL	[88]	Accuracy	20NG, Reuters-2157, etc.

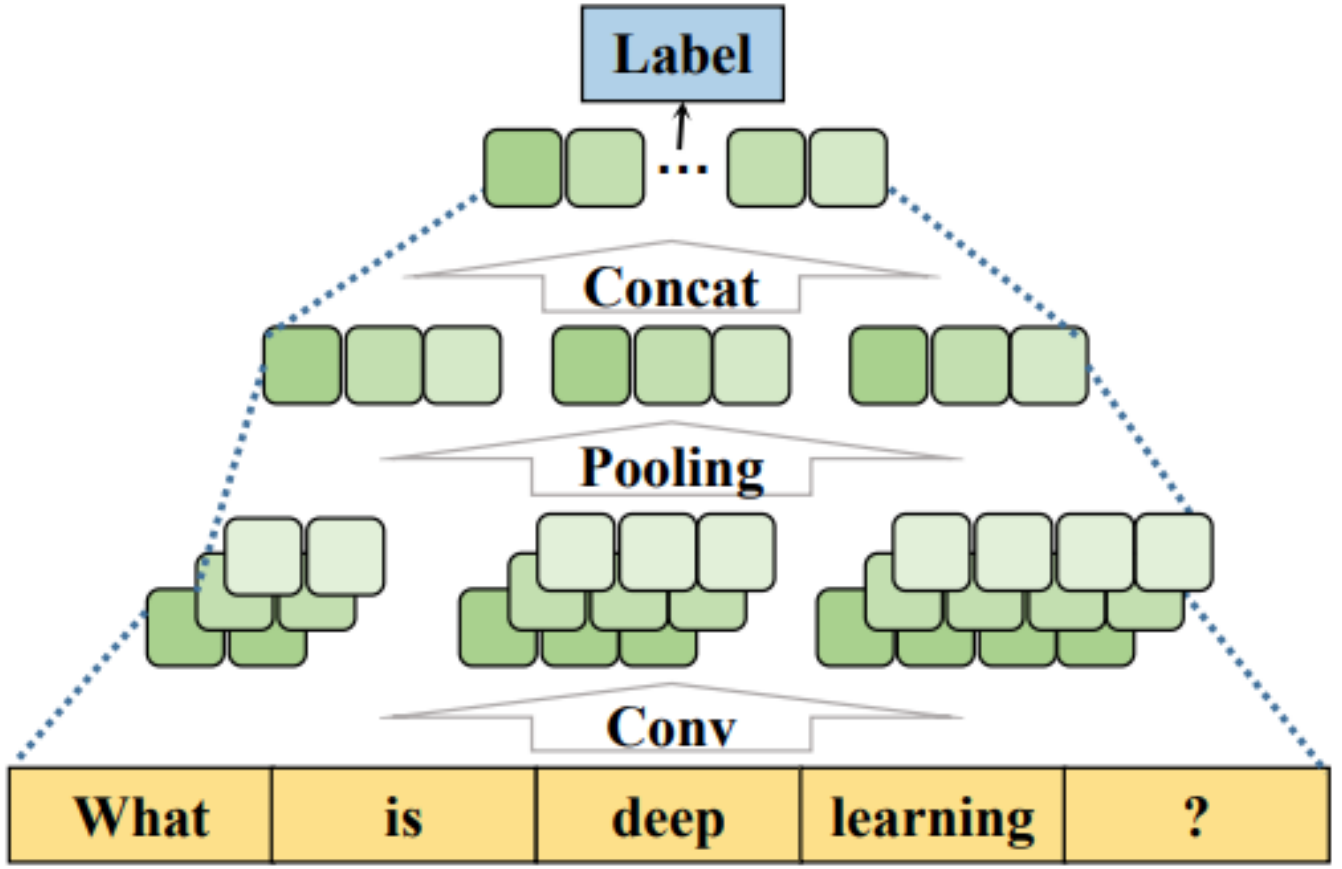


Fig. 10. The the architecture of the Convolutional neural network (CNN).

卷积神经网络(CNN)架构图

2 基于CNN的算法-TextCNN

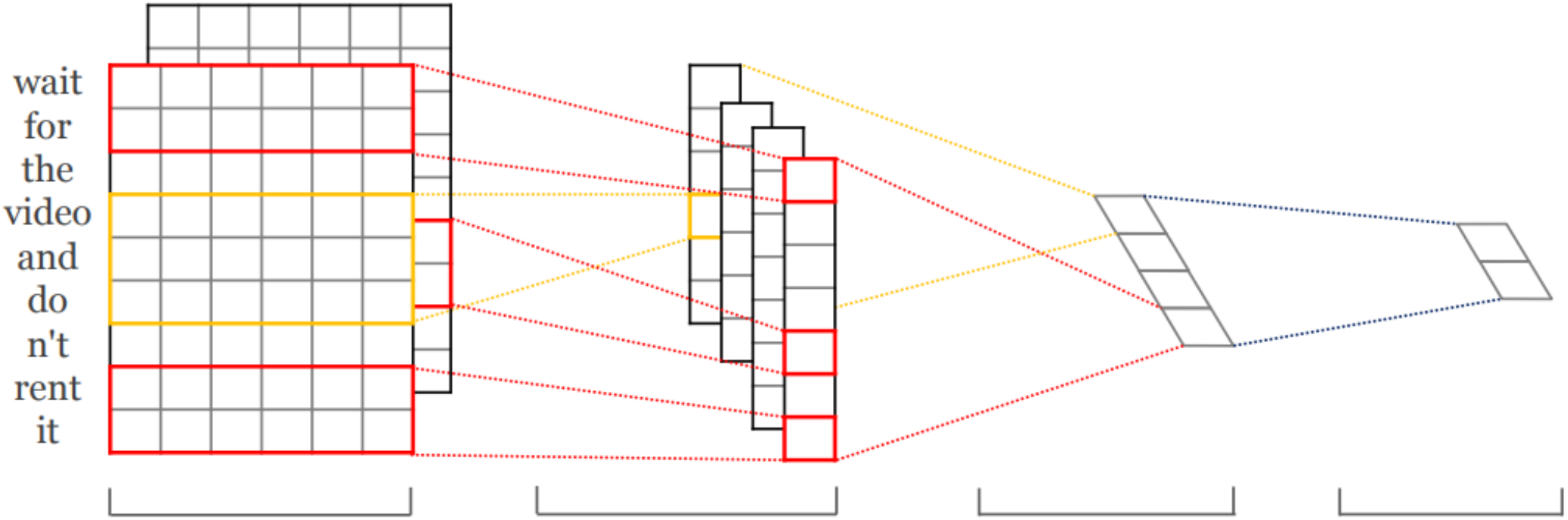
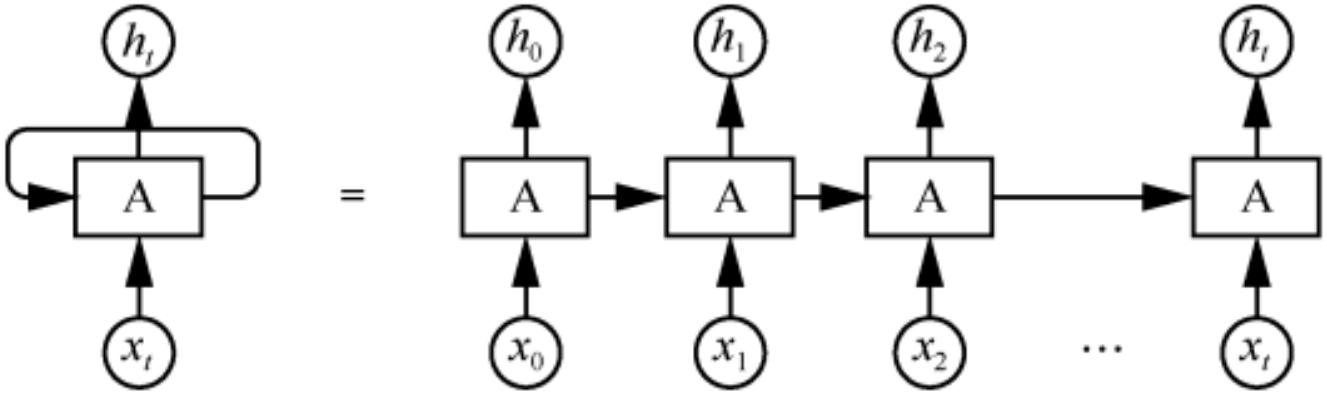


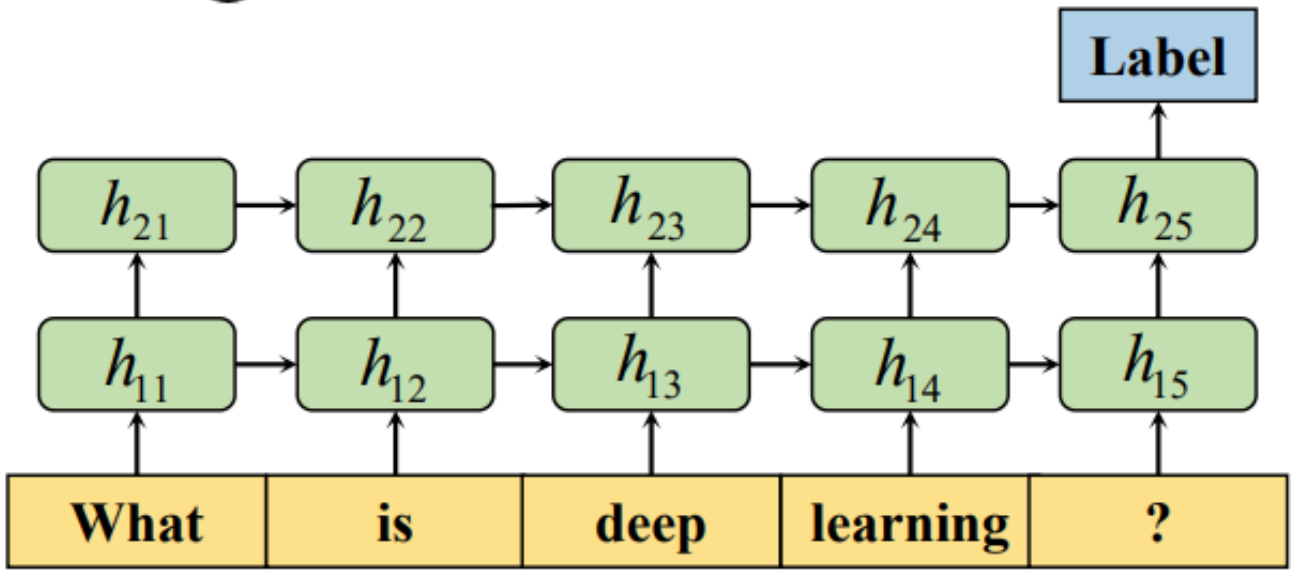
图 2 TextCNN 结构
Figure 2 TextCNN structure

KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.

基于RNN的算法

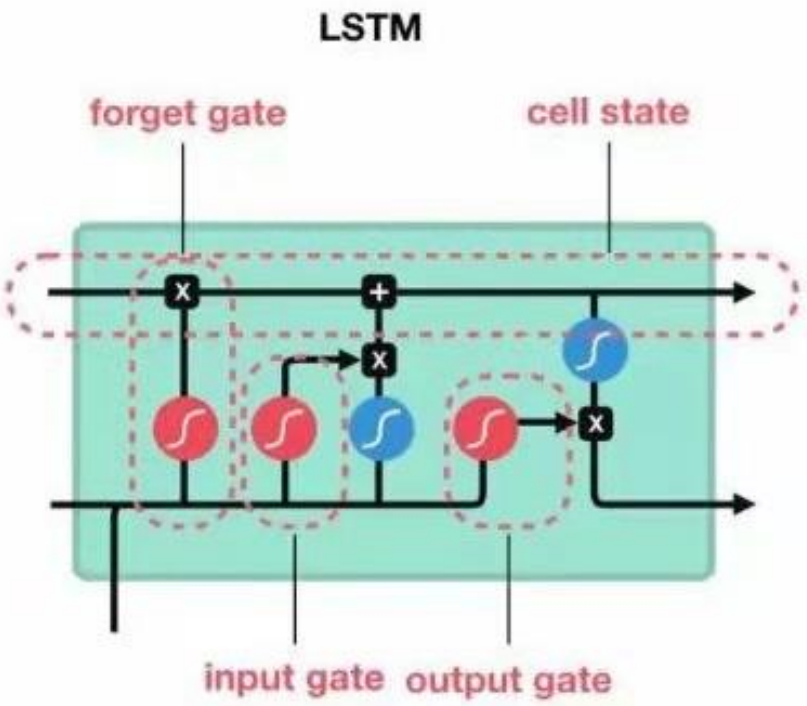


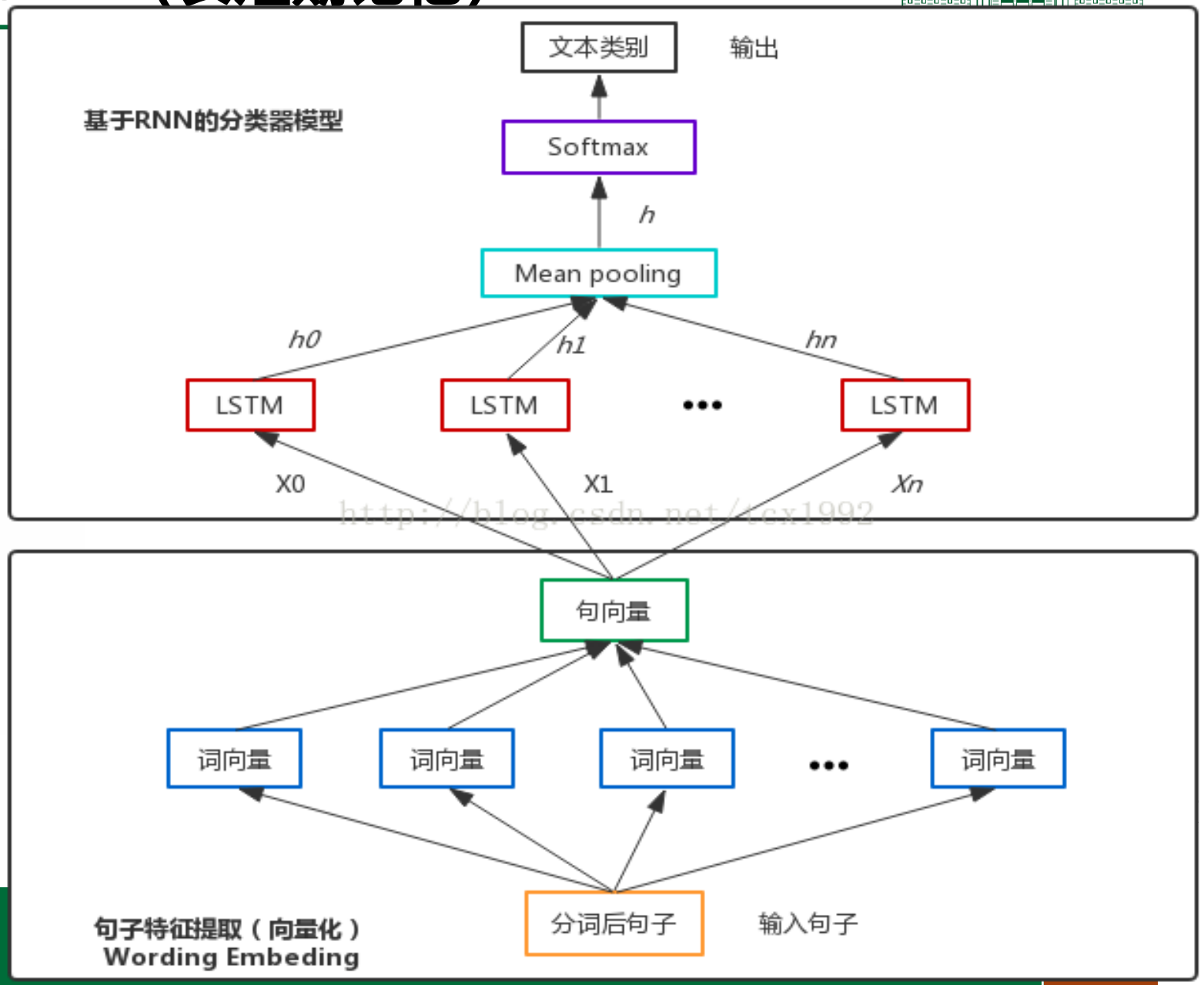
RNN结构图



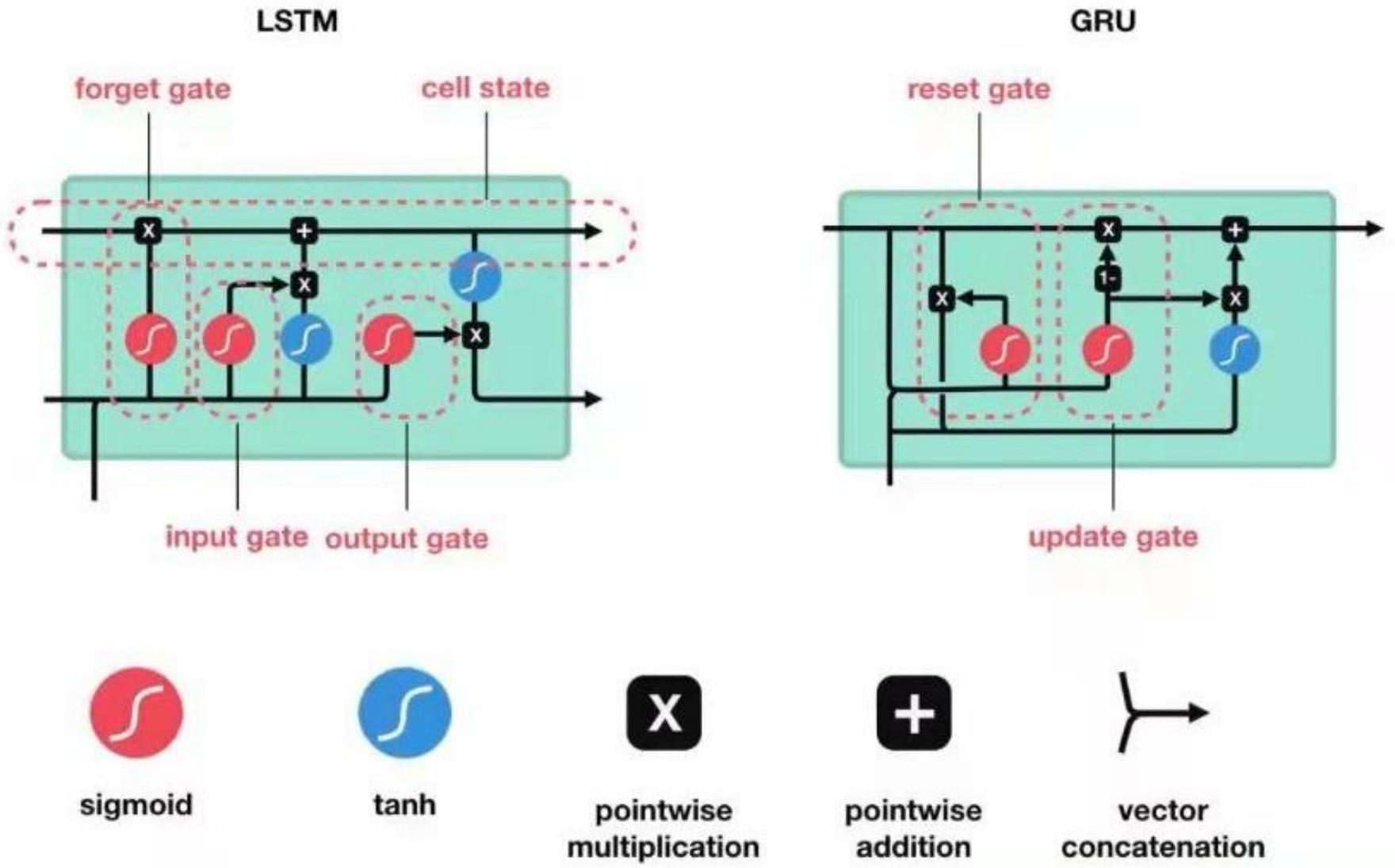
使用简单样本进行文本分类的RNN模型

3 基于RNN的算法 - LSTM（长短期记忆）





3 基于RNN的算法 - GRU（门控循环单元）



3 基于RNN的算法 - textRNN

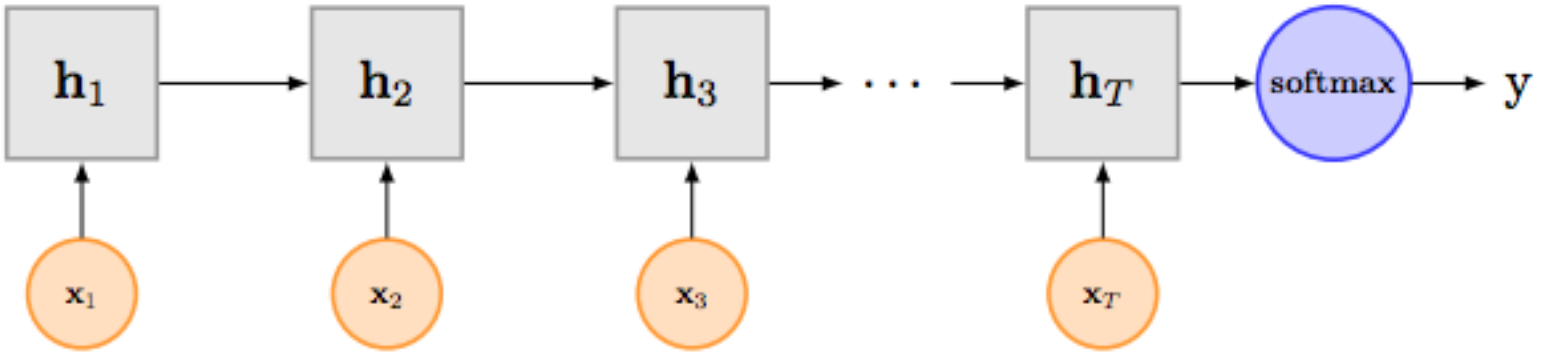


Figure 1: Recurrent Neural Network for Classification

P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning" arXiv preprint arXiv:1605.05101,2016.

4 CNN与RNN的比较

相同点

- 传统神经网络的扩展。
- 前向计算产生结果，反向计算模型更新。
- 每层神经网络横向可以多个神经元共存,纵向可以有多层神经网络连接。

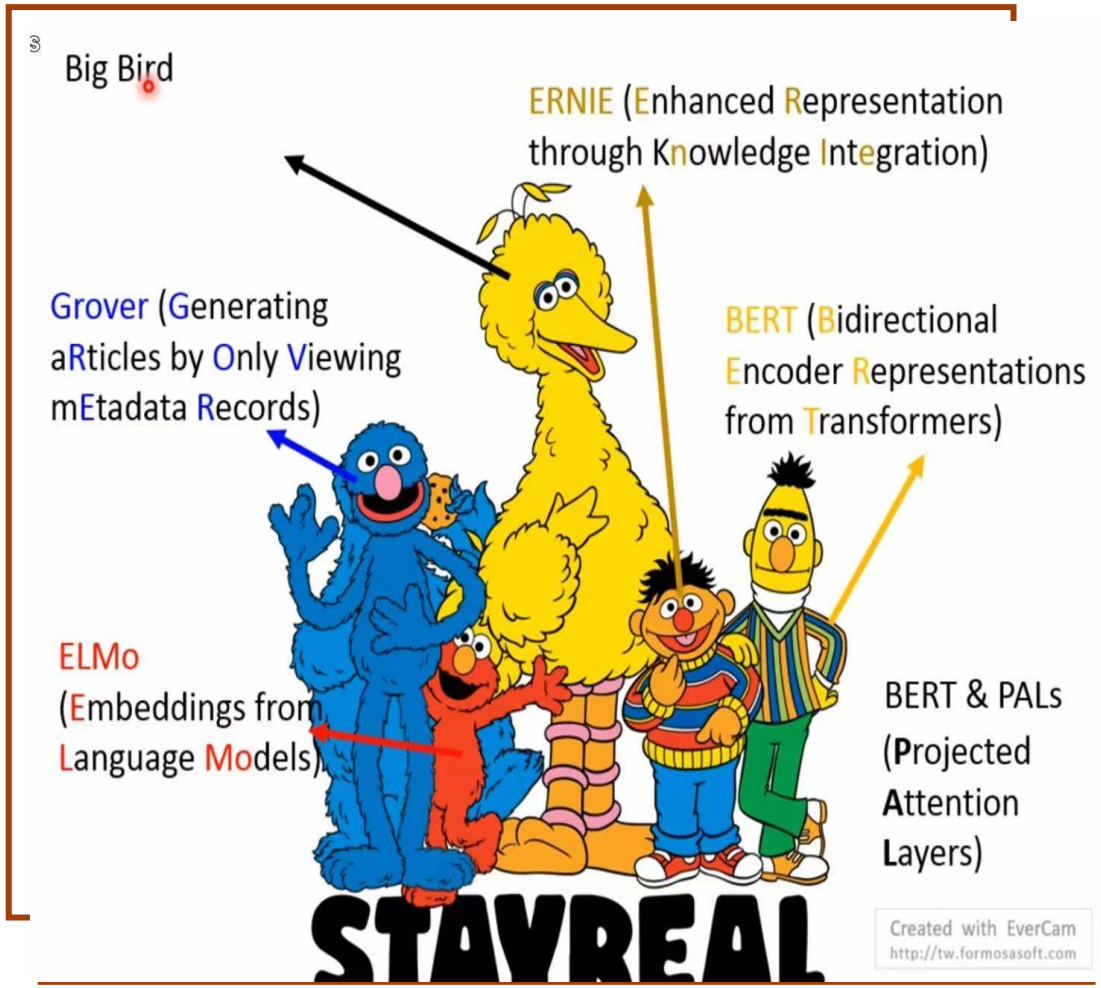
不同点

- NCNN空间扩展，神经元与特征卷积；
- RNN时间扩展，神经元与多个时间输出计算。
- CNN用于静态输出；RNNNN可以用于描述时间上连续状态的输出，有记忆功能。
- CNN高级100+深度；RNN深度有限。

4+ 预训练语言模型

介绍+对比

1 预训练模型（pre-trained model）



- 预训练模型的优点：
  - 获取的词向量是动态的；
  - 基于上下文的；
  - 拥有更多的语义信息；
  - 对下游任务友好；
- 预训练模型举例：
  - GPT
  - ElMo
  - Bert
  - XLNet
  - ERNIE

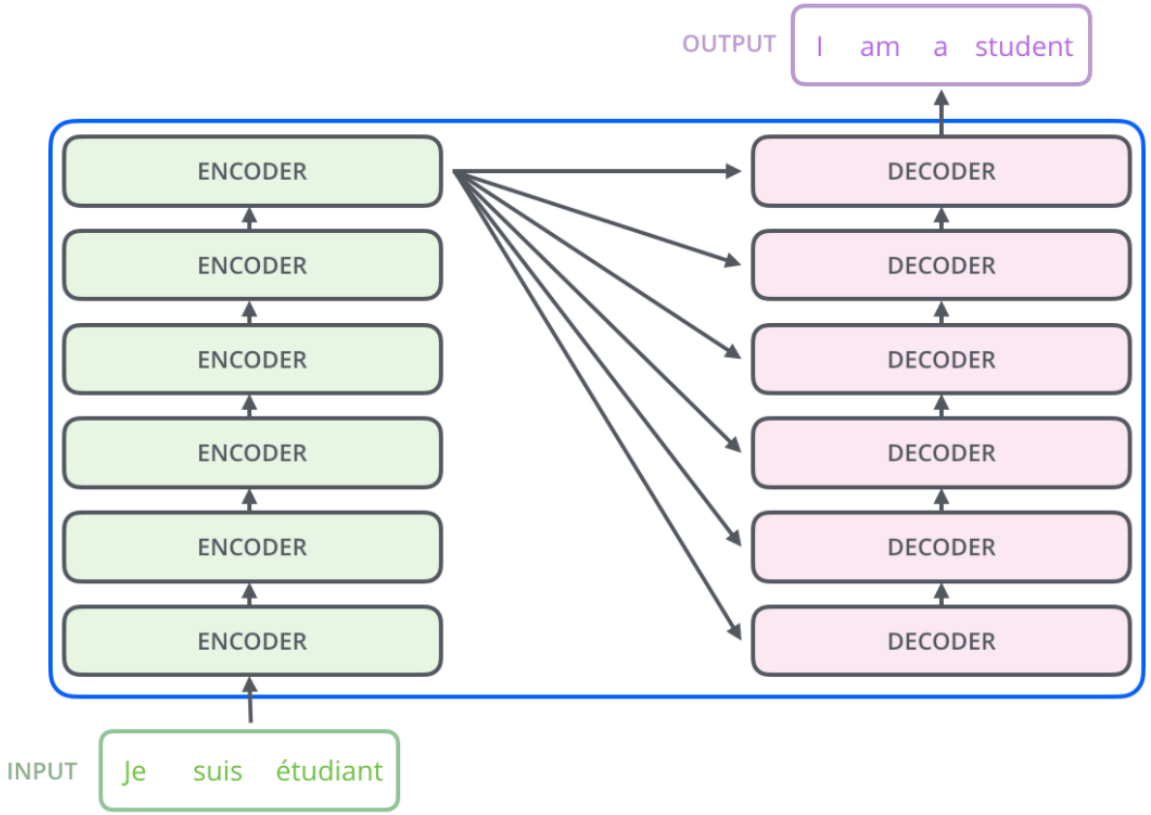
预训练模型分类

自回归语言模型

- 定义：单向预测
- 代表：ELMo/GPT1.0/GPT2.0/XLNet
- 优点：对文本序列的联合概率进行建模，天然适用于生成类NLP
- 缺点：无法获得包含上下文的双向表征

自编码语言模型

- 定义：mask单词，然后对其预测
- 代表：Bert/Bert-Based model
- 优点：利用了上下文信息得到双向表征
- 缺点：引入了独立性假设（Mask），产生联合概率偏差



Transformer的优缺点：

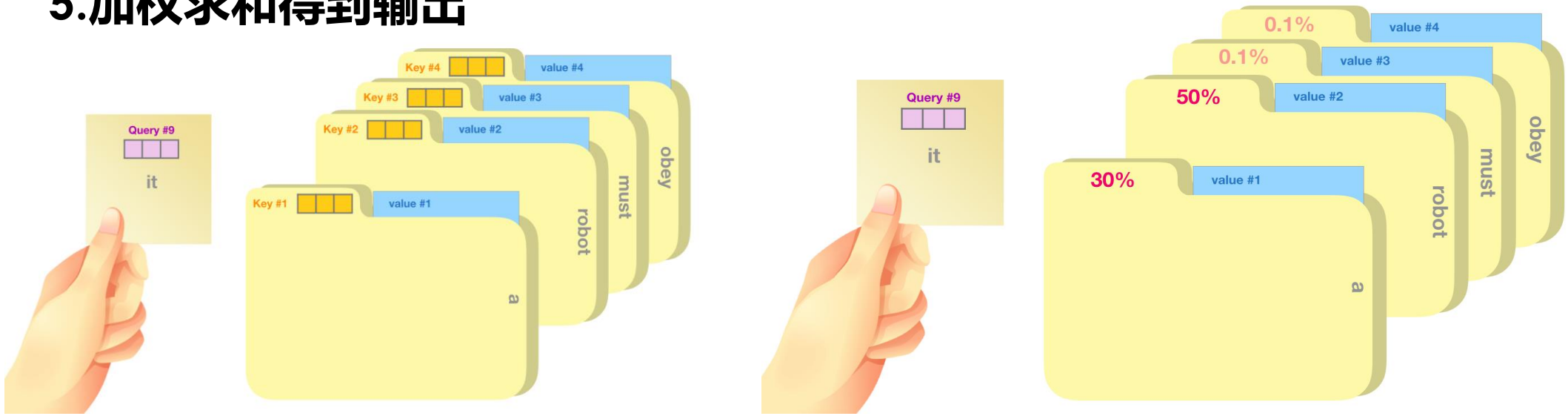
- 解决长序列信息捕获问题；
- 解决梯度相关相关问题
- 模型基本都特大
- self-attention机制可能过分关注序列信息而忽略了局部信息

2.1self-attention

The animal didn't cross the street because it was too tired

计算方式：

1. Word Embedding+Position Embedding
2.获得各自的Q、K、V向量
3.计算score（Q、K点积）
4.计算相似度（scale+softmax）
5.加权求和得到输出

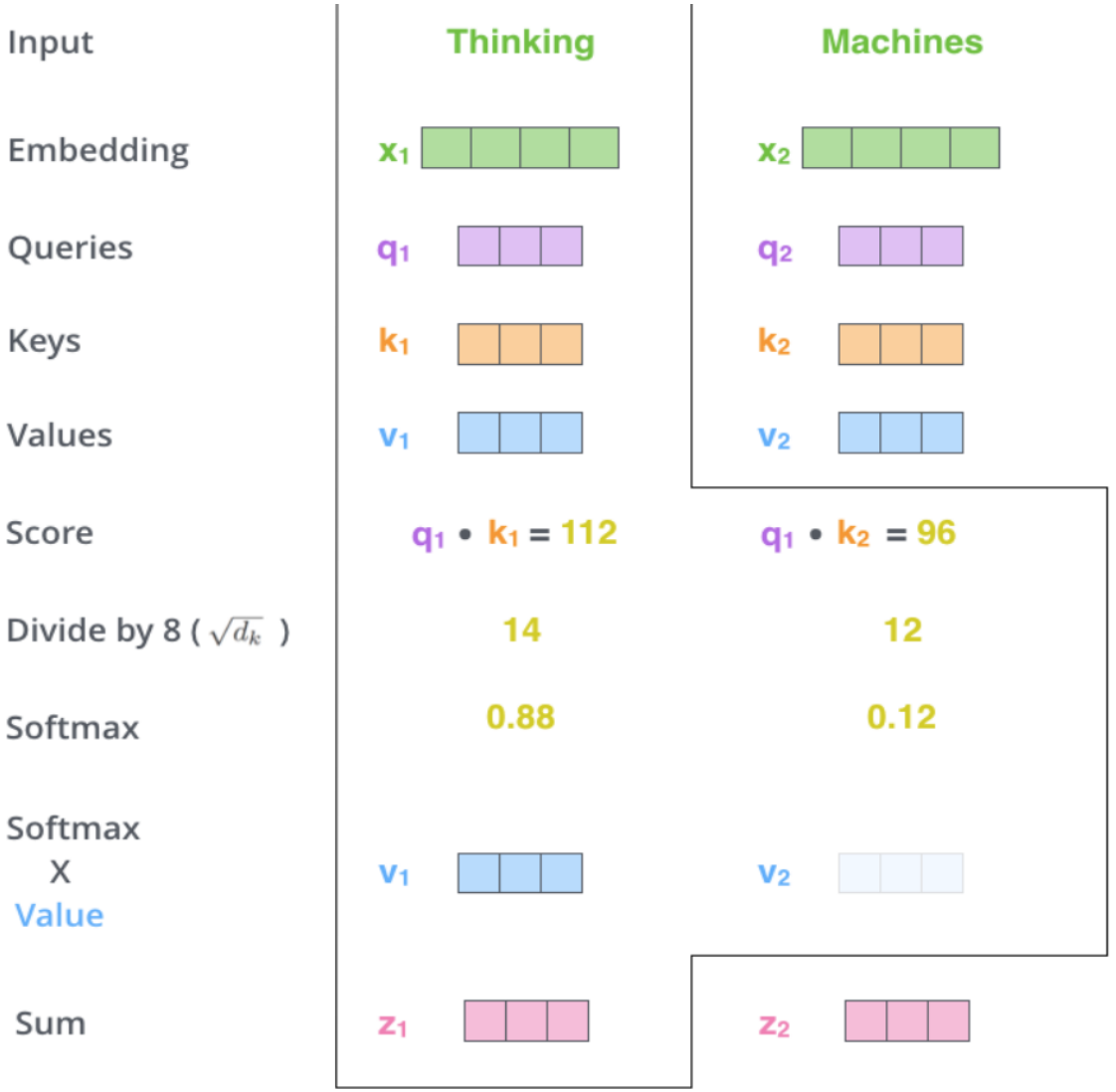


self-attention计算过程

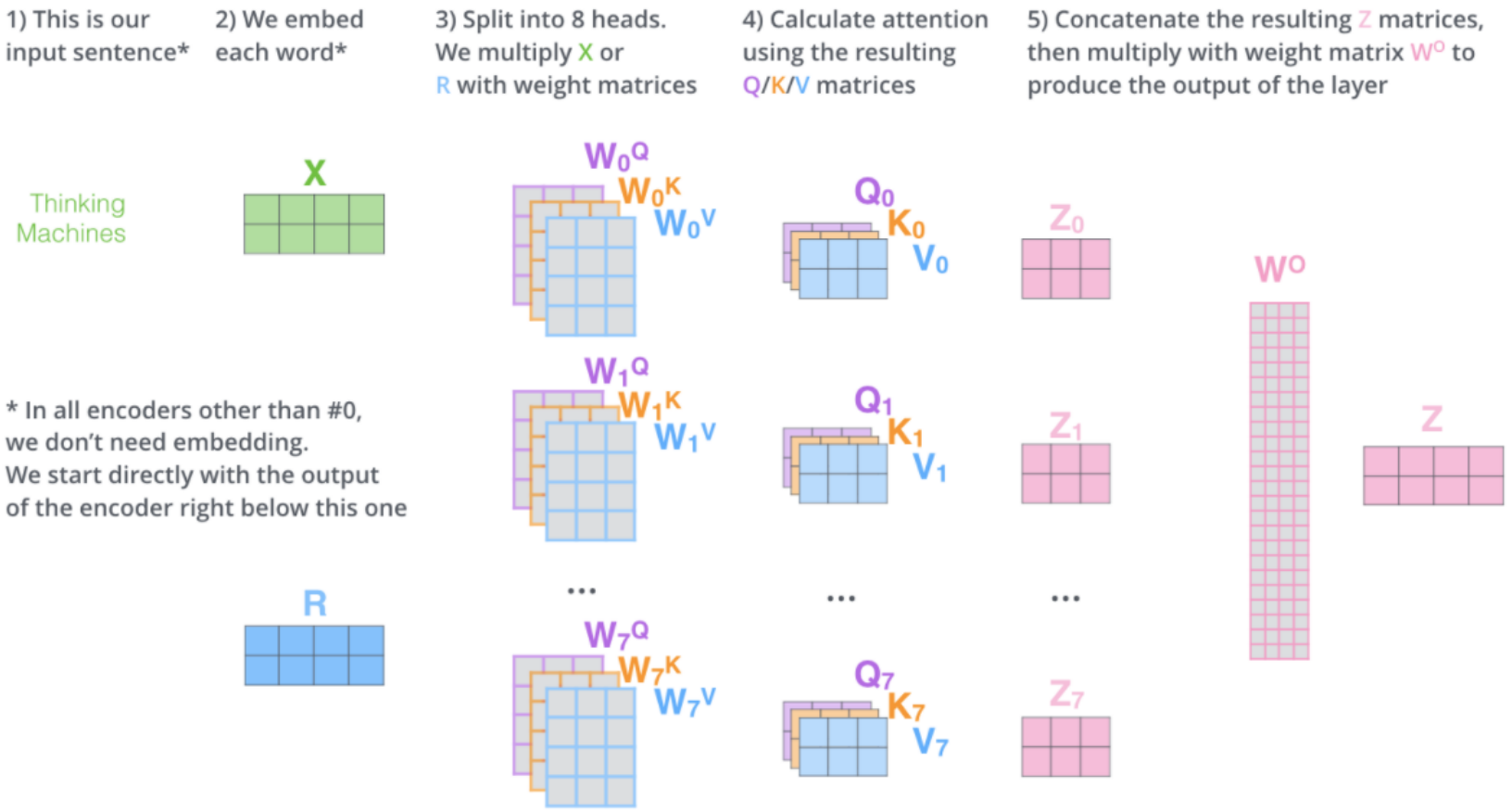
问题:

- 可学习机制少
- 表达能力有限

输出将作为FNN的输入，以进一步获取更加高级的特征



2.2多头注意力机制

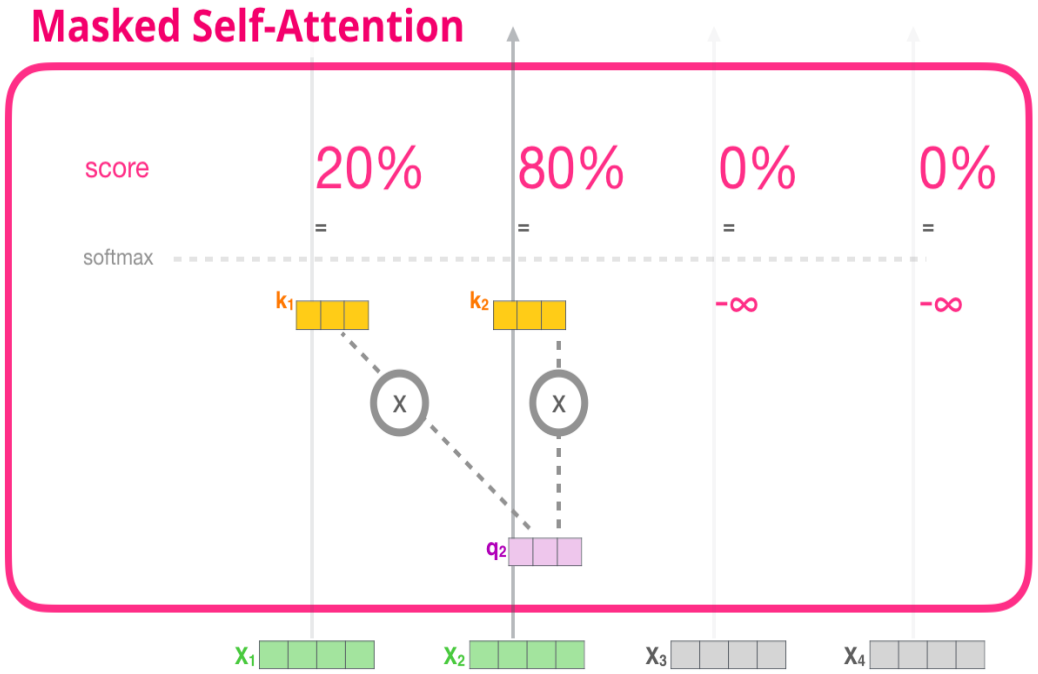


优点:

- 赋予了模型更多的子空间表达
- 扩展了模型关注其他位置的能力

split

2.3 Edcoder-side



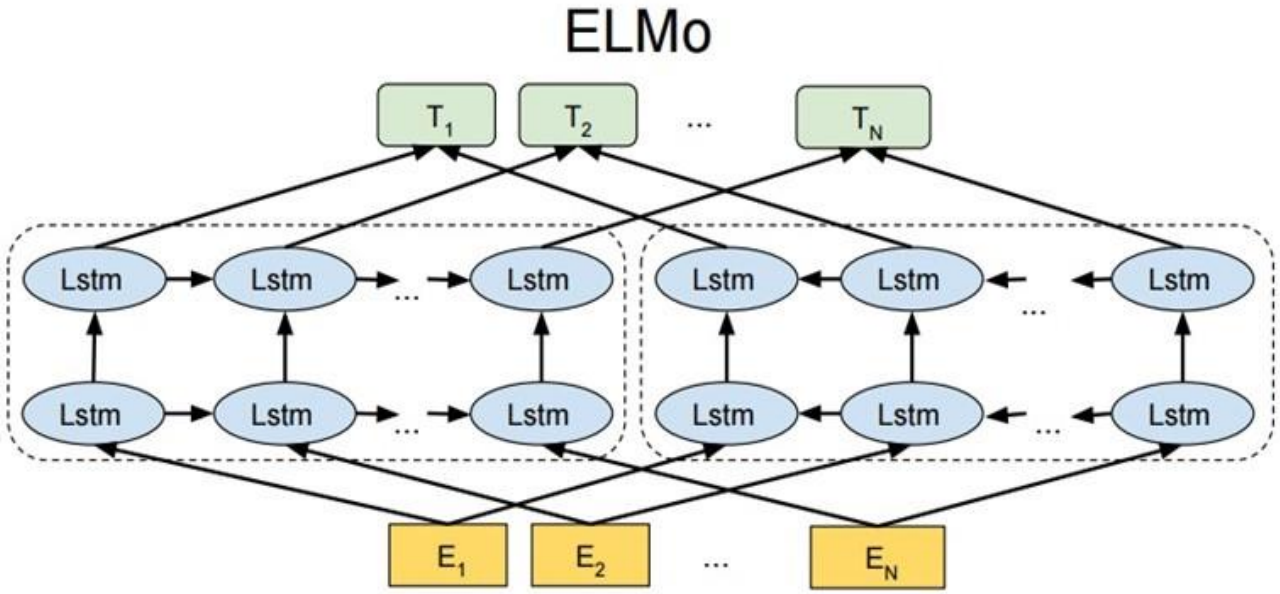
与Encoder区别

- Encoder-Decoder Attention层
- self-attention计算不同（需要mask后方token）
- K，V源于Encoder的顶层

Linear+Softmax

- 将解码向量通过线性层（/NN），输出与词典大小的向量；
- 再经过softmax得到的值，相当于预测的下一个词的概率

ElMo

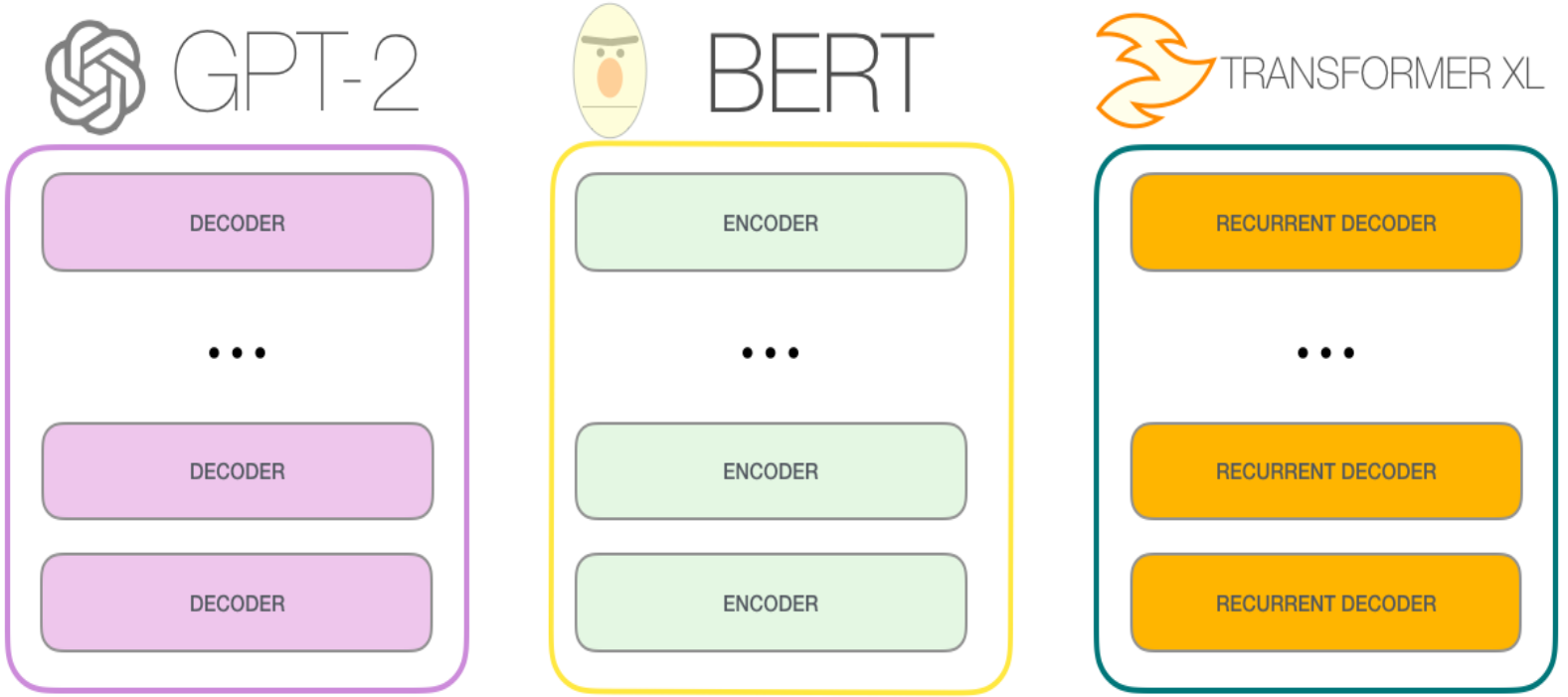


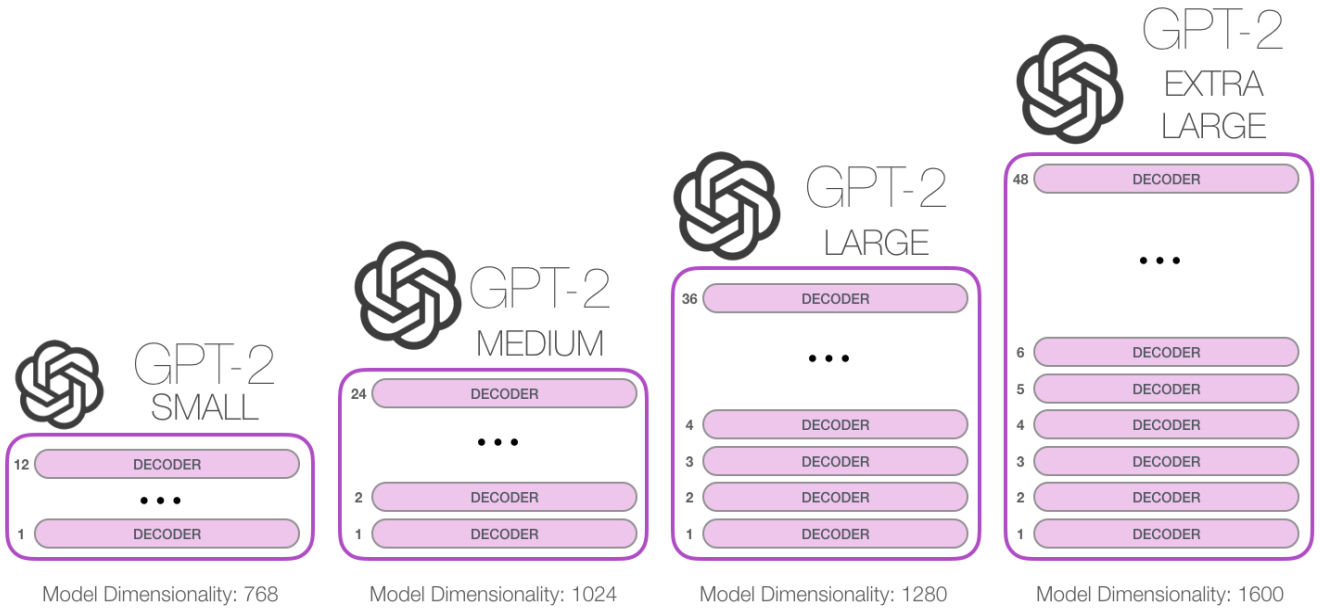
问题：

- 仅仅是单向处理的叠加
- 模型不能深度编码
- LSTM的特征能力较差

基本单元：Bi-LSTM

Transformer-Based model





基本结构：

Trans-Decoder

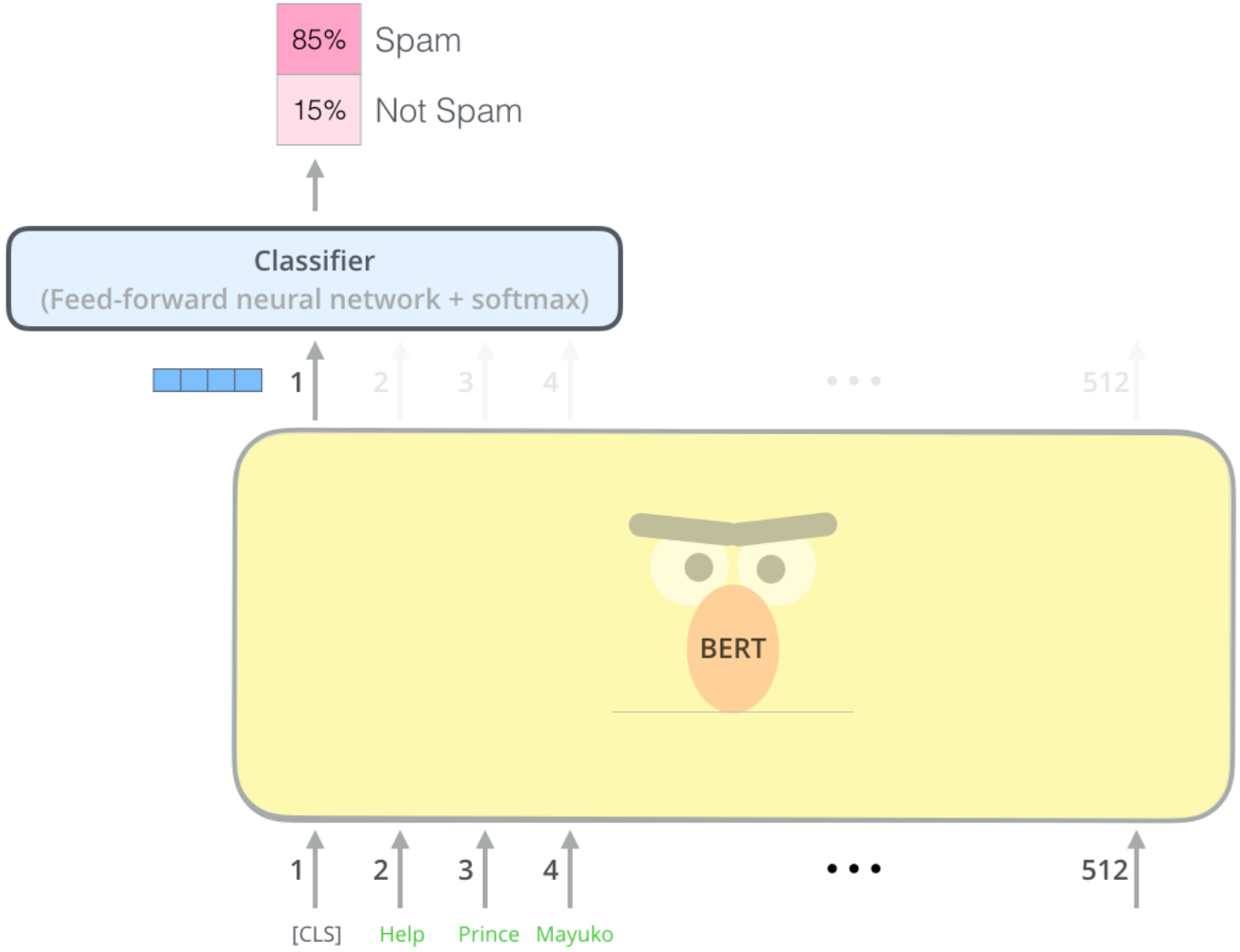
优点：

- 特征提取能力更强的LSTM
- 更深的结构

Text Classification

ELMo和GPT输出向量作为下游任务输入，加上分类器和Softmax层即可实现文本分类

4.2Bert


85%
Spam
15%
Not Spam
Classifier
(Feed-forward neural network + softmax)
1
2
3
4
512
BERT
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

4.3 XLNet

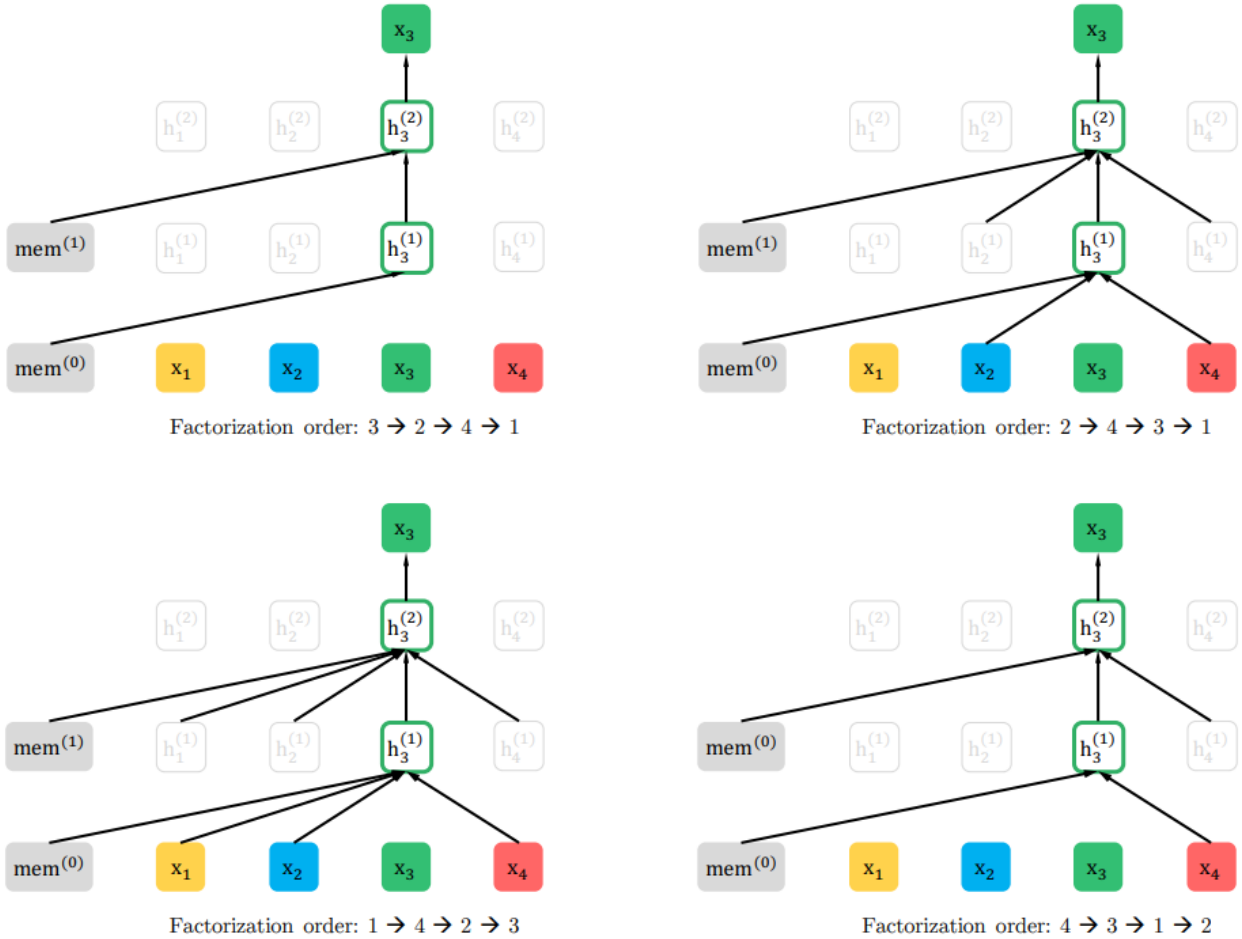


Figure 4: Illustration of the permutation language modeling objective for predicting x_3 given the same input sequence x but with different factorization orders.

排列语言模型

- 固定预测位置不变
- 使用随机序列单向预测

4 文本分类前沿进展

模型、数据

这是一个过渡页

最近的研究主要集中在

- 模型改进
- 文本数据增强
- 可解释性
- 多任务文本分类等

AAAI接收发表的文本分类的论文较多，2021年有11篇

CNN的的优缺点：

- 卷积可以捕获局部信息——n-gram
- 池化提取去全局信息；
- 池化会损失掉位置信息
- 长序列信息捕获问题；

Attention+CNN?

Transformer的优缺点：

- 解决长序列信息捕获问题；
- 解决梯度相关相关问题
- 模型基本都特大
- self-attention机制可能过分关注序列信息而忽略了局部信息

5 Attentive Convolutional Transformer

变与不变:

- 保留多头注意力机制
- 卷积注意力机制——卷积捕获n-gram信息
- 卷积核空间（convolutional filter space）——将transformer的token级别的信息提升到了短语级别
- 全局注意力机制——结合全局、局部、位置信息

主要思想:

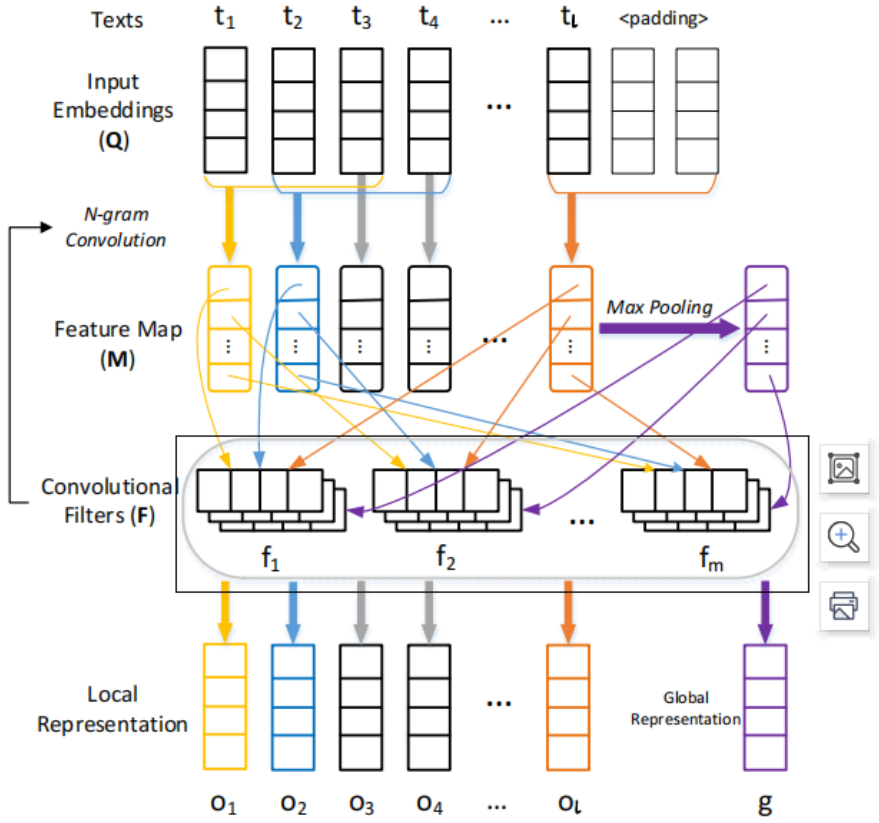
- 卷积核可以捕获n-gram信息，并且自身也嵌入了n-gram信息

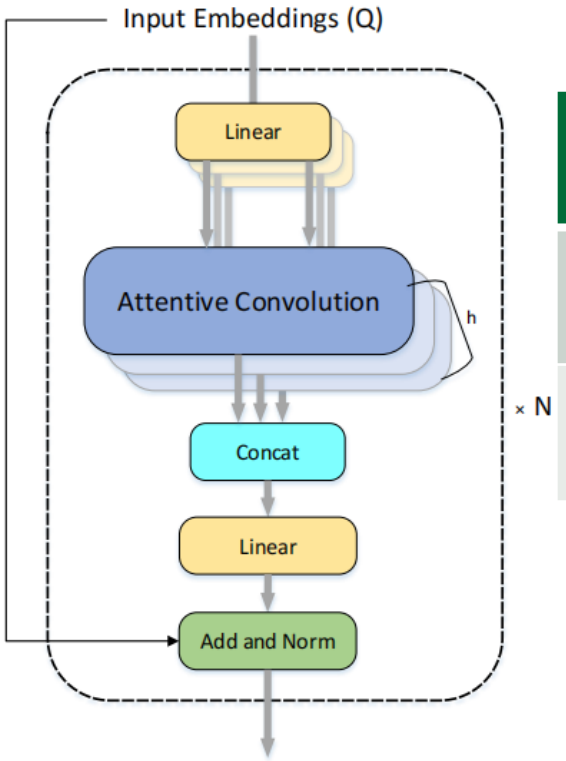
结论:

- 比Transformer、CNN、RNN更好
- 模型是Trans的1/2

论文地址 2021 AAAI

5.1 attentive convolution mechanism





与普通attention对照：

	Attention	ACM
权重	Q*K	M(特征图)
值	V	F（卷积核）

卷积核学习到了n-gram语义信息?

Figure 1: Left: attentive convolution mechanism. Outputs are obtained by combining convolutional filters attentively utilizing feature map as attention weights. Right: multi-head multi-layer structure of ACT. h and N indicate number of attentive convolution heads and layers respectively.

5.1global attention mechanism

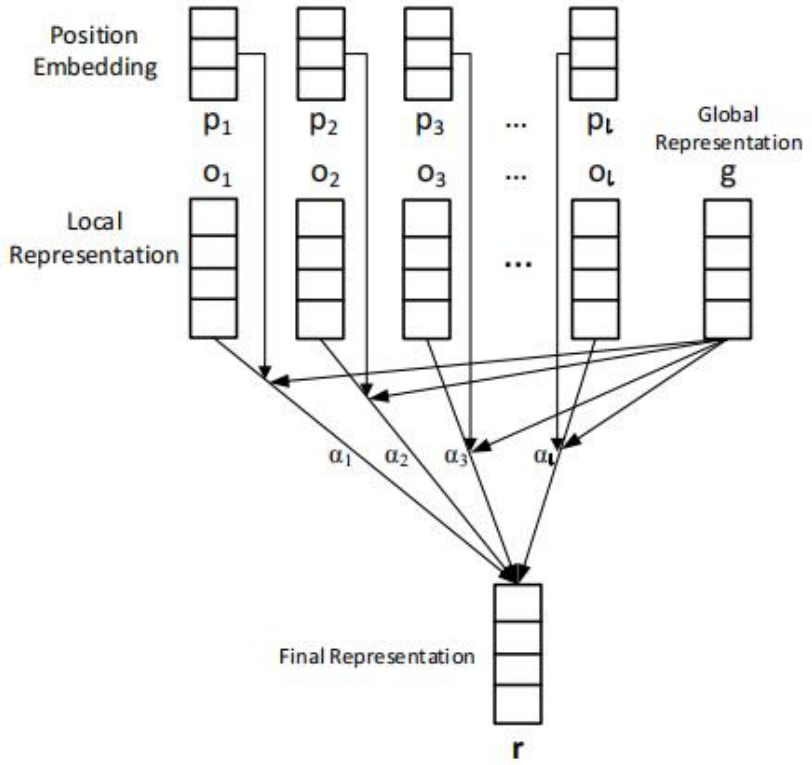


Figure 2: Global attention mechanism. Attention weights α_i are calculated based on local representation $\mathbf{o}_i$, global representation $\mathbf{g}$, and position embedding $\mathbf{p}_i$ of each token.

与普通attention对照：

	Attention	GAM
权重	Q*K	f(O,P,g)
值	V	O

O:局部信息
P：位置信息
g:全局信息

把Attention作为一种工具，基于主要信息，用于嵌入想要的信息

More Model

之前的模型的共同点：

- 均在文本本身也就是文本空间进行不断改进

Text space+?

如:

- 文本本身的统计信息？ （统计）
- 是否可以模拟人类的认知过程？ （认知科学）
- 符号神经网络（符号主义+NN）

启示

- 融合传统和深度的优点；
- 模拟人类认知才能更加接近智能
- 无论什么方法，用作工具来模拟人类思维过程

AAAI 2021

6.1AGN(Adaptive Gate Network)

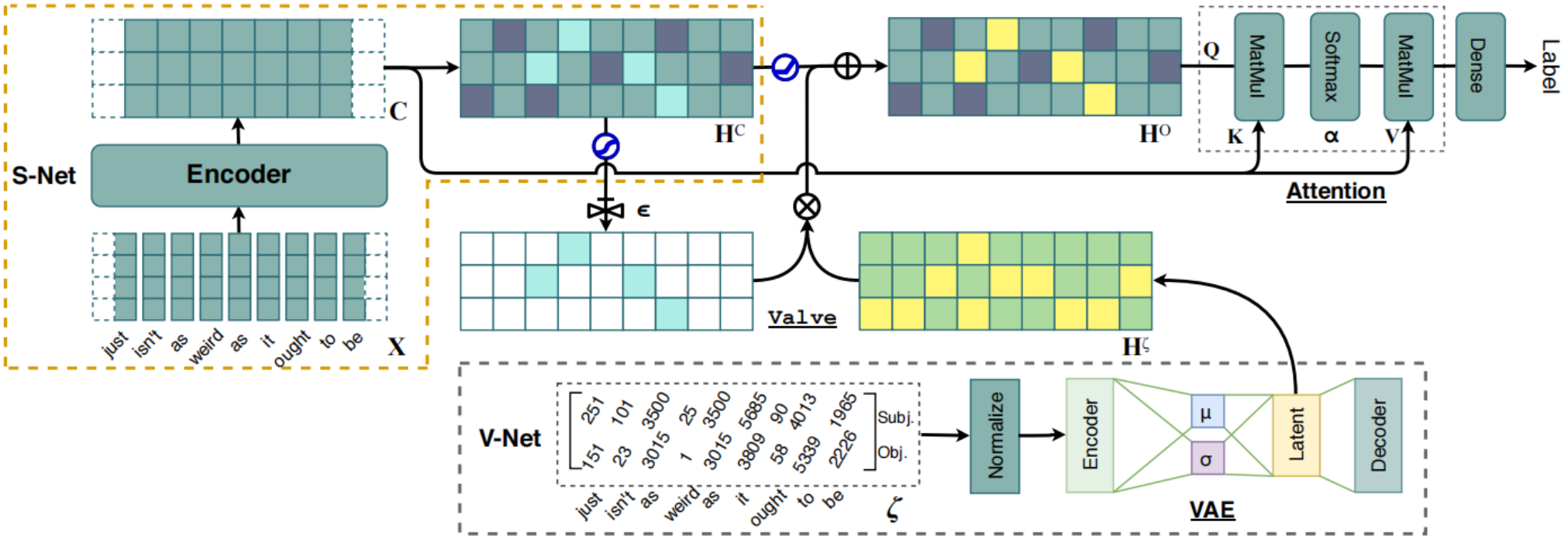


Figure 1: The generic framework of the proposed AGN. The *subj.* and *obj.* are labels of the Subj dataset.

- V-Net：变分编码器用于提取统计信息的全局表示
- S-Net：语义表征映射网络-使用Bert、RNN等特征提取器
- Valve:自适应门机制——选择性加入统计信息

2021 AAAI

V-Net

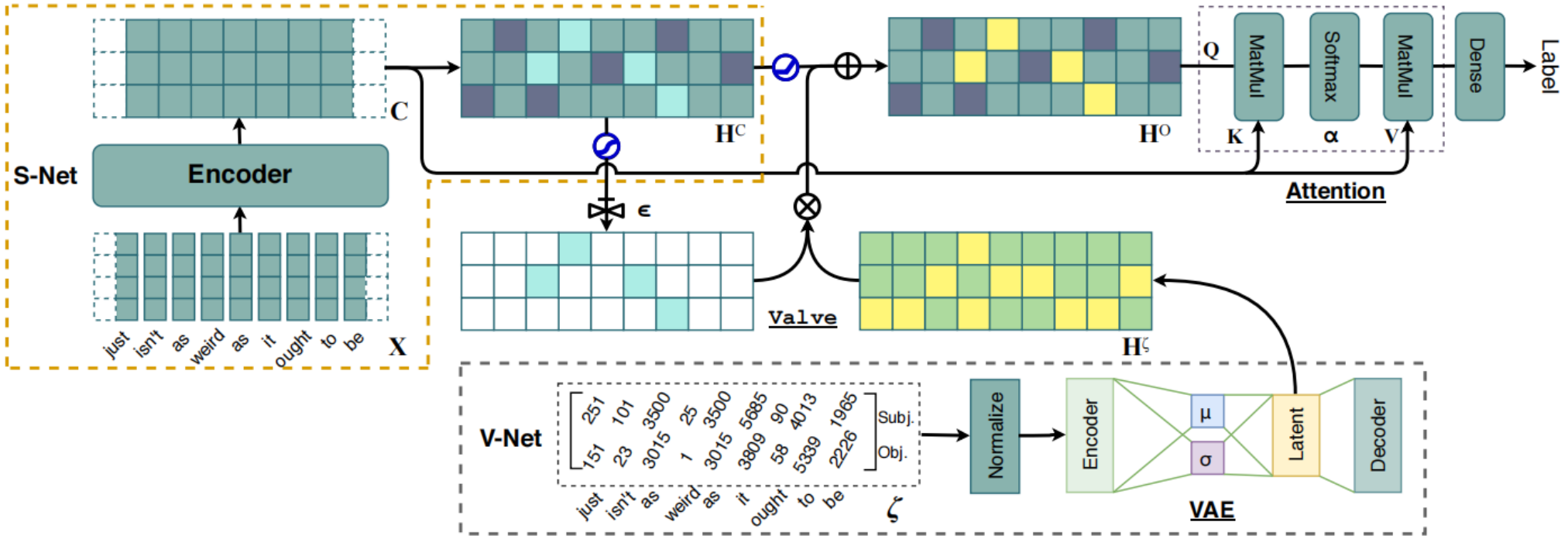


Figure 1: The generic framework of the proposed AGN. The *subj.* and *obj.* are labels of the Subj dataset.

TCoL向量 —VAE→ 连续稠密，适应文本空间的向量

S-Net

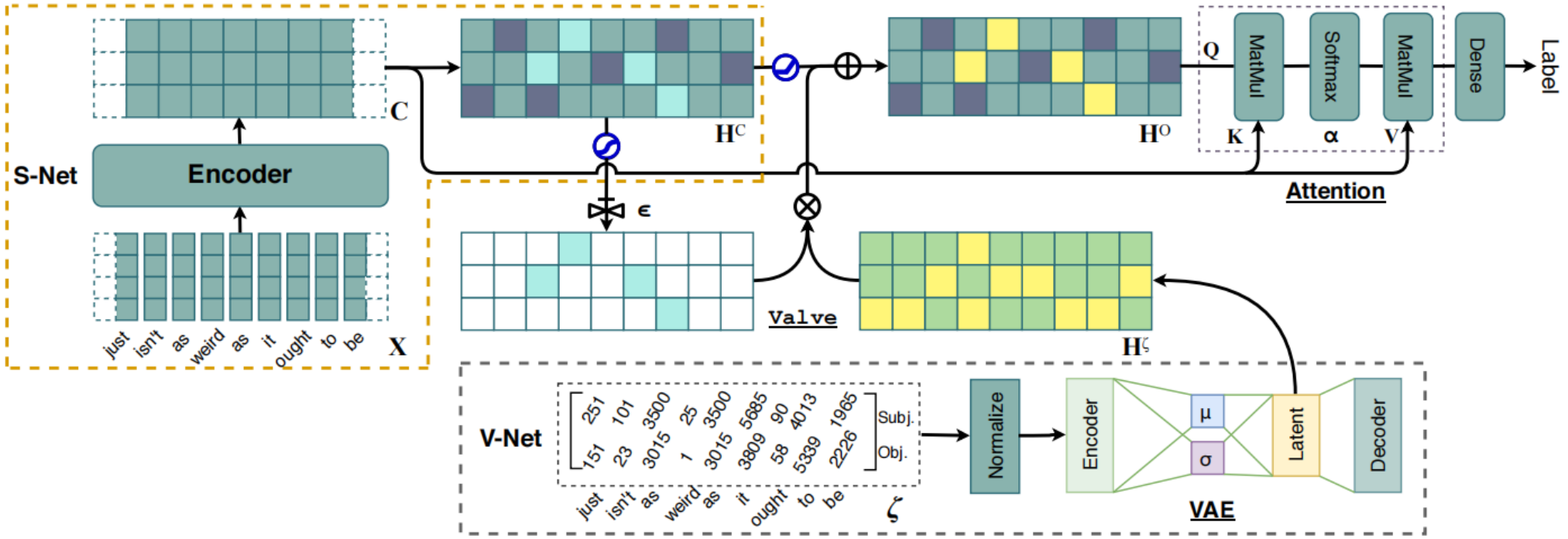


Figure 1: The generic framework of the proposed AGN. The *subj.* and *obj.* are labels of the Subj dataset.

V-Net:提取文本信息，并通过一个NN+softmax得到特征的置信度评估

Valve

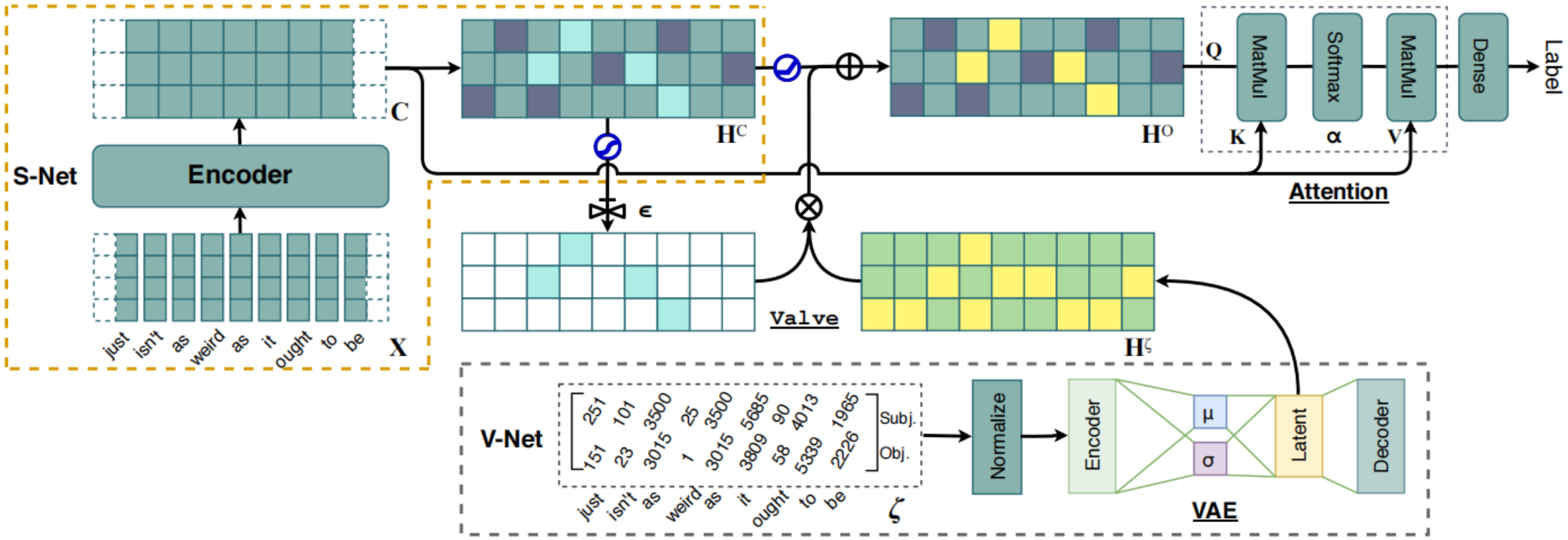


Figure 1: The generic framework of the proposed AGN. The *subj.* and *obj.* are labels of the Subj dataset.

基于AdaGate函数和置信度，将语义信息和统计信息自适应地融合

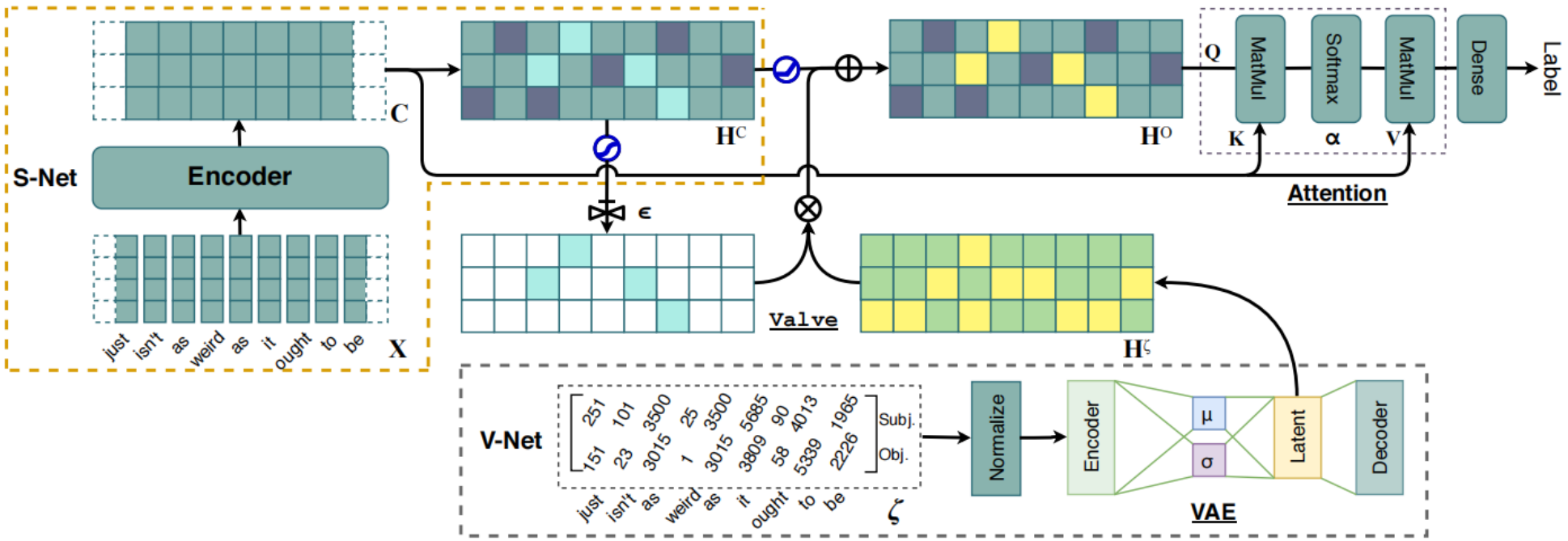
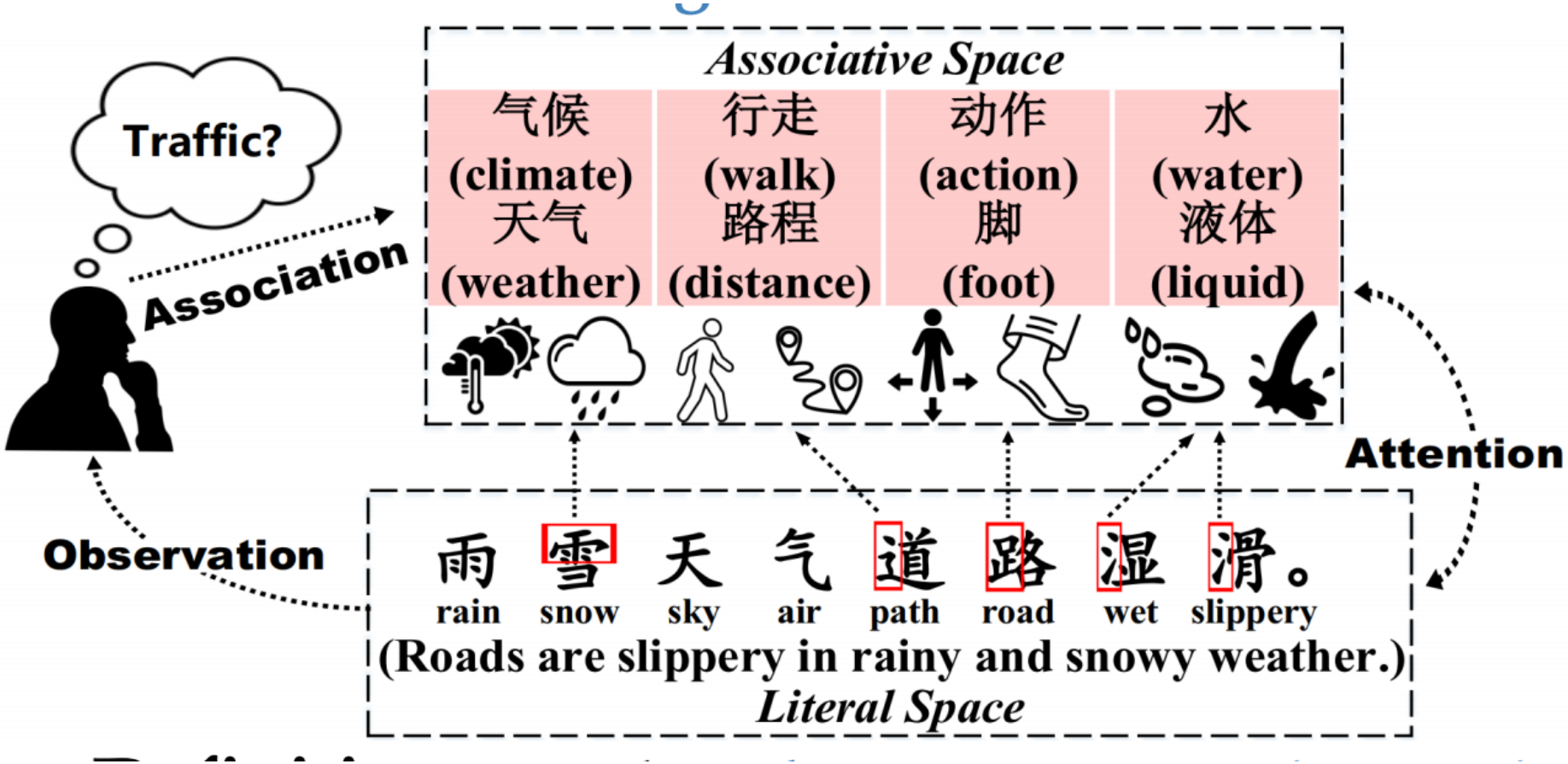


Figure 1: The generic framework of the proposed AGN. The *subj.* and *obj.* are labels of the Subj dataset.

$$Attention(\mathbf{H}^O, \mathbf{C}) = \text{softmax}(\mathbf{H}^O\mathbf{C}^\intercal)\mathbf{C}. \quad (12)$$

Note that if we reject all statistical information (i.e., $\epsilon = 0$), Eqn. (12) will become self-attention (Vaswani et al. 2017) as $\mathbf{H}^O = \mathbf{C}$.

6.2RAM(Radical-guided Associative Model)

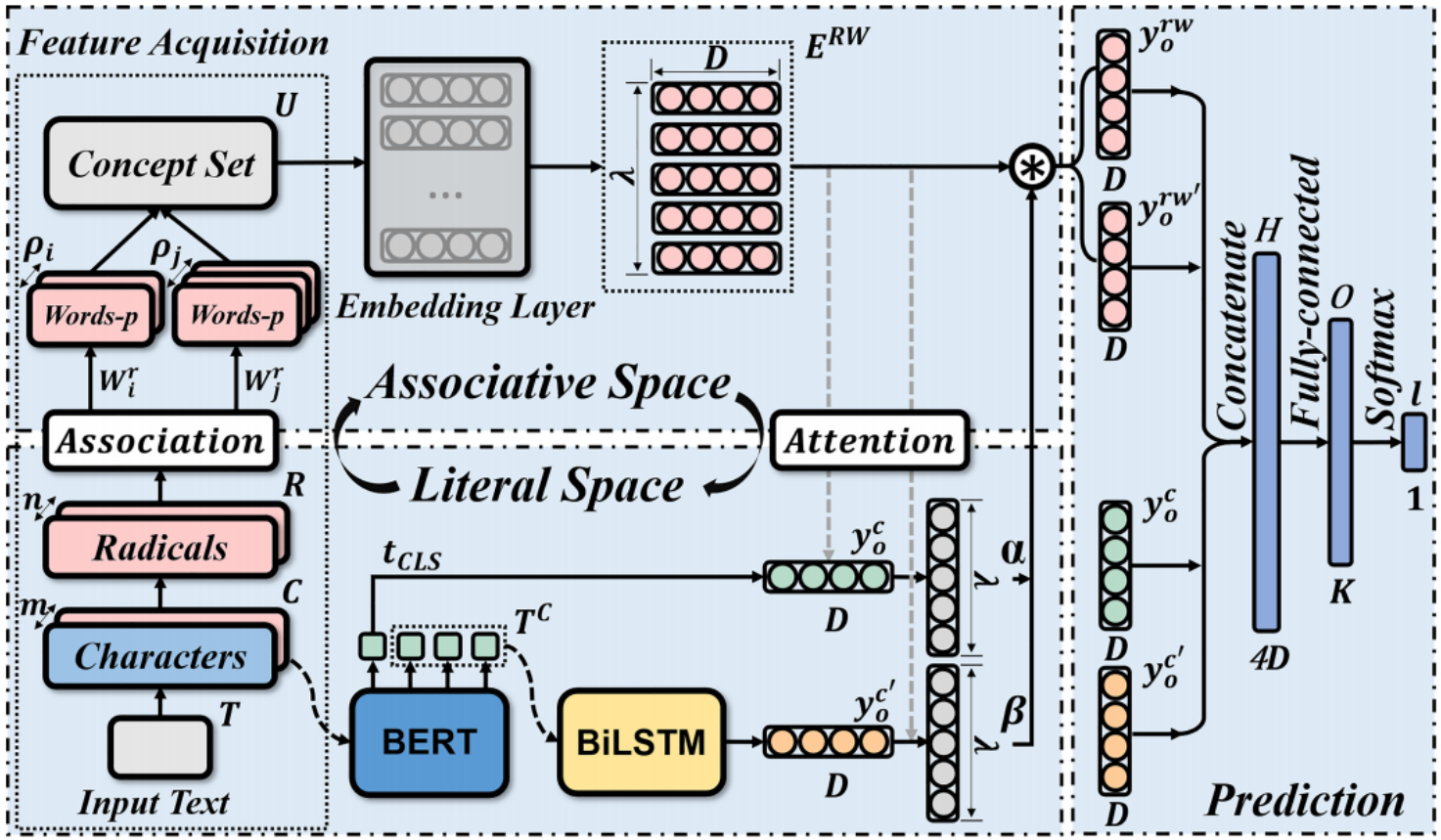


联想
帮助人类理解文本

结论:
比Bert等效果更好

2021 AAAI

基本结构



- **Feature Acquisition**——特征提取（字符级别）
- **Literal Space Modeling**——提取文本空间特征
- **Associative Space Modeling**——获取联想空间的表示
- **Prediction**——融合特征

Figure 2: The overall architecture of our Radical-guided Associative Model (RAM).

Feature Acquistion

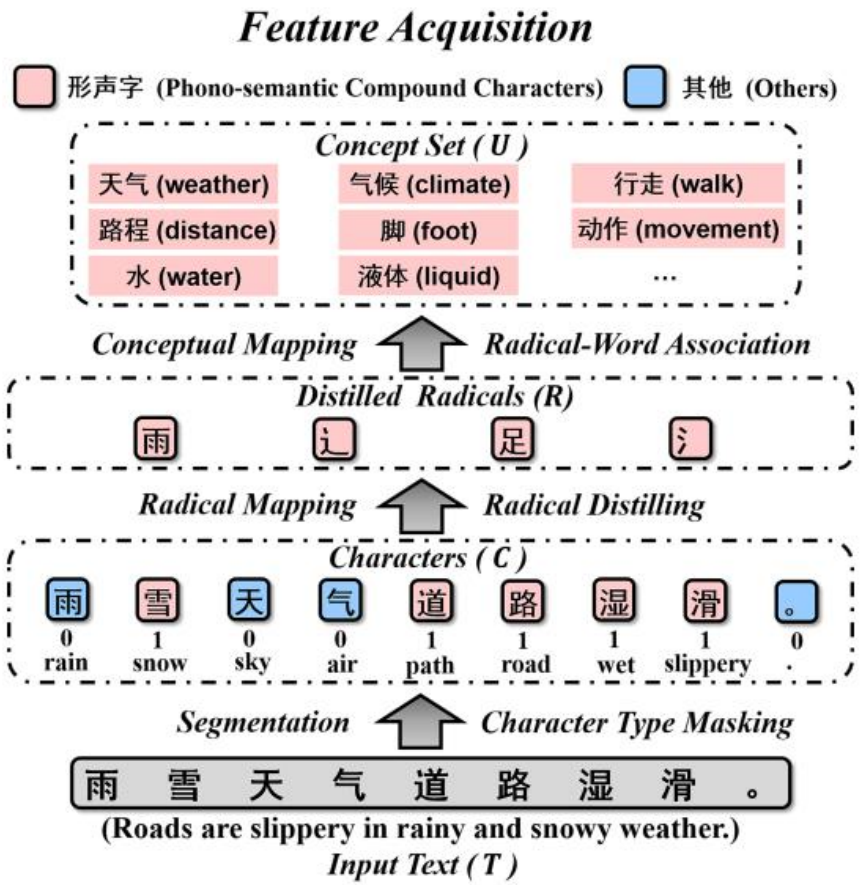


Figure 3: An intuitive illustration of the *Feature Acquisition* process for Chinese text.

1. 分割为单独的字
2. 根据字典，标注mask掉纯词根的字（偏旁、部首）
3. 蒸馏：将为Mask的字，基于字典查找词根（部首**radical**）
4. 联想：基于字典和词根，查找一定数量的相关词

Literal Space Modeling

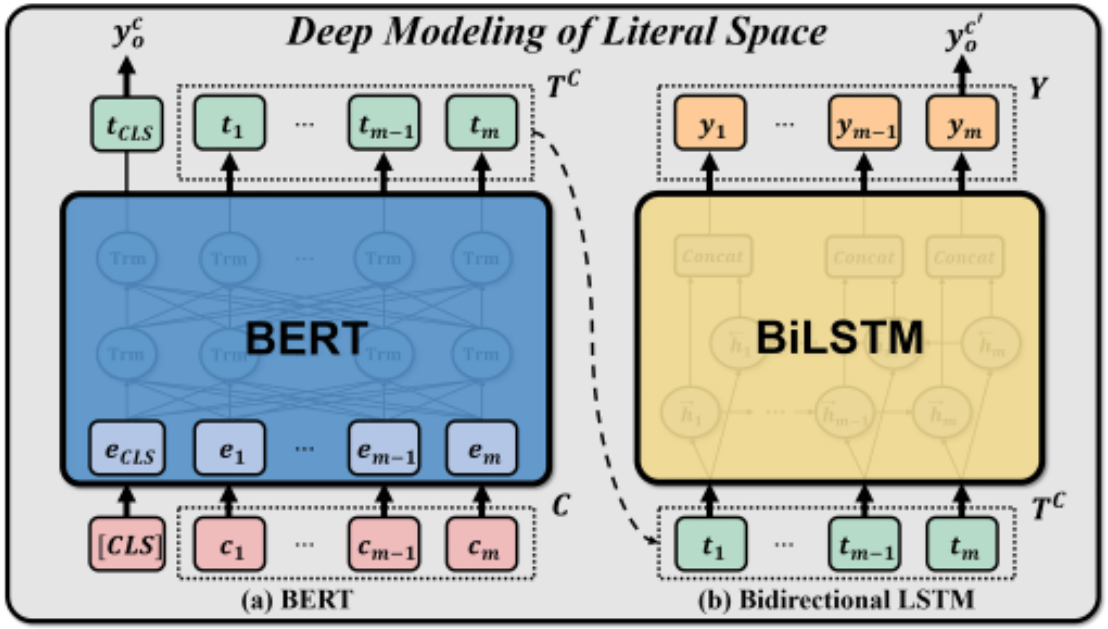


Figure 4: Diagrammatic sketch of *Literal Space* modeling.

- 使用Bert初步提取文本信息
- 使用Bi-LSTM进行步提取上下文依赖信息

Associative Space Modeling

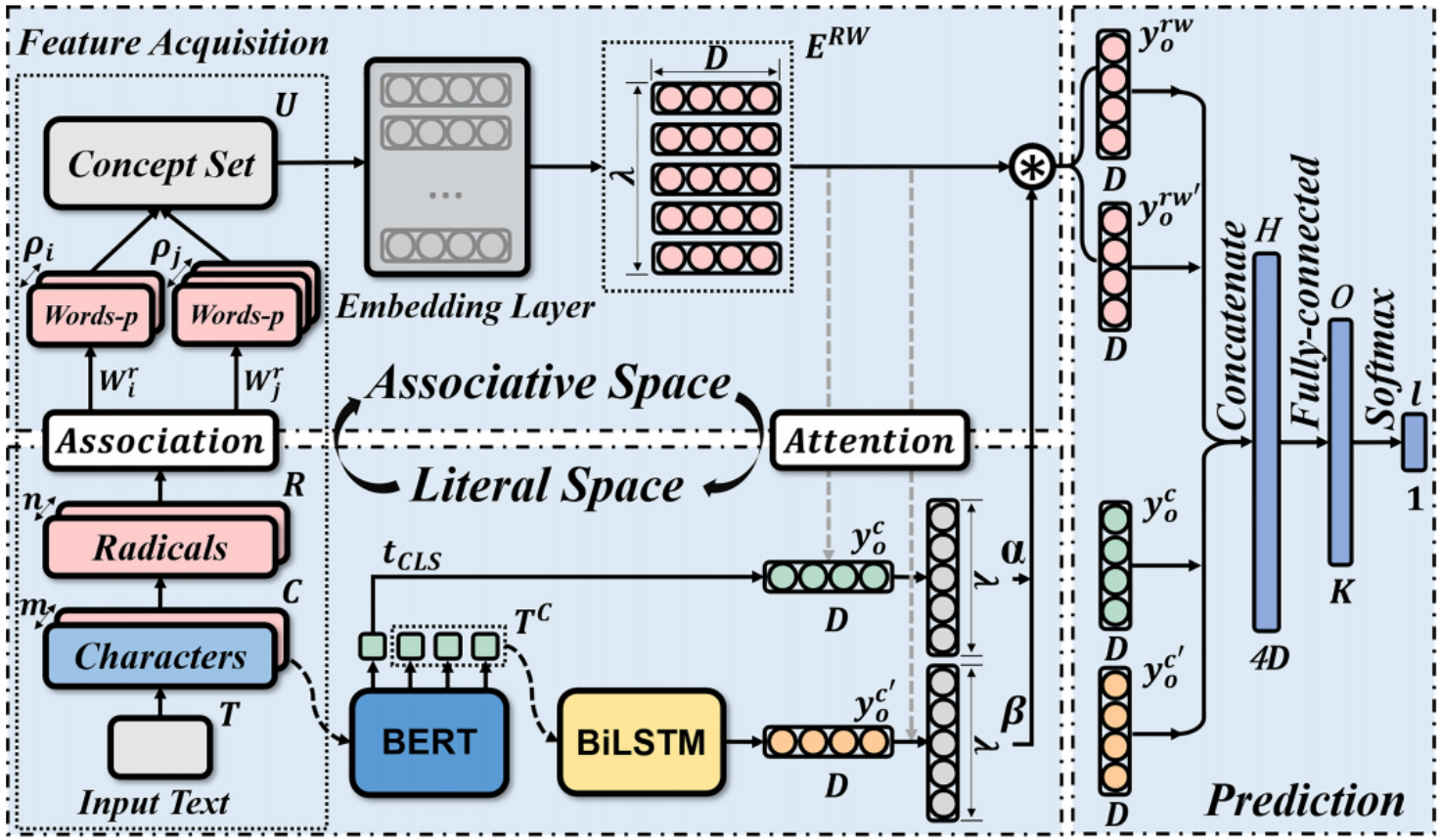


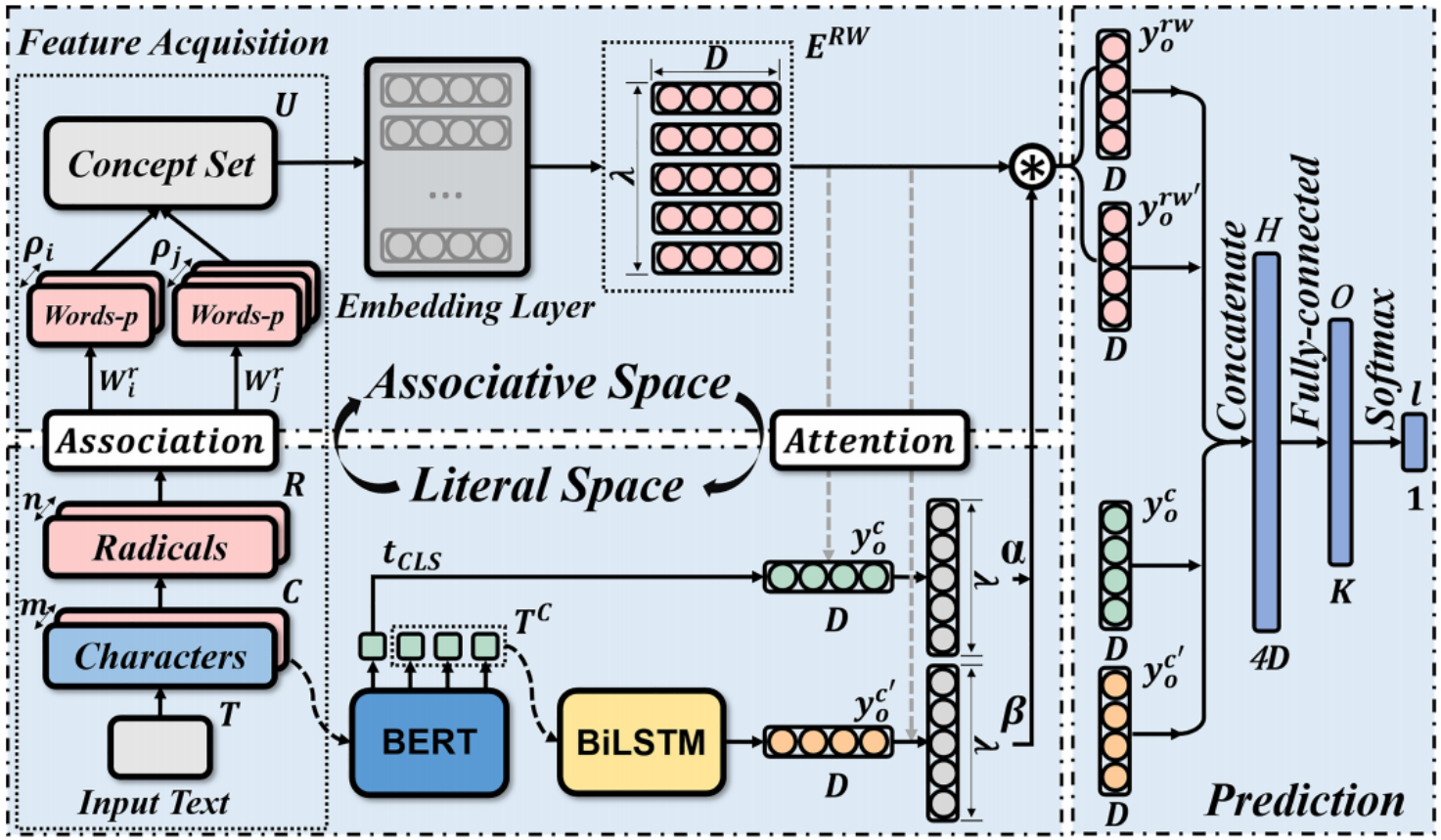
Figure 2: The overall architecture of our Radical-guided Associative Model (RAM).

通过基于分布假设的预训练的Embedding 层+注意力机制进一步表征我们的联想词，同时加强所需要的信息特征

Attention:

1. 文本空间的输出作为查询q，联想空间的embedding作为k和v
2. 得到联想空间的attention权重α、β

预测



1. 基于Attention机制获得嵌入联想空间信息的特征
2. 将之与文本空间特征结合
3. 通过NN+softmax进行文本分类

Figure 2: The overall architecture of our Radical-guided Associative Model (RAM).

EDA——文本分类数据增强方法

- - Synonym Replacement（同义词替换）：随机找n个非停用词的词，再随机从其同义词中挑选一个进行替换
- - Random Insertion（随机插入）：随机找一个句子中的非停用词，再随机找一个其同义词，再最忌插入；重复n次
- - Random Swap（随机交换）：随机选择两个词，并交换其位置；执行n次
- -Random Deletion（随机删除）：依概率p，随机删除句子中的每一个词

优点:

- 小数据集上性能提升快
- 简单容易操作
- 50%的数据集可以做100%数据集的效果
- 增强的数据保留了标签信息

5 总结与展望

挑战

多语种的文本分类

源语言数据的特征空间与目标语言数据之间缺乏重叠；各国语言包含不同的语言学特征，这无疑加大了跨语言文本分类的难度

数据标注瓶颈

网络上存在大量杂乱无章的无标签数据，而依赖人工标注的成本高、效率低

深度学习的可解释性

训练过程难以复现，隐语义和输出结果的可解释性较差，模型的改进与优化没有明确的指引，加深了研究人员调参难度

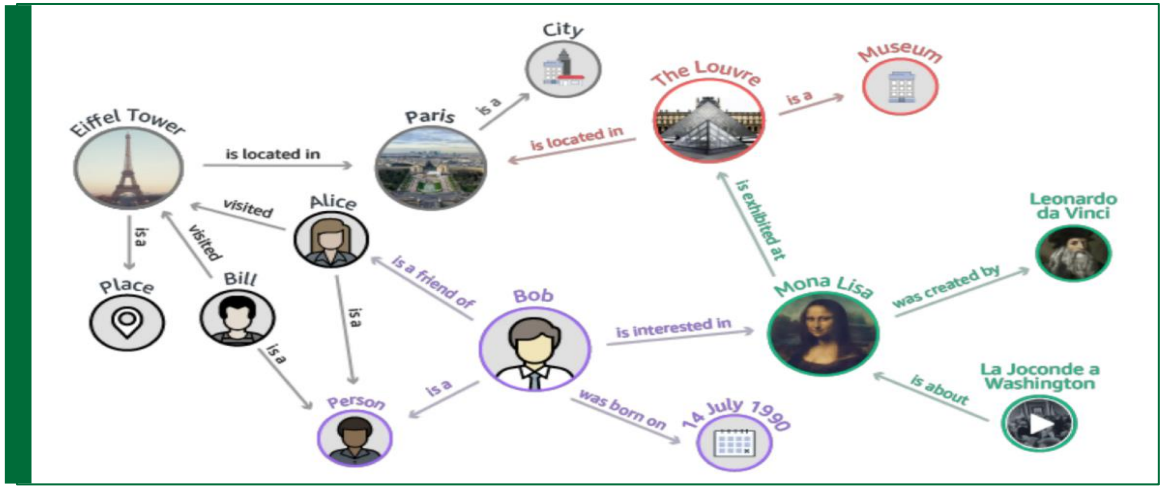
期望

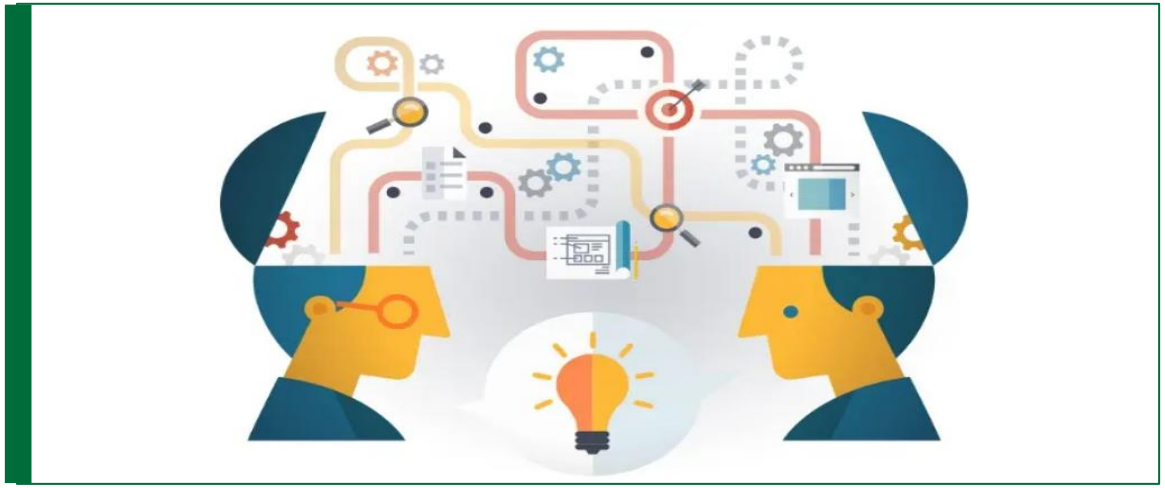
- **传统文本分类方法的优化**：机器学习模型的改进，传统的机器学习算法、特征提取方法与深度学习模型的融合

- **结构化外部知识的引入**：引入知识库、知识图谱等，优化文本表示和预训练的语言模型，进而提升文本分类的性能

- **多任务联合学习**:自然语言处理中，很多任务具有较强的内部关联性，采用此类方法，优化现有深度学习模型

- **迁移学习**：以较小的训练数据量，在保证模型收敛速度的同时达到较好的模型效果





更多的人工智能，才有更多的人工智能

6 demo展示

数据收集

爬取数据

url：https://www.haodf.com/bingcheng/list-mianyineike.html

数据处理

噪声

数据量大

分词停词

	大类	小类	内容
0	介入医学科	介入科	血管瘤并血小板减少，卡梅综合症 你好谭医生，麻烦您帮我开两瓶他克莫司药膏
1	介入医学科	介入科	原发性肝癌，晚期 2021年9月底发现病情，有乙肝病史，肝内两个巨型
2	介入医学科	介入科	血管瘤 血管瘤，血管畸形 表皮红色斑点增多，包块增高，想下周去做手术
3	介入医学科	介入科	胸椎突变 胸椎病变，计划3.23日下午进行穿刺活检检查。
4	介入医学科	介入科	血管畸形 左耳后血管瘤，住院七天，打了五针硬化，现在没有明显
...	...	...	...
1488	骨科	骨质疏松科	十二指肠球部粘膜下隆起性质判断 尊敬的刘主任，\r\n\r\n您好！抱歉打扰您了。\r\n\r\n
1489	骨科	骨质疏松科	脚后跟干裂 脚后跟干裂
1490	骨科	骨质疏松科	金**来你处会诊患有白癜风，计三次。 月初问诊配药。这半个月发现严重 疫情期间，面诊困难，急需帮助
1491	骨科	骨质疏松科	肝癌手术后一个月治疗，肝硬化治疗 2021年7月发现肝硬化，抗病毒治疗，每3个月复查
1492	骨科	骨质疏松科	头皮屑超级多，头顶有炎症 面部有脂溢性皮炎，近几年头顶经 请问用药建议如何，是否需要去医院皮...

624698 rows × 3 columns

特征工程

TFIDF

Word2Vec

Sentence embedding

```
[[0.         0.         0.         ... 0.         0.         2.17073171]
 [0.         0.         0.         ... 0.         0.         2.10344828]
 [0.         0.         0.         ... 0.         0.         1.8       ]
 ...
 [0.         0.         0.         ... 0.         0.         1.93939394]
 [0.         0.         0.         ... 0.         0.         1.94736842]
 [0.         0.         0.         ... 0.         0.         1.75      ]]
```

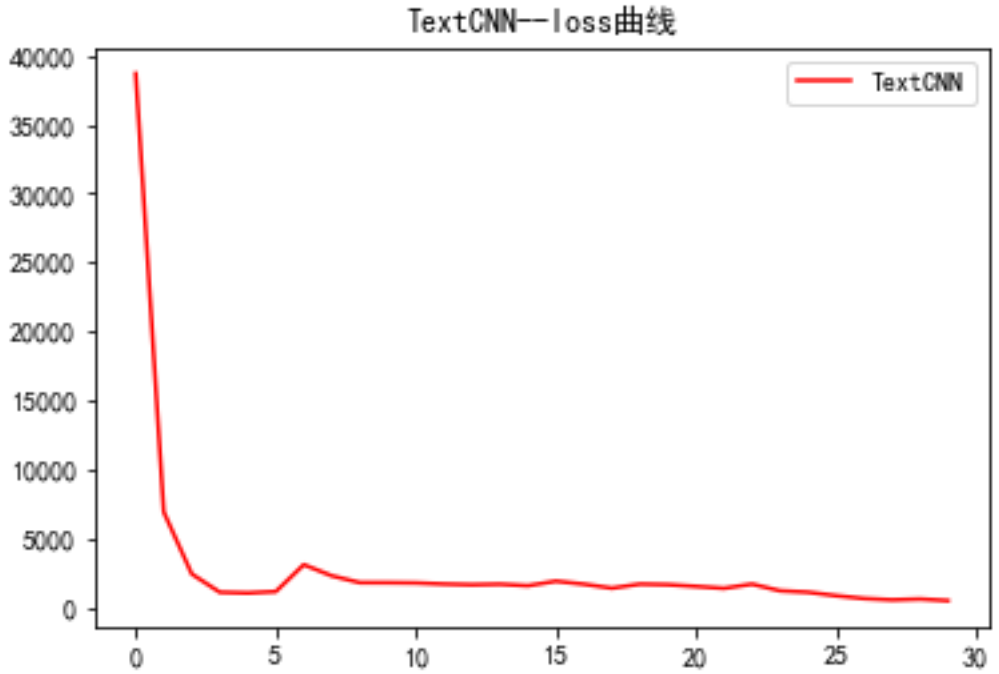
LDA

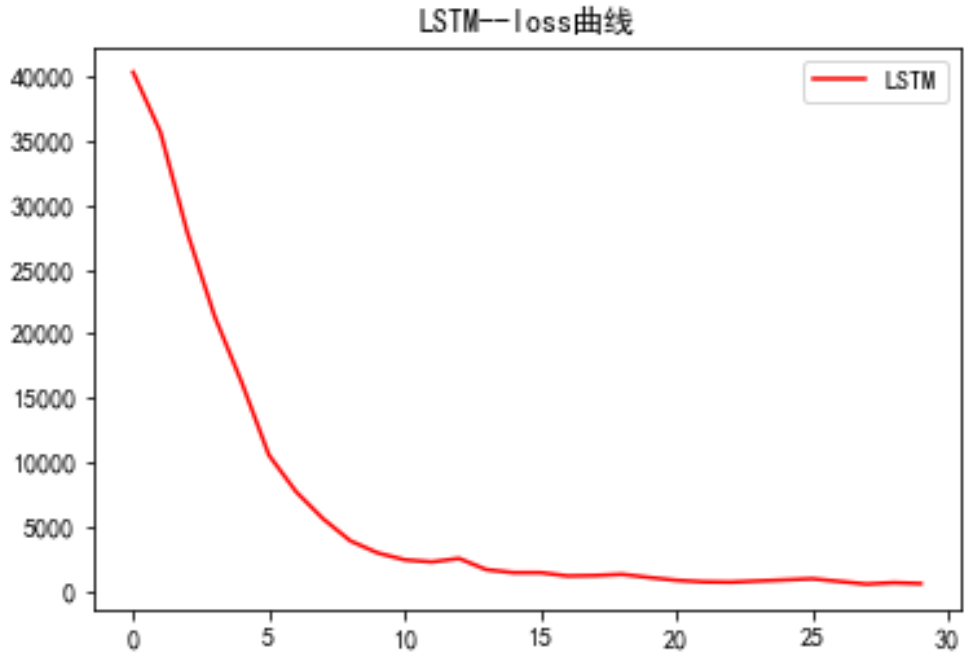
传统模型

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

模型	运行时间	准确率（w2v）	准确率
朴素贝叶斯	9.1	8.13	84.42
SVM	3004.4		90.46
KNN	11.2	8.44	89.99
决策树	98.4	9.68	85.42
随机森林	5.6	9.88	91.17
adaBoost	315.1	11.01	24.57

深度学习模型





模型	准确率
TextCNN	0.904
LSTM	0.857

预训练模型

模型	训练集准确率	验证集准确率
Bert	0.99524	0.99666
XLNet	0.82693	0.85639
Ernie	0.93631	0.89753

注：XLNet\Ernie待更新；更多模型正在试验

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

谢谢观看
敬请各位老师批评指正

答辩人：文本分类研究小组
导　师：张华平