NLP模型调研

一、NLP模型的定义

NLP(Natural Language Processing)自然语言处理，研究人与计算机交互的语言问题的一门学科。机器能够理解并解释人类写作与说话方式的能力。NLP实现机器翻译，聊天机器人，情感分类和语义搜索。

常见的nlp模型：

1. 循环神经网络 (RNN): RNN是一种序列模型，常用于处理具有顺序关系的文本数据，但容易受到梯度消失或梯度爆炸问题的影响。
2. 长短时记忆网络 (LSTM): LSTM是一种改进的RNN变种，用于解决梯度消失问题，适用于长序列数据的建模。
3. 门控循环单元 (GRU): GRU是另一种改进的RNN变种，与LSTM类似，但具有更少的门控单元。
4. 卷积神经网络 (CNN): CNN最初用于图像处理，但也可用于NLP任务，如文本分类和命名实体识别，通过卷积操作捕获局部特征。
5. 词袋模型 (Bag of Words, BoW): BoW是一种简单的文本表示方法，将文本视为词汇的无序集合，用于文本分类和情感分析。
6. 词嵌入 (Word Embeddings): 词嵌入模型如Word2Vec、GloVe和fastText将词汇映射到低维连续向量空间，以捕获词汇语义信息。
7. 递归神经网络 (Recursive Neural Network, RvNN): RvNN用于处理树形结构的文本数据，如句法树。
8. 转换器模型 (Transformer): Transformer是一种革命性的NLP模型，用于各种NLP任务，如机器翻译、文本生成和情感分析。
9. BERT (Bidirectional Encoder Representations from Transformers): BERT是一种预训练的Transformer模型，通过双向上下文捕获了更丰富的语义信息，用于多种NLP任务的微调。
10. GPT (Generative Pre-trained Transformer): GPT是一系列预训练的Transformer模型，特别擅长生成文本，如对话生成和文章创作。
11. XLNet: XLNet是基于Transformer的模型，结合了BERT和自回归性能，用于多任务NLP。
12. RoBERTa (A Robustly Optimized BERT Pretraining Approach): RoBERTa是对BERT的改进版本，通过更大的数据和训练技巧提高了性能。

13. T5 (Text-to-Text Transfer Transformer): T5模型将NLP任务转化为文本到文本的问题，并在大规模数据上进行训练，可用于各种文本任务。
14. BERT Variants: 有许多BERT的变种，如DistilBERT、ALBERT和TinyBERT，它们是对BERT的轻量级或改进版本。
15. BERT-for-domain (领域特定BERT): 针对特定领域或任务的BERT变种，如BioBERT用于生物医学领域。

NLP常用开发包：

- NLTK (Natural Language Toolkit): NLTK是一个广泛用于NLP的库，提供了各种文本处理工具和语料库，包括分词、词性标注、句法分析等。
- gensim: 是用于主题建模和词向量学习的库，包括Word2Vec、Doc2Vec等模型的实现。
- tensorflow，谷歌基于DistBelief进行研发的第二代人工智能学习系统，是一个用于构建和训练深度学习模型的强大框架，可以用于自定义NLP模型的开发。
- PyTorch: 是另一个深度学习框架，提供了灵活的工具和库，用于NLP模型的构建和训练。
- jieba，中文分词工具

实现算法通用的步骤：

1）论文的阅读，最新算法的研究

2）算法的大概方向的评估和确定

3）训练数据收集，清洗以及数据预处理

4）算法实现，系统设计，参数调优，模型升级

参考资料论文；软件模块设计架构；更改网格参数；模型算法升级，错误样本再训练

5）模型效果评估与部署

准确率；召回率；本地调用/封装成服务

文本分类操作流程：

（1）使用类似jieba的中文分词库对整个训练集进行中文分词

（2）统计词汇的出现频率，删除部分低频词

（3）根据禁用词库滤除禁用词

（4）根据生成的词库对训练集进行分词处理，只保留词库中已有的词汇，词汇之间用空格隔开

（5）分词后使用word2vec训练词向量模型（该步骤是可选项，可以不做）

（6）搭建文本训练模型，例如TextCNN（若进行了第五步操作，使用训练好的词向量weights初始化embdding层参数）

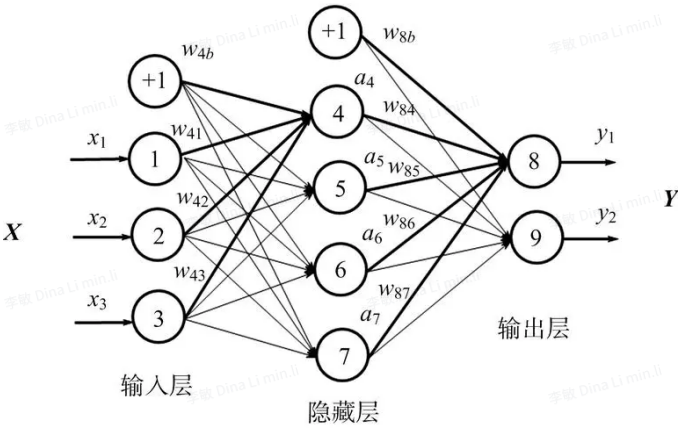
（7）训练模型，调参，有更高的需求的可以改进模型

二、卷积神经网络原理

（一）卷积神经网络的特点：

1. **稀疏交互**(sparse interactions): 也叫稀疏权重(sparse weights)、稀疏连接(sparse connectivity)

在传统神经网络中，网络层之间输入与输出的连接关系可以由一个权值参数矩阵来表示。对于全连接网络，任意一对输入与输出神经元之间都产生交互，形成稠密的连接结构。这里面的交互是指每个单独的参数值，该参数值表示了前后层某两个神经元节点之间的交互。



在卷积神经网络中，卷积核尺度远小于输入的维度，这样每个输出神经元仅与前一层特定局部区域内的神经元存在连接权重（即产生交互），我们称这种特性为稀疏交互。

稀疏交互的物理意义：通常图像、文本、语音等现实世界中的数据都具有局部的特征结构， 我们可以先学习局部的特征， 再将局部的特征组合起来形成更复杂和抽象的特征。

2. **参数共享**(parameter sharing)

参数共享是指在同一个模型的不同模块中使用相同的参数。卷积运算中的参数共享让网络只需要学一个参数集合，而不是对于每一位置都需要学习一个单独的参数集合。

参数共享的物理意义：使得卷积层具有**平移等变性**。在第三个特点中会谈到。

显然，我们可以看到，卷积神经网络在存储大小和统计效率方面极大地优于传统的使用矩阵乘法的神经网络。

3. **等变表示**(equivariant representations)

假如图像中有一只猫，那么无论它出现在图像中的任何位置，我们都应该将它识别为猫，也就是说神经网络的输出对于平移变换来说应当是等变的。特别地，当函数f(x)与g(x)满足f(g(x))=g(f(x))时，我们称f(x)关于变换g具有等变性。在猫的图片上先进行卷积，再向右平移l像素的输出，与先将图片向右平移l像素再进行卷积操作的输出结果是相等的。

（二）卷积(convolution)操作

卷积神经网络使用的数据可以是一维、二维和三维的，对于这三种数据，每种都可以是单通道或者是多通道的。一维常用于序列模型，NLP模型，二维卷积 Conv2d也用于NLP，三维卷积常用于医学领域（CT音像），视频处理领域（检测动作及人物行为）。

（三）池化(pooling)操作

池化也叫做亚采样、下采样(downsampling)或子采样(subsampling)，主要针对非重叠区域。包括均值池化（mean pooling）、最大池化（max pooling）。池化操作的本质是降采样。例如，我们可以利用最大池化将4×4的矩阵降采样为2×2的矩阵。

均值池化通过对邻域内特征数值求平均来实现，能够抑制由于邻域大小受限造成估计值方差增大的现象，特点是对背景的保留效果更好。均值池化是求窗口中元素的均值。

最大池化则通过取邻域内特征的最大值来实现，能够抑制网络参数误差造成估计均值偏移的现象，特点是更好地提取纹理信息。最大池化是求窗口中元素的最大值。

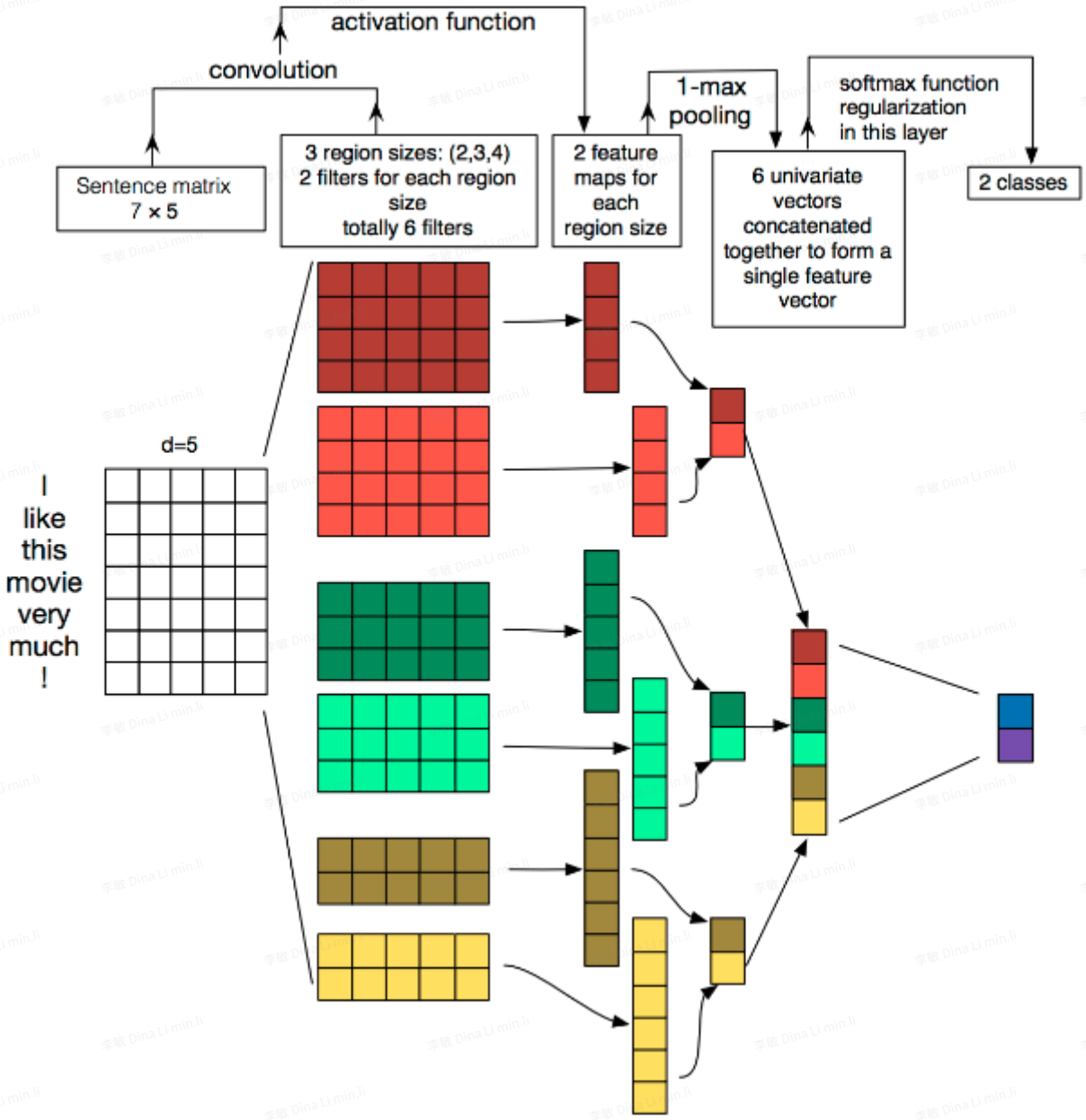
三、卷积神经网络在文本分类上的应用

TextCNN(Convolutional Neural Network for Text Classification)

Yoon Kim在论文(2014 EMNLP) Convolutional Neural Networks for Sentence Classification 提出TextCNN。

将**卷积神经网络CNN**应用到**文本分类**任务，利用多个不同size 的 kernel 来提取句子中的关键信息（类似于多窗口大小的ngram），从而能够更好地捕捉局部相关性。

（一）TextCNN的详细过程原理：



（1）第一层为输入层。

输入层是一个 n×k 的矩阵，其中 n 为一个句子中的单词数， k 是每个词对应的词向量的维度。也就是说，输入层的每一行就是一个单词所对应的 k 维的词向量。另外，这里为了使向量长度一致对原句子进行了padding操作。

（2）第二层为卷积层。

卷积核的宽和该词矩阵的宽相同，该宽度即为词向量大小，且卷积核只会在高度方向移动。由于卷积核和word embedding的宽度一致，一个卷积核对于一个sentence，卷积后得到的结果是一个vector，其shape=(sentence_len - filter_window_size + 1, 1)。

一个卷积核经过卷积操作只能得到一个scalar，将相同filter_window_size卷积出来的num_filter个scalar组合在一起，组成这个filter_window_size下的feature_vector。我们会使用多个filter_window_size（原因是，这样不同的kernel可以获取不同范围内词的关系，获得的是纵向的差异信息，也就是在一个句子中不同范围的词出现会带来什么信息。）

（3）池化层

在经过max-pooling操作后得到的就是一个Scalar。

将所有filter_window_size下的feature_vector也组合成一个single vector，作为最后一层softmax的输入。

(4) softmax层

使用Softmax激活函数输出每个类别的概率。

（二）TextCNN的超参数调参

默认配置

Description	Values	remark
输入词向量	word2vec	词向量表征的选取(如选word2vec还是GloVe)
filter大小	(3,4,5)	一个合理的值范围在1~10。若语料中的句子较长，可以考虑使用更大的卷积核。
每个size下的filter个数	100	feature map特征图个数：主要考虑的是当增加特征图个数时，训练时间也会加长，因此需要权衡好。
激活函数	ReLU	f(x) = max(0,x) 深度学习中最常用的激活函数之一，简单的非线性函数， 优点：非线性、稀疏激活、计算速度快 缺点：死亡ReLU问题
池化策略	1-max pooling	1-max pooling表现最佳
Dropout rate	0.5	是指在训练期间随机关闭一部分神经元的比例。这是一种**正则化**技巧，旨在防止过拟合并提高模型的泛化能力。通常，合理的 Dropout rate 取值范围是从 0.2 到 0.5
L2 正则化	0.01	L2**正则化**通过添加一个惩罚项（权重衰减项）到损失函数中，鼓励模型的权重保持较小的值。这有助于使模型的参数更平滑，降低了模型对训练数据的过度依赖，从而提高了泛化能力。

Reference

https://www.huaxiaozhuan.com/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/chapters/5_CNN.html

https://blog.csdn.net/weixin_50295745/article/details/126333612