



(21) 申请号 201911054827.X

(22) 申请日 2019.10.31

(65) 同一申请的已公布的文献号
申请公布号 CN 110879934 A

(43) 申请公布日 2020.03.13

(73) 专利权人 杭州电子科技大学
地址 310018 浙江省杭州市下沙高教园区2号大街

(72) 发明人 傅啸

(74) 专利代理机构 杭州君度专利代理事务所
(特殊普通合伙) 33240
专利代理师 杨舟涛

(51) Int.Cl.
G06F 40/117 (2020.01)
G06F 40/289 (2020.01)
G06N 3/0464 (2023.01)
G06N 3/08 (2023.01)

(56) 对比文件

CN 109582794 A, 2019.04.05

CN 109657246 A, 2019.04.19

CN 110032980 A, 2019.07.19

CN 110287180 A, 2019.09.27

EP 3144860 A2, 2017.03.22

WO 2019149135 A1, 2019.08.08

WO 2019200806 A1, 2019.10.24

CONNEAU A. Very deep convolutional

networks for text classification.

《Conference of the European Chapter of the Association for Computational Linguistics》. 2017, 第1107-1116页.

涂文博; 袁贞明; 俞凯. 无池化层卷积神经网络的中文分词方法. 计算机工程与应用. 2018, (第02期), 第120-126页.

审查员 张盈盈

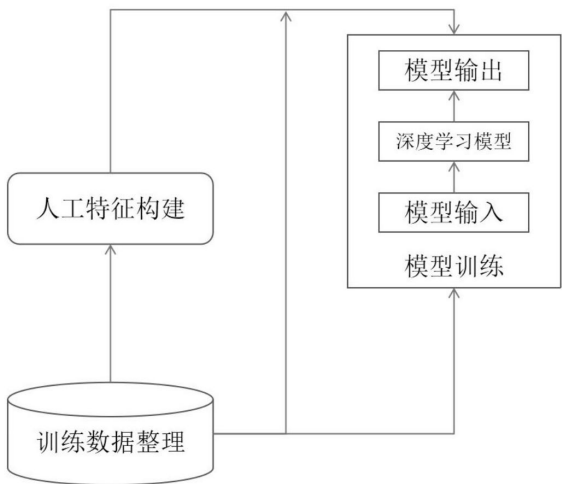
权利要求书2页 说明书5页 附图2页

(54) 发明名称

一种基于Wide&Deep深度学习模型的文本预测方法

(57) 摘要

本发明公开了一种基于Wide&Deep深度学习模型的文本预测方法。本发明旨在解决深度学习模型中对人工特征的充分利用以及Wide&Deep模型预测时的计算消耗问题。本发明的实施方案是：首先对文本数据进行标注，其次构建Wide特征，然后对基于Wide&Deep的深度学习模型添加辅助训练任务，最后对文本内容进行预测。本发明提及的人工特征作为深度学习模型训练过程中的附加任务，为模型提供了更多的有效信息，提高了模型学习效率。并且，在文本内容预测时不需要计算Wide特征，从而提升了模型性能。不同宽度的卷积核卷积后，可以通过一次卷积block获得不同距离上的信息，这有助于提升模型性能。同时，减少调整不同卷积核的时间，通过模型学习能自动适应不同的卷积核宽度。



1.一种基于Wide&Deep深度学习模型的文本预测方法,其特征在于包括如下步骤:

步骤1:文本数据的预处理,进行对文本数据进行标注;

步骤2:人工特征提取,构建wide特征;

步骤3:构建Wide&Deep深度学习模型;

步骤4:对Wide&Deep深度学习模型进行训练;

步骤5:通过模型对文本内容进行预测;

步骤1具体实现如下:

1-1.收集文本数据集 $X = \{x_1, x_2, \dots, x_N\}$, N 为总的训练样本数量;其中, $x_i \in \{x_1, x_2, \dots, x_N\}$, x_i 表示第 i 个待处理的文本数据;

1-2.利用人工对文本数据进行标记,针对文本数据集中每一个文本数据 x_i ,标记对应的 $y_i, y_i \in \{0, 1\}$;标记文本数据的目的在于根据业务需要利用人工经验对文本数据进行分析,其中,1表示和业务相关,0表示不相关;

1-3.利用Jieba分词器对文本数据集 X 进行分词,汇总得到所有文本数据的词表 $W, W = \{w_1, w_2, \dots, w_D\}$,其中 D 为词表的总长度; w_k 表示词表中第 k 个词;则文本数据 x_i 表示为 $x_i = \{w_{k_1}, w_{k_2}, \dots, w_{k_n}\}$,其中 k_n 表示文本数据 x_i 中第 n 个词在词表 W 中的序号,从而用词表序号重新表示文本数据 $x_i = \{k_1, k_2, \dots, k_n\}$;

步骤2具体实现如下:

2-1.统计文本数据集 X 中 $y_i = 1$ 对应的所有文本数据 x_i 中的词,得到出现频率最高的与业务相关的词,并排序获得序列I;以及 $y_i = 0$ 对应的所有文本数据 x_i 中的词,得到出现频率最高的与业务不相关的词,并排序获得序列II;

2-2.排除序列I和序列II中的语法型词汇,包括代词、冠词、连词的停用词;

2-3.将序列I和序列II中出现频率最高的文本特定词、 j 个高频词形成的组合作为wide特征,其中 j 范围为2-4;

所述的排序通过计算如下两个公式进行:

$$w_{k_n} = \frac{\#\{(w_{k_n} \in x_i, y_i = 1)\}}{\#\{y_i = 1\}};$$

$$w'_{k_n} = \frac{\#\{(w_{k_n} \in x_i, y_i = 0)\}}{\#\{y_i = 0\}};$$

对 w_{k_n} 和 w'_{k_n} 分别进行排序获得序列I和序列II,获取序列I中出现频率最大的5个词

$$w_{f_1}, w_{f_2}, \dots, w_{f_5}, w_f = \begin{cases} 0, & \text{其它} \\ 1, & w_f \in x_i \end{cases}, \text{以及序列II中出现频率最大的5个词 } w'_{f_1}, w'_{f_2}, \dots, w'_{f_5},$$

$$w'_f = \begin{cases} 0, & \text{其它} \\ 1, & w'_f \in x_i \end{cases};$$

2-4.获得最终的wide特征 $w_{f_1}, w_{f_2}, \dots, w_{f_5}, w'_{f_1}, w'_{f_2}, \dots, w'_{f_5}$;

步骤3具体实现如下:

3-1. 对文本数据 $x_i = \{w_{k_1}, w_{k_2}, \dots, w_{k_n}\}$ 的长度进行固定;

当文本数据的长度小于预设的固定长度时,使用“PAD”进行填充达到固定长度;

当文本数据的长度大于预设的固定长度时,丢弃超过预设的固定长度的部分;

如果文本数据中的词不在词表W中,使用“UNK”代替;

3-2. 建立每个词的embedding向量;将词表W中的所有词转化为预设的固定长度的稠密向量;

3-3. 获取一维卷积block;使用多个不同宽度的卷积核同时对embedding向量进行卷积;将每个卷积核的卷积结果进行拼接,得到一维卷积block,每次卷积后保证序列长度不变;

3-4. 多次叠加一维卷积;

利用3-3定义的一维卷积block对wide特征进行处理,得到新的embedding向量,在每次卷积后使用窗口大小为4,步长为4的最大池化;每一次最大池化操作后长度缩小1/4;反复进行4次卷积和最大池化操作;获取最终序列长度变为64,特征长度为8;

3-5. 文本表示;将序列长度为64,特征长度为8的向量拉直成一个序列长度为512维的向量,得到整个文本序列的512维特征向量的表示;

3-6. 构建损失函数;使用全连接层将512维特征向量映射成 $(16+1)*2$ 维的向量,再通过softmax函数,转化为 $(16+1)*2$ 个文本数据标记 y_i 的预测概率值,其中包括:1个文本数据标记的预测概率值,以及16个wide特征的预测概率值;损失函数表示为:

$$loss = y_i \log p_i + (1 - y_i) \log(1 - p_i) + \sum_{k=1}^{16} [f_{i,k} \log q_{i,k} + (1 - f_{i,k}) \log(1 - q_{i,k})]$$

其中, y_i 为文本数据标记, p_i 为文本数据标记 y_i 的预测概率值, $f_{i,k}$ 为wide特征值, $q_{i,k}$ 为wide特征的预测概率值, i 为数据标记的索引,表示第 i 个文本数据。

2. 根据权利要求1所述的一种基于Wide&Deep深度学习模型的文本预测方法,其特征在于步骤4通过梯度下降算法,求得步骤3-6中损失函数的最小化来训练模型。

3. 根据权利要求2所述的一种基于Wide&Deep深度学习模型的文本预测方法,其特征在于步骤5对需要预测的文本内容,利用步骤3计算出文本数据标记 y_i 的预测值,得到文本内容与业务的相关程度。

一种基于Wide&Deep深度学习模型的文本预测方法

技术领域

[0001] 本发明涉及深度学习模型的一种基于Wide&Deep深度学习模型的文本预测方法，主要是深度学习模型的训练架构的设计方法。

背景技术

[0002] 目前最主要的深度学习模型包括端到端的学习、人工特征、Wide&Deep模型。在Wide&Deep模型中，一般是先通过人工整理出有效的特征作为wide特征，以加速深度学习模型的训练速度，提高模型的性能。然后构建深度神经网络模型得到deep特征。最后计算wide特征和deep特征进行文本内容预测。但是，提取wide特征和deep特征是两个独立的过程，不能充分利用对方构建深度学习过程中的完整信息。并且在模型预测时需要先计算wide特征，再计算deep特征，使得模型的时效性较差。

发明内容

[0003] 本发明旨在解决深度学习模型中对人工特征的充分利用以及Wide&Deep模型预测时的计算消耗问题。提出了一种基于Wide&Deep深度学习模型的文本预测方法。

[0004] 本发明首先对文本数据进行标注，其次构建wide特征，然后对基于Wide&Deep的深度学习模型添加辅助训练任务，最后对文本内容进行预测。

[0005] 本发明解决其技术问题所采用的技术方案如下：

[0006] 步骤1：文本数据的预处理。

[0007] 1-1. 收集文本数据集 $X = \{x_1, x_2, \dots, x_N\}$ ， N 为总的训练样本数量。其中， $x_i \in \{x_1, x_2, \dots, x_N\}$ ， x_i 表示第 i 个待处理的文本数据。

[0008] 1-2. 利用人工对文本数据进行标记，针对文本数据集中每一个文本数据 x_i ，标记对应的 y_i ， $y_i \in \{0, 1\}$ 。标记文本数据的目的在于根据业务需要利用人工经验对文本数据进行分析，其中，1表示和业务相关，0表示不相关。

[0009] 1-3. 利用Jieba分词器对文本数据集 X 进行分词，汇总得到所有文本数据的词表 W ， $W = \{w_1, w_2, \dots, w_D\}$ ，其中 D 为词表的总长度。 w_k 表示词表中第 k 个词。则文本数据 x_i 可以表示为 $x_i = \{w_{k_1}, w_{k_2}, \dots, w_{k_n}\}$ ，其中 k_n 表示文本数据 x_i 中第 n 个词在词表 W 中的序号，从而可用词表序号重新表示文本数据 $x_i = \{k_1, k_2, \dots, k_n\}$ 。

[0010] 步骤2：人工特征提取，构建wide特征。

[0011] 2-1. 统计文本数据集 X 中 $y_i = 1$ 对应的所有文本数据 x_i 中的词，得到出现频率最高的与业务相关的词，并排序获得序列I；以及 $y_i = 0$ 对应的所有文本数据 x_i 中的词，得到出现频率最高的与业务不相关的词，并排序获得序列II。

[0012] 2-2. 排除序列I和序列II中的语法型词汇，包括代词、冠词、连词等停用词。

[0013] 2-3. 将序列I和序列II中出现频率最高的文本特定词、 j 个高频词形成的组合作为wide特征，其中 j 范围为2-4，一般取值为2。

[0014] 所述的排序通过计算如下两个公式进行：

$$[0015] \quad w_{k_n} = \frac{\#\{(w_{k_n} \in x_i, y_i = 1)\}}{\#\{y_i = 1\}};$$

$$[0016] \quad w'_{k_n} = \frac{\#\{(w_{k_n} \in x_i, y_i = 0)\}}{\#\{y_i = 0\}};$$

[0017] 对 w_{k_n} 和 w'_{k_n} 分别进行排序获得序列I和序列II, 获取序列I中出现频率最大的5个词

$$w_{f_1}, w_{f_2}, \dots, w_{f_5}, w_f = \begin{cases} 0, & \text{其它} \\ 1, & w_f \in x_i \end{cases}, \text{以及序列II出现频率最大的5个词 } w'_{f_1}, w'_{f_2}, \dots, w'_{f_5},$$

$$w'_f = \begin{cases} 0, & \text{其它} \\ 1, & w'_f \in x_i \end{cases}.$$

[0018] 2-4. 获得最终的wide特征 $w_{f_1}, w_{f_2}, \dots, w_{f_5}, w'_{f_1}, w'_{f_2}, \dots, w'_{f_5}$ 。

[0019] 步骤3: 构建Wide&Deep深度学习模型。具体过程如下:

[0020] 3-1. 对文本数据 $x_i = \{w_{k_1}, w_{k_2}, \dots, w_{k_n}\}$ 的长度进行固定;

[0021] 当文本数据的长度小于预设的固定长度时, 使用“PAD”进行填充达到固定长度。

[0022] 当文本数据的长度大于预设的固定长度时, 丢弃超过预设的固定长度的部分;

[0023] 如果文本数据中的词不在词表W中, 使用“UNK”代替。

[0024] 比如预设的固定长度为1024个词, 如果原始词序列小于1024个, 则使用“PAD”进行填充, 如果长度超过1024则丢弃超过长度的部分。

[0025] 3-2. 建立每个词的embedding向量。将词表W中的所有词转化为预设的固定长度的稠密向量, 比如用128维向量。

[0026] 3-3. 获取一维卷积block。使用多个不同宽度的卷积核同时对embedding向量进行卷积。将每个卷积核的卷积结果进行拼接, 得到一维卷积block, 每次卷积后保证序列长度不变。

[0027] 比如使用宽度为1*128, 3*128, 5*128, 7*128的卷积核, 同时对embedding向量进行卷积;

[0028] 3-4. 多次叠加一维卷积。

[0029] 利用3-3定义的一维卷积block对wide特征进行处理, 得到新的embedding向量, 在每次卷积后使用窗口大小为4, 步长为4的最大池化。每一次最大池化操作后长度缩小1/4; 反复进行4次卷积和最大池化操作; 获取最终序列长度变为64, 特征长度为8。

[0030] 3-5. 文本表示。将序列长度为64, 特征长度为8的向量拉直成一个序列长度为512维的向量, 得到整个文本序列的512维特征向量的表示。

[0031] 3-6. 构建损失函数。使用全连接层将512维特征向量映射成(16+1)*2维的向量, 再通过softmax函数, 转化为(16+1)*2个文本数据标记 y_i 的预测概率值, 其中包括: 1个文本数据标记的预测概率值, 以及16个wide特征的预测概率值。损失函数可以表示为:

$$[0032] \quad loss = y_i \log p_i + (1 - y_i) \log(1 - p_i) + \sum_{k=1}^{16} [f_{i,k} \log q_{i,k} + (1 - f_{i,k}) \log(1 - q_{i,k})]$$

[0033] 其中, y_i 为文本数据标记, p_i 为文本数据标记 y_i 的预测概率值, $f_{i,k}$ 为wide特征值,

$q_{i,k}$ 为wide特征的预测概率值, i 为数据标记的索引, 表示第 i 个文本数据。

[0034] 步骤4: 模型训练。

[0035] 通过梯度下降算法, 求得步骤3-6中损失函数的最小化来训练模型。

[0036] 步骤5: 模型预测。

[0037] 对需要预测的文本内容, 利用步骤3计算出文本数据标记 y_i 的预测值, 得到文本内容与业务的相关程度。

[0038] 本发明有益效果如下:

[0039] 本发明提及的人工特征作为深度学习模型训练过程中的附加任务, 为模型提供了更多的有效信息, 提高了模型学习效率。并且, 在文本内容预测时不需要计算wide特征, 从而提升了模型性能。

[0040] 不同宽度的卷积核卷积后, 可以通过一次卷积block获得不同距离上的信息, 这有助于提升模型性能。同时, 减少调整不同卷积核的时间, 通过模型学习能自动适应不同的卷积核宽度。

附图说明

[0041] 图1为本发明架构图;

[0042] 图2为本发明训练模块的整体架构;

[0043] 图3为本发明输入模块的具体流程;

[0044] 图4为输入模块的具体流程示意图;

具体实施方式

[0045] 下面结合附图和实施例对本发明作进一步说明。

[0046] 如图1-4所示, 一种基于Wide&Deep深度学习模型的文本预测方法, 首先对文本数据进行标注, 其次构建wide特征, 然后对基于Wide&Deep的深度学习模型添加辅助训练任务, 最后对文本内容进行预测。具体实现步骤如下:

[0047] 步骤1: 文本数据的预处理。

[0048] 1-1. 收集文本数据集 $X = \{x_1, x_2, \dots, x_N\}$, N 为总的训练样本数量。其中, $x_i \in \{x_1, x_2, \dots, x_N\}$, x_i 表示第 i 个待处理的文本数据。

[0049] 1-2. 利用人工对文本数据进行标记, 针对文本数据集中每一个文本数据 x_i , 标记对应的 y_i , $y_i \in \{0, 1\}$ 。标记文本数据的目的在于根据业务需要利用人工经验对文本数据进行分析, 其中, 1表示和业务相关, 0表示不相关。

[0050] 1-3. 利用Jieba分词器对文本数据集 X 进行分词, 汇总得到所有文本数据的词表 W , $W = \{w_1, w_2, \dots, w_D\}$, 其中 D 为词表的总长度。 w_k 表示词表中第 k 个词。则文本数据 x_i 可以表示为 $x_i = \{w_{k_1}, w_{k_2}, \dots, w_{k_n}\}$, 其中 k_n 表示文本数据 x_i 中第 n 个词在词表 W 中的序号, 从而可用词表序号重新表示文本数据 $x_i = \{k_1, k_2, \dots, k_n\}$ 。

[0051] 步骤2: 人工特征提取, 构建wide特征。

[0052] 2-1. 统计文本数据集 X 中 $y_i = 1$ 对应的所有文本数据 x_i 中的词, 得到出现频率最高的与业务相关的词, 并排序获得序列 I ; 以及 $y_i = 0$ 对应的所有文本数据 x_i 中的词, 得到出现

频率最高的与业务不相关的词,并排序获得序列Ⅱ。

[0053] 2-2.排除序列I和序列Ⅱ中的语法型词汇,包括代词、冠词、连词等停用词。

[0054] 2-3.将序列I和序列Ⅱ中出现频率最高的文本特定词、j个高频词形成的组合作为wide特征,其中j范围为2-4,一般取值为2。

[0055] 所述的排序通过计算如下两个公式进行:

$$[0056] \quad w_{k_n} = \frac{\#\{(w_{k_n} \in x_i, y_i = 1)\}}{\#\{y_i = 1\}};$$

$$[0057] \quad w'_{k_n} = \frac{\#\{(w_{k_n} \in x_i, y_i = 0)\}}{\#\{y_i = 0\}};$$

[0058] 对 w_{k_n} 和 w'_{k_n} 分别进行排序获得序列I和序列Ⅱ,获取序列I中出现频率最大的5个词

$w_{f_1}, w_{f_2}, \dots, w_{f_5}$, $w_f = \begin{cases} 0, & \text{其它} \\ 1, & w_f \in x_i \end{cases}$, 以及序列Ⅱ出现频率最大的5个词 $w'_{f_1}, w'_{f_2}, \dots, w'_{f_5}$,

$$w'_f = \begin{cases} 0, & \text{其它} \\ 1, & w'_f \in x_i \end{cases}.$$

[0059] 2-4.获得最终的wide特征 $w_{f_1}, w_{f_2}, \dots, w_{f_5}, w'_{f_1}, w'_{f_2}, \dots, w'_{f_5}$ 。

[0060] 步骤3:构建Wide&Deep深度学习模型。具体过程如下:

[0061] 3-1.对文本数据 $x_i = \{w_{k_1}, w_{k_2}, \dots, w_{k_n}\}$ 的长度进行固定;

[0062] 当文本数据的长度小于预设的固定长度时,使用“PAD”进行填充达到固定长度。

[0063] 当文本数据的长度大于预设的固定长度时,丢弃超过预设的固定长度的部分;

[0064] 如果文本数据中的词不在词表W中,使用“UNK”代替。

[0065] 比如预设的固定长度为1024个词,如果原始词序列小于1024个,则使用“PAD”进行填充,如果长度超过1024则丢弃超过长度的部分。

[0066] 3-2.建立每个词的embedding向量。将词表W中的所有词转化为预设的固定长度的稠密向量,比如用128维向量。

[0067] 3-3.获取一维卷积block。使用多个不同宽度的卷积核同时对embedding向量进行卷积。将每个卷积核的卷积结果进行拼接,得到一维卷积block,每次卷积后保证序列长度不变。

[0068] 比如使用宽度为1*128,3*128,5*128,7*128的卷积核,同时对embedding向量进行卷积;

[0069] 3-4.多次叠加一维卷积。

[0070] 利用3-3定义的一维卷积block对wide特征进行处理,得到新的embedding向量,在每次卷积后使用窗口大小为4,步长为4的最大池化。每一次最大池化操作后长度缩小1/4;反复进行4次卷积和最大池化操作;获取最终序列长度变为64,特征长度为8。

[0071] 3-5.文本表示。将序列长度为64,特征长度为8的向量拉直成一个序列长度为512维的向量,得到整个文本序列的512维特征向量的表示。

[0072] 3-6.构建损失函数。使用全连接层将512维特征向量映射成(16+1)*2维的向量,再

通过softmax函数,转化为 $(16+1)*2$ 个文本数据标记 y_i 的预测概率值,其中包括:1个文本数据标记的预测概率值,以及16个wide特征的预测概率值。损失函数可以表示为:

$$[0073] \quad loss = y_i \log p_i + (1 - y_i) \log(1 - p_i) + \sum_{k=1}^{16} [f_{i,k} \log q_{i,k} + (1 - f_{i,k}) \log(1 - q_{i,k})]$$

[0074] 其中, y_i 为文本数据标记, p_i 为文本数据标记 y_i 的预测概率值, $f_{i,k}$ 为wide特征值, $q_{i,k}$ 为wide特征的预测概率值, i 为数据标记的索引,表示第 i 个文本数据。

[0075] 步骤4:模型训练。

[0076] 通过梯度下降算法,求得步骤3-6中损失函数的最小化来训练模型。

[0077] 步骤5:模型预测。

[0078] 对需要预测的文本内容,利用步骤3计算出文本数据标记 y_i 的预测值,得到文本内容与业务的相关程度。

[0079] 如图1所示,本发明方法能包装成3个部分,其中,第一个部分为训练数据整理,主要是收集和清洗与任务相关的数据。第二个部分为wide特征构建,主要是通过行业知识,对数据进行加工,产生特征词典。第三个部分为模型训练,主要是利用训练数据和wide特征构建深度学习模型。

[0080] 本发明核心的模型训练部分又包含3个阶段,如图1所示,包括:(1)模型输入,一般的是对文本数据进行结构化表示,具体如图3所示。(2)深度学习模型,由多个卷积block构成,每个卷积block结构如图4所示。(3)模型输出,本发明中包含文本数据标记的预测值和wide特征值。

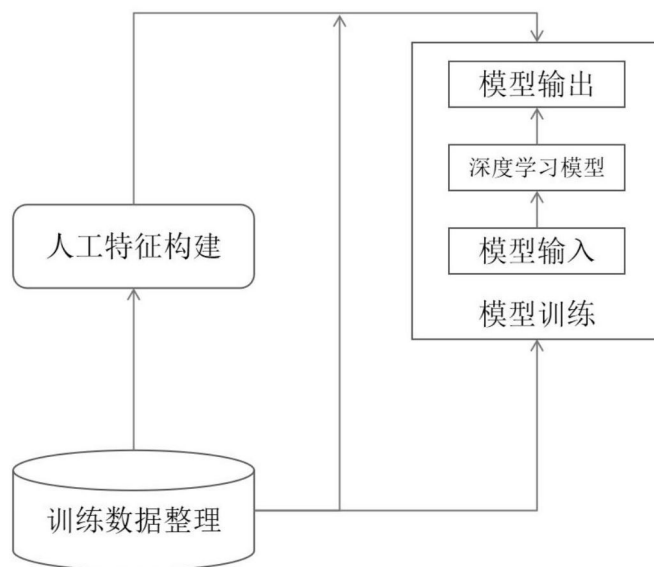


图1

$$\text{预处理} \rightarrow \text{embedding} \rightarrow [\text{卷积block} \times \text{max pool}]_{\times 3} \rightarrow \begin{bmatrix} p, 1-p \\ q_{i,k}, q_{i,k} \\ \vdots \\ q_{16}, q_{16} \end{bmatrix}$$

图2

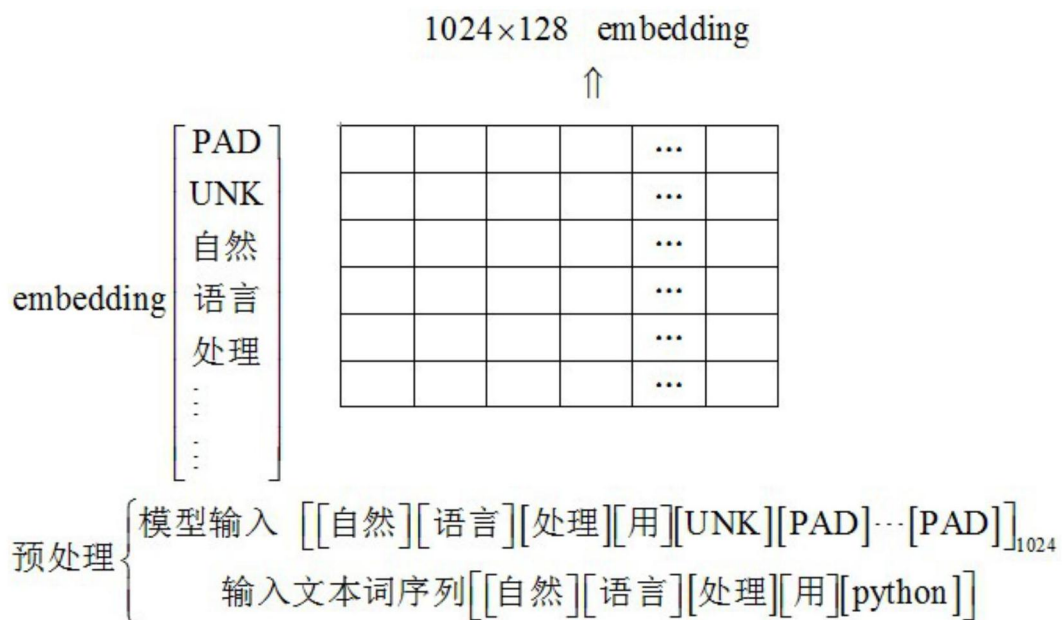


图3

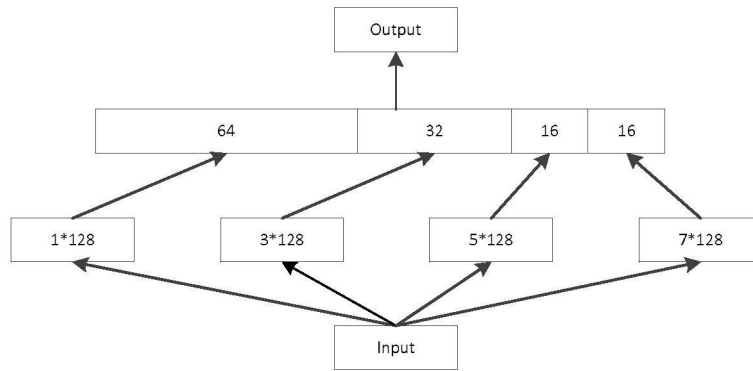


图4