TextCNN算法介绍

鲲鹏创新实践课：鲲鹏应用数据分析与管理实战

版权所有© 2020 华为技术有限公司

前言

- 本章主要介绍TextCNN在情感分析的应用，在文本分类任务中，我们也可以将文本当作一维图像，从而可以用一维卷积神经网络来捕捉临近词之间的关联。本节将介绍将卷积神经网络应用到文本分析的开创性工作之一：textCNN。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

目标

- 学完本课程后，您将能够：
  - 理解卷积神经网络(CNN)的原理和结构；
  - 了解TextCNN的应用场景；
  - 理解TextCNN的原理和结构；
  - 创建一个用于文本分类的TextCNN模型。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

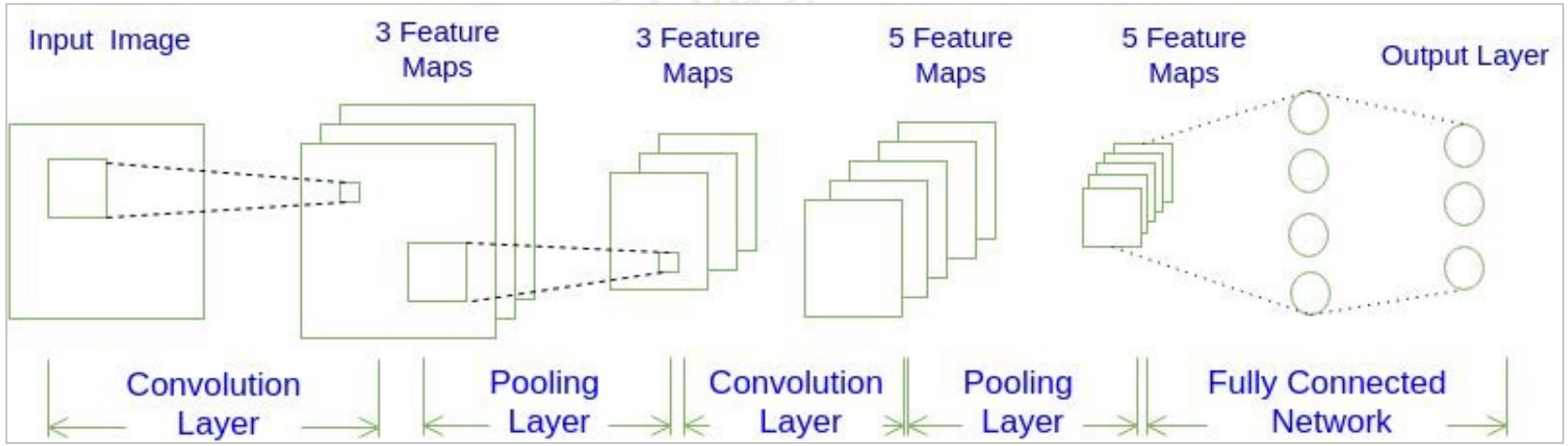
目录

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

卷积神经网络介绍

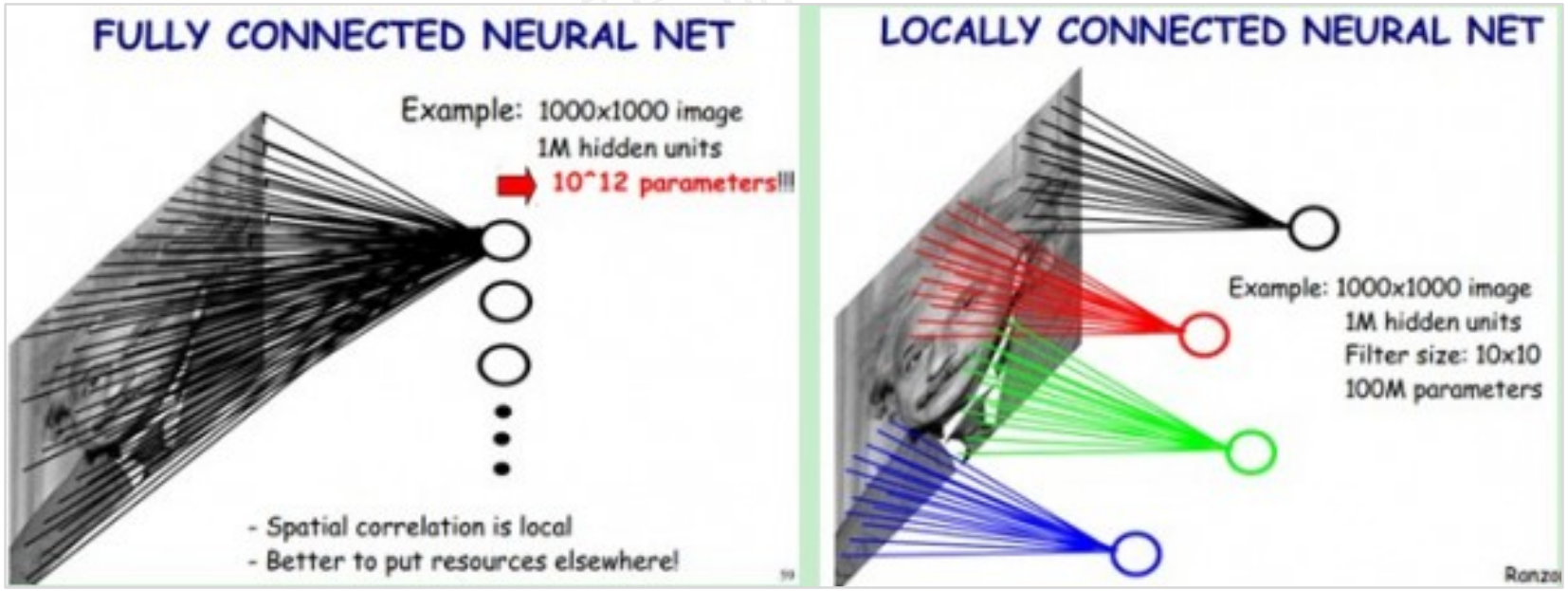
- 卷积神经网络（Convolutional Neural Network, CNN）是一种前馈神经网络，它的人工神经元可以响应一部分覆盖范围内的周围单元，对于二维图像处理有出色表现。它包括卷积层（convolution layer），池化层（pooling layer）和全连接层（fully connected layer）。



版权所有© 2020 华为技术有限公司

CNN核心思想 - 局部感知

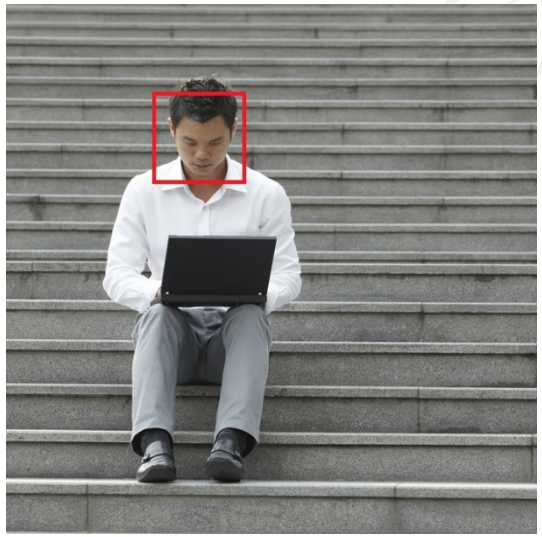
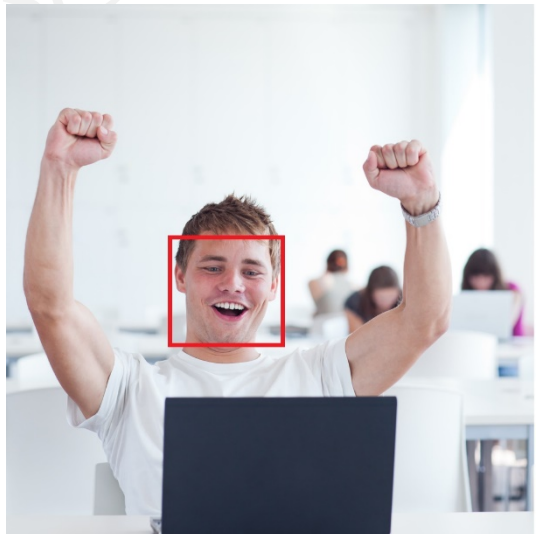
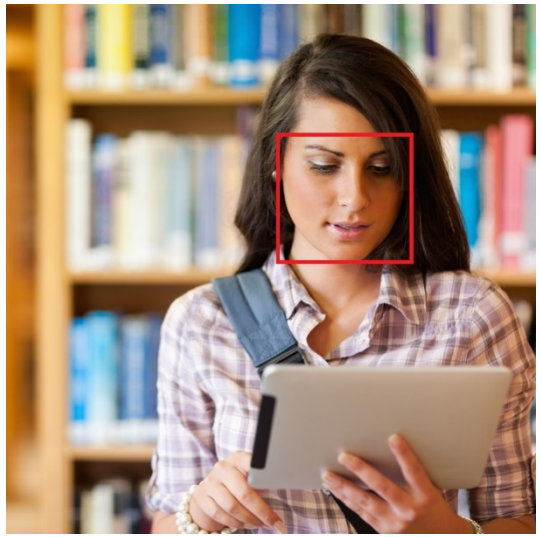
- 局部感知：一般认为人对外界的认知是从局部到全局的，图像像素点的空间联系也是局部的像素联系较为紧密，而距离较远的像素相关性则较弱。
- 实现方式：使kernel的尺寸远小于输入图像的尺寸。



版权所有© 2020 华为技术有限公司

CNN核心思想 - 参数共享

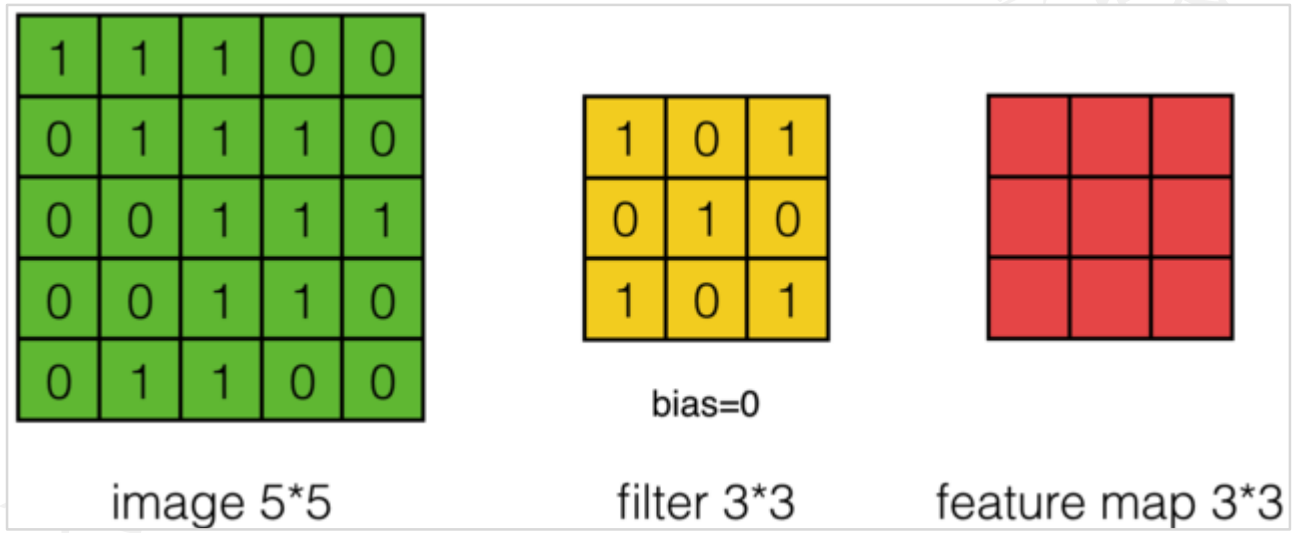
- 优点:
  - 解决图像位置不变性的问题。
  - 减少计算和内存需求。
- 实现:
  - 用参数相同的kernel去扫描整副图像

课程版权属于银鹏创新实践课

版权所有© 2020 华为技术有限公司

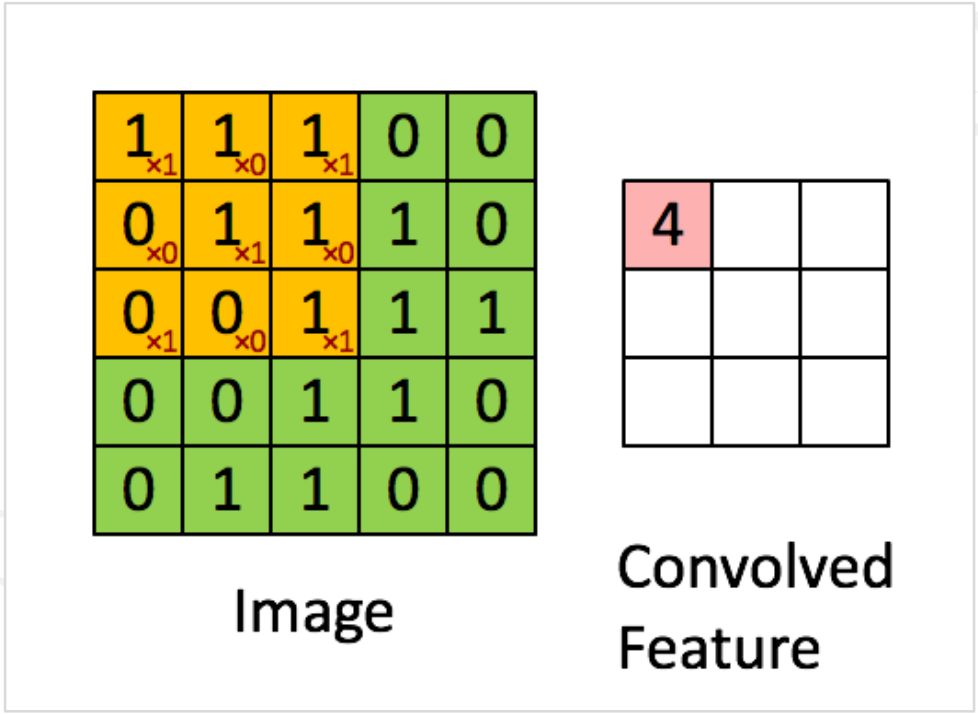
单卷积核计算 (1)



版权所有© 2020 华为技术有限公司

单卷积核计算 (2)

- 卷积效果演示：

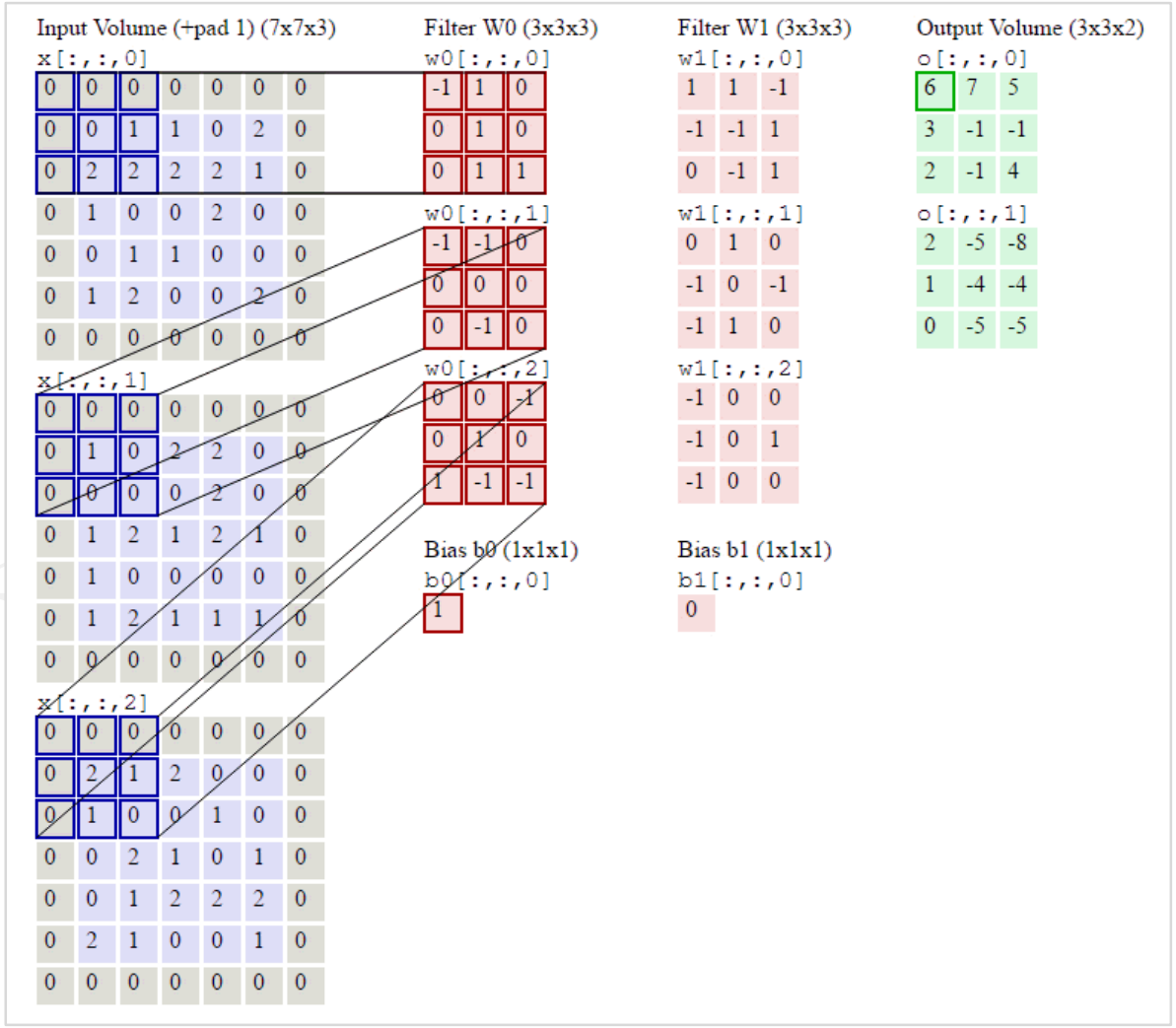


浙大 实践课

版权所有© 2020 华为技术有限公司

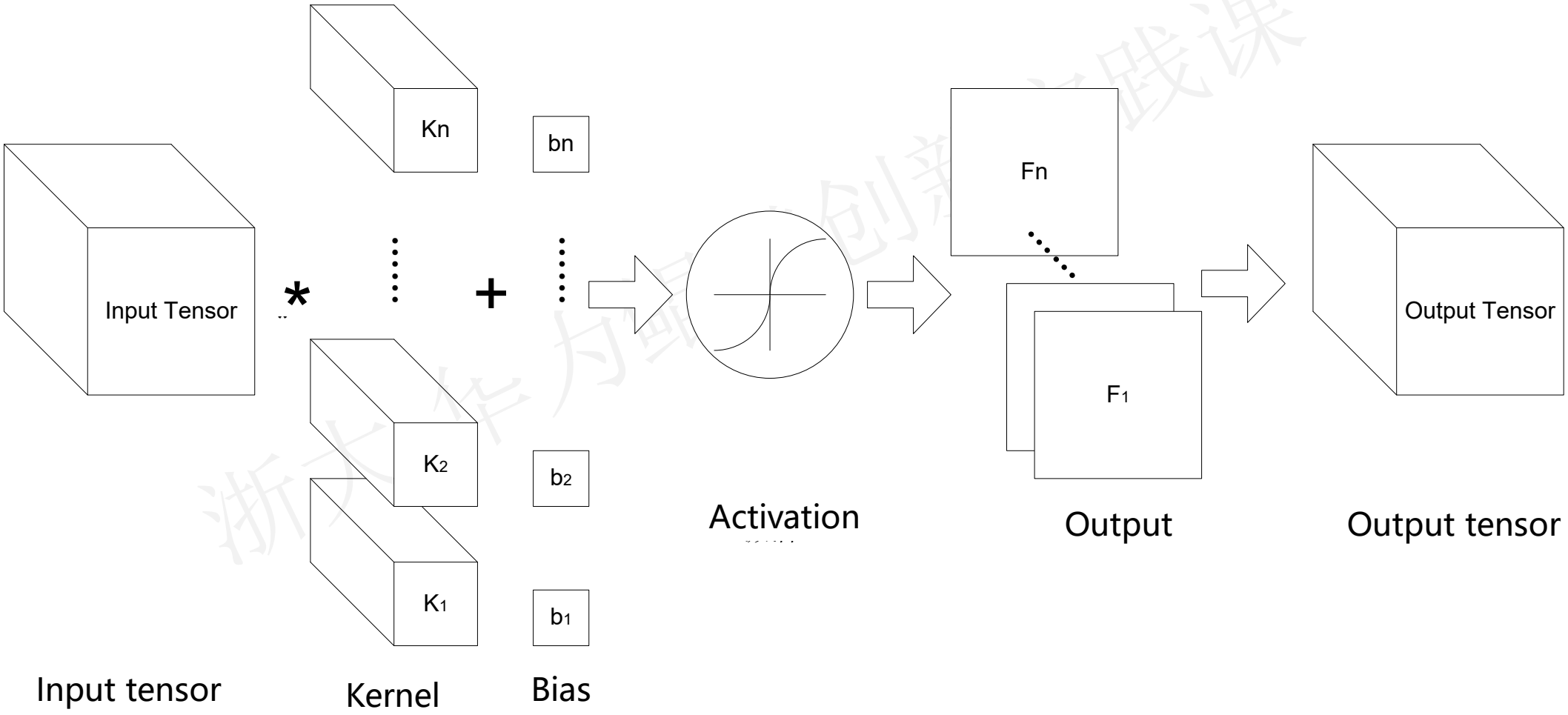
多卷积核运算

- 输入图像：
- 5*5*3
- Padding = 1
- 卷积核2个
- 3*3*3
- Stride = 2
- 特征图：
- 3*3*2



版权所有© 2020 华为技术有限公司

卷积层

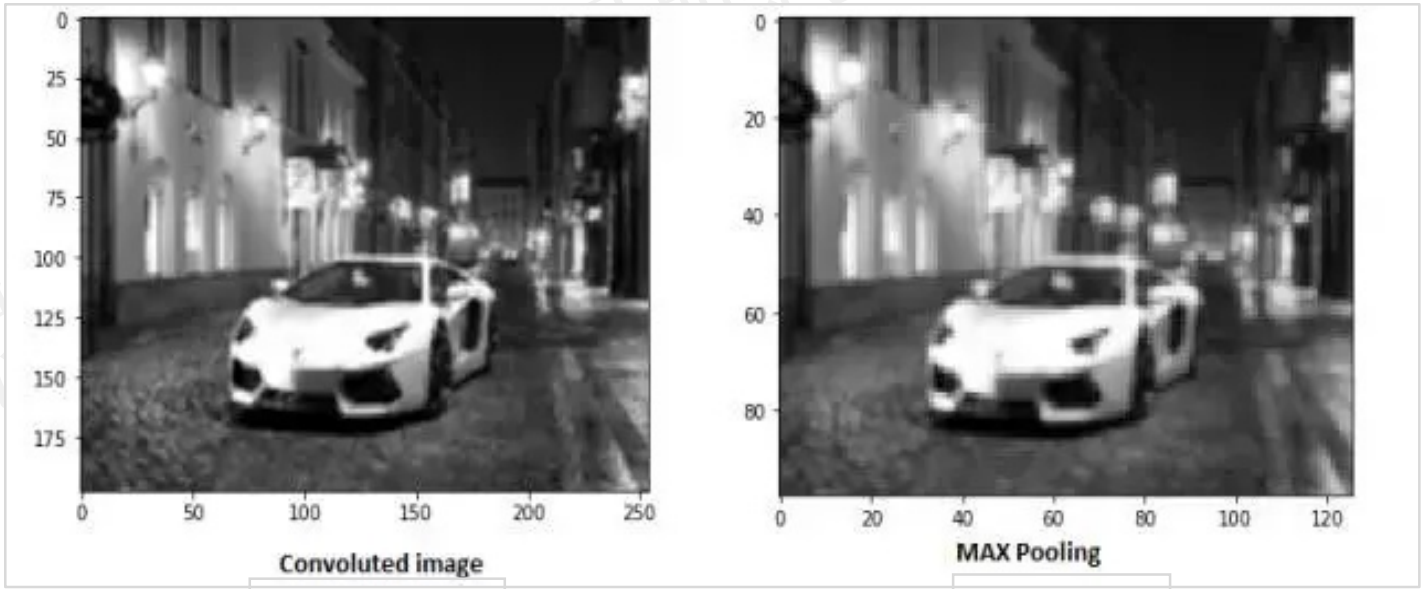


浙大华为创新实践课

版权所有© 2020 华为技术有限公司

池化层 (1)

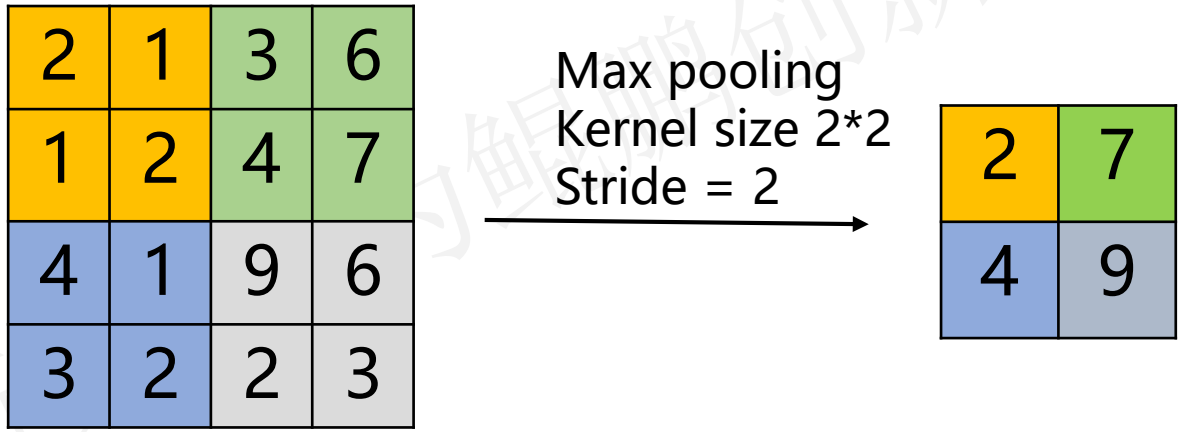
- 池化的目的是减少图像特征图（feature map）的空间尺寸。
- 有时模型太大，我们需要减少训练参数的数量，它被要求在随后的卷积层之间周期性地引进池化层。
- 池化层一般分为最大池化（max pooling）和平均池化（mean pooling）。
- 池化层的最常见形式是最大池化。



版权所有© 2020 华为技术有限公司

池化操作

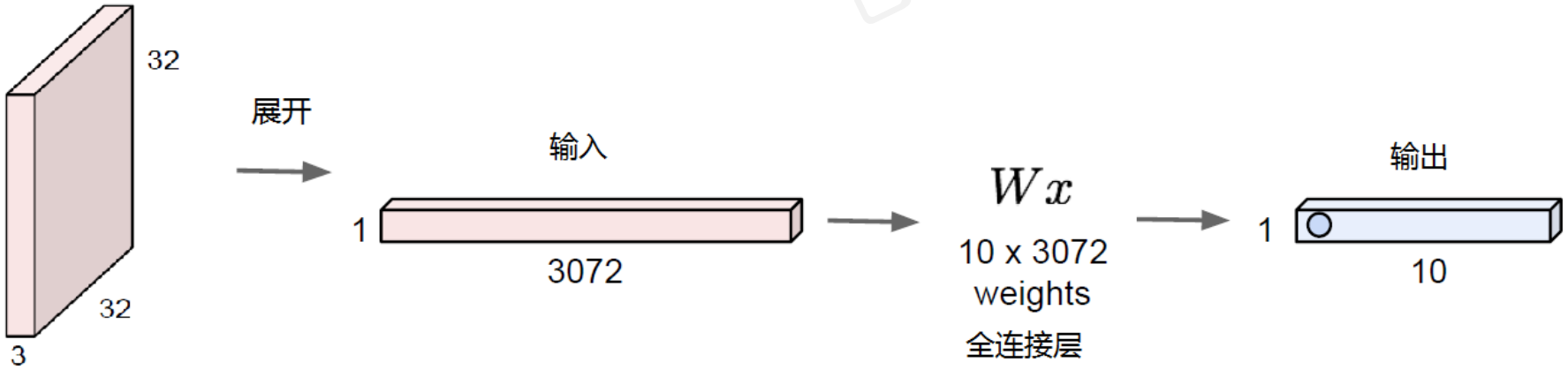
- 在这里，我们把步长定为 2，池化尺寸也为 2。最大池化也应用在调整卷积的输出尺寸中。最大池化操作后，4*4 的图片经过池化操作，输出的大小变成了 2*2。



版权所有© 2020 华为技术有限公司

全连接层

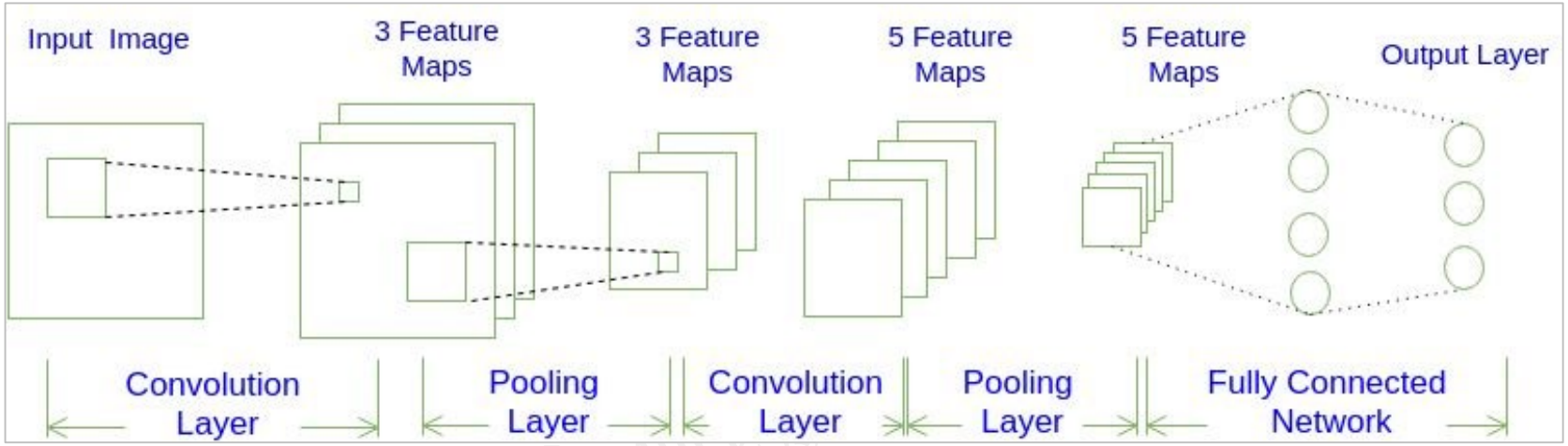
- 全连接层是整个网络的“**分类器**“，可以用来将最后得到的特征映射到线性可分的空间。
- 使用的Softmax激活函数将最终的输出量化。

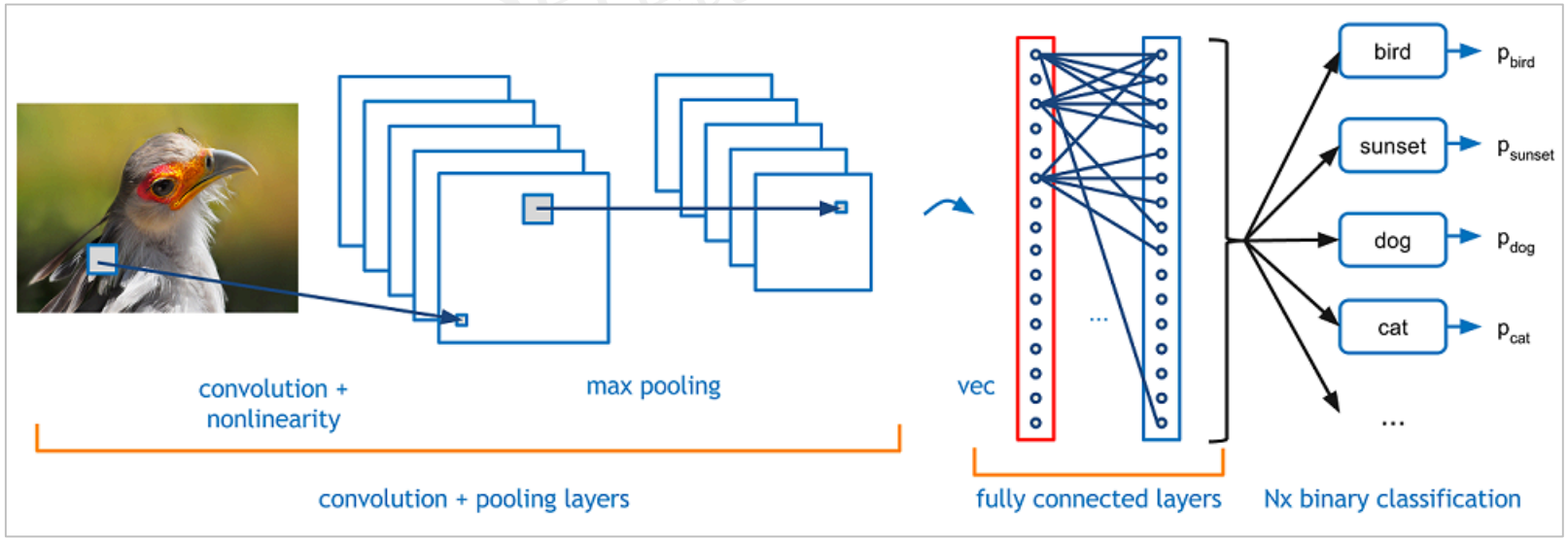


创新实践课

版权所有© 2020 华为技术有限公司

卷积神经网络的应用 – 计算机视觉





版权所有© 2020 华为技术有限公司

目录

浙大-华为鲲鹏创新实践课

 版权所有© 2020 华为技术有限公司

HUAWEI

文本分类问题的步骤

- 文本分类问题是将给定的文本根据问题定义分成两个或多个类别的任务。
- 传统的解决文本分类问题的步骤通常为：特征提取、特征选择、向量表示以及构建分类器。
- 基于深度学习的文本分类问题，步骤通常为：特征提取、构建模型、调参优化，我们通常使用CNN、RNN等模型处理。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

TextCNN简介（1）

- TextCNN使用卷积神经网络（CNN）用预先训练好的词向量对句子级别的文本分类做了一系列实验。
  - 实验表明，一个简单的CNN，只需少量超参数调整和静态向量（词向量不参与训练），就可以在多个基准测试中取得出色的结果。通过微调特定学习任务的词向量可以进一步提高性能。
  - 论文还对模型架构进行简单的修改，能同时使用特定任务的词向量和静态向量。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

TextCNN简介 (2)

- 在文本分类问题中，我们经常关注局部信息。
- 文本中的局部信息通常指同时出现的词。
- CNN具有提取这种信息的能力，将CNN应用到文本分类领域的模型就是TextCNN。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

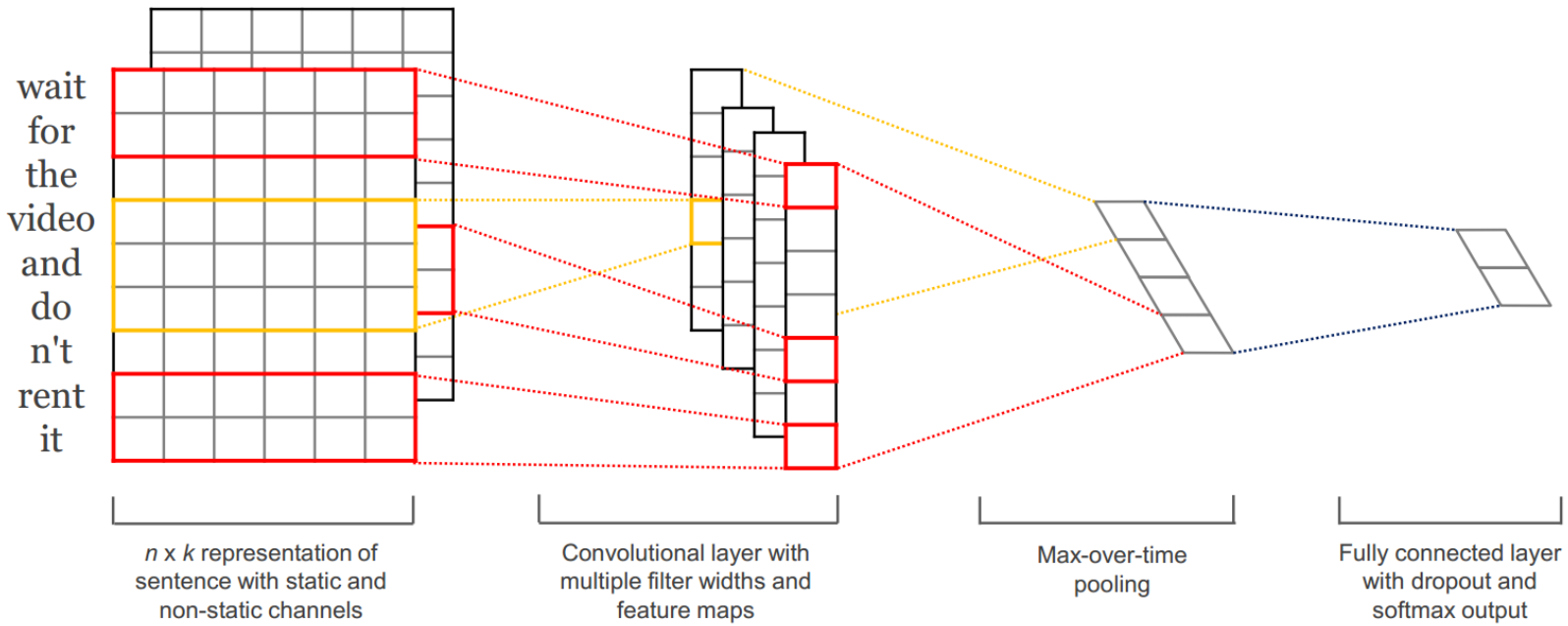
目录

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

TextCNN

- TextCNN是一种使用卷积神经网络对文本进行分类的模型。
- 文本在经过词嵌入后被送入神经网络，与CNN不同的是,TextCNN的卷积操作用在一维数据上。
- TextCNN提取的特征类似N-gram模型，是词的局部关系。



版权所有© 2020 华为技术有限公司

TextCNN嵌入层 (1)

- 文本数据需要向量化才能导入神经网络。
- 通常来说，向量化是词级别的，我们首先要对文本进行预处理，分词，停用词登记等操作。
- 在进行完分词后，需要使用词嵌入(word embedding)方法将词转换为向量，常用的方法为word2vec。
- TextCNN使用预训练的词向量模型做为嵌入层。
- 好的词嵌入是NLP任务性能的关键。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

TextCNN嵌入层 (2)

- word2vec是一种将稀疏的词向量表示转换成稠密的词向量表示的方法。
- Word2vec相比one-hot表示的好处是降低计算量，并且能表示词之间的相关性。
- word2vec的结果如下图：

版权所有© 2020 华为技术有限公司

嵌入层的几种结构

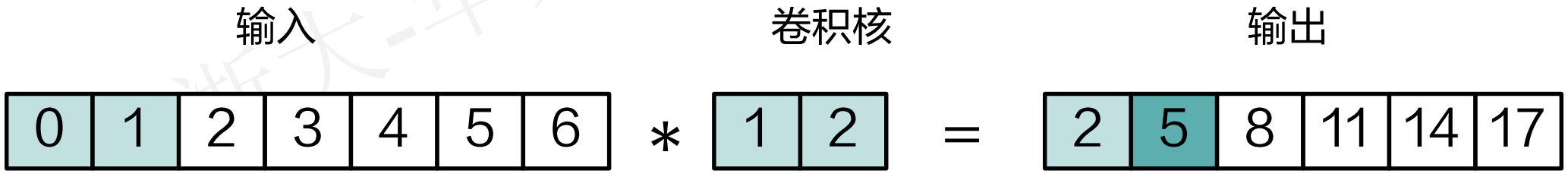
- CNN-rand：作为一个基础模型，嵌入层所有词向量被随机初始化，然后模型整体进行训练。
- CNN-static：模型使用预训练的word2vec初始化嵌入层，对于那些在预训练的word2vec没有的单词，随机初始化。然后固定嵌入层，更新网络的其余部分。
- CNN-non-static：训练的时候，整个嵌入层跟随整个网络一起训练。
- CNN-multichannel：嵌入层有两个通道，一个通道为static，一个为non-static。然后整个网络更新时只有一个通道更新参数。两个通道都是使用预训练的word2vec初始化的。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

TextCNN卷积层 (1)

- TextCNN是一种使用卷积神经网络对文本进行分类的模型。
- 文本在经过词嵌入后被送入神经网络，与CNN不同的是，TextCNN的卷积操作是在一维数据上。
- 一维卷积的操作如下，提取的仍然是局部特征：

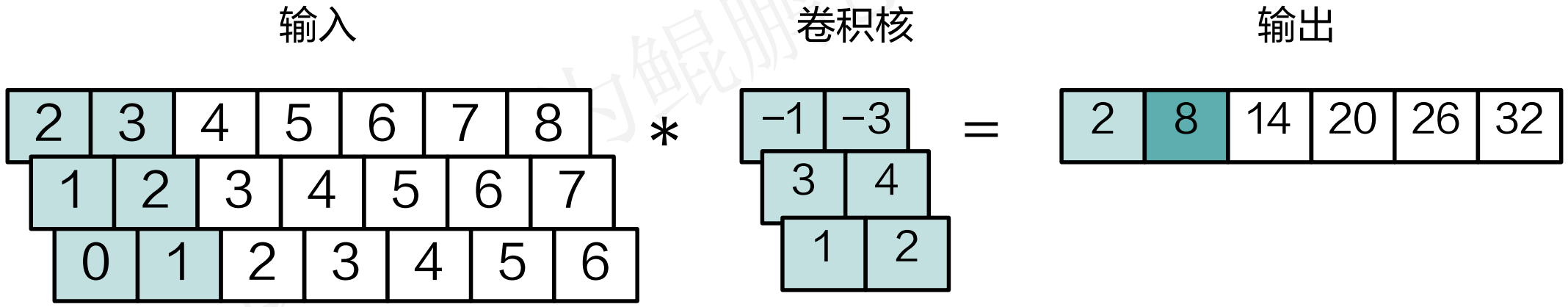


浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司

TextCNN卷积层 (2)

- 在一维卷积过程中，也涉及到多通道的情况。
- 一维卷积的多通道计算如下：

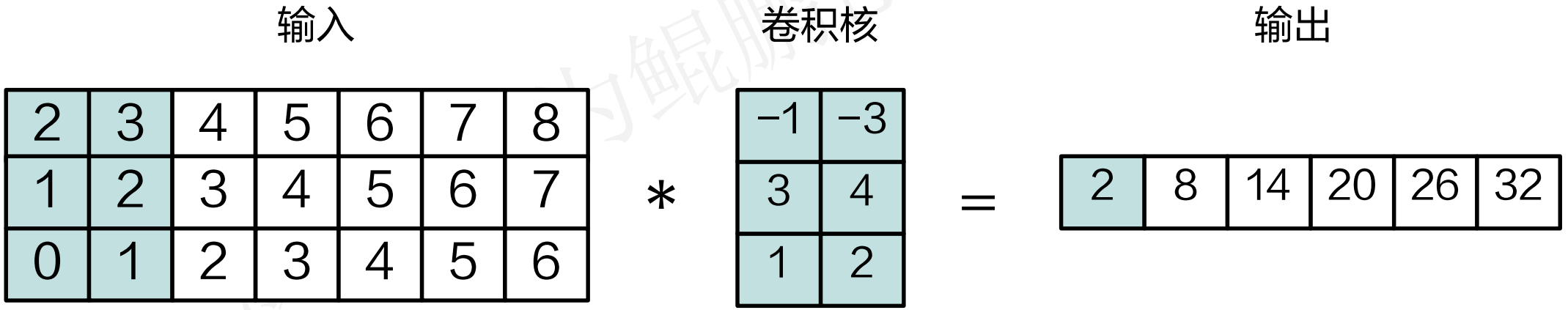


内部鲲鹏创新实践课

版权所有© 2020 华为技术有限公司

TextCNN卷积层 (3)

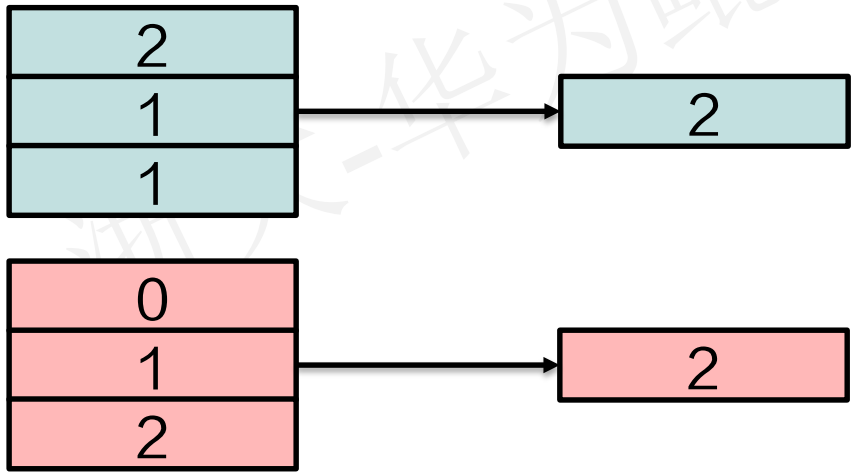
- 多输入通道的一维互相关运算以等价的单输入通道的二维互相关运算呈现。这里核的高等于输入的高。
- TextCNN用于文本分类时，一维卷积操作的方向是从上到下，而不是词向量的方向。



版权所有© 2020 华为技术有限公司

TextCNN – 时序最大池化层

- TextCNN使用的池化方法为max-over-time pooling，既时序最大池化。
- 假设输入包含多个通道，各通道由不同时间步上的数值组成，各通道的输出即该通道所有时间步中最大的数值。因此，时序最大池化层的输入在各个通道上的时间步数可以不同。



浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司

TextCNN的Softmax层

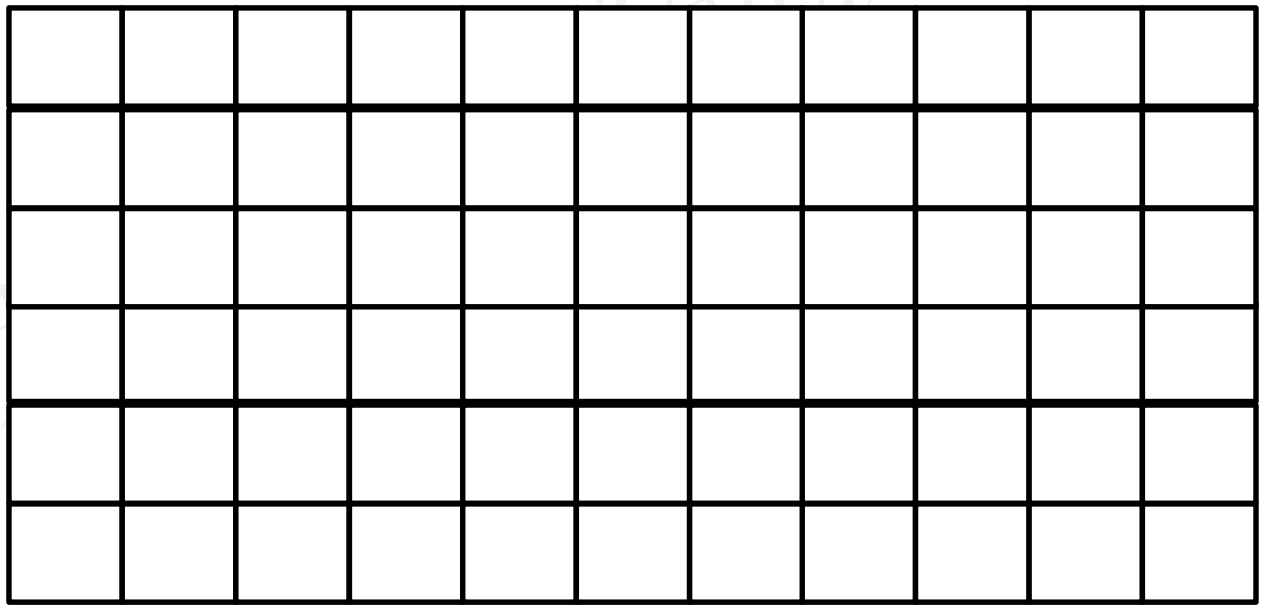
- 将 max-pooling的结果拼接起来, 送入到softmax当中, 得到各个类别的概率。
  - 预测阶段：到这里整个TextCNN的流程结束。
  - 训练阶段：此时会根据预测label以及实际label来计算损失函数，计算出softmax 函数、max-pooling 函数、激活函数以及卷积核函数四个函数当中参数需要更新的梯度，来依次更新这四个函数中的参数，完成一轮训练。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

正面/负面情感预测 – 示例 (1)

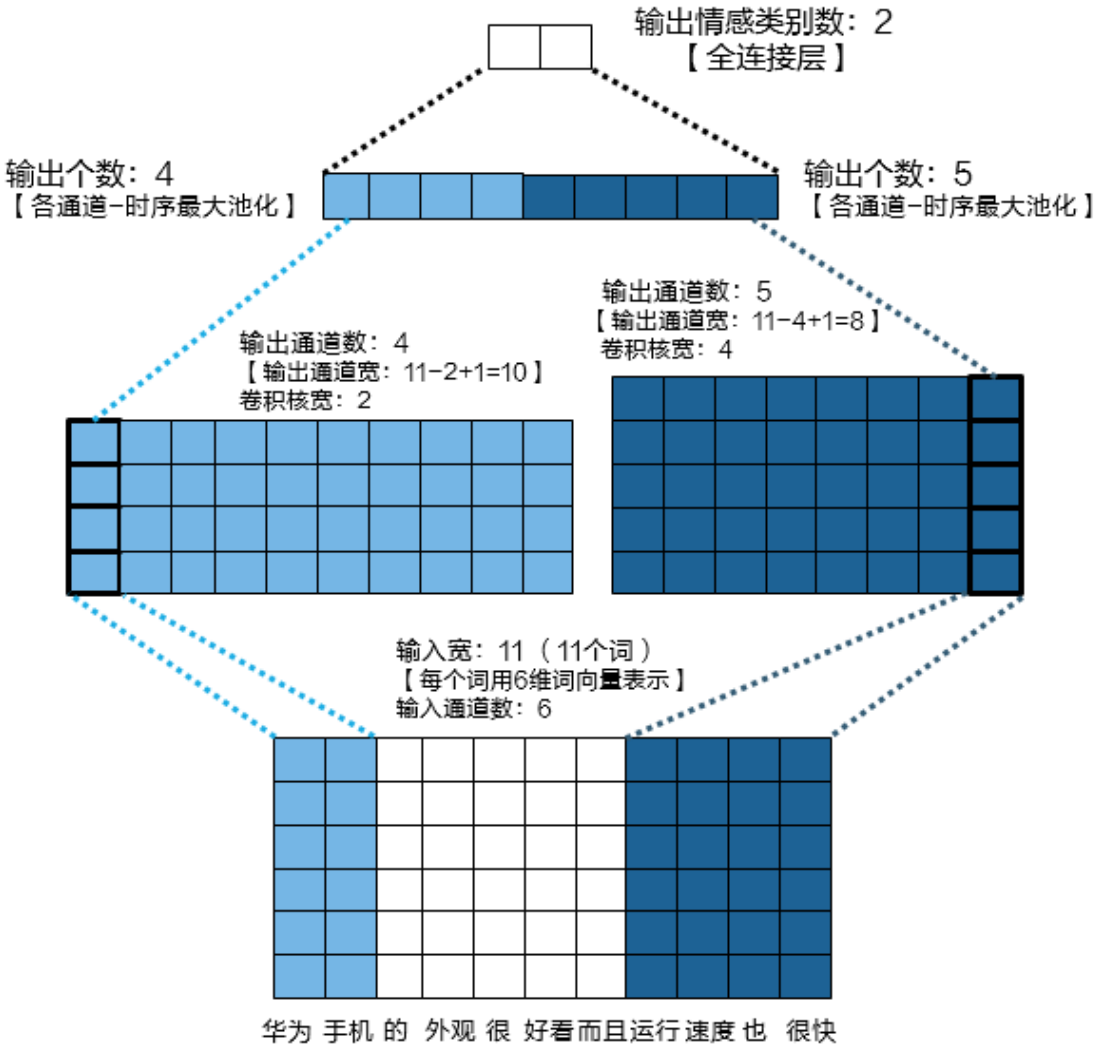
- 输入是一个有11个词的句子，每个词用6维词向量表示：
  - 因此输入序列的宽为11，输入通道数为6。
- 给定2个一维卷积核，核宽分别为2和4，输出通道数分别设为4和5。



华为 手机 的 外观 很 好看 而且 运行 速度 也 很快

版权所有© 2020 华为技术有限公司

正面/负面情感预测 – 示例 (2)



版权所有© 2020 华为技术有限公司

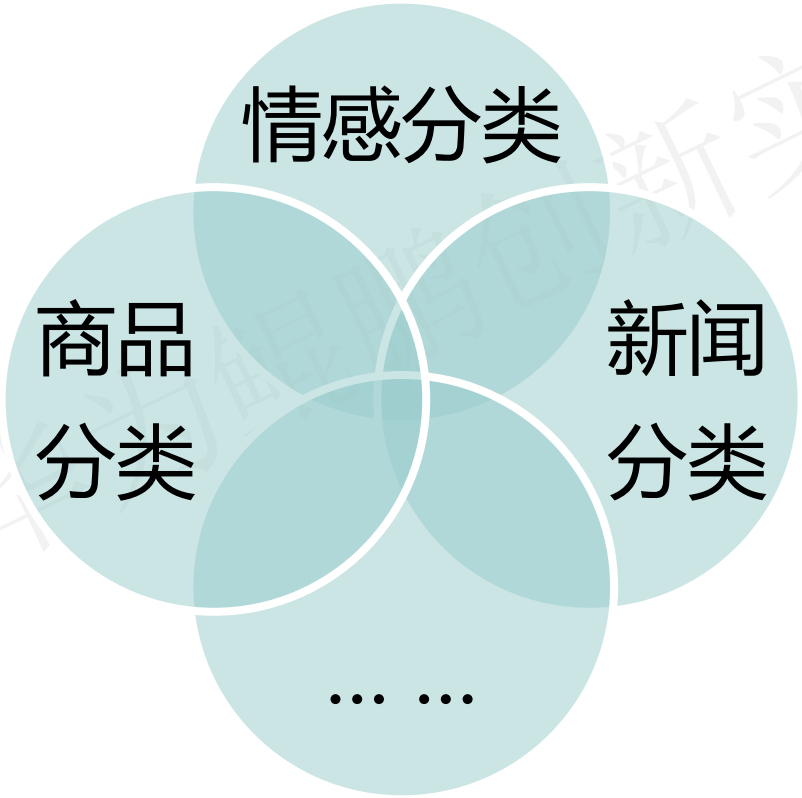
目录

浙大-华为鲲鹏创新实践课

 版权所有© 2020 华为技术有限公司

HUAWEI

TextCNN应用



浙大-华为智能计算创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

TextCNN用于文本分类的分析

- TextCNN的流程：先将文本分词做embedding得到词向量, 将词向量经过一层卷积, 一层max-pooling, 最后将输出外接softmax 来做n分类。
- TextCNN 的优势：模型简单, 训练速度快，效果不错。
- TextCNN的缺点：模型可解释型不强，在调优模型的时候，很难根据训练的结果去针对性的调整具体的特征，因为在TextCNN中没有类似GBDT模型中特征重要度(feature importance)的概念, 所以很难去评估每个特征的重要性。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

本章总结

- TextCNN是使用CNN处理文本问题的一种方法。CNN是一种端到端的，可以提取局部信息的模型，常用于图像处理。TextCNN可以提取文本中的局部信息，在文本分类任务中也有一定的效果。

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI

思考题

1. 以下哪些是TextCNN的结构？（　）

   A. 卷积层

   B. 时序最大池化层

   C. 全连接层

   D. 权重共享层

2. TextCNN的卷积方式为？（　）

   A. 在单个词向量上进行卷积

   B. 从上到下，从左到右进行卷积

   C. 对整个输入矩阵进行卷积

   D. 在输入矩阵上，沿词向量顺序进行卷积

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司

学习推荐

- Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

浙大-华为鲲鹏创新实践课

版权所有© 2020 华为技术有限公司
HUAWEI


谢 谢
www.huawei.com