**数智2022第二阶段考核**

**学院：计算机学院  
 专业：计算机类**

**姓名：李金强**

**学号：3121004863**

**摘要：**文本分类是自然语言处理的必备入门任务，在日常生活中应用非常广泛。在这个文档中，我将会简要概述通过运用卷积神经网络和残差网络训练，使用tensorflow框架进行初步的文本分类。其中还包括使用BPE算法、Embedding嵌入层的使用、LSTN和CNN模型等等，同时对在模型参数拟合过程中出现的过拟合现象进行处理。

**关键词：**卷积神经网络、fasttext模型、过拟合

**Abstract**: Text classification is an essential introductory task for natural language processing, and it is widely used in daily life. In this document, I will give a brief overview of initial text classification using the tensorflow framework by training with convolutional neural networks and residual networks. It also includes the use of the BPE algorithm, the use of the Embedding embedding layer, the LSTN and CNN models, etc., and the overfitting phenomenon that occurs in the process of model parameter fitting is dealt with.

**Keywords**: convolutional neural network, fasttext model, overfitting

1. 任务实现路线
   1. **前期预备工作**

**·** 根据python版本下载安装了对应tensorflow以及keras

· 对第一轮考核的学习内容进行查缺补漏，夯实机器学习的基础

· 明确实现文本分类需要的过程步骤，制定所需的学习内容及学习计划

**1.2 任务内容的学习**

· 线性代数的学习，便于对机器学习进行深层度的学习理解

·tensorflow基础语法的使用，包括向量运算、矩阵运算，以及目标函数、损失函数和优化函数的表达

· 对卷积神经网络和残差网络进行学习，为文本分类打下基础

· 查阅fasttext的相关学习资料，自身能够训练词向量，在保持高精度的情况下很大程度上加快了训练速度和测试速度

· 了解多种损失函数的形式与主要用途，选择合适的损失函数减少误差

**1.3 处理任务出现的问题**

· 模型选取不恰当，出现了较高的错误率，通过改进模型降低错误率

· 参数选择不恰当，训练时间长以及错误率高，通过不断调试选择错误率相对较低的参数

· 后期出现了过拟合，对训练次数进行调整，防止出现过拟合和欠拟合问题

**1.4 后期进行总结**

· 对近期的学习内容作出回顾与总结

· 明确下一阶段的学习计划与目标

· 与师兄进行学习交流

**二、任务的学习**

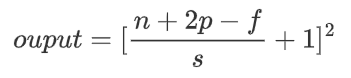
**2.1 卷积神经网络**

**2.1.1 简要概述**

卷积神经网络具有表征学习能力，能够按其阶层对输入信息进行平移不变分类，由输入层、卷积层、激活函数、池化层和全连接层组成，可用于分类、特征提取、关键点定位等领域。

**2.1.2 各层次工作原理**

**卷积层**类似于图像空间处理的卷积操作，设置一个小区域的滤波器，将输入的图像按照滤波器的大小来划分区域，然后将滤波器在输入图像的每一个区域上滑动，对于每一个区域产生一个特征值。特征值的大小ouput由输入图像宽度n、滤波器宽度f、卷积步长s、填充0的宽度p组成，



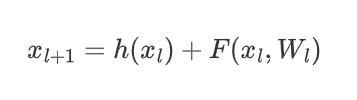
**池化层**可缩减模型大小，提高计算速度。选取一个小区域在特征图滑动，取个小区域的min值或max值代表这一区域的特征值。**全连接层**将卷积层、池化层的特征提取出来，计算属于某个类的概率。

**2.2 残差网络**

**2.2.1 残差网络的实现**

网络的一层通常看作是y=H(x),而残差网络的一个残差块可表示为H(x)=F(x)+x，即F(x)=H(x)-x。在单位映射中，y=x是观测值，H(x)是预测值，F(x)对应着残差。

残差网络由残差块组成，残差块又分成直接映射部分和残差部分。若是直接映射，是残差部分，则一个残差块为：

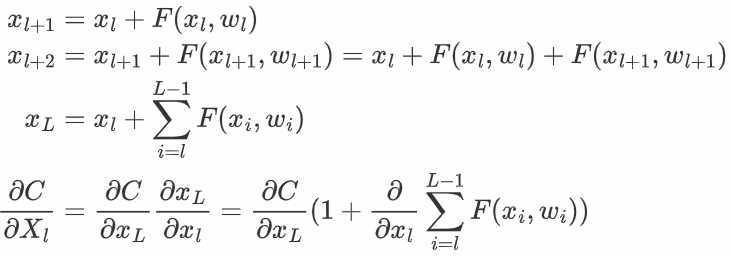


**2.2.2 残差网络优化问题**

· **解决梯度消失**：

在很深的网络层，由于参数初始化一般更靠近0，这会使得训练过程中更新浅层网络的参数时，容易使得随着网络的深入而导致梯度的消失，浅层的参数无法更新。

在残差网络的链式求导中，由于公式右边有1的存在，同时将链式求导的连乘转为连加状态，保证了该节点参数更新不会发生梯度消失或梯度爆炸的现象。



· **解决网络退化问题**

随着网络深度的增加，模型会产生退化现象。它不是由过拟合产生的，而是由冗余的网络层学习不是恒等映射的参数造成的。

要让网络不退化，我们就要做到恒等映射。由于在h(x)=F(x)+x中，F(x)称为残差项，我们需要让F(x)=0。如果该层网络只经过线性变换，随机初始化权重一般偏向于0，相比于h(x)=x,模型会更快得到F(x)=0。当网络自己决定哪些网络层为冗余层时，使用ResNet的网络很大程度上解决了学习恒等映射的问题。

**2.3 BPE算法**

**2.3.1 实现原理**

字节对编码是一种数据压缩算法，用来在固定大小的词表中实现可变长度的子词。首先将词分成单个字符，然后依次用另一个字符替换频率最高的一对字符，直到循环结束。

**2.3.2 算法过程**

· 构建BPE算法的词表，确定subwork词表大小等参数

· 在每个单词末尾添加后缀<\w>，统计每个单词出现的频率。

· 将所有单词拆分为单个字符，用所有单个字符建立最初的词典， 统计每个字符的最高频率。此时subword的粒度是字符。

· 挑出频次最高的字符对。重复第二步直到词表中的单词数达到设 定量或下一个最高频数为1。

**2.3.3 BPE优缺点**

BPE算法能够有效地平衡词典的大小和编码步骤数。随着合并次数的增加，词表的大小通常先增加后减小。若迭代次数太少，大部分还是字母；若迭代次数太多，又会重新变成原来的词语。所以词表大小应该选一个中间值。

**2.4 fasttext算法**

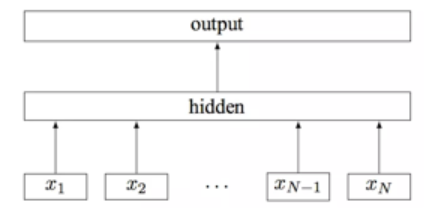
**2.4.1 算法概述**

fastext是一个快速文本分类的算法，与基于神经网络的分类算法相比，fasttext在保持高精度的情况下加快了训练速度和测试速度。此外，fasttext不需要预训好的词向量，能够自己训练词向量

**2.4.2 模型架构**

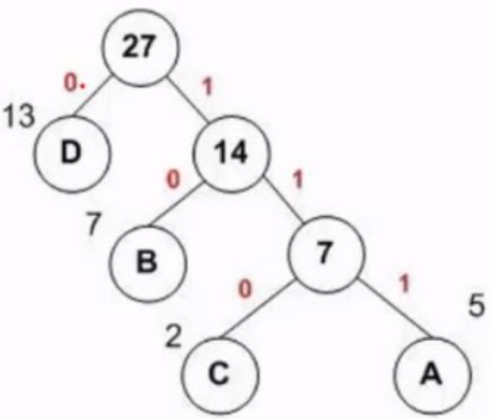
fasttext用来预测标签。该模型输入一个词的序列，输出这个词序列属于不同类别的概率。序列中的词和词组组成特征向量，特征向量通过线性变换映射到中间层，中间层再映射到标签。fasttext在预测标签中使用了非线性激活函数，但在中间层不使用非线性激活函数。

其中表示一个文本中的n-gram向量，每个特征是词向量的平均值。



**2.4.3 层次softmax**

对于有大量类别的数据集，fasttext使用了一个分层分类器。不同的类别被整合进树形结构中。在某些文本分类任务中类别很多，计算线性分类器的复杂度高。为了改善运行时间，fasttext模型使用了层次softmax技巧。层次softmax层次技巧在对标签进行编码时，能够极大地缩小模型预测目标的数量。



**2.4.4 N-gram特征**

1. Gram是一种基于统计语言模型的算法。它的基本思想是将文本里面的内容按照字节进行大小为N的滑动窗口操作，形成了长度是N的字节片段序列。

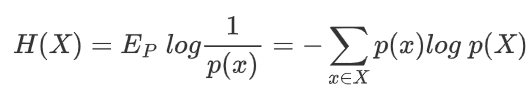
每一个字节片段称为gram，对所有gram的出现频度进行统计，并且按照事先设定好的阈值进行过滤，形成关键gram列表，也就是这个文本的向量特征空间，列表中的每一种gram就是一个特征向量维度。

**2.5 Cross-entropy交叉熵**

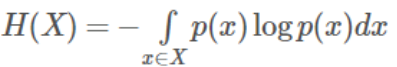
**2.5.1 交叉熵预备知识**

交叉熵是深度学习中用来求目标与预测值之间的差距。

熵是信息量的期望值，它是一个随机变量的确定性的度量。熵越大，变量的取值越不确定。对于一个随机变量x而言，其所有可能取值的信息量的期望为(E[I(x)])。熵的定义为：

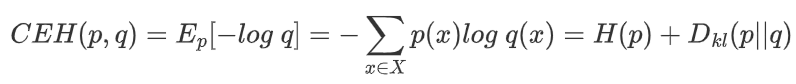


如果p(x)是连续型随机变量，则熵的定义为：



**2.5.2 交叉熵模型**

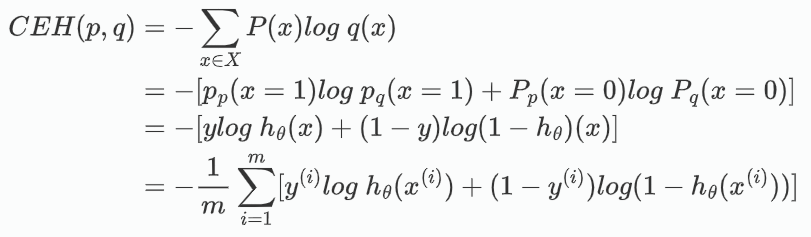
假设有两个分布p，q，则它们在给定样本集上的交叉熵定义为：



当p已知时，把H(p)看作一个常数，交叉熵反映了分布p，q的相似程度。最小化交叉熵等于最小化KL距离。它们都将在p=q（KL的距离为0）时取得的最小值H(p)。

p：真实样本分布，服从参数为p的分布，即X~B(1,p)

q：待估计的模型，服从参数为q的分布，即X~B(1,q)



**三、任务问题的处理**

**3.1 过拟合问题**

过拟合问题可能是由于过于强调拟合原始数据，而丢失了算法的本质—预测新数据。或许这个过拟合的模型能够很好地适应我们的训练集，但在预测中的表现却不是特别好。x的次数越高，拟合得越好，但相应的预测能力就可能会变差。

面对过拟合问题时，我们可以使用多种方法通过控制变量法测试模型是否为过拟合。同时对模型的参数进行不断调试，选择错误率相对较低的参数。

**3.2 卷积层的数量**

在一开始模型构建时，模型中的卷积层数量较少，导致了错误率较高。后来，在了解到卷积层可以减少网络的参数数量但能达到同等水平的表现力时，我就增加了卷积层的数量。叠加小型滤波器来加深网络的好处为减少参数数量，扩大感受野，而且通过叠加层，将ReLu等激活函数夹在卷积层的中间，进一步提高了网络的表现力。

**3.3 卷积核的选择**

多个小的卷积核叠加使用要远比一个大的卷积核单独使用效果要好，降低了参数个数和计算的复杂度。当时应该注意的是，不能选择过小的卷积核。在面对特别稀疏的数据时，过小的卷积核可能无法表示其特征，如果采用较大的卷积核则会增加复杂度。在初始时，我选择了较大的卷积核，导致计算速度慢，后来选用了小而深的卷积核，模型的性能相对得到了提升。

**四、收获与感悟**

**4.1 收获**

· 加强了对高等数学的求偏导和线性代数中矩阵相关知识的学习

· 学习了python语言和python语言的一些常用内置函数

· 了解了机器学习中的线性回归模型与softmax模型

· 学会subword中的BPE原理,能够对文本进行分类

· 对CNN和fasttext模型有了初步的认识，增强了对文本分类的能力

**4.2 感悟**

数智深度学习的两轮考核也缓缓落下了帷幕。在这里，我首先要感谢一路走来师兄及身边同学对我的支持鼓励和帮助。由于我在一轮考核中未参与答辩，本不能进入二轮考核的，我非常感谢詹培林师兄能够给我这次机会站在答辩的舞台上。在我得知我有机会继续参加二轮考核时，我的内心是兴奋的，同时我也暗暗发誓一定要抓住这次难得的机会，绝对不能辜负师兄们的殷切期望和自身向往深度学习的梦想。

在工作室报名招新时，我同时通过了数智的面试和topview的笔试及面试，但是在我的一番思索后，我毅然决然放弃了topview大数据开发的机会，坚定走上了深度学习的道路。尽管此时我拥有了开发的基础而对深度学习的专业知识了解甚少，但是我的内心已经坚定走上了深度学习的道路。在其他同学进行一轮考核的任务时，我才从零基础开始学习python。虽然我的起步比其他同学要晚，但是我时刻告诫自己“我命由我不由天”。很遗憾的是，我还是未能在一轮结束前从零基础到完成所有的任务。在二轮考核的最初两个星期，我自责自身为什么不能完成一轮的学习。由于我的内心真的不愿放弃数智的深度学习组，我还是鼓起勇气向詹培琳师兄申请了一个机会——也是给我的梦想一个机会。

这时离二轮结束仅剩下十一天，我开始了机器学习的学习及文本分类的模型方法。在这十一天中，除了上课做作业，我的大部分时间都拿来了学习机器学习的内容。虽然在这段日子身体稍许疲惫，但是内心感到无比的充实，依然充满了活力，只是时间不够用。或许我的文档不如其他同学完美，或许我的答辩不如其他同学优秀。但是，经过这段时间的学习，我对考研有着更加无比坚定的决心，对深度学习也有更加深沉且坚定的热爱，绝对不会出现半途而废或遇到难题退缩的情况。如果我有幸能够进入工作室，我将会继续全身心投入学习，和师兄及其他同学进行探索讨论，一起汲取深度学习领域的知识，达到共同进步的目的。

1. **参考文献**
2. 山下隆义.图解深度学习[M].北京.人民邮电出版社
3. Salman Khan.卷积神经网络与计算机视觉[M].北京.机械工业出版社
4. Armand Joulin.Bag of Tricks for Efficent Text Classification[R]
5. 刘建平.卷积神经网络(CNN)模型结构[OL]