

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 7 月 9 日 (09.07.2020)



(10) 国际公布号
WO 2020/140403 A1

(51) 国际专利分类号:
G06F 16/35 (2019.01)

(21) 国际申请号: PCT/CN2019/092531

(22) 国际申请日: 2019 年 6 月 24 日 (24.06.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910007705.9 2019 年 1 月 4 日 (04.01.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 金戈(JIN, Ge); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安

金融中心 23 楼, Guangdong 518000 (CN)。 徐亮(XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市精英专利事务所(SHENZHEN TALENT PATENT SERVICE); 中国广东省深圳市福田区深南中路 6009 号绿景广场 B 栋 20 层 B, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: TEXT CLASSIFICATION METHOD AND APPARATUS, COMPUTER DEVICE AND STORAGE MEDIUM

(54) 发明名称: 文本分类方法、装置、计算机设备及存储介质

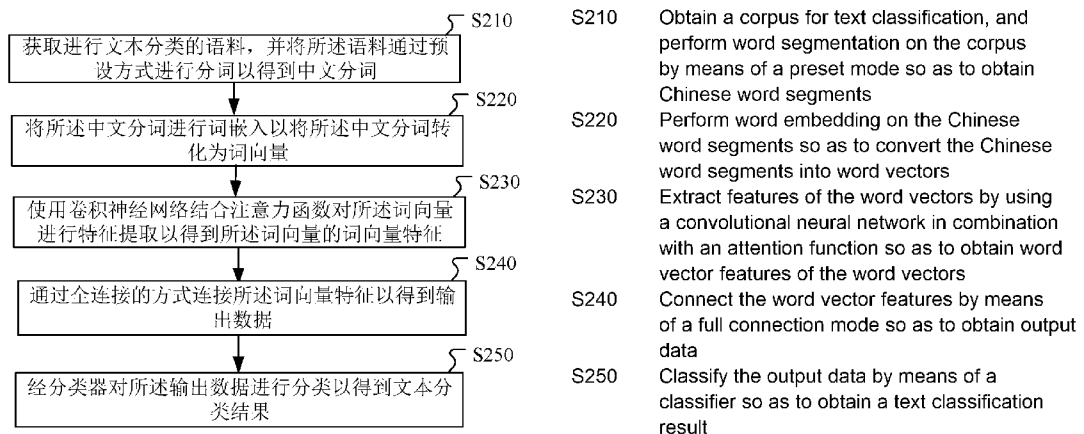


图 2

(57) Abstract: A text classification method and apparatus, a computer device and a computer readable storage medium, belonging to the technical field of text classification. The method comprises: obtaining a corpus for text classification, and performing word segmentation on the corpus by means of a preset mode so as to obtain Chinese word segments (S210); performing word embedding on the Chinese word segments so as to convert the Chinese word segments into word vectors (S220); extracting features of the word vectors by using a convolutional neural network in combination with an attention function so as to obtain word vector features of the word vectors (S230); connecting the word vector features by means of a full connection mode so as to obtain output data (S240); and classifying the output data by means of a classifier so as to obtain a text classification result (S250).

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

- (84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种文本分类方法、装置、计算机设备及计算机可读存储介质, 属于文本分类技术领域, 该方法获取进行文本分类的语料, 并将所述语料通过预设方式进行分词以得到中文分词 (S210), 将所述中文分词进行词嵌入以将所述中文分词转化为词向量 (S220), 使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征 (S230), 通过全连接的方式连接所述词向量特征以得到输出数据 (S240), 经分类器对所述输出数据进行分类以得到文本分类结果 (S250)。

文本分类方法、装置、计算机设备及存储介质

本申请要求于2019年1月4日提交中国专利局、申请号为201910007705.9、申请名称为“文本分类方法、装置、计算机设备及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及文本分类技术领域，尤其涉及一种文本分类方法、装置、计算机设备及计算机可读存储介质。

背景技术

传统基于卷积神经网络的文本分类模型，也就是 TextCNN，英文为 Text Convolutional Neural Network，一般包括输入层、词嵌入层、卷积层、池化层、连接层及输出层，通过各层对文本语料进行逐层处理以实现对文本分类。但有时由于进行文本分类的语料数据比较大而使得 TextCNN 模型进行文本分类效率较低。

发明内容

本申请实施例提供了一种文本分类方法、装置、计算机设备及计算机可读存储介质，能够解决传统技术中文本分类效率比较低的问题。

第一方面，本申请实施例提供了一种文本分类方法，所述方法包括：获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词；将所述中文分词进行词嵌入以将所述中文分词转化为词向量；使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征；通过全连接的方式连接所述词向量特征以得到输出数据；经分类器对所述输出数据进行分类以得到文本分类结果。

第二方面，本申请实施例还提供了一种文本分类装置，其中，所述装置包括：获取单元，用于获取进行文本分类的语料，并将所述语料通过预设方式进

行分词以得到中文分词；转化单元，用于将所述中文分词进行词嵌入以将所述中文分词转化为词向量；提取单元，用于使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征；连接单元，用于通过全连接的方式连接所述词向量特征以得到输出数据；分类单元，用于经分类器对所述输出数据进行分类以得到文本分类结果。

第三方面，本申请实施例还提供了一种计算机设备，其包括存储器及处理器，所述存储器上存储有计算机程序，所述处理器执行所述计算机程序时实现所述文本分类方法。

第四方面，本申请实施例还提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时使所述处理器执行所述文本分类方法。

附图说明

为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1为本申请实施例提供的文本分类方法的应用场景示意图；

图2为本申请实施例提供的文本分类方法的流程示意图；

图3为本申请实施例提供的文本分类方法中词向量示意图；

图4为本申请实施例提供的文本分类方法的另一个流程示意图；

图5为图4中的文本分类方法对应的模型示意图；

图6为本申请实施例提供的文本分类方法对应的另一个模型示意图；

图7为本申请实施例提供的文本分类方法对应的第三个模型示意图；

图8为本申请实施例提供的文本分类装置的示意性框图；

图9为本申请实施例提供的文本分类装置的另一个示意性框图；以及

图10为本申请实施例提供的计算机设备的示意性框图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

请参阅图 1，图 1 为本申请实施例提供的文本分类方法的应用场景示意图。所述应用场景包括：

(1) 终端。图 1 所示终端上安装有应用程序，研发人员通过终端实现执行文本分类方法的步骤，所述终端可以为笔记本电脑、平板电脑或者台式电脑等电子设备，图 1 中所示的终端应用环境也可以更换为服务器等计算机设备。若图 1 中的应用环境为服务器，服务器可以为服务器集群或者云服务器。服务器集群又可以采用分布式系统，分布式系统的服务器又可以包括主服务器和从服务器，以使主服务器使用获得的语料执行文本分类方法的步骤。

图 1 中的各个主体工作过程如下：终端获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词，将所述中文分词进行词嵌入以将所述中文分词转化为词向量，使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征，通过全连接的方式连接所述词向量特征以得到输出数据；经分类器对所述输出数据进行分类以得到文本分类结果。

需要说明的是，图 1 中仅仅示意出台式电脑作为终端，在实际操作过程中，终端的类型不限于图 1 中所示，所述终端还可以为手机、笔记本电脑或者平板电脑等电子设备，上述文本分类方法的应用场景仅仅用于说明本申请技术方案，并不用于限定本申请技术方案。

图 2 为本申请实施例提供的文本分类方法的示意性流程图。该文本分类方法应用于图 1 中的终端中以完成文本分类方法的全部或者部分功能。

请参阅图 2，图 2 是本申请实施例提供的文本分类方法的流程示意图。如图 2 所示，该方法包括以下步骤 S210-S250：

S210、获取进行文本分类的语料，并将所述语料通过预设方式进行分词以

得到中文分词。

其中，本申请实施例中的文本分类是指基于卷积神经网络分类模型对文本进行的分类。基于卷积神经网络的文本分类模型，英文为 Text Convolutional Neural Network，简写为 TextCNN，称为 TextCNN 网络结构或者 TextCNN 网络模型，TextCNN 是用来做文本分类的卷积神经网络，也就是利用卷积神经网络对文本进行分类。

分词，是指对中文文本进行分词，英文为 Chinese Word Segmentation，指的是将一个汉字序列切分成一个个单独的词，分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。对中文文本分类需要先对中文文本分词，对中文文本分词，有很多开源的中文分词工具，例如最常用的 Jieba 分词，还有 word 分词及盘古分词等。分词还包括做进一步的处理，去除掉一些高频词汇和低频词汇，去掉一些无意义的符号等。具体地，在使用 TextCNN 模型将文本进行分类前，需要将待分类文本进行预处理以将获得的待分类语料进行分词，从而获得中文分词，将获得的中文分词进一步转化为词向量。

具体地，终端获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词，所述语料可以通过爬取网络上指定网站上的预设语料，爬取规则可以根据实际需要预先设置，比如，爬取规则为某一网页的语料，也可以是爬取的某一主体的相关语料。所述语料还可以是通过语料数据库提供的语料，比如某一网站积累的用户数据等。本申请实施例的应用场景为文本分类，比如文本还可以包括新闻标题分类，模型输入为文本词向量信息，输出为文本分类结果。

S220、将所述中文分词进行词嵌入以将所述中文分词转化为词向量。

其中，词嵌入，英文为 Word Embedding，是一种词的类型表示，具有相似意义的词具有相似的表示，是将词汇映射到实数向量的方法总称，词嵌入所在的结构层称为词嵌入层，或者简称为嵌入层，英文为 Embedding layer。词嵌入是一类技术，是指单个词在预定义的向量空间中被表示为实数向量，每个单词都映射到一个向量。请参阅图 3，图 3 为本申请实施例提供的文本分类方法中词向量示意图。如图 3 所示，假如在一个文本中包含“猫”“狗”及“爱情”等若

干单词,而这若干单词映射到向量空间中,“猫”对应的向量为(0.1, 0.2, 0.3),“狗”对应的向量为(0.2, 0.2, 0.4),“爱情”对应的映射为(-0.4, -0.5, -0.2)(本数据仅为示意)。像这种将文本 $X\{x_1, x_2, x_3, x_4, x_5 \cdots x_n\}$ 映射到多维向量空间 $Y\{y_1, y_2, y_3, y_4, y_5 \cdots y_n\}$, 这个映射过程就叫做词嵌入。之所以希望把每个单词都变成一个向量,目的还是为了方便计算,比如“猫”,“狗”,“爱情”三个词。对于我们人而言,可以知道“猫”和“狗”表示的都是动物,而“爱情”是表示的一种情感,但是对于机器而言,这三个词都是用 0 和 1 表示成二进制的字符串而已,无法对其进行计算。而通过词嵌入这种方式将单词转变为词向量,机器便可对单词进行计算,通过计算不同词向量之间夹角余弦值 $\cos$ 而得出单词之间的相似性,比如,在图 3 中,由于 $\cos \alpha < \cos \beta$,可知“猫”与“狗”更相似,猫与“爱情”差异较大。其中,两个向量之间的夹角可以通过如下方式计算:向量夹角的公式是 $\cos \theta = \text{向量 } a \cdot \text{向量 } b / |\text{向量 } a| \times |\text{向量 } b|$, 其中,注意是点乘,比如,在 Python 中可以使用 `Python.numpy` 来计算向量的夹角。

具体地,将文本语料通过分词得到中文分词后进而转化为预训练的词向量,也就是将输入的自然语言经过分词后编码成词向量,为预训练词向量准备。具体实施时,可以使用预训练好的词向量,也可以直接在训练 TextCNN 的过程中训练出一套词向量,不过使用预训练好的词向量比在训练 TextCNN 的过程中训练出一套词向量快 100 倍不止。如果使用预训练好的词向量,又分为 Static 方法和 No-static 方法,Static 方法是指在训练 TextCNN 过程中不再调节词向量的参数, No-static 方法在训练过程中调节词向量的参数,所以 No-static 方法的结果比 Static 方法的结果要好。

进一步地,按照预设数量批次的所述中文分词对嵌入层进行调节,其中,批次,英文为 Batch。也就是还可以不在每一个 Batch(批)中都调节 Embedding 层(嵌入层),而是每个 100 个 Batch 的中文分词调节一次嵌入层,这样可以减少训练的时间,又可以微调词向量。

更进一步地,使用预设词向量字典将所述中文分词进行词嵌入以将所述中文分词转化为词向量,也就是使用训练好的预设词向量字典将所述中文分词进

行词嵌入以将所述中文分词转化为词向量，相比将所述中文分词直接进行词嵌入，可以提高将所述中文分词转化为词向量的效率。也就是可以使用训练好的预设词向量字典将所述中文词向量对应的语料进行词嵌入以将所述语料转化为词向量。在一个实施例中，词向量可以采用 Word2Vec 预训练词向量，即每个词汇都有对应的向量表示，此类向量表示能够以数据形式表达词汇信息，词向量维度可以为 300。其中，Word2vec，英文为 Word to vector，是一款用于训练词向量的软件工具，用来产生词向量的相关模型，词向量的自动训练可以通过 Python 中的 Gensim 库实现。

S230、使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征。

其中，卷积神经网络，英文为 Convolutional Neural Networks，简称为 CNN，是一类包含卷积或者相关计算且具有深度结构的前馈神经网络（Feedforward Neural Networks），是深度学习（Deep Learning）的代表算法之一。由于卷积神经网络能够进行平移不变分类（英文为 Shift-Invariant Classification），因此也被称为“平移不变人工神经网络（英文为 Shift-Invariant Artificial Neural Networks，简称为 SIANN）。

注意力，又称为注意力结构或者注意力机制，英文为 Attention Mechanism，卷积神经网络中的注意力主要用于决定卷积神经网络需要关注输入的哪部分并分配有限的信息处理资源给重要的部分以聚焦卷积神经网络处理输入数据的针对性，提高卷积神经网络对输入数据的处理效率。

具体地，终端在 TextCNN 模型中加入了注意力结构，通过改善 TextCNN 模型对数据的处理针对性，提升 TextCNN 模型的训练效率和文本分类效率，从而实现使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征。比如，对于 TextCNN 模型加入两种注意力机制，一类为 Word-wise 注意力结构，也就是词注意力结构，Word-wise 注意力结构作用于词嵌入层，根据词向量输入，训练得到相应注意力值，针对词嵌入层输出的词向量添加注意力以提高卷积神经网络对词向量处理的针对性，一类为 Filter-wise 注意力结构，也就是通道注意力结构，Filter-wise 注意力结构则作用于卷积层

之后，根据卷积通道输出，训练得到相应注意力值，以使池化后的输出数据在进行全连接时，提高全连接过程中对全连接对象的针对性，通过提高文本分类过程中的针对性从而提高整个文本分类的效率。

S240、通过全连接的方式连接所述词向量特征以得到输出数据。

其中，全连接是指连接所有的特征，将输出数据送给分类器，比如 Softmax 分类器，在 TextCNN 模型中，全连接是指输出层的神经元和输入层的每个神经元都连接，也就是卷积神经网络中的全连接层的每一个结点都与上一层的所有结点相连，用来把前边提取到的特征综合起来。全连接层，英文为 Fully connected layer，是 TextCNN 模型中的一个层。

具体地，在 TextCNN 结构中，经过全连接层之前的逐层处理，比如经过多个卷积层和池化层后，一般会连接着 1 个或 1 个以上的全连接层，全连接层中的每个神经元与其前一层的所有神经元进行全连接，全连接层可以整合卷积层或者池化层中具有类别区分性的局部信息，进而全连接层将前面经过多次卷积后或者池化后的高度抽象化的特征进行整合，然后进行归一化以对各种特征的分类情况输出一个概率，发送至全连接层之后的分类器，以使分类器 (Classifier) 可以根据全连接得到的概率进行分类。并且，实际中为了提高网络的学习能力，可以拼接多个全连接层。

S250、经分类器对所述输出数据进行分类以得到文本分类结果。

具体地，终端上的 TextCNN 模型中的分类器接收全连接层发送的经全连接之后的输出数据，将所述输出数据经分类器进行分类以最终得到对文本的分类结果，其中，分类器可以采用 softmax 分类器。

本申请实施例在实现文本分类时，通过获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词，将所述中文分词进行词嵌入以将所述中文分词转化为词向量，使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征，通过全连接的方式连接所述词向量特征以得到输出数据；经分类器对所述输出数据进行分类以得到文本分类结果，从而在基于卷积神经网络的文本分类模型中加入了注意力，以聚焦文本处理的针对性，能够有效提升文本分类模型的训练效率和文本分类效率。

请参阅图 4 和图 5, 图 4 为本申请实施例提供的文本分类方法的另一个流程示意图, 图 5 为图 4 中的文本分类方法对应的模型示意图。如图 4 和图 5 所示, 该方法包括以下步骤 S410-S490:

S410、获取进行文本分类的语料, 并将所述语料通过预设方式进行分词以得到中文分词;

S420、将所述中文分词进行词嵌入以将所述中文分词转化为词向量。

具体地, 步骤 S410 和图 2 中的步骤 S210 相同, 步骤 S420 和图 2 中的步骤 S220 相同, 将图 2 中的步骤 S210 和 S420 均通过应用的方式包含于此, 在此不再赘述。其中, 图 5 中的输入层就是用于获取进行分类的文本语料, 比如对新闻标题分类, 图 5 所示的文本分类模型输入为文本词向量信息, 输出为文本分类结果。图 5 中的词嵌入层就是用于将输入的自然语言语料进行词嵌入以将所述自然语言的语料编码成词向量。

S430、使用第一注意力函数对所述词向量分配注意力权重以得到调整后的词向量。

其中, 注意力, 又称为注意力机制, 或者注意力模型, 或者注意力结构, 英文为 Attention Model。自然语言处理中的注意力模型, 借鉴了人类的注意力概念, 一般来说, 视觉注意力是人类视觉所特有的大脑信号处理机制, 人类视觉通过快速扫描全局图像, 获得需要重点关注的目标区域, 也就是一般所说的注意力焦点, 而后对这一区域投入更多注意力资源, 以获取更多所需要关注目标的细节信息, 而抑制其他无用信息, 人类视觉注意力极大地提高了视觉信息处理的效率与准确性, 本申请实施例中的注意力从本质上讲和人类的选择性视觉注意力类似, 核心目标也是从众多信息中选择出对当前任务目标更关键的信息。

进一步地, 注意力模型可以表现为一种函数, 比如 $y=f(x)$, $y=f(x)$ 可以为线性关系, 比如, $y=wx+b$, 其中, y 表示输出, x 表示输入, w 和 b 分别表示 x 和 y 线性关系的参数, w 和 b 可以分别在训练过程中得到调整。

具体地, 在卷积神经网络基础上添加第一注意力函数, 使用第一注意力函数对所述词向量分配注意力权重以得到调整后的词向量, 所述第一注意力函数

主要是用于对词向量的权重进行分配，对各词向量分配注意力权重，以重点突出需要关注的词向量。由于第一注意力函数是用于对词向量的权重进行分配，第一注意力函数又可以称为词注意力，英文为 Word-wise，又可以称为 Word-wise 注意力或者 Word-wise 机制，Word-wise 注意力用于提炼词注意力信息，即根据输入词汇分配注意力权重，并进行自动学习的过程，在自动学习中确定词的注意力权重，并输出至下一步运算。

Word-wise 注意力所在的层称为 Word-wise 注意力结构层，或者称为 Word-wise 注意力层，或者 Word-wise 注意力结构层，Word-wise 注意力层用于根据输入词汇分配注意力权重，并输出分配权重后的词向量至下一步运算。

进一步地，请参阅图 5，如图 5 所示，Word-wise 注意力结构层添加于词嵌入层与卷积层之间，Word-wise 注意力结构层的输入来自于词嵌入层，Word-wise 注意力结构层可以建立两层全连接层并以 Softmax 函数输出注意力权重，该输出用于调整词嵌入层的输出，并将注意力权重调整后的词嵌入层输出的数据输入至卷积层。比如，针对“我喜欢吃苹果”这样的描述，针对“我”、“喜欢”、“苹果”通过神经网络的自学习分配不同的权重，权重如何调整，也是神经网络在训练的过程中进行自适应和自学习从而实现动态调整。卷积神经网络以及注意力结构的建立均通过 Python 中的 Tensorflow 库实现。

在一实施例中，可将 Word-wise 注意力结构层添加于首层卷积层前方，以将 Word-wise 结构层的提炼信息输入至卷积层。比如，Word-wise 注意力根据词嵌入层的词向量输入，通过建立一层或者两层的全连接隐层，并通过 Softmax 函数输出。进一步地，词注意力结构层的输出为注意力权重，需要将词嵌入层输出的词向量与注意力结构层的输出进行点乘，以完成所述词向量的注意力权重调整，并将调整后的词向量输出输入至卷积层以完成后续运算。其中，Softmax 函数，或称归一化指数函数，是逻辑函数的一种推广，它能将一个含任意实数的 K 维向量 $\mathbf{z}$ “压缩”到另一个 K 维实向量 $\sigma(\mathbf{z})$ 中，使得每一个元素的范围都在 $(0,1)$ 之间，并且所有元素的和为 1，Softmax 函数实际上是有限项离散概率分布的梯度对数归一化。点乘，也叫向量的内积、数量积，点乘的结果是一个数，比如，向量 $\mathbf{a} \cdot$ 向量 $\mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos\langle \mathbf{a}, \mathbf{b} \rangle$ ， $\cos\langle \mathbf{a}, \mathbf{b} \rangle$ 表示向量 $\mathbf{a}$ 和向量 $\mathbf{b}$ 夹角的

余弦值，将向量用坐标表示（三维向量），若向量 $a=(a_1, b_1, c_1)$ ，向量 $b=(a_2, b_2, c_2)$ ，则向量 $a \cdot b = a_1a_2 + b_1b_2 + c_1c_2$ 。

S440、使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征。

其中，卷积层，英文为 Convolutional layer，主要是用一个采样器从输入数据中采集关键数据内容，卷积层最大的特点是局部感知和权重共享，从而实现通过卷积提取文本的不同特征，一般来说，卷积层、池化层及全连接层都属于卷积神经网络，并且是三种不同类型的隐藏层。

具体地，首先建立词嵌入层实现训练文本向词向量的转化，随后建立卷积神经网络以形成卷积层，通过卷积层进行文本特征提取，其中，卷积层可以包括多层卷积核。比如，卷积层含有高度为 1、3、5 的卷积核各 128 通道，也就是高度为 1 行、3 行、5 行的卷积核各 128 通道。在本实施例中，卷积层的输出将输入至后续激活层与池化层。

请继续参阅图 5，卷积层主要是通过卷积以提取不同的 N-gram 特征。输入的语句或者文本，通过词嵌入层后，会转变成一个二维矩阵，假设文本的长度为 $|T|$ ，词向量的大小为 $|d|$ ，则该二维矩阵的大小为 $|T|*|d|$ ，卷积的工作就是对这一个 $|T|*|d|$ 的二维矩阵进行的。卷积核的大小一般设定为 $n*|d|$ ， n 是卷积核的长度， $|d|$ 是卷积核的宽度，这个宽度和词向量的维度是相同的，也就是卷积只是沿着文本序列进行的， n 可以有多种选择，比如 2、3、4、5 等。对于一个 $|T|*|d|$ 的文本，如果选择卷积核 kernel 的大小为 $2*|d|$ ，则卷积后得到的结果是 $|T-2+1|*1$ 的一个向量。在 TextCNN 网络中，需要同时使用多个不同类型的 kernel，同时每个 size 的 kernel 又可以有多个。如果我们使用的 kernel size 大小为 2、3、4、5 $*|d|$ ，每个种类的 size 又有 128 个 kernel，则卷积网络一共有 4*128 个卷积核。其中 N-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。

更进一步地，在训练卷积神经网络的过程中，卷积神经网络的损失函数为交叉熵，训练方法为 ADAM，学习率为 0.001，其中，ADAM，英文为 Adaptive Moment Estimation，是自适应矩估计。同时，在训练神经网络时，需要设置学习率控制参数更新的速度，其中，学习率，英文为 Learning rate，又称为学习速率，用于控制模型的学习进度。神经网络的训练通过 Python 中的 Tensorflow 库实现。

S450、使用激活函数对所述词向量特征进行激活。

其中，激活函数是用来加入非线性因素的。其中，常用的激活函数包括：Sigmoid 函数、Tanh 函数及 ReLU 函数等。目前大部分的卷积神经网络中，基本上都是采用了 ReLU 函数。使用激活函数来加入非线性因素的层在 TextCNN 模型中称为激活层。

具体地，由于线性模型的表达力不够，在神经网络中，对于图像，主要采用了卷积的方式来处理，也就是对每个像素点赋予一个权值，这个操作是线性的。但是对于使用的训练样本来说，不一定是线性可分的，为了解决这个问题，可以进行线性变化或者引入非线性因素，解决线性模型所不能解决的问题，因此，TextCNN 模型中使用激活函数对所述词向量特征进行激活，也就是使用激活函数对所述调整后的词向量的词向量特征进行调整以在卷积层提取的词向量特征中加入非线性因素，从而提高词向量处理的准确性。

S460、对激活后的词向量特征进行池化。

其中，池化，英文为 Pooling，是指使用卷积神经网络对输入的特征进行压缩以提取主要特征并将数据量变小，池化操作通常被用在卷积神经网络中。用于进行池化的层在 TextCNN 模型中称为池化层，英文为 Pooling layer。池化层用于降低卷积层输出的特征向量，同时改善结果以有效控制过拟合。

具体地，在卷积神经网络中，池化层往往在卷积层后面，通过池化来降低卷积层输出的特征向量，同时改善结果以有效控制过拟合。最常见的池化操作为平均池化 Mean Pooling 和最大池化 Max Pooling。池化层的形式可以为 Max Pooling，也就是最大池化层，最大池化即取局部接收域中值最大的点，Max Pooling 能减小卷积层参数误差造成估计均值的偏移误差，更多的保留纹理信息。

一个最大池化层从一块特征中选取最大值。和卷积层一样，池化层也是通过窗口（块）大小和步幅尺寸进行参数化，比如，在一个 10×10 特征矩阵上以 2 的步幅滑动一个 2×2 的窗口，然后选取每个窗口的 4 个值中的最大值，得到一个 5×5 特征矩阵。池化层通过只保留最突出的信息来减少表征的维度。池化层的输出将接入全连接层实现全局特征提取，并以 Softmax 函数完成最终输出。

S470、使用第二注意力函数对所述词向量特征分配注意力权重以得到第一词向量特征。

其中，第二注意力函数用于调整各通道的池化层输出，对各通道分配注意力权重，并将权重调整后的池化层输出的数据输入全连接层进行后续计算。由于第二注意力函数用于调整各通道的池化层输出，因此也称为通道注意力，英文为 Filter-wise，也可以称为 Filter-wise 注意力，或者 Filter-wise 注意力结构层。使用通道注意力调整各通道的池化层输出的层在 TextCNN 模型中称为通道注意力层，也可以称为 Filter-wise 注意力层。Filter-wise 注意力层用于卷积通道输出，训练得到相应注意力值，Filter-wise 注意力层添加与卷积层之后，可以添加于池化层与全连接层之间，该部分的输出来自于池化层输出，并可以通过建立两层全连接层以 Softmax 函数输出注意力权重。卷积神经网络以及注意力的建立均可以通过 Python 中的 Tensorflow 库实现。

具体地，Filter-wise 机制与 Word-wise 机制具有类似的结构，但是作用于模型的不同部分，对不同的对象分配权重，起不同的作用。Filter-wise 机制提炼通道注意力信息。由于卷积神经网络含有多个通道，因此 Filter-wise 机制能够显著提升模型训练效率。具体实施中，可将 Filter-wise 机制添加于通道最终输出，通过全连接隐层以及 Softmax 函数计算注意力权重，并将计算结果与通道输出进行点乘计算，以得到权重调整后的通道输出，以实现此处使用第二注意力函数对池化后的词向量特征分配注意力权重以得到第一词向量特征，并进行后续计算。

S480、通过全连接的方式连接所述第一词向量特征以得到输出数据。

具体地，通过全连接的方式连接所述第一词向量特征以得到输出数据，在 TextCNN 模型中，也就是通过全连接层将前面经过多次卷积后高度抽象化的特征

进行整合，然后可以进行归一化，对各种分类情况都输出一个概率，之后的分类器可以根据全连接得到的概率进行分类，分类器可以为 Classifier 分类器。比如，Fully-connected layer 在 Max-pooling layer 后再拼接一层，将该层的输出作为输出结果。实际中为了提高网络的学习能力，可以拼接多个全连接层。

S490、经分类器对所述输出数据进行分类以得到文本分类结果。

具体地，步骤 S490 和图 2 中的步骤 S250 相同，将图 2 中的步骤 S250 通过应用的方式包含于此，在此不再赘述。经分类器对所述输出数据进行分类以得到文本分类结果，文本分类结果经过输出层输出，输出层用于输出文本分类结果。

本申请实施例中对于 TextCNN 网络结构加入了两种注意力，一类为 Word-wise 注意力，一类为 Filter-wise 注意力。其中，Word-wise 注意力作用于词嵌入层输出的词向量，根据词嵌入层的词向量输入，训练得到词向量相应的注意力值。而 Filter-wise 注意力则作用于卷积层的输出，根据卷积通道的输出数据，训练得到输出数据的相应注意力值，Word-wise 注意力和 Filter-wise 注意力与 TextCNN 模型的其他部分共同进行训练，而不需要进行额外训练计算。实际操作中，以 TextCNN 分类结果交叉熵为损失函数，以 ADAM 为优化方法，进行训练。本申请实施例通过在 TextCNN 模型中加入了针对词的词注意力与针对通道的通道注意力，能够有效提升 TextCNN 模型的训练效率。本申请实施例的 TextCNN 模型应用于文本分类，如新闻标题分类，TextCNN 模型输入为文本词向量信息，输出为文本分类结果，实践中，Word-wise 与 Filter-wise 机制的加入能够提升 TextCNN 模型训练效率，尤其是 Filter-wise 机制具有显著效果。

在一个实施例中，所述使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征的步骤包括：

使用卷积神经网络对所述词向量进行特征提取以得到所述词向量的第二词向量特征；

使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征。

具体地,请参阅图 6,图 6 为该实施例提供的文本分类方法对应的模型示意图。如图 6 所示,在进行文本分类的过程中,通过输入层获取进行文本分类的语料,并将所述语料通过预设方式进行分词以得到中文分词,将所述中文分词通过词嵌入层进行词嵌入以将所述中文分词转化为词向量,通过卷积层使用卷积神经网络对所述词向量进行特征提取以得到所述词向量的第二词向量特征,通过 Filter-wise 注意力结构层使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征,然后通过全连接层使用全连接的方式连接所述词向量特征以得到输出数据,最后通过输出层经分类器对所述输出数据进行分类以得到文本分类结果。其中,第二注意力函数是用于调整各通道的池化层输出,因此也称为通道注意力,英文为 Filter-wise,也可以称为 Filter-wise 注意力。使用通道注意力调整各通道的池化层输出的层在 TextCNN 模型中称为通道注意力层,也可以称为 Filter-wise 注意力层,或者 Filter-wise 注意力结构层。

在一个实施例中,所述使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征的步骤之前,还包括:

使用激活函数对所述第二词向量特征进行激活;

对激活后的所述第二词向量特征进行池化。

具体地,请参阅图 7,图 7 为本实施例提供的文本分类方法对应的模型示意图。如图 7 所示,在进行文本分类的过程中,通过输入层获取进行文本分类的语料,并将所述语料通过预设方式进行分词以得到中文分词,将所述中文分词通过词嵌入层进行词嵌入以将所述中文分词转化为词向量,通过卷积层使用卷积神经网络对所述词向量进行特征提取以得到所述词向量的第二词向量特征,通过激活层使用激活函数对卷积层输出的所述第二词向量特征加入非线性进行调整,从而提高词向量处理的准确性,然后通过池化层降低激活层输出的经过非线性调整的第二词特征向量,再通过 Filter-wise 注意力结构层使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征,然后通过全连接层使用全连接的方式连接所述词向量特征以得到输出数据,最后通过输出层经分类器对所述输出数据进行分类以得到文本分类结果。其中,

第二注意力函数是用于调整各通道的池化层输出，因此也称为通道注意力，英文为 Filter-wise，也可以称为 Filter-wise 注意力。使用通道注意力调整各通道的池化层输出的层在 TextCNN 模型中称为通道注意力层，也可以称为 Filter-wise 注意力层，或者 Filter-wise 注意力结构层。

需要说明的是，上述各个实施例所述的文本分类方法，可以根据需要将不同实施例中包含的技术特征重新进行组合，以获取组合后的实施方案，但都在本申请要求的保护范围之内。

请参阅图 8，图 8 为本申请实施例提供的文本分类装置的示意性框图。对应于上述文本分类方法，本申请实施例还提供一种文本分类装置。如图 8 所示，该文本分类装置包括用于执行上述文本分类方法的单元，该装置可以被配置于终端或者服务器等计算机设备中。具体地，请参阅图 8，该文本分类装置 800 包括获取单元 801、转化单元 802、提取单元 803、连接单元 804 及分类单元 805。

其中，获取单元 801，用于获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词；

转化单元 802，用于将所述中文分词进行词嵌入以将所述中文分词转化为词向量；

提取单元 803，用于使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征；

连接单元 804，用于通过全连接的方式连接所述词向量特征以得到输出数据；

分类单元 805，用于经分类器对所述输出数据进行分类以得到文本分类结果。

请参阅图 9，图 9 为本申请实施例提供的文本分类装置的另一个示意性框图。如图 9 所示，在该实施例中，所述提取单元 803 包括：

第一分配子单元 8031，用于使用第一注意力函数对所述词向量分配注意力权重以得到调整后的词向量；

提取子单元 8032，用于使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征；

激活子单元 8033，用于使用激活函数对所述词向量特征进行激活；

池化子单元 8034，用于对激活后的词向量特征进行池化。

第二分配子单元 8035, 用于使用第二注意力函数对所述词向量特征分配注意力权重以得到第一词向量特征;

所述连接单元 804, 用于通过全连接的方式连接所述第一词向量特征以得到输出数据。

请继续参阅图 9, 如图 9 所示, 在另一个实施例中, 所述提取单元 803 包括:

提取子单元 8032, 用于使用卷积神经网络对所述词向量进行特征提取以得到所述词向量的第二词向量特征;

激活子单元 8033, 用于使用激活函数对所述第二词向量特征进行激活;

池化子单元 8034, 用于对激活后的所述第二词向量特征进行池化。

第二分配子单元 8035, 用于使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征。

需要说明的是, 所属领域的技术人员可以清楚地了解到, 上述文本分类装置和各单元的具体实现过程, 可以参考前述方法实施例中的相应描述, 为了描述的方便和简洁, 在此不再赘述。

同时, 上述文本分类装置中各个单元的划分和连接方式仅用于举例说明, 在其他实施例中, 可将文本分类装置按照需要划分为不同的单元, 也可将文本分类装置中各单元采取不同的连接顺序和方式, 以完成上述文本分类装置的全部或部分功能。

上述文本分类装置可以实现为一种计算机程序的形式, 该计算机程序可以在如图 10 所示的计算机设备上运行。

请参阅图 10, 图 10 是本申请实施例提供的一种计算机设备的示意性框图。该计算机设备 1000 可以是台式机电脑或者服务器等计算机设备, 也可以是其他设备中的组件或者部件。

参阅图 10, 该计算机设备 1000 包括通过系统总线 1001 连接的处理器 1002、存储器和网络接口 1005, 其中, 存储器可以包括非易失性存储介质 1003 和内存存储器 1004。

该非易失性存储介质 1003 可存储操作系统 10031 和计算机程序 10032。该计算机程序 10032 被执行时, 可使得处理器 1002 执行一种上述文本分类方法。

该处理器 1002 用于提供计算和控制能力，以支撑整个计算机设备 1000 的运行。

该内存储器 1004 为非易失性存储介质 1003 中的计算机程序 10032 的运行提供环境，该计算机程序 10032 被处理器 1002 执行时，可使得处理器 1002 执行一种上述文本分类方法。

该网络接口 1005 用于与其它设备进行网络通信。本领域技术人员可以理解，图 10 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备 1000 的限定，具体的计算机设备 1000 可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。例如，在一些实施例中，计算机设备可以仅包括存储器及处理器，在这样的实施例中，存储器及处理器的结构及功能与图 10 所示实施例一致，在此不再赘述。

其中，所述处理器 1002 用于运行存储在存储器中的计算机程序 10032，以实现本申请实施例的文本分类方法。

应当理解，在本申请实施例中，处理器 1002 可以是中央处理单元（Central Processing Unit, CPU），该处理器 1002 还可以是其他通用处理器、数字信号处理器（Digital Signal Processor, DSP）、专用集成电路（Application Specific Integrated Circuit, ASIC）、现成可编程门阵列（Field-Programmable Gate Array, FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中，通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

本领域普通技术人员可以理解的是实现上述实施例的方法中的全部或部分流程，是可以通过计算机程序来完成，该计算机程序可存储于一计算机可读存储介质。该计算机程序被该计算机系统中的至少一个处理器执行，以实现上述文本分类方法的实施例的步骤。

因此，本申请实施例还提供一种计算机可读存储介质。该计算机可读存储介质存储有计算机程序，该计算机程序被处理器执行时使处理器执行以上各实施例中所描述的文本分类方法的步骤。

所述计算机可读存储介质可以是U盘、移动硬盘、只读存储器(Read-Only Memory, ROM)、磁碟或者光盘等各种可以存储计算机程序的计算机可读存储介质。

本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

以上所述,仅为本申请的具体实施方式,但本申请明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到各种等效的修改或替换,这些修改或替换都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以权利要求的保护范围为准。

权利要求书

1. 一种文本分类方法，包括：

获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词；

将所述中文分词进行词嵌入以将所述中文分词转化为词向量；

使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征；

通过全连接的方式连接所述词向量特征以得到输出数据；

经分类器对所述输出数据进行分类以得到文本分类结果。

2. 根据权利要求 1 所述文本分类方法，其中，所述使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征的步骤包括：

使用第一注意力函数对所述词向量分配注意力权重以得到调整后的词向量；

使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征。

3. 根据权利要求 2 所述文本分类方法，其中，所述使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征的步骤之后，还包括：

使用第二注意力函数对所述词向量特征分配注意力权重以得到第一词向量特征；

所述通过全连接的方式连接所述词向量特征以得到输出数据的步骤包括：

通过全连接的方式连接所述第一词向量特征以得到输出数据。

4. 根据权利要求 3 所述文本分类方法，其中，所述使用第二注意力函数对所述词向量特征分配注意力权重以得到第一词向量特征的步骤之前，还包括：

使用激活函数对所述词向量特征进行激活。

5. 根据权利要求 4 所述文本分类方法，其特征在于，所述使用激活函数对所述词向量特征进行激活的步骤之后，还包括：

对激活后的词向量特征进行池化。

6. 根据权利要求 1 所述文本分类方法，其中，所述使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征的步骤

包括：

使用卷积神经网络对所述词向量进行特征提取以得到所述词向量的第二词向量特征；

使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征。

7. 根据权利要求 6 所述文本分类方法，其中，所述使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征的步骤之前，还包括：

使用激活函数对所述第二词向量特征进行激活；

对激活后的所述第二词向量特征进行池化。

8. 根据权利要求 1 所述文本分类方法，其中，所述将所述中文分词进行词嵌入以将所述中文分词转化为词向量的步骤包括：

使用预设词向量字典将所述中文分词进行词嵌入以将所述中文分词转化为词向量。

9. 一种文本分类装置，包括：

获取单元，用于获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词；

转化单元，用于将所述中文分词进行词嵌入以将所述中文分词转化为词向量；

提取单元，用于使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征；

连接单元，用于通过全连接的方式连接所述词向量特征以得到输出数据；

分类单元，用于经分类器对所述输出数据进行分类以得到文本分类结果。

10. 根据权利要求 9 所述文本分类装置，其中，所述提取单元包括：

第一分配子单元，用于使用第一注意力函数对所述词向量分配注意力权重以得到调整后的词向量；

提取子单元，用于使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征。

11. 一种计算机设备，其中，所述计算机设备包括存储器以及与所述存储器相连的处理器；所述存储器用于存储计算机程序；所述处理器用于运行所述存储器中存储的计算机程序，以执行如下步骤。

获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词；

将所述中文分词进行词嵌入以将所述中文分词转化为词向量；

使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征；

通过全连接的方式连接所述词向量特征以得到输出数据；

经分类器对所述输出数据进行分类以得到文本分类结果。

12. 根据权利要求 11 所述计算机设备，其中，所述使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征的步骤包括：

使用第一注意力函数对所述词向量分配注意力权重以得到调整后的词向量；

使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征。

13. 根据权利要求 12 所述计算机设备，其中，所述使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征的步骤之后，还包括：

使用第二注意力函数对所述词向量特征分配注意力权重以得到第一词向量特征；

所述通过全连接的方式连接所述词向量特征以得到输出数据的步骤包括：

通过全连接的方式连接所述第一词向量特征以得到输出数据。

14. 根据权利要求 13 所述计算机设备，其中，所述使用第二注意力函数对所述词向量特征分配注意力权重以得到第一词向量特征的步骤之前，还包括：

使用激活函数对所述词向量特征进行激活。

15. 根据权利要求 14 所述计算机设备，其中，所述使用激活函数对所述词向量特征进行激活的步骤之后，还包括：

对激活后的词向量特征进行池化。

16. 根据权利要求 11 所述计算机设备，其中，所述使用卷积神经网络结合

注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征的步骤包括：

使用卷积神经网络对所述词向量进行特征提取以得到所述词向量的第二词向量特征；

使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征。

17. 根据权利要求 16 所述计算机设备，其中，所述使用第二注意力函数对所述第二词向量特征分配注意力权重以得到所述词向量的词向量特征的步骤之前，还包括：

使用激活函数对所述第二词向量特征进行激活；

对激活后的所述第二词向量特征进行池化。

18. 根据权利要求 11 所述计算机设备，其中，所述将所述中文分词进行词嵌入以将所述中文分词转化为词向量的步骤包括：

使用预设词向量字典将所述中文分词进行词嵌入以将所述中文分词转化为词向量。

19. 一种计算机可读存储介质，其中，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时使所述处理器执行如下步骤：

获取进行文本分类的语料，并将所述语料通过预设方式进行分词以得到中文分词；

将所述中文分词进行词嵌入以将所述中文分词转化为词向量；

使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征；

通过全连接的方式连接所述词向量特征以得到输出数据；

经分类器对所述输出数据进行分类以得到文本分类结果。

20. 根据权利要求 19 所述计算机可读存储介质，其中，所述使用卷积神经网络结合注意力函数对所述词向量进行特征提取以得到所述词向量的词向量特征的步骤包括：

使用第一注意力函数对所述词向量分配注意力权重以得到调整后的词向量；

使用卷积神经网络对所述调整后的词向量进行特征提取以得到词向量特征。

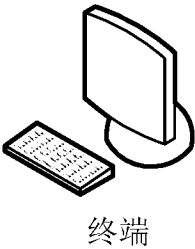


图 1

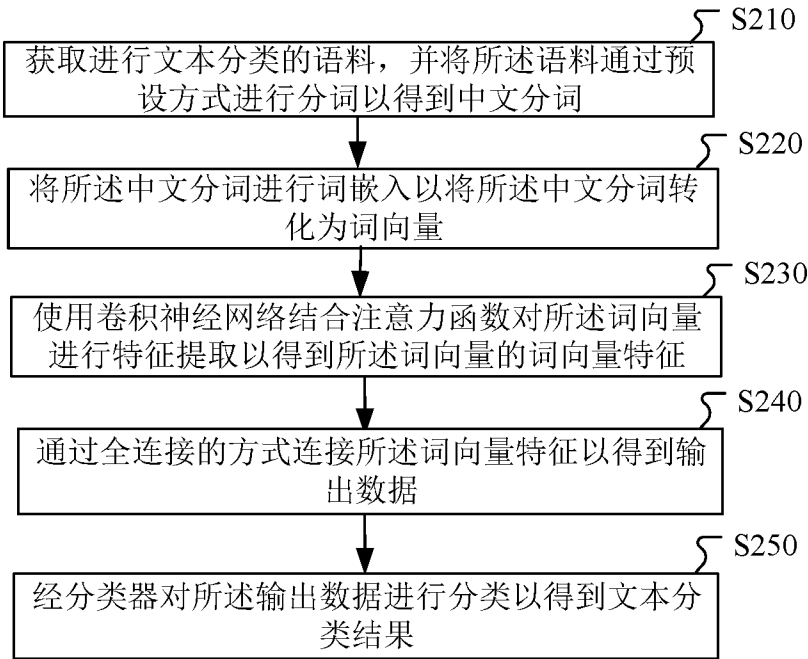


图 2

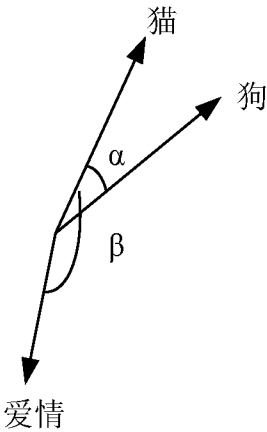


图 3

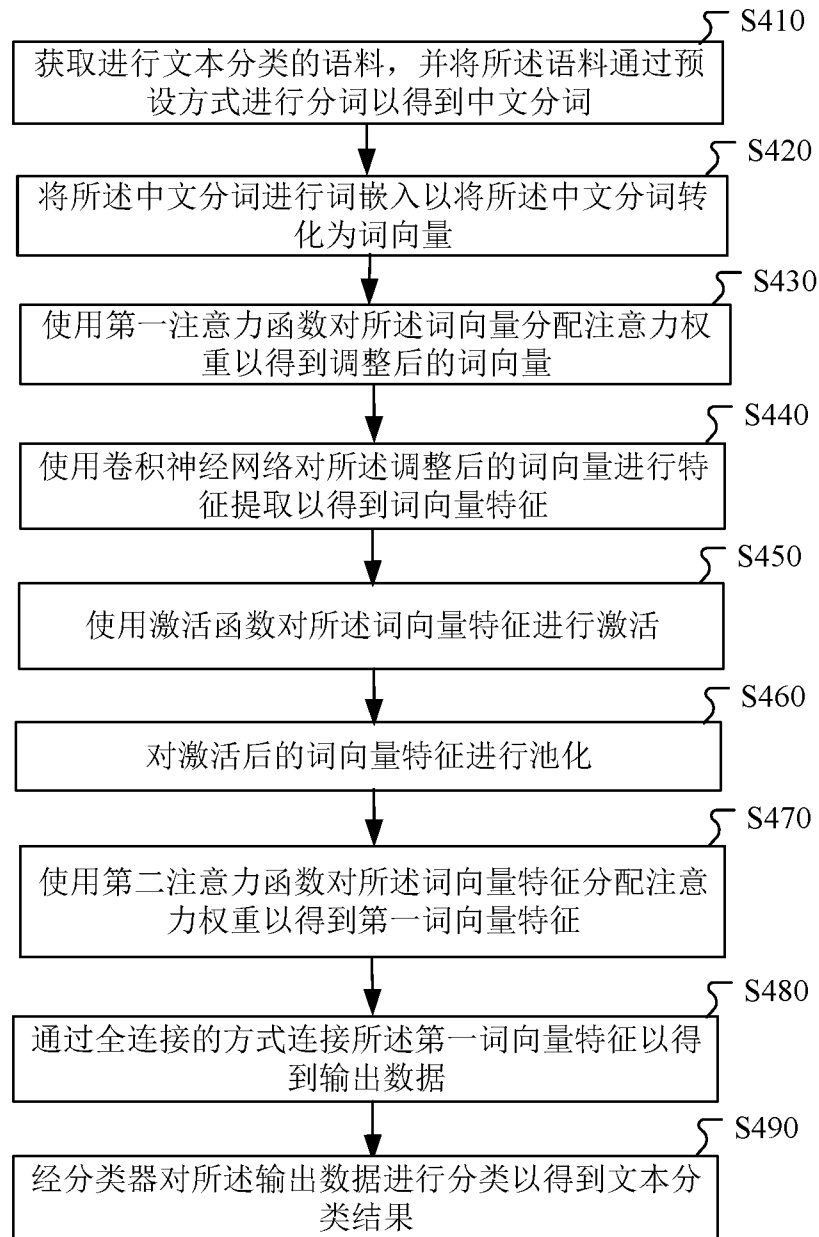


图 4

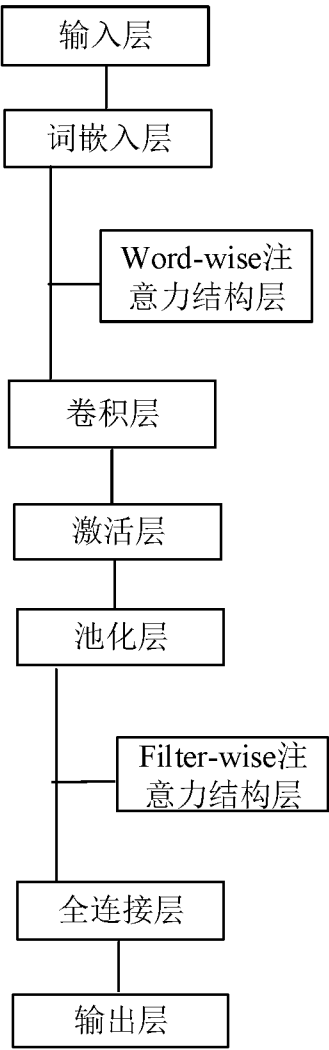


图 5

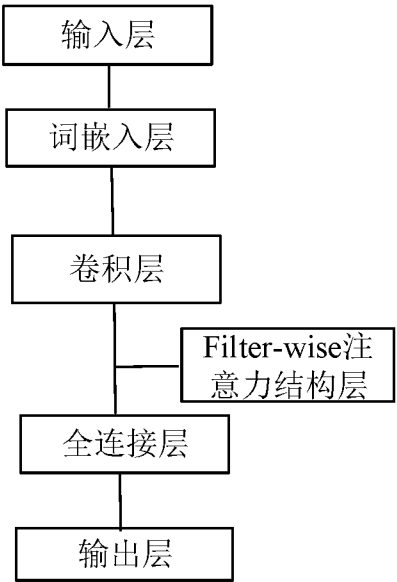


图 6

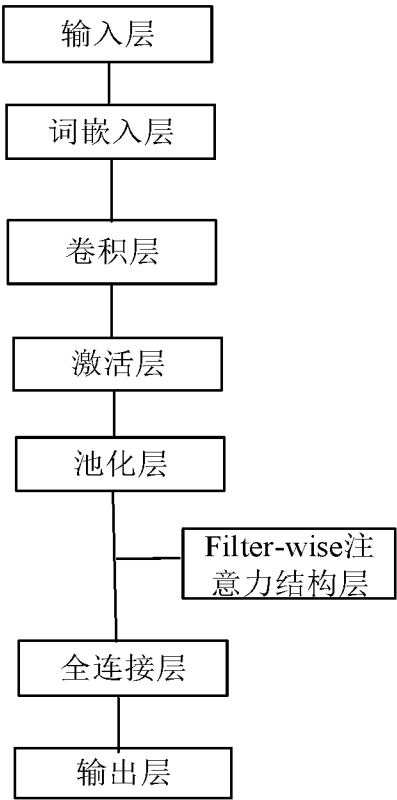


图 7

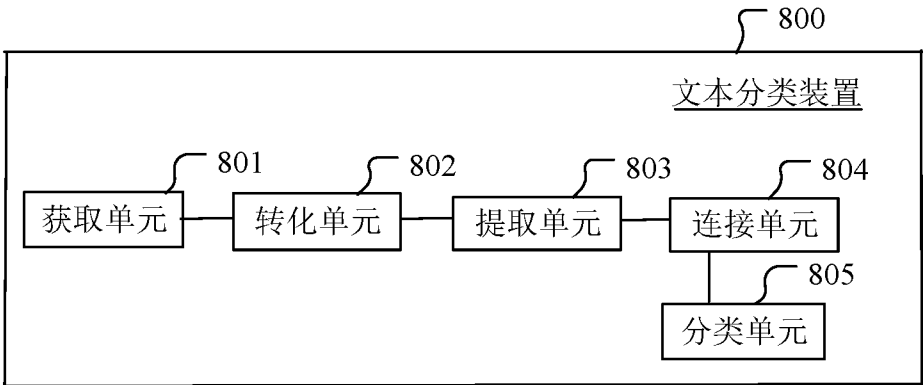


图 8

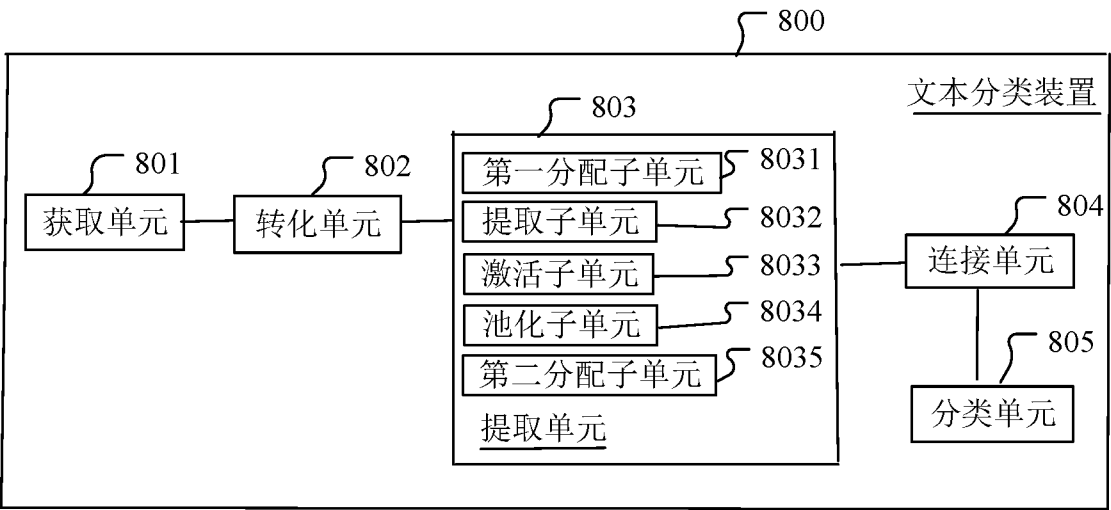


图 9

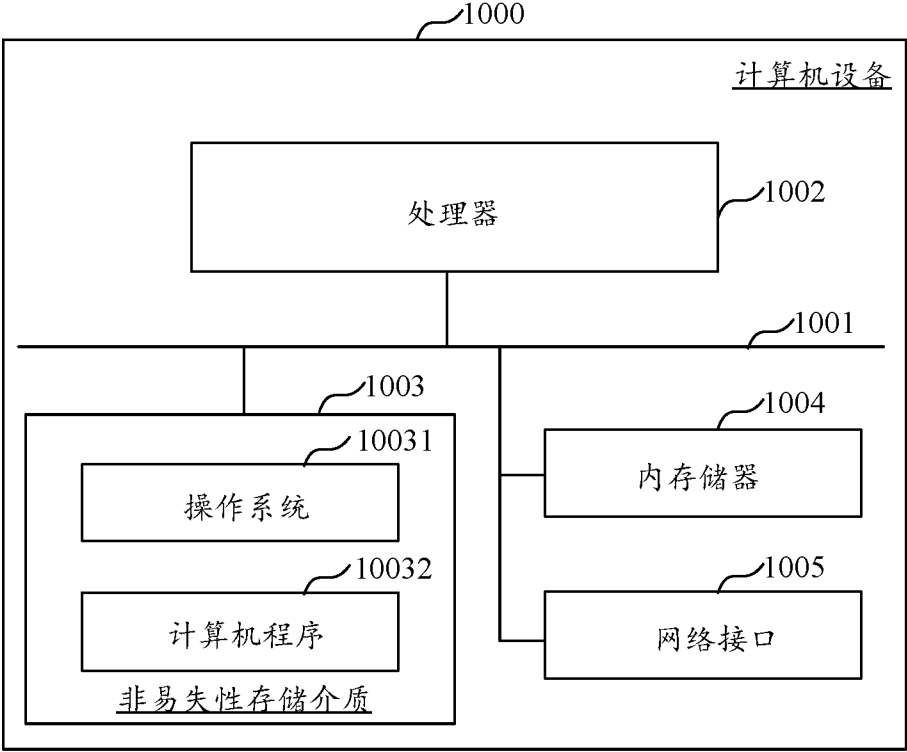


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/092531

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/35(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE: 文本, 分类, 分词, 向量, 全连接, 权重, 激活, 池化, 词嵌入, text, sort+, classif+, participle, vector, connect+, weight, activate, embed

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109857860 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 07 June 2019 (2019-06-07) claims 1-10, and figures 1-10	1-20
Y	CN 107885853 A (TONGJI UNIVERSITY) 06 April 2018 (2018-04-06) description, paragraphs [0049]-[0113], and figure 1	1-20
Y	CN 108415977 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 17 August 2018 (2018-08-17) description, paragraphs [0009]-[0016], and figure 1	1-20
A	CN 108614875 A (BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 02 October 2018 (2018-10-02) entire document	1-20
A	US 2013318124 A1 (FUJITSU LIMITED) 28 November 2013 (2013-11-28) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

28 August 2019

Date of mailing of the international search report

27 September 2019

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/092531

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109857860	A	07 June 2019	None			
CN	107885853	A	06 April 2018	None			
CN	108415977	A	17 August 2018	None			
CN	108614875	A	02 October 2018	None			
US	2013318124	A1	28 November 2013	WO	2012108006	A1	16 August 2012
				EP	2674874	A1	18 December 2013
				JP	WO2012108006	A1	03 July 2014

国际检索报告

国际申请号

PCT/CN2019/092531

A. 主题的分类

G06F 16/35 (2019.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, WPI, EPDOC, CNKI, IEEE: 文本, 分类, 分词, 向量, 全连接, 权重, 激活, 池化, 词嵌入, text, so-rt+, classif+, participle, vector, connect+, weight, activate, embed

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 109857860 A (平安科技深圳有限公司) 2019年 6月 7日 (2019 - 06 - 07) 权利要求1-10, 附图1-10	1-20
Y	CN 107885853 A (同济大学) 2018年 4月 6日 (2018 - 04 - 06) 说明书第[0049]-[0113]段, 附图1	1-20
Y	CN 108415977 A (华南理工大学) 2018年 8月 17日 (2018 - 08 - 17) 说明书第[0009]-[0016]段, 附图1	1-20
A	CN 108614875 A (北京邮电大学) 2018年 10月 2日 (2018 - 10 - 02) 全文	1-20
A	US 2013318124 A1 (FUJITSU LIMITED) 2013年 11月 28日 (2013 - 11 - 28) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 8月 28日

国际检索报告邮寄日期

2019年 9月 27日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

周亚楠

电话号码 86-(10)-53961530

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/092531

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	109857860	A	2019年 6月 7日	无			
CN	107885853	A	2018年 4月 6日	无			
CN	108415977	A	2018年 8月 17日	无			
CN	108614875	A	2018年 10月 2日	无			
US	2013318124	A1	2013年 11月 28日	W0	2012108006	A1	2012年 8月 16日
				EP	2674874	A1	2013年 12月 18日
				JP	W02012108006	A1	2014年 7月 3日