北京理工大学本科毕业论文

|  |  |
| --- | --- |
| **题目：** | 深度学习在情感分析中的应用 |

|  |  |
| --- | --- |
| **学 院:** | 计算机科学与技术 |
| **专 业:** |  |
| **学 号:** |  |
| **学生姓名:** |  |
| **指导教师:** |  |
| **日 期:** | 2021.04 |

摘 要

情感分析或观点挖掘是对人们对产品、服务、组织、个人、问题、事件、话题及其属性的观点、情感、情绪、评价和态度的计算研究。该领域的开始和快速发展与社交媒体的发展相一致，如评论、论坛、博客、微博、推特和社交网络，因为这是人类历史上第一次拥有如此海量的以数字形式记录的观点数据。早在 2000 年，情感分析就成为 NLP 中最活跃的研究领域之一。它在数据挖掘、Web 挖掘、文本挖掘和信息检索方面得到了广泛的研究。实际上，因其对商业和社会的整体重要性，它已经从计算机科学扩展到管理学和社会学，如营销、金融、政治学、传播学、健康科学，甚至历史学。这种发展原因在于观点是几乎所有人类活动的核心，是人类行为的重要影响因素。我们的信念、对现实的感知，以及我们所做的决策在很大程度上依赖于别人看到和评价世界的方式。因此，我们在做决策的时候，通常会寻求别人的意见。不只是个人，组织也是如此。

现有研究已经产生了可用于情感分析多项任务的大量技术，包括监督和无监督方法。在监督方法中，早期论文使用所有监督机器学习方法（如支持向量机、最大熵、朴素贝叶斯等）和特征组合。无监督方法包括使用情感词典、语法分析和句法模式的不同方法。现有多本综述书籍和论文，广泛地涵盖了早期的方法和应用。

大约十年前，深度学习成为强大的机器学习技术，在很多应用领域产生了当前最优的结果，包括计算机视觉、语音识别、NLP 等。近期将深度学习应用到情感分析也逐渐变得流行。本文首先概述深度学习，然后对基于深度学习的情感分析进行综述。

本文全面描述深度学习实现情绪以前分析，主要介绍了系我们希望通过深度学习进行情感预测，尝试超越传统机器学习算法的情感分析表现，详细解析了解决问题所需要完成的功能，并对系统的开发过程，开发工具和所运用的技术加以解释。从系统需求分析、概要设计、详细设计、功能模块实现这四个角度对系统的开发过程进行了具体的阐述，并对系统的主要特点进行了介绍。

**关键词**：python, pandas, tensorflow，tensorflow matplotlib

**Abstract**

Sentiment analysis or opinion mining is a computational study of people's opinions, emotions, emotions, evaluations, and attitudes about products, services, organizations, individuals, problems, events, topics and themes. The development is consistent, such as comments, forums, blogs, Weibo, Twitter and social networks, because this is the first time in human history that there is such a large amount of opinion data recorded in digital form. As early as 2000, sentiment analysis became one of the most active research fields in NLP. It has been extensively studied in data mining, web mining, text mining and information retrieval. To management and sociology, such as marketing, finance, political science, communication, health science, and even history. The reason for this development is that viewpoints are at the core of almost all human activities and are important factors influencing human behavior. Our beliefs, our perception of reality, and the decisions we make depend to a large extent on the way others see and evaluate the world. Therefore, when we make decisions, we usually seek the opinions of others. in this way.

Existing research has produced a large number of techniques that can be used for a variety of tasks in sentiment analysis, including supervised and unsupervised methods. Among the supervised methods, early papers used all supervised machine learning methods (such as support vector machines, maximum entropy, naive Bayes, etc.) and feature combinations.

About ten years ago, deep learning became a powerful machine learning technology, producing the latest and best results in many application fields, including computer vision, speech recognition, NLP, etc. This article first outlines deep learning, and then summarizes sentiment analysis based on deep learning..

**Key words:** ython, pandas, tensorflow，tensorflow matplotlib

目 录

[深度学习在情感分析中的应用 I](#_Toc70922385)

[1 系统开发概述 1](#_Toc70922386)

[1.1 系统开发背景与发展 1](#_Toc70922387)

[1.2 深度学习 1](#_Toc70922388)

[1.3 系统目的 2](#_Toc70922389)

[1.4 论文组织结构 6](#_Toc70922390)

[2 开发技术及工具介绍 8](#_Toc70922391)

[2.1 tensorflow开发框架 8](#_Toc70922392)

[2.2 keras 10](#_Toc70922393)

[2.3 NumPy 10](#_Toc70922394)

[2.4 Jupyter Notebook 11](#_Toc70922395)

[3 情感分析系统系统分析与设计 12](#_Toc70922396)

[3.1 系统需求分析 12](#_Toc70922397)

[3.2 系统设计效果 13](#_Toc70922398)

[3.3 系统角色功能设计 13](#_Toc70922399)

[4 情感分析系统系统详细设计 16](#_Toc70922400)

[4.1 系统结构设计 16](#_Toc70922401)

[4.2 系统概念模型 17](#_Toc70922402)

[4.3 系统数据表逻辑结构设计 20](#_Toc70922403)

[4.3.1 数据表介绍 20](#_Toc70922404)

[4.3.2 数据表设计 20](#_Toc70922405)

[5 情感分析系统系统应用程序设计 23](#_Toc70922406)

[5.1 前台系统设计 23](#_Toc70922407)

[5.1.1 主页面 23](#_Toc70922408)

[5.1.2 个人信息页面 23](#_Toc70922409)

[5.1.3 密码修改页面 24](#_Toc70922410)

[5.1.4 活动详情页面 24](#_Toc70922411)

[5.2 后台系统设计 25](#_Toc70922412)

[5.2.1 登录主页面 25](#_Toc70922413)

[5.2.2 学校注册页面 26](#_Toc70922414)

[5.2.3 个人信息页面 27](#_Toc70922415)

[5.2.4 密码修改页面 28](#_Toc70922416)

[5.2.5 组织部门设置页面 28](#_Toc70922417)

[5.2.6 组织部门信息页面 29](#_Toc70922418)

[5.2.7 部门建设页面 30](#_Toc70922419)

[5.2.8 活动发布页面 31](#_Toc70922420)

[5.2.9 活动进展页面 32](#_Toc70922421)

[结论 34](#_Toc70922422)

[参考文献 35](#_Toc70922423)

[致谢 36](#_Toc70922424)

# 

# 1 系统开发概述

## 1.1 系统开发背景与发展

情感分析或观点挖掘是对人们对产品、服务、组织、个人、问题、事件、话题及其属性的观点、情感、情绪、评价和态度的计算研究。

该领域的开始和快速发展与社交媒体的发展相一致，如评论、论坛、博客、微博、推特和社交网络，因为这是人类历史上第一次拥有如此海量的以数字形式记录的观点数据。

早在 2000 年，情感分析就成为 NLP 中最活跃的研究领域之一。它在数据挖掘、Web 挖掘、文本挖掘和信息检索方面得到了广泛的研究。

这种发展原因在于观点是几乎所有人类活动的核心，是人类行为的重要影响因素。我们在做决策的时候，通常会寻求别人的意见。

然而，由于各种网站的出现，在网上查找和监测意见网站并提取其中的信息仍然是一项艰巨的任务。

现有的研究已经为情绪分析的各种任务产生了大量的技术，包括有监督和无监督的方法。近年来，将深度学习应用于情绪分析也变得非常流行。

## 1.2 深度学习

* 神经网络，深度学习
* Word Embedding: Word2Vec (CBOW-小数据集，Skip-Gram-大数据集)、GloVe
* AutoEncoder / Denoising AutoEncoder

自编码器神经网络是一个三层神经网络，其目标是使输出值近似等价于输入值。目标是学习输入的特征。

降噪自编码器（DAE）是自编码器的扩展，其中输入被破坏。思想是强制隐藏层发现更鲁棒的特征，并阻止自编码器简单地学习恒等变换。也就是说，模型应该在存在噪声时仍能重构输入。这种技术也体现在情感分析中，例如从文档中删除或添加一些文字不应该改变文档的语义。

* Convolutional Neural Network

卷积神经网络（CNN）是一种特殊的前向神经网络，最初应用于计算机视觉领域。它的设计灵感来自人类视觉皮层。视觉皮层包含许多细胞，这些细胞负责在视野的小而重叠的亚区域（称为感受野）中检测光线。

卷积层可以提取局部特征，而不关心他们的位置如何。

* Recurrent Neural Network

循环神经网络（RNN）在处理序列信息方面很受欢迎。理论上，RNN可以利用任意长序列的信息，但在实际应用中，由于梯度消失和梯度爆炸，标准RNN只能回顾几个时间步。

由于标准RNN不足，开发了 双向RNN、Deep Bidirectional RNN、Long Short Term Memory (LSTM)

LSTM：4个层交互，2个状态 hidden state and cell state，3个门 forget/input/output。通常用于顺序数据或树形结构的数据，处理数据的远程依赖关系。树形结构的LSTM- Tree-LSTM / 简化 - GRU。

* Attention Mechanism

注意机制受到人类视觉注意机制的启发，人类的视觉注意力能够以“高分辨率”聚焦在图像的某个区域，同时以“低分辨率”感知周围的图像，然后随着时间的推移调整焦点。

* Memory Network

针对QA问题引入了记忆网络 (MemNN)。End-to-End Memory Network (MemN2N)

* Recursive Neural Network

递归神经网络(RecNN)是一种通常用于从数据中学习有向无环图结构(例如，树结构)的神经网络。给定句子的结构表示(例如，解析树)，RecNN通过组合记号来产生短语的表示，最终生成整个句子，从而以自下而上的方式递归地生成父表示

## 1.3 系统目的

##### 1.3.1 情感分析任务

主要在三个粒度级别研究情感分析：document level、sentence level、aspect level。

**document level**：将观点鲜明的文本（如产品评论）分类为整体**pos/neg**的观点。**a single entity**。

**sentence level**：对文档内单独的语句进行分类。然而，单独的语句不能假定为观点鲜明的。**subjectivity classification** 判断主观/客观。若主观，判断**pos/neg**。三分类问题：**pos/neg/中立**。

**aspect level**：提取和总结人们对**实体**表达的意见和实体的方面/特征。由**aspect extraction**, **entity extraction**, 和 \**aspect sentiment classification \**几个子任务组成。

举例：“iPhone的音质很棒，但是它的电池很烂”，实体提取应该把“iPhone”识别为实体，方面提取应该识别“音质”和“电池”是两个方面，方面情感分类应将对iPhone音质的情感分为正面，对iPhone电池的情感分为负面。

除此之外，情感分析还做emotion analysis, sarcasm detection, multilingual sentiment analysis这些方向。

* **Document Level Sentiment Classification**

determine whether the document (e.g., a full online review) conveys an overall positive or negative opinion.

文本表示：BoW（忽略词序，几乎无法编码单词语义） - Bag-of-N-Grams（数据稀疏/高维）- word embedding。除此，也可以直接用 BoW 学习密集文档向量。

情感分类：传统监督学习/神经网络。利用数据特征（如产品评论）。

| 工作 | 方法 |
| --- | --- |
| Moraes et al.34 | 使用Bow向量，发现在文档级别上ANN优于SVM |
| 35 | Bow缺点太多，提出Paragraph Vector，通过预测上下文单词来学习的。 |
| 36 | 研究情感分类的领域适应问题。Bow学习密集向量。基于Stacked Denoising Autoencoder。同时使用有/无标注数据。 |
| 37 | 引入半监督的autoencoder。修改损失函数达到特定于任务的表示。 |
| 38 | BoW-CNN。Seq-CNN，通过连接多个单词的one-hot向量来保持单词的顺序信息。 |
| 39 | 提出考虑句子关系的NN来学习文档表示。CNN/LSTM学习句子表示，GRU编码句子语义及内在关系。 |
| 40 | 在review分类中同时考虑user和product信息，捕获全局线索。 |
| 41 42 | 考虑user和product信息，使用词和句级别的注意力机制。 |
| 43 | 提出 cached LSTM 模型。模型记忆被划分为不同遗忘率的几组。low rate：全局语义特征。high rate：局部特征。 |
| 44 | 2层注意力机制。词/句子。 |
| 45 | 视为一个机器理解问题。多层交互注意力机制。文档和伪aspect问题交互，学习aspect-aware的文档表示。 |
| 46 | cross-lingual 情感分类。将情感信息从资源丰富的语言适配到资源贫乏的语言。使用2个层次结构化的attention-based LSTM。 |
| 47 | 迁移学习，cross-domain情感分类，使用对抗记忆网络。2个任务：情感分类、领域二分类。 |

* **Sentence level sentiment classification**

determine the sentiment expressed in a single given sentence.

subjectivity classification：主观/客观

polarity classification：积极/消极

在现有深度学习模型中，一般视为**三分类任务**：positive、neural、negative。

同文档，**句子表示**也很重要。句子通常较短，可使用句法和语义信息(如parse trees, opinion lexicons, and pos tags)。其他信息，如评论评分、社会关系和跨域信息。

早期使用解析树，现在用CNN/RNN加上embedding。

| 工作 | 方法 |
| --- | --- |
| 49 50 51 | 使用各种RecursiveNN去得到embedding。 |
| 52 53 | 提出Dynamic CNN，使用dynamic K-Max pooling。CNN-rand/static/non-static/-multichannel。 |
| 54 | Character to Sentence CNN，没有embedding，直接把字符one-hot，CNN训练。 |
| 55 56 57 | LSTM。BiLSTM。CNN-LSTM。 |
| 58 | joint CNN & RNN，短文本。利用CNN生成的粗粒度局部特征和RNN学习的长距离相关性。 |
| 59 | LSTM & CNN。word2vec 和 linguistic embeddings。 |
| 60 | 将句法知识（如pos）编码在一个树状结构的LSTM中增强表示。 |
| 61 | 模型融合。金融微博和新闻的细粒度情感分类。 |
| 62 | 弱监督CNN。2步：用总体review rating学习句子表示，labels微调 |
| 63 | 上下文敏感词汇，加权和模型，BiLSTM，学习情感强度。 |
| 64 | 学习generalized sentence embeddings for cross-domain。分别设计CNN，从已/未标记的数据中联合学习两个隐藏的特征表示。 |
| 65 | recurrent random walk network。opinionated tweets。利用用户的推文和社交关系。 |
| 66 | CNN。从人类阅读的眼动数据中自动提取认知特征。。。 |
| 67 | linguistically regularized LSTM。整合 情感词典/否定词/强度词。 |

* **Aspect level sentiment classification**

aspect level sentiment classification considers both the sentiment and the target information, as a sentiment always has a target.

具有挑战，很难对目标与其周围上下文词的语义关联性进行建模。

三个重要任务：

* + **目标上下文**的表示：用之前的方法。
  + 与其上下文正确交互的**目标**表示：学习目标嵌入，类似词嵌入。
  + 识别指定目标的重要情感上下文(词)：常用注意力机制，没有主流方法。

| 工作 | 方法 |
| --- | --- |
| 68 | Adaptive RecursiveNN。 |
| 69 |  |

* **Aspect extraction and categorization**

要进行aspect level的情感分类，需要有方面（或目标），可以手动给定或自动提取。

举例：“图像非常清晰”，“图像”是一个aspect term（or sentiment target）。方面分类：将相同的方面表达式分组到一个类别中。方面术语“图像”、“照片”和“图片”可以分组为一个方面类别，称为图像。

##### 其他情感分析相关任务

* Opinion expression extraction：意见表达提取，旨在识别句子或文档中的情感表达。
* Sentiment composition：
* Opinion holder extraction：是识别谁持有意见的任务。
* Temporal opinion mining：时态意见挖掘，预测未来观点。随着时间推移，观点会发生改变。
* sentiment analysis with word embedding：嵌入词的情感分析。词嵌入在情感分析模型中起着重要作用。
* Sarcasm analysis：讽刺分析。
* Emotion analysis：情绪分析。主要的情绪包括爱、喜悦、惊讶、愤怒、悲伤和恐惧。情绪的概念与情感密切相关。
* Multimodal data for sentiment analysis：多模态数据的情感分析。如承载文本、视觉和听觉信息的数据，已经被用来帮助情感分析，提供额外的情感信号。
* resource-poor languages and multilingual sentiment analysis：资源贫乏语言与多语言情感分析。
* 其他：Sentiment Intersubjectivity、Lexicon Expansion、Financial Volatility Prediction、Opinion Recommendation、Stance Detection

## 1.4 论文组织结构

第一章、系统开发概述。介绍该系统的开发环境，包括系统设计背景，系统发展前景，系统设计的目标和意义,也从技术方面、经济方面、操作方面进行了系统的可行性分析。

第二章、开发技术及工具介绍。由于该系统前后台由两个不同的框架开发，所以用到的开发技术和开发工具也比较丰富，包括

termcolor

pandas

tensorflow

torch>=0.4.0

torchvision>=0.2.1

numpy

nltk

keras

scikit-learn

scipy>=1.0.1

scikit-image>=0.13.0

matplotlib>=1.5.1

tqdm>=4.28.1

jupyter

jellyfish

Flask==1.1.2

flask-wtf==0.14.2。

第三章、情感分析系统系统分析与设计。从系统需求分析、系统设计效果和系统设计角色功能三方面进行介绍。

第四章、情感分析系统系统详细设计。描述了系统结构图，系统概念模型，同时详细介绍了各数据表的设计结构。

第五章、情感分析系统系统应用程序设计。叙述了各个页面的功能，相关的操作界面和主要代码。

# 

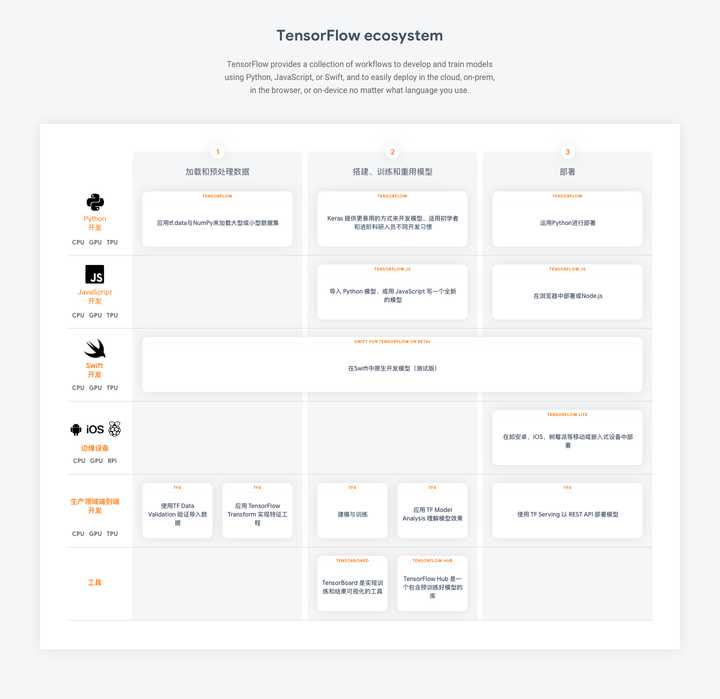
# 2 开发技术及工具介绍

## 2.1 tensorflow开发框架

TensorFlow 是端到端的开源机器学习平台。提供全面，灵活的专业工具，使个人开发者轻松创建机器学习应用，助力研究人员推动前沿技术发展，支持企业建立稳健的规模化应用。



从2015年发布以来，TensorFlow在全球已有4100万的下载。作为全球使用广泛的机器学习框架，TensorFlow 已经度过三周年，TensorFlow 逐渐成为了端到端的成熟平台，有着完整的生态体系。随着 TensorFlow 2.0 的发布，标志着 TensorFlow 全新时代到来，它更为简单易用，更为灵活强大，更为生产环境可用，希望 TensorFlow 可以成为适合所有用户的机器学习平台，可以帮助每一个人利用机器学习解决现实世界的难题。



机器学习主要分为训练和部署两个步骤。

在训练阶段，TensorFlow不仅支持Python，更提供对Swift语言和JS语言的支持，你可以选择你熟悉的语言来进行开发。

在部署阶段，TensorFlow模型可以跑在不同的平台，支持服务器端部署的TensorFlow Serving， 支持Android，iOS和嵌入式设备等端侧平台部署的TensorFlow Lite，支持浏览器和Node 服务器部署的TensorFlow.js，以及包括C语言，Java 语言，Go语言，C#语言，Rust和R等多种语言

## 2.2 keras

[Keras](https://keras.io/" \t "_blank) 是由纯 Python 编写的神经网络库，专注于深度学习，运行在 TensorFlow 或 Theano 之上。  
[TensorFlow](https://www.tensorflow.org/" \t "_blank) 和 [Theano](http://deeplearning.net/software/theano/) 是当前比较流行的两大深度学习库，但是对初学者来说相对有些复杂。  
Keras 使用简单，结构清晰，底层计算平台可基于 TensorFlow 或 Theano 之上，功能强大。  
Keras 可运行于 Python 2.7 或 3.5 环境，完美结合于 GPU 和 CPU，基于 MIT license 发布。  
Keras 由 Google 工程师 [François Chollet](https://www.linkedin.com/in/fchollet) 开发和维护。  
以下是 Keras 的设计原则：

* 模块化（Modularity）：一个模型（model）可以理解为一个独立的序列（sequence）或图，模型之间是相互独立的，可以自由组合。
* 极简主义（Minimalism）：每个模块都应该尽量的简洁。每一段代码都应该在初次阅读时都显得直观易懂。没有黑魔法，因为它将给迭代和创新带来麻烦。
* 易扩展（Extensibility）：添加新模块超级简单的容易，只需要仿照现有的模块编写新的类或函数即可。创建新模块的便利性使得Keras更适合于先进的研究工作。
* 与 Python 协作：Keras 没有单独的模型配置文件类型（作为对比，caffe有），模型由 Python 代码描述，使其更加紧凑，易调试和易扩展。

## 2.3 NumPy

NumPy是Python中科学计算的基础包。它是一个Python库，提供多维数组对象，各种派生对象（如掩码数组和矩阵），以及用于数组快速操作的各种API，有包括数学、逻辑、形状操作、排序、选择、输入输出、离散傅立叶变换、基本线性代数，基本统计运算和随机模拟等等。

NumPy包的核心是 ndarray 对象。它封装了python原生的同数据类型的 n 维数组，为了保证其性能优良，其中有许多操作都是代码在本地进行编译后执行的。

NumPy数组 和 原生Python Array（数组）之间有几个重要的区别：

* NumPy 数组在创建时具有固定的大小，与Python的原生数组对象（可以动态增长）不同。更改ndarray的大小将创建一个新数组并删除原来的数组。
* NumPy 数组中的元素都需要具有相同的数据类型，因此在内存中的大小相同。 例外情况：Python的原生数组里包含了NumPy的对象的时候，这种情况下就允许不同大小元素的数组。
* NumPy 数组有助于对大量数据进行高级数学和其他类型的操作。通常，这些操作的执行效率更高，比使用Python原生数组的代码更少。
* 越来越多的基于Python的科学和数学软件包使用NumPy数组; 虽然这些工具通常都支持Python的原生数组作为参数，但它们在处理之前会还是会将输入的数组转换为NumPy的数组，而且也通常输出为NumPy数组。换句话说，为了高效地使用当今科学/数学基于Python的工具（大部分的科学计算工具），你只知道如何使用Python的原生数组类型是不够的 - 还需要知道如何使用 NumPy 数组。

## 2.4 Jupyter Notebook

Jupyter Notebook是基于网页的用于交互计算的应用程序。其可被应用于全过程计算：开发、文档编写、运行代码和展示结果。

简而言之，Jupyter Notebook是以网页的形式打开，可以在网页页面中直接编写代码和运行代码，代码的运行结果也会直接在代码块下显示的程序。如在编程过程中需要编写说明文档，可在同一个页面中直接编写，便于作及时的说明和解释。

网页应用即基于网页形式的、结合了编写说明文档、数学公式、交互计算和其他富媒体形式的工具。简言之，网页应用是可以实现各种功能的工具

文档即Jupyter Notebook中所有交互计算、编写说明文档、数学公式、图片以及其他富媒体形式的输入和输出，都是以文档的形式体现的。

这些文档是保存为后缀名为.ipynb的JSON格式文件，不仅便于版本控制，也方便与他人共享。此外，文档还可以导出为：HTML、LaTeX、PDF等格式

Jupyter Notebook的主要特点

① 编程时具有语法高亮、缩进、tab补全的功能。

② 可直接通过浏览器运行代码，同时在代码块下方展示运行结果。

③ 以富媒体格式展示计算结果。富媒体格式包括：HTML，LaTeX，PNG，SVG等。

④ 对代码编写说明文档或语句时，支持Markdown语法。

⑤ 支持使用LaTeX编写数学性说明。

# 3 情感分析系统实验

# 3.1 实验介绍和数据读取

我们希望通过深度学习进行情感预测，尝试超越传统机器学习算法的情感分析表现。

我们主要探索下列方面：

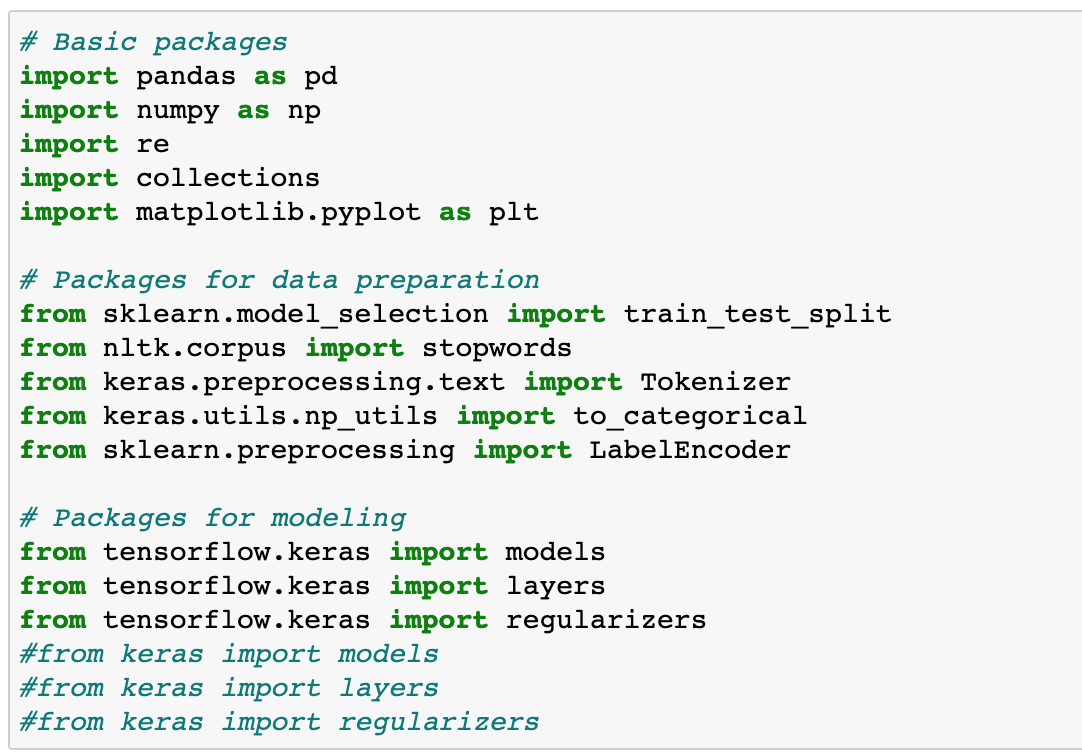
\*使用适应深度学习模型

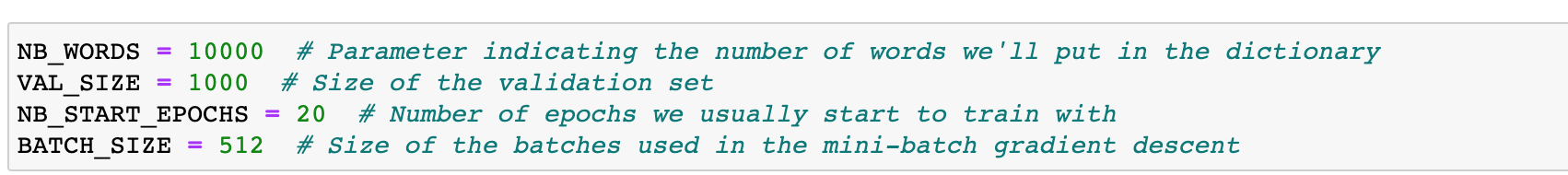
\*识别并处理过度拟合

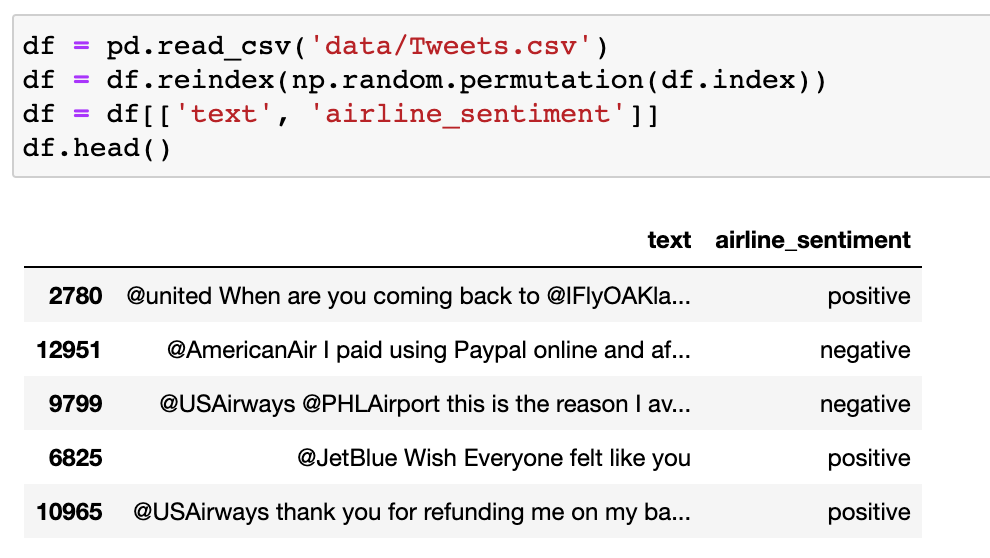
\*使用单词嵌入

\*建立在预先训练的模型上

我们使用推文数据读取csv并执行随机化数据。 好的做法是在训练和测试集之间拆分之前先对数据进行混洗。 这样，情感等级在训练集和测试集上平均分配。







我们只将text列作为输入，而airen\_sentiment列作为输出目标。

# 3.2 Data preparation

数据清洗是指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查[数据一致性](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E4%B8%80%E8%87%B4%E6%80%A7/9895547" \t "_blank)，处理无效值和[缺失值](https://baike.baidu.com/item/%E7%BC%BA%E5%A4%B1%E5%80%BC/2336661" \t "_blank)等。与问卷审核不同，录入后的数据清理一般是由计算机而不是人工完成。数据清洗(Data cleaning)– 对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

数据清洗从名字上也看的出就是把“脏”的“洗掉”，指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。因为数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。我们要按照一定的规则把“脏数据”“洗掉”，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。数据清洗是与问卷审核不同，录入后的数据清理一般是由计算机而不是人工完成 。

一致性检查

一致性检查(consistency check)是根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发现超出正常范围、逻辑上不合理或者相互矛盾的数据。例如，用1-7级量表测量的变量出现了0值，体重出现了负数，都应视为超出正常值域范围。SPSS、SAS、和Excel等计算机软件都能够根据定义的取值范围，自动识别每个超出范围的变量值。具有逻辑上不一致性的答案可能以多种形式出现：例如，许多调查对象说自己开车上班，又报告没有汽车；或者调查对象报告自己是某品牌的重度购买者和使用者，但同时又在熟悉程度量表上给了很低的分值。发现不一致时，要列出问卷序号、记录序号、变量名称、错误类别等，便于进一步核对和纠正。

无效值和缺失值的处理

由于调查、编码和录入误差，数据中可能存在一些无效值和缺失值，需要给予适当的处理。常用的处理方法有：估算，整例删除，变量删除和成对删除。

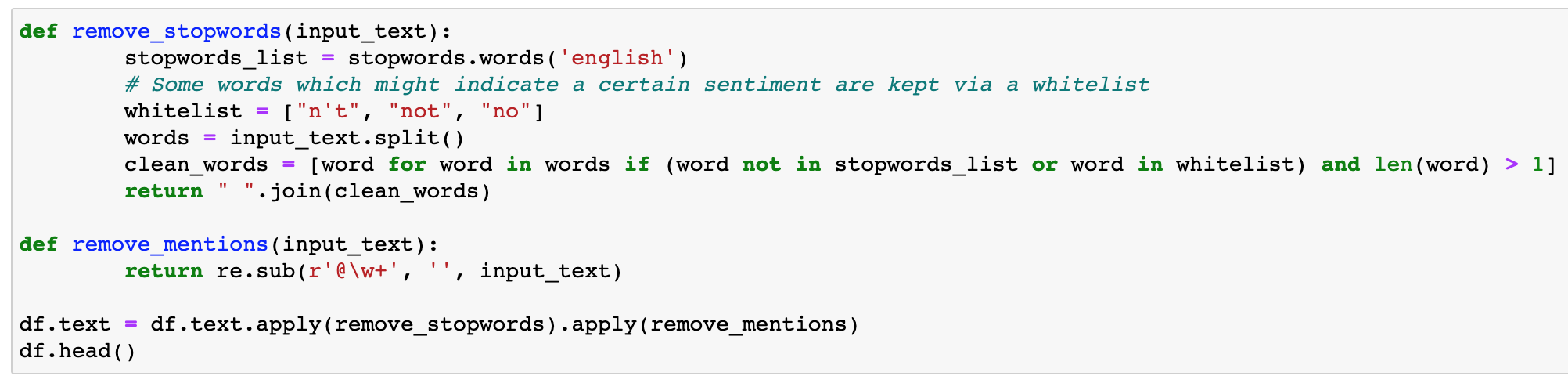
估算(estimation)。最简单的办法就是用某个变量的样本均值、中位数或众数代替无效值和缺失值。这种办法简单，但没有充分考虑数据中已有的信息，误差可能较大。另一种办法就是根据调查对象对其他问题的答案，通过变量之间的相关分析或逻辑推论进行估计。例如，某一产品的拥有情况可能与家庭收入有关，可以根据调查对象的家庭收入推算拥有这一产品的可能性。

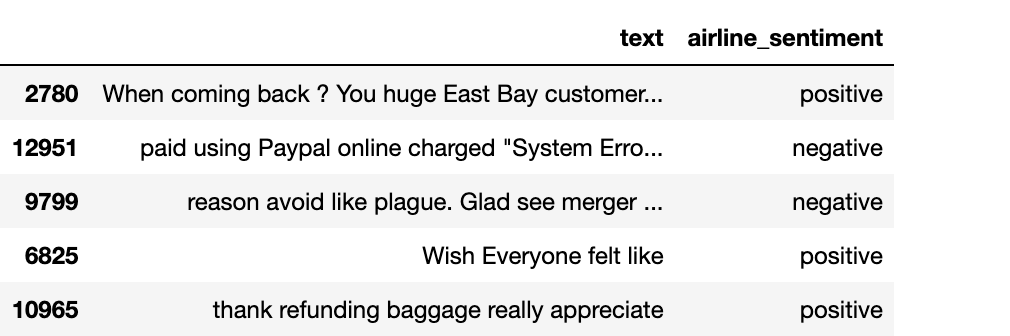
整例删除(casewise deletion)是剔除含有缺失值的样本。由于很多问卷都可能存在缺失值，这种做法的结果可能导致有效样本量大大减少，无法充分利用已经收集到的数据。因此，只适合关键变量缺失，或者含有无效值或缺失值的样本比重很小的情况。

变量删除(variable deletion)。如果某一变量的无效值和缺失值很多，而且该变量对于所研究的问题不是特别重要，则可以考虑将该变量删除。这种做法减少了供分析用的变量数目，但没有改变样本量。

成对删除(pairwise deletion)是用一个特殊码(通常是9、99、999等)代表无效值和缺失值，同时保留数据集中的全部变量和样本。但是，在具体计算时只采用有完整答案的样本，因而不同的分析因涉及的变量不同，其有效样本量也会有所不同。这是一种保守的处理方法，最大限度地保留了数据集中的可用信息。

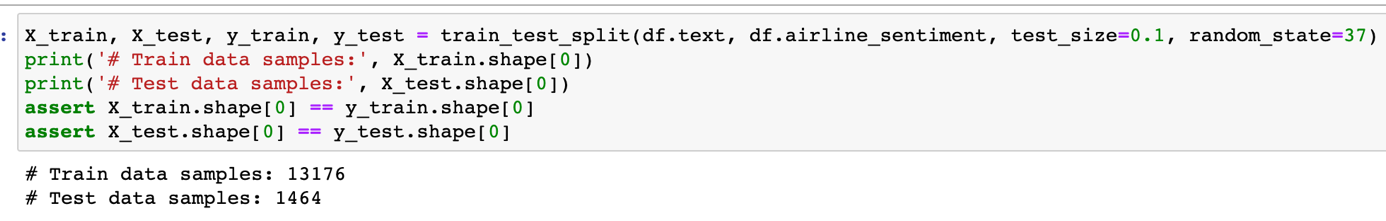
采用不同的处理方法可能对分析结果产生影响，尤其是当缺失值的出现并非随机且变量之间明显相关时。因此，在调查中应当尽量避免出现无效值和缺失值，保证数据的完整性





训练数据测试数据集拆分

对模型性能的评估需要在单独的测试集中进行。 这样，我们可以估计模型的概括程度。 这是通过scikit-learn的train\_test\_split方法完成的。



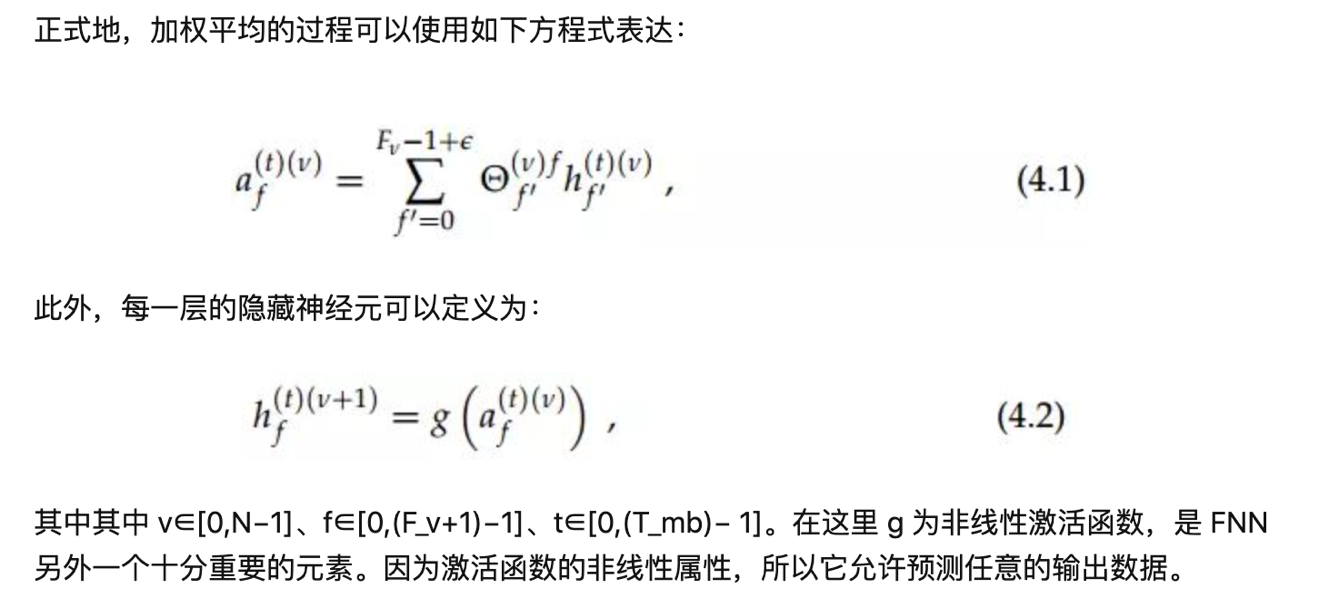
# 3.3 Deep learning模型

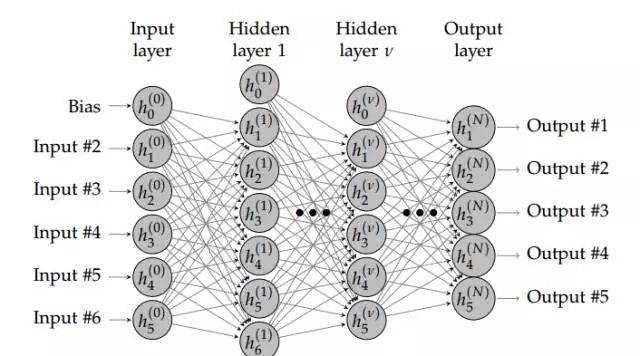
**前馈神经网络**

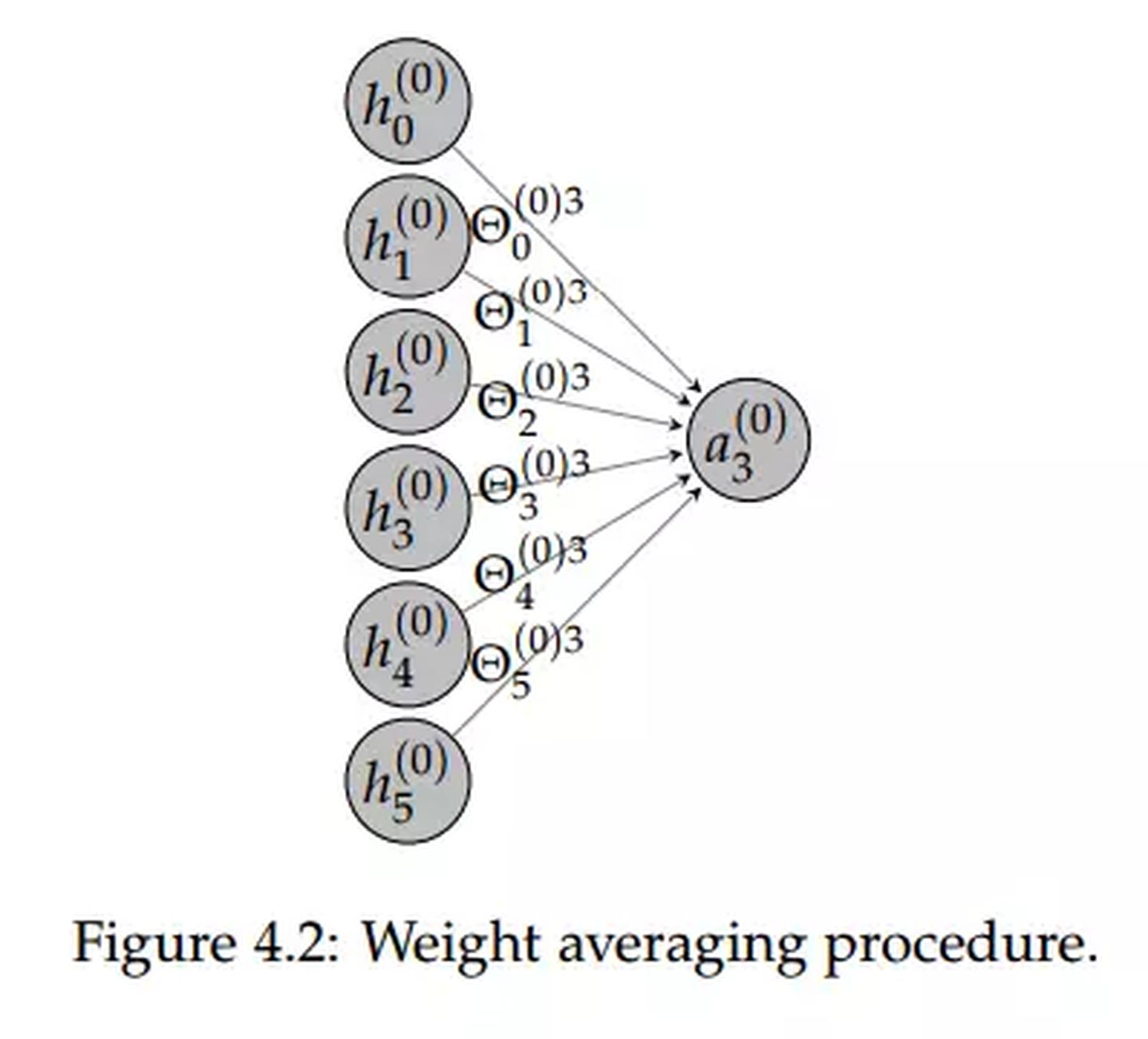
常规前馈神经网络（FNN）不考虑输入数据可能具备的任何特定结构。尽管如此，它仍是非常强大的机器学习工具，尤其是与先进的正则化技术一起使用时。这些正则化技术帮助解决人们处理「深度」网络时遇到的训练问题：神经网络有大量隐藏层，隐藏层非常难以训练（梯度消失和过拟合问题）。有 N + 1 层（N − 1 个隐藏层）的神经网络。浅层网络架构仅使用一个隐藏层。深度学习需要使用多个隐藏层，通常包含同样数量的隐藏神经元。数量大约是输入和输出变量数量的平均值。

FNN 由一个输入层、一个（浅层网络）或多个（深层网络，因此叫作深度学习）隐藏层，和一个输出层构成。每个层（除输出层以外）与下一层连接。这种连接是 FNN 架构的关键，具有两个主要特征：加权平均值和激活函数。

加权平均过程，即将前一层给神经元的激励值和对应的权重矩阵相乘而得出后一个神经元的输入值，这一过程展示在下图中，我们可以说前一层神经元的加权和就是后一层神经元的输入。





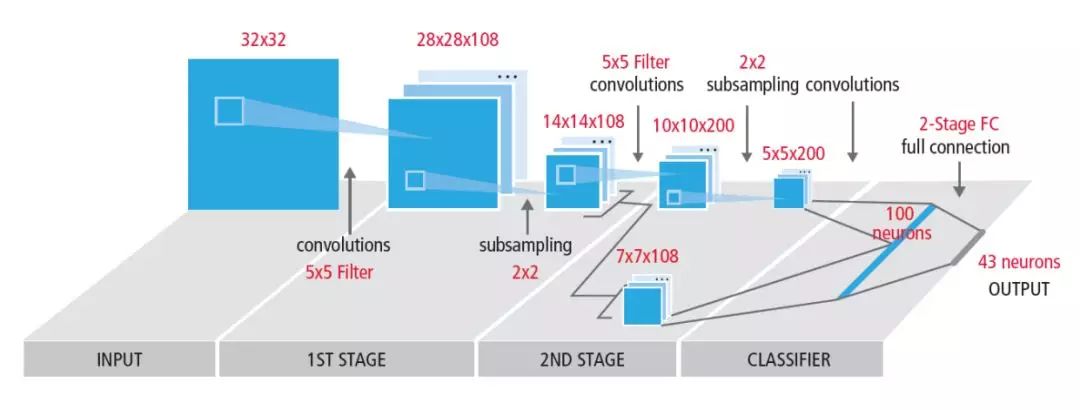


**卷积神经网络**

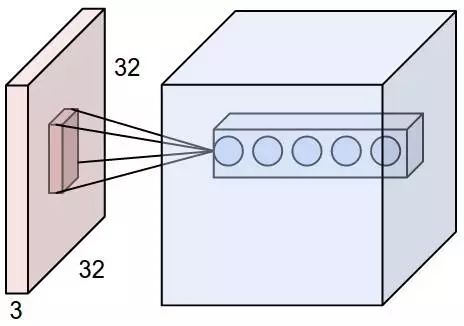
CNN 非常擅长处理图像数据，如下图所示，它们一般由若干个卷积和池化操作组成，通常跟随着一个或多个全连接层（与传统的 FNN 层相似）。

相比于全连接神经网络，卷积网络每一个单元都只会和上一层部分单元相连接。一般每个卷积层的单元都可以组织成一个三维张量，即矩阵沿第三个方向增加一维数据。例如 Cifar-10 数据集的输入层就可以组织成 32×32×3 的三维张量，其中 32×32 代表图片的尺寸或像素数量，而 3 代表 RGB 三色通道。

卷积神经网络中最重要的就是卷积层，卷积层试图将神经网络中的每一小块进行更加深入的分析，从而得出抽象程度更高的特征。一般来说通过卷积层处理的神经元结点矩阵会变得更深，即神经元的组织在第三个维度上会增加。



为了理解卷积层，下图展示了卷积核或滤波器（filter）将当前层级上的一个子结点张量转化为下一层神经网络上的一个长和宽都为 1，深度不限的结点矩阵。下图输入是一个 32×32×3 的张量，中间的小长方体为卷积核，一般可以为 3×3 或 5×5 等，且因为要计算乘积，那么卷积核的第三个维度必须和其处理的图像深度（即输入张量第三个维度 3）相等。最右边的矩形体的深度为 5，即前面使用了五个卷积核执行卷积操作。这五个卷积核有不同的权重，但每一个卷积层使用一个卷积核的权重是一样的，所以下图五层特征中每一层特征都是通过一个卷积核得出来的，也就是该层共享了权重。



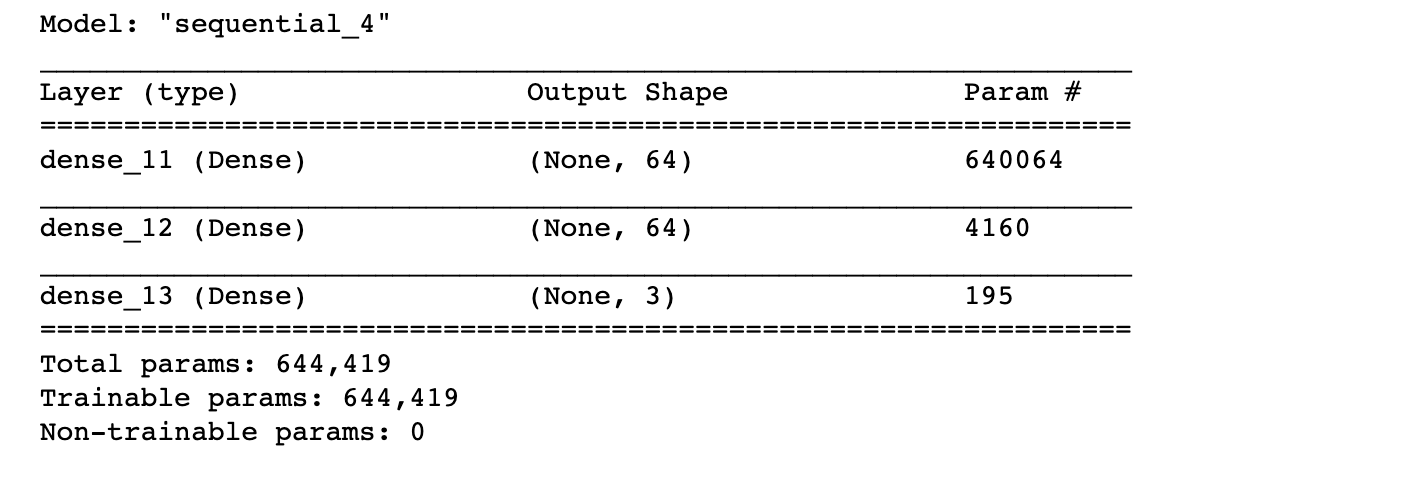
# 我们从一个模型开始，该模型具有2个紧密连接的层，其中包含64个隐藏元素。 第一层的input\_shape等于我们在字典中允许的单词数，并且我们为其创建了一个热编码特征。

# 由于我们需要预测3个不同的情感类别，因此最后一层包含3个隐藏元素。 softmax激活功能可确保三个概率的总和为1。

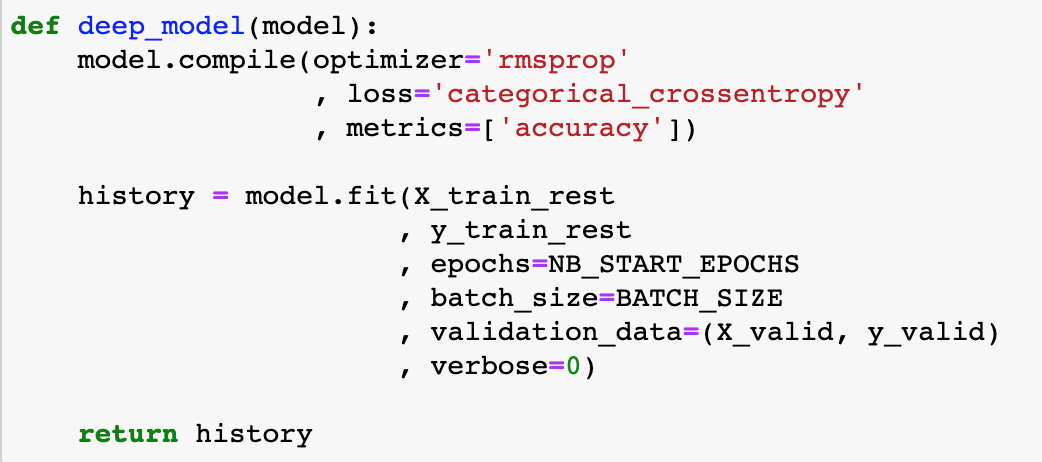
# 在第一层中，我们需要估计640064权重。 这取决于（nb个输入\* nb个隐藏元素）+ nb个偏置项，或者（10000 x 64）+ 64 = 640064

# 在第二层中，我们估计（64 x 64）+ 64 = 4160权重

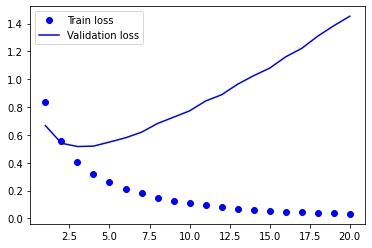
# 在最后一层，我们估计（64 x 3）+ 3 = 195个权重



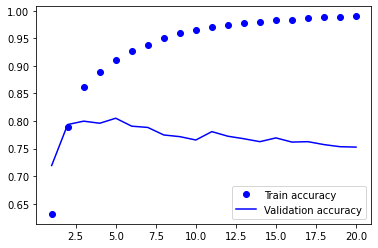
由于该项目是一个多类别的单标签预测，因此我们将categorical\_crossentropy用作损失函数，将softmax用作最终激活函数。 我们将模型拟合到其余火车数据上，并在验证集上进行验证。 我们运行了预定的时期，然后将看到模型何时开始过拟合。



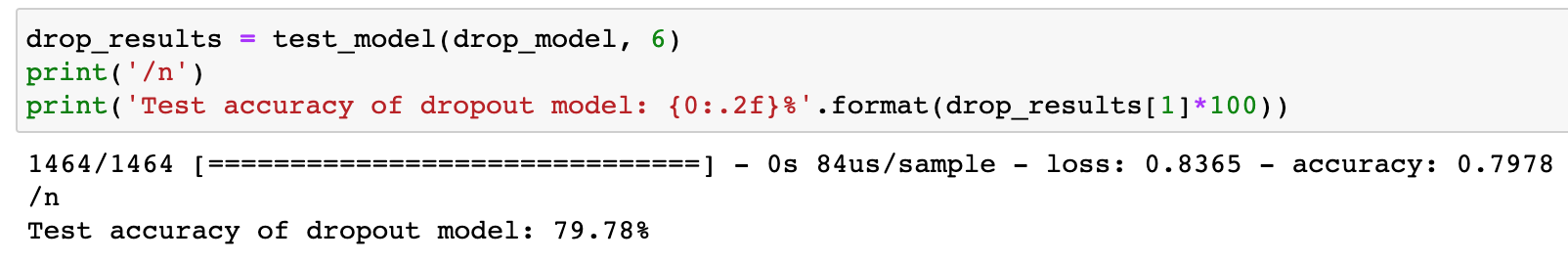
从这里我们可以看到，验证损失从第4代开始增加。训练损失继续降低，这是正常的，因为模型经过训练以尽可能拟合训练数据。



就像验证损失一样，验证准确性会在较早的时期达到顶峰。 之后，它会稍微下降。 综上所述，我们可以说该模型从时代4开始就开始过拟合。



在测试集结果：



# 结论

正如我们在上面看到的那样，具有dropout层的模型在测试数据上表现最佳。 但是，这比我们以前的内核中使用LogisticRegression和Countvectorizer实现的结果持平。 但是在那里，输入数据的变换与此处有所不同。

深度学习机器学习模型在情感分析，因为其比文本提供了更多的信息。深度学习模型把输入映射到一些特征空间，来自多模态数据的不同形式的输入也可以被这些模型投射到一些联合潜在空间或表征。因此，使用深度学习模型处理多模态数据的趋势不断增长

参考文献

# [1] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors.[J]. 1986, 323(6088):399-421.

# [2] Cover T M, Hart P E. Nearest neighbor pattern classification. IEEE Trans Inf Theory IT-13(1):21-27[J]. IEEE Transactions on Information Theory, 1967, 13(1):21-27.

# [3] Daral N. Histograms of Oriented Gradients for Human Detection[J]. Proc. of CVPR, 2005, 2005.

# [3.1] Histograms of Oriented Gradients for Human Detection. Dalai,N,B.Triggs. Computer Vision and Pattern Recognition, 2005.CVPR 2005.IEEE Computer Society Conference on . 2005

# [4] Kazemi V, Sullivan J. One Millisecond Face Alignment with an Ensemble of Regression Trees[C] Computer Vision and Pattern Recognition. IEEE, 2014:1867-1874.

# [5] David J. Hand and Robert J. Till（ 2001）. A Simple Generalization of the Area Under the ROC Curve for Multiple Class Classification Problems . Machine Learning , 45(2), 171 – 186 .

# 周志华，机器学习，清华大学出版社，2016

# 李航，统计学习方法，清华大学出版社，2012

# Scikit-learn，https://scikit-learn.org/stable/index.html

# Qcon 2017 feature engineering by Gabriel Moreira

# Thomas M.Cover, JoyA. Thomas. Elementsof InformationTheory. 2006

# Christopher M.Bishop. Pattern Recognition and Machine Learning. Springer-Verlag. 2006

致谢

完成这个系统，我得到了许多人的帮助：

首先，我要感谢母校给了我学习的机会，感谢计算机学院给我提供了一个学习计算机专业知识的环境，使我度过了四年充实的学习生活。

其次，感谢学校带课的老师们，他们在传承者知识，接力者培养我，使我打下了扎实的计算机技术功底，在系统开发过程中发挥了巨大作用。

另外，感谢朋友、同学在毕业设计期间基于我的鼓励和帮助，正是有他们与我同行，我再用有坚定的决心，解决一个个难题，完成系统的设计。

最后，我要特别感谢何亨老师在我整个系统开发过程中的悉心指导，他认真负责、一丝不苟的工作作风深深的感染了我，他的技术思想给了我无尽的启迪。

感谢所有被本文引用或参考的文章的作者，你们富有开拓性的工作是本文进行深入研究的基础。