# Best Practices for Document Classification with Deep Learning 深入学习文档分类的最佳实践

原文链接：  
原文链接：

Text classification describes a general class of problems such as predicting the sentiment of tweets and movie reviews, as well as classifying email as spam or not.  
文本分类描述了一类常见的问题，例如预测推特和电影评论的情绪，以及是否将电子邮件分类为垃圾邮件。

Deep learning methods are proving very good at text classification, achieving state-of-the-art results on a suite of standard academic benchmark problems.  
事实证明，深度学习方法非常擅长文本分类，在一系列标准学术基准问题上取得了最先进的结果。

In this post, you will discover some best practices to consider when developing deep learning models for text classification.  
在这篇文章中，您将发现在为文本分类开发深度学习模型时需要考虑的一些最佳实践。

After reading this post, you will know:  
读完这篇文章，你会知道：

* The general combination of deep learning methods to consider when starting your text classification problems.  
  在开始文本分类问题时要考虑的深度学习方法的一般组合。
* The first architecture to try with specific advice on how to configure hyperparameters.  
  第一个尝试使用关于如何配置超参数的特定建议的架构。
* That deeper networks may be the future of the field in terms of flexibility and capability.  
  在灵活性和能力方面，更深层次的网络可能是该领域的未来。

Let’s get started.  
我们开始吧。

Best Practices for Document Classification with Deep Learning Photo by , some rights reserved.  
文档分类的最佳实践与深入学习图片由，保留一些权利。

## Overview 概述

This tutorial is divided into 5 parts; they are:  
本教程分为5个部分，分别是：

1. Word Embeddings + CNN = Text Classification  
   Word Embeddings+CNN=文本分类
2. Use a Single Layer CNN Architecture  
   使用单层CNN架构
3. Dial in CNN Hyperparameters  
   拨入CNN超参数
4. Consider Character-Level CNNs  
   考虑字符级CNN
5. Consider Deeper CNNs for Classification  
   考虑更深层次的cnn进行分类

### Need help with Deep Learning for Text Data? 需要帮助深入学习文本数据吗？

Take my free 7-day email crash course now (with code).  
现在就参加我7天的免费电子邮件速成课程（带代码）。

Click to sign-up and also get a free PDF Ebook version of the course.  
点击注册，也可以获得免费的PDF电子书版本的课程。

[Start Your FREE Crash-Course Now](https://machinelearningmastery.lpages.co/leadbox/144855173f72a2%3A164f8be4f346dc/5655638436741120/)

## 1. Word Embeddings + CNN = Text Classification 一。Word Embeddings+CNN=文本分类

The modus operandi for text classification involves the use of a word embedding for representing words and a Convolutional Neural Network (CNN) for learning how to discriminate documents on classification problems.  
文本分类的方法包括使用单词嵌入来表示单词，以及使用卷积神经网络（CNN）来学习如何根据分类问题来区分文档。

Yoav Goldberg, in his primer on deep learning for natural language processing, comments that neural networks in general offer better performance than classical linear classifiers, especially when used with pre-trained word embeddings.  
Yoav Goldberg在他的《自然语言处理的深度学习入门》一书中评论说，神经网络通常比经典的线性分类器提供更好的性能，特别是在与预先训练的单词嵌入一起使用时。

The non-linearity of the network, as well as the ability to easily integrate pre-trained word embeddings, often lead to superior classification accuracy.  
网络的非线性，以及易于集成预先训练好的单词嵌入的能力，往往导致更好的分类精度。

— , 2015.  
-2015年。

He also comments that convolutional neural networks are effective at document classification, namely because they are able to pick out salient features (e.g. tokens or sequences of tokens) in a way that is invariant to their position within the input sequences.  
他还评论说，卷积神经网络在文档分类方面是有效的，也就是说，卷积神经网络能够以其在输入序列中的位置不变的方式来识别显著特征（例如标记或标记序列）。

Networks with convolutional and pooling layers are useful for classification tasks in which we expect to find strong local clues regarding class membership, but these clues can appear in different places in the input. […] We would like to learn that certain sequences of words are good indicators of the topic, and do not necessarily care where they appear in the document. Convolutional and pooling layers allow the model to learn to find such local indicators, regardless of their position.  
具有卷积层和池层的网络对于分类任务是有用的，在分类任务中，我们期望找到关于类成员的强局部线索，但是这些线索可以出现在输入中的不同位置。[……]我们希望了解，某些单词序列是本专题的良好指标，并不一定关心它们在文件中的位置。卷积层和汇集层允许模型学习查找此类本地指标，而不管它们的位置如何。

— , 2015.  
-2015年。

The architecture is therefore comprised of three key pieces:  
因此，该架构由三个关键部分组成：

1. Word Embedding: A distributed representation of words where different words that have a similar meaning (based on their usage) also have a similar representation.  
   单词嵌入：单词的一种分布式表示，其中具有相似含义的不同单词（基于其用法）也具有相似的表示。
2. Convolutional Model: A feature extraction model that learns to extract salient features from documents represented using a word embedding.  
   卷积模型：一种特征提取模型，学习从使用字嵌入表示的文档中提取显著特征。
3. Fully Connected Model: The interpretation of extracted features in terms of a predictive output.  
   全连接模型：根据预测输出解释提取的特征。

Yoav Goldberg highlights the CNNs role as a feature extractor model in his book:  
Yoav Goldberg在他的书中强调了CNNs作为特征提取模型的作用：

… the CNN is in essence a feature-extracting architecture. It does not constitute a standalone, useful network on its own, but rather is meant to be integrated into a larger network, and to be trained to work in tandem with it in order to produce an end result. The CNNs layer’s responsibility is to extract meaningful sub-structures that are useful for the overall prediction task at hand.  
…CNN本质上是一个特征提取架构。它本身并不构成一个独立的、有用的网络，而是要融入一个更大的网络，并接受与之协同工作的培训，以便产生最终结果。CNNs层的职责是提取有意义的子结构，这些子结构对于当前的总体预测任务是有用的。

— Page 152, , 2017.  
-第152页，2017年。

The tying together of these three elements is demonstrated in perhaps one of the most widely cited examples of the combination, described in the next section.  
这三个元素的结合在下一节描述的最广泛引用的组合示例中可能会得到证明。

## 2. Use a Single Layer CNN Architecture 2。使用单层CNN架构

You can get good results for document classification with a single layer CNN, perhaps with differently sized kernels across the filters to allow grouping of word representations at different scales.  
使用单层CNN可以获得很好的文档分类结果，也许在过滤器中使用不同大小的内核，以允许在不同的尺度上对单词表示进行分组。

Yoon Kim in his study of the use of pre-trained word vectors for classification tasks with Convolutional Neural Networks found that using pre-trained static word vectors does very well. He suggests that pre-trained word embeddings that were trained on very large text corpora, such as the freely available word2vec vectors trained on 100 billion tokens from Google news may offer good universal features for use in natural language processing.  
Yoon Kim在研究用卷积神经网络将预先训练好的词向量用于分类任务时发现，使用预先训练好的静态词向量效果非常好。他建议，在非常大的文本语料库上训练的预先训练的单词嵌入，例如在Google news的1000亿个标记上训练的免费word2vec向量，可以为自然语言处理提供良好的通用特性。

Despite little tuning of hyperparameters, a simple CNN with one layer of convolution performs remarkably well. Our results add to the well-established evidence that unsupervised pre-training of word vectors is an important ingredient in deep learning for NLP  
尽管超参数的调整很小，但是一个简单的CNN加上一层卷积就表现得非常好。我们的研究结果进一步证明了无监督预训练是NLP深度学习的重要组成部分

— , 2014.  
-2014年。

He also discovered that further task-specific tuning of the word vectors offer a small additional improvement in performance.  
他还发现，进一步对单词向量进行特定于任务的调整可以在性能上提供一个小的额外改进。

Kim describes the general approach of using CNN for natural language processing. Sentences are mapped to embedding vectors and are available as a matrix input to the model. Convolutions are performed across the input word-wise using differently sized kernels, such as 2 or 3 words at a time. The resulting feature maps are then processed using a max pooling layer to condense or summarize the extracted features.  
Kim描述了使用CNN进行自然语言处理的一般方法。句子被映射到嵌入向量，并可用作模型的矩阵输入。卷积是通过使用不同大小的内核（例如一次2个或3个单词）跨输入单词执行的。然后，使用最大池层对生成的特征映射进行处理，以压缩或汇总提取的特征。

The architecture is based on the approach used by Ronan Collobert, et al. in their paper ““, 2011. In it, they develop a single end-to-end neural network model with convolutional and pooling layers for use across a range of fundamental natural language processing problems.  
该体系结构基于Ronan Collobert等人在其论文“2011”中使用的方法。其中，他们开发了一个单端到端的神经网络模型，该模型具有卷积层和池层，可用于一系列基本的自然语言处理问题。

Kim provides a diagram that helps to see the sampling of the filters using differently sized kernels as different colors (red and yellow).  
Kim提供了一个图表，帮助您将不同大小的内核用作不同颜色（红色和黄色）来查看过滤器的采样。

An example of a CNN Filter and Polling Architecture for Natural Language Processing. Taken from “Convolutional Neural Networks for Sentence Classification”, 2014.  
自然语言处理的CNN过滤器和轮询架构的示例。摘自《句子分类的卷积神经网络》，2014年。

Usefully, he reports his chosen model configuration, discovered via grid search and used across a suite of 7 text classification tasks, summarized as follows:  
有用的是，他报告了他选择的模型配置，通过网格搜索发现，并在一组7个文本分类任务中使用，总结如下：

* Transfer function: rectified linear.  
  传递函数：线性校正。
* Kernel sizes: 2, 4, 5.  
  内核大小：2，4，5。
* Number of filters: 100  
  过滤器数量：100
* Dropout rate: 0.5  
  辍学率：0.5
* Weight regularization (L2): 3  
  权重正则化（L2）：3
* Batch Size: 50  
  批量：50
* Update Rule: Adadelta  
  更新规则：Adadelta

These configurations could be used to inspire a starting point for your own experiments.  
这些配置可以用来激发你自己实验的起点。

## 3. Dial in CNN Hyperparameters 三。拨入CNN超参数

Some hyperparameters matter more than others when tuning a convolutional neural network on your document classification problem.  
在文档分类问题上调整卷积神经网络时，一些超参数比其他参数更重要。

Ye Zhang and Byron Wallace performed a sensitivity analysis into the hyperparameters needed to configure a single layer convolutional neural network for document classification. The study is motivated by their claim that the models are sensitive to their configuration.  
叶章和拜伦华莱士对配置单层卷积神经网络进行文档分类所需的超参数进行了敏感性分析。这项研究的动机是他们声称这些模型对它们的结构很敏感。

Unfortunately, a downside to CNN-based models – even simple ones – is that they require practitioners to specify the exact model architecture to be used and to set the accompanying hyperparameters. To the uninitiated, making such decisions can seem like something of a black art because there are many free parameters in the model.  
不幸的是，基于CNN的模型（即使是简单的模型）的一个缺点是，它们要求实践者指定要使用的确切模型架构，并设置伴随的超参数。对于不了解情况的人来说，做出这样的决定似乎是一门黑色艺术，因为模型中有许多自由参数。

— , 2015.  
-2015年。

Their aim was to provide general configurations that can be used for configuring CNNs on new text classification tasks.  
他们的目的是提供可以用于在新的文本分类任务上配置cnn的通用配置。

They provide a nice depiction of the model architecture and the decision points for configuring the model, reproduced below.  
它们很好地描述了模型架构和配置模型的决策点，如下所示。

Convolutional Neural Network Architecture for Sentence Classification Taken from “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification“, 2015.  
句子分类的卷积神经网络结构摘自“句子分类的卷积神经网络的敏感性分析（和从业者指南）”，2015年。

The study makes a number of useful findings that could be used as a starting point for configuring shallow CNN models for text classification.  
这项研究提出了一些有用的发现，可以作为配置浅CNN模型进行文本分类的起点。

The general findings were as follows:  
一般结论如下：

* The choice of pre-trained word2vec and GloVe embeddings differ from problem to problem, and both performed better than using one-hot encoded word vectors.  
  预训练的word2vec和globe嵌入的选择因问题而异，并且两者的性能都优于使用一个热编码的词向量。
* The size of the kernel is important and should be tuned for each problem.  
  内核的大小很重要，应该针对每个问题进行调整。
* The number of feature maps is also important and should be tuned.  
  功能图的数量也很重要，应该进行调整。
* The 1-max pooling generally outperformed other types of pooling.  
  1-max池通常优于其他类型的池。
* Dropout has little effect on the model performance.  
  辍学对模型性能影响不大。

They go on to provide more specific heuristics, as follows:  
它们继续提供更具体的启发式方法，如下所示：

* Use word2vec or GloVe word embeddings as a starting point and tune them while fitting the model.  
  使用word2vec或GloVe word嵌入作为起点，在拟合模型时对它们进行优化。
* Grid search across different kernel sizes to find the optimal configuration for your problem, in the range 1-10.  
  在不同的内核大小上进行网格搜索，以找到问题的最佳配置，范围为1-10。
* Search the number of filters from 100-600 and explore a dropout of 0.0-0.5 as part of the same search.  
  搜索100-600个过滤器的数量，并在同一搜索中搜索0.0-0.5的辍学率。
* Explore using tanh, relu, and linear activation functions.  
  探索使用tanh，relu和线性激活函数。

The key caveat is that the findings are based on empirical results on binary text classification problems using single sentences as input.  
关键的警告是，这些发现是基于对使用单句作为输入的二元文本分类问题的经验结果。

I recommend reading the full paper to get more details:  
我建议阅读全文以了解更多细节：

* , 2015.  
  2015年。

## 4. Consider Character-Level CNNs 四。考虑字符级CNN

Text documents can be modeled at the character level using convolutional neural networks that are capable of learning the relevant hierarchical structure of words, sentences, paragraphs, and more.  
文本文档可以使用卷积神经网络在字符级建模，卷积神经网络能够学习单词、句子、段落等的相关层次结构。

Xiang Zhang, et al. use a character-based representation of text as input for a convolutional neural network. The promise of the approach is that all of the labor-intensive effort required to clean and prepare text could be overcome if a CNN can learn to abstract the salient details.  
张翔，等。使用基于字符的文本表示作为卷积神经网络的输入。这种方法的承诺是，如果CNN能够学习抽象出显著的细节，那么清理和准备文本所需的所有劳动密集型工作都可以克服。

… deep ConvNets do not require the knowledge of words, in addition to the conclusion from previous research that ConvNets do not require the knowledge about the syntactic or semantic structure of a language. This simplification of engineering could be crucial for a single system that can work for different languages, since characters always constitute a necessary construct regardless of whether segmentation into words is possible. Working on only characters also has the advantage that abnormal character combinations such as misspellings and emoticons may be naturally learnt.  
…深层ConvNets不需要词汇知识，此外，从先前的研究得出结论，ConvNets不需要关于语言的句法或语义结构的知识。这种工程的简化对于一个可以为不同语言工作的系统来说是至关重要的，因为不管分词是否可能，字符总是构成一个必要的结构。只处理字符还有一个优点，就是可以自然地学习不正常的字符组合，如拼写错误和表情符号。

— , 2015.  
-2015年。

The model reads in one-hot encoded characters in a fixed-sized alphabet. Encoded characters are read in blocks or sequences of 1,024 characters. A stack of 6 convolutional layers with pooling follows, with 3 fully connected layers at the output end of the network in order to make a prediction.  
该模型以固定大小的字母表读入一个热编码字符。编码字符以1024个字符的块或序列读取。接下来是一个由6个带池的卷积层组成的堆栈，在网络的输出端有3个完全连接的层，以便进行预测。

Character-based Convolutional Neural Network for Text Classification Taken from “Character-level Convolutional Networks for Text Classification“, 2015.  
文本分类的基于字符的卷积神经网络摘自“文本分类的字符级卷积网络”，2015年。

The model achieves some success, performing better on problems that offer a larger corpus of text.  
该模型取得了一定的成功，在提供更大文本语料库的问题上表现得更好。

… analysis shows that character-level ConvNet is an effective method. […] how well our model performs in comparisons depends on many factors, such as dataset size, whether the texts are curated and choice of alphabet.  
分析表明，字符级转换网是一种有效的方法。[…]我们的模型在比较中的表现如何取决于许多因素，例如数据集大小、文本是否被整理和字母表的选择。

— , 2015.  
-2015年。

Results using an extended version of this approach were pushed to the state-of-the-art in a follow-up paper covered in the next section.  
结果使用这种方法的扩展版本在下一节的后续文章中被推到最新水平。

## 5. Consider Deeper CNNs for Classification 5个。考虑更深层次的cnn进行分类

Better performance can be achieved with very deep convolutional neural networks, although standard and reusable architectures have not been adopted for classification tasks, yet.  
尽管标准和可重用的结构还没有被用于分类任务，但是使用非常深的卷积神经网络可以获得更好的性能。

Alexis Conneau, et al. comment on the relatively shallow networks used for natural language processing and the success of much deeper networks used for computer vision applications. For example, Kim (above) restricted the model to a single convolutional layer.  
Alexis Conneau等人评论了用于自然语言处理的较浅的网络，以及用于计算机视觉应用的更深入的网络的成功。例如，Kim（上）将模型限制在单个卷积层。

Other architectures used for natural language reviewed in the paper are limited to 5 and 6 layers. These are contrasted with successful architectures used in computer vision with 19 or even up to 152 layers.  
本文所讨论的用于自然语言的其他体系结构仅限于5层和6层。这与计算机视觉中使用的19层甚至152层的成功架构形成对比。

They suggest and demonstrate that there are benefits for hierarchical feature learning with very deep convolutional neural network model, called VDCNN.  
他们提出并证明了用深卷积神经网络模型（VDCNN）进行分层特征学习的好处。

… we propose to use deep architectures of many convolutional layers to approach this goal, using up to 29 layers. The design of our architecture is inspired by recent progress in computer vision […] The proposed deep convolutional network shows significantly better results than previous ConvNets approach.  
…我们建议使用许多卷积层的深层架构来实现这一目标，最多使用29层。我们架构的设计灵感来自计算机视觉的最新进展[……]所提出的深卷积网络显示出明显优于以前的ConvNets方法的效果。

Key to their approach is an embedding of individual characters, rather than a word embedding.  
他们方法的关键是嵌入单个字符，而不是嵌入单词。

We present a new architecture (VDCNN) for text processing which operates directly at the character level and uses only small convolutions and pooling operations.  
我们提出了一种新的文本处理架构（VDCNN），它直接在字符层操作，只使用小卷积和池操作。

— , 2016.  
-，2016年。

Results on a suite of 8 large text classification tasks show better performance than more shallow networks. Specifically, state-of-the-art results on all but two of the datasets tested, at the time of writing.  
在一组8个大文本分类任务上的结果显示，比更浅的网络表现出更好的性能。具体来说，在编写本文时，除了两个测试数据集之外，其他所有数据集的最新结果。

Generally, they make some key findings from exploring the deeper architectural approach:  
一般来说，他们通过探索更深层的架构方法得出了一些关键的发现：

* The very deep architecture worked well on small and large datasets.  
  这种非常深入的体系结构在小型和大型数据集上都很有效。
* Deeper networks decrease classification error.  
  较深的网络减少了分类误差。
* Max-pooling achieves better results than other, more sophisticated types of pooling.  
  最大池比其他更复杂类型的池获得更好的结果。
* Generally going deeper degrades accuracy; the shortcut connections used in the architecture are important.  
  通常深入会降低准确性；架构中使用的快捷连接非常重要。

… this is the first time that the “benefit of depths” was shown for convolutional neural networks in NLP.  
……这是第一次在NLP中显示卷积神经网络的“深度效益”。

— , 2016.  
-，2016年。

## Further Reading 进一步阅读

This section provides more resources on the topic if you are looking go deeper.  
如果您想深入了解，本节将提供更多有关此主题的资源。

* , 2015.  
  2015年。
* , 2014.  
  2014年。
* , 2011.  
  2011年。
* , 2016.  
  ，2016年。
* , 2015.  
  2015年。
* , 2015.  
  2015年。

Have you come across some good resources on deep learning for document classification? Let me know in the comments below.  
你有没有找到一些关于文档分类深度学习的好资源？请在下面的评论中告诉我。

## Summary

In this post, you discovered some best practices for developing deep learning models for document classification.

Specifically, you learned:

* That a key approach is to use word embeddings and convolutional neural networks for text classification.
* That a single layer model can do well on moderate-sized problems, and ideas on how to configure it.
* That deeper models that operate directly on text may be the future of natural language processing.

Do you have any questions? Ask your questions in the comments below and I will do my best to answer.