基于 Oracle-MNIST 数据集的图像分类算法实验

Yue-Cheng Cao[1]
[1] Beijing University of Posts and Telecommunications

Abstract

为更好地弘扬中华文化，提高甲骨文研究者的文字识别效率，本次实验针对甲骨文 Oracle-MINST 数据集，采用卷积神经网络 (CNN)、支持向量机 (SVC)、多层感知机 (MLP) 等方法进行分类和识别任务，在不同网络结构和参数下进行运行测试和结果分析。本次实验丰富了提供文献中所列算法的实践情况并进行结果展示，在 SVC 算法中对于错误样本分布情况进行输出，结合数据集的特点进行分析并总结展望，文末对签名转换的新方法进行初步尝试。

1. 引言

人类智慧的一个重要方面是其认识外界事物的能力，这种能力在不断的学习中得到增强，人日常中的很多活动都离不开对外界事物的分类和识别过程。人能够区分花鸟鱼虫，这实际上是一个分类的过程，大脑对于之前记忆进行保存，即使遇到新的品种，也可以进行分类和判断。所谓模式识别 (pattern recognition)，就是一种规律，能够根据事物所具备的特征或形状进行分类和判别，寻找内部的规律性的过程。在图像分类领域，模式识别与计算机视觉相结合，目标在于针对已有的数据集训练一个模型，该模型能够对新输入的样本进行分类并得到结果输出，并优化模型以提高识别的准确率。

甲骨文距今已有 3000 多年的历史，其流传至今，成为中华民族宝贵的文化，甲骨文的研究对于文字起源和中国古代文化的研究具有重要意义。由于刻划复杂、噪声严重以及甲片信息缺失等问题，学者的识别难度较大。随着深度学习技术的不断发展，可以采用神经网络等人工智能方法来进行甲骨文的分类识别。

针对已构建的 Oracle-MINST 数据集，其中包括 30222 张甲骨文 28×28 的灰度图片，共分为大、日、月、牛、翌、田、勿、矢、巳、木十类。在所有甲骨文图片中，随机选取 27222 张作为训练集，每类选择 300 张图片作为测试集。

现有的模式识别与计算机视觉领域中，识别 MINST、CIFAR-10 等经典数据集的技术已相对成熟，然而由于 Oracle-MINST 数据集中的甲骨文图像噪声干扰较多同时相同类别的不同图像之间的差异较大，给人和机器的识别带来较大的难度。

本文选用 Oracle-MINST 数据集，主要采用了卷积神经网络 (CNN)、支持向量机 (SVC)、多层感知机 (MLP) 等成熟的图像分类算法，在不同的参数或网络条件下进行效果测试，进行结果分析，挖掘数据集的特征并进行探讨。本文主要工作如下：

- 采用卷积神经网络 (CNN) 方法对数据集进行训练，并计算训练损失；参考已有的网络结构，搭建新的网络并进行效果测试，将结果进行整理。
- 采用支持向量机 (SVC) 算法对原始图像进行训练，在测试集中进行效果测试，并且修改参数值和函数类型，进行效果测试，将结果整理为折线图；同时输出测试集中错误的 id，对图像分类中的错误信息进行分析，找到不同类的文字图像错误情况，将结果整理为表格进行展示。
- 采用多层感知机 (MLP) 对图像进行分类测试，修改参数，将结果整理在表格中。
- 对文献 https://arxiv.org/abs/2204.07953 中提到的签名转换的方法进行尝试，对其代码结构进行分析讲解，提出疑问，以视频录制的形式进行展示。
- 针对研究过程中对于模式识别、甲骨文识别、以及数据集、算法进行个人观点总结，整理报告。

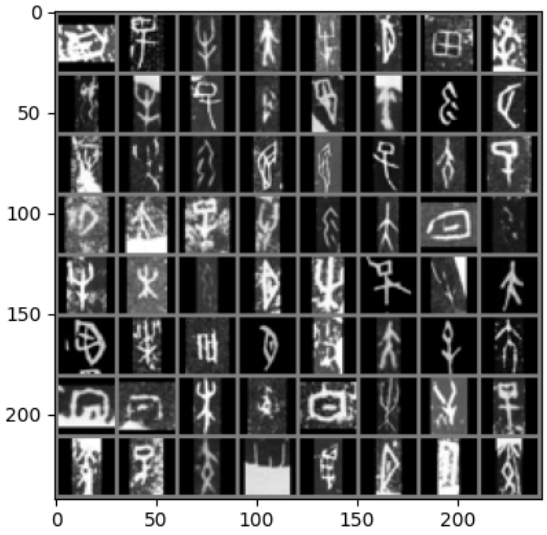

图 1. 甲骨文数据集图像输出

2. 相关工作

此部分主要针对甲骨文数据集 Oracle-MINST 的分类方法、常见的手写数据集相关信息、在其他经典手写数据集上的方法调研三部分进行论述。

2.1. Oracle-MINST 识别研究

目前，针对甲骨文 Oracle-MINST 数据集的研究还比较有限，采用构建完整的数据集，在现有的深度神经网络中进行手写甲骨文上进行训练，可以达到 94.9% 的识别率，但将其直接应用于甲骨文拓片识别，准确率仅有 2.1%，在《基于无监督结构-纹理分离网络的甲骨文字符识别》(Unsupervised Structure-Texture Separation Network for Oracle Character Recognition) 一文首次在甲骨文字符识别中应用无监督域自适应技术，引入迁移学习思想，为机器学习开拓了新的应用领域，将甲骨文拓片的识别率从 2.1% 提升至 47.1%. 该算法还可以实现手写甲骨文和拓片甲骨文之间的转换，实现图像噪声的模拟，对于甲骨文的修复工作具有重要意义。

2.2. 经典数据集介绍

为完成甲骨文数据集上的识别任务，本文主要采用几种经典图像分类数据集的分类方法进行迁移。调研过程以 MINST 数据集为例进行介绍。MNIST (Mixed National Institute of Standards and Tech nology Database) 中包含 70 000 张已经过预处理的手写数字灰度图片，每张图片都由 28*28 个像素点组成，包括数字 0-9，数据集分为训练集和测试集，成为图像识别领域的入门数据集之一。

2.3. 常用图像分类算法

现代很多领域都需要应用图像识别技术，提高企业的工作效率。基于神经网络的图像识别系统是一种新型模式识别系统，首先对数据进行预处理并提取特征新型，最后采用 BP 神经网络进行分类；基于小波矩的图像识别方法对于只经过平面变换的样本具有较好的分辨率，但对于噪声较为敏感。

深度学习的不断发展使得深层神经网络训练方法逐渐普及，包括线性感知器算法、卷积神经网络算法 (CNN)、循环神经网络算法 (RNN)、长短时记忆网络算法 (LSTM) 等。

文献 [3] 中提到一种基于 CNN 和粒子群优化 SVM 的手写数字识别研究方法，针对传统卷积神经网络手写体数字识别中 Softmax 因指数函数运算而易产生计算溢出以及较高的计算机硬件需求问题, 提出了基于卷积神经网络特征提取的支持向量机手写体数字识别方法。同时, 为了提高手写体数字的识别精度, 设计了基于 K-CV 意义下适应度函数的粒子群优化 SVM 参数方法。基于 Semeion 及 MNIST 手写体数字集的实验仿真表明, 文章所设计的方法与传统方法相比能够获得更高的识别率。

3. 算法简介

3.1. 卷积神经网络 (CNN)

卷积神经网络（convolution 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，对于大型图像处理有出色表现。卷积神经网络由一个或多个卷积层（convolution layer）和末端的全连接层组成，同时也包括关联权重和池化层（pooling layer）。这一结构使得卷积神经网络能够利用输入数据的二维结构，并且也可以使用反向传播算法进行训练。

卷积层的主要作用是通过对卷积核大小的设置，由浅入深不断对前一层传输的数据进行特征提取。由于设置了共享权值，在同一特征图中神经元使用同一组卷积核，可以减少训练参数。其中，卷积核数值在初始化后由后续的网络训练确定，图片各像素点的值与卷积核的乘积加入偏置后经过激活函数的运算即可得到图片的一个特征映射：

$$a_j^L = f(\sum_i a_j^{L-1} w_{ij}^L + b_j^L)$$

其中 a_j^L 表示 L 层卷积后第 j 个神经元的输出；x_{ij}^L 表示卷积核；b_j^L 表示偏置。f 为神经元激活函数。

池化是卷积神经网络中的一个重要操作，能够减少图片中冗余特征，同时保持特征的局部不变性。图片经过卷积处理后，每个 n×n 邻域内的像素点采用最大池化（MaxPooling）的方法变为一个像素

$$a_j^{L+1} = down(a_j^L)$$

其中 $down$ 为下采样函数，该层运算不包含可学习的权值和阈值。全连接层可以整合前向传来的具有类别区分性的局部信息；同时，可以增强网络的非线性映射能力；限制网络规模的大小。

3.2. 支持向量机 (SVC)

支持向量机是建立在统计学理论基础上的一种数据挖掘算法，其工作机理是寻找一个满足分类要求的最优分类超平面，使得该超平面在保证分类精度的同时，能够使超平面两侧的空白区域最大化。理论上，支持向量机能够实现对线性可分数据的最优分类。其原理示意图如2。

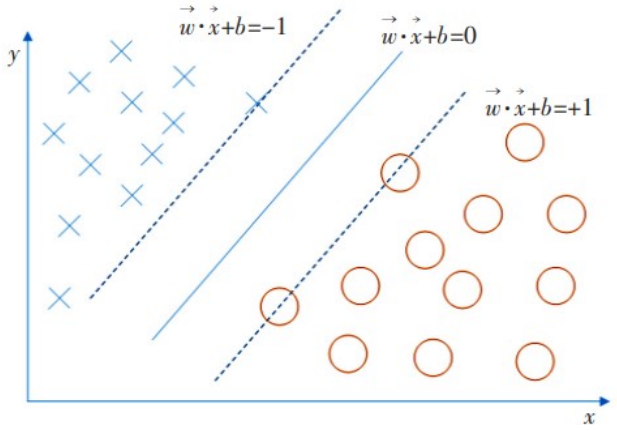

图 2. 支持向量机示意图

支持向量机中的核心参数包括核函数、和惩罚参数 C，参数 C 用于权衡” 训练样本的正确分类” 与” 决策函数的边际最大化” 两个不可同时完成的目标，希望找出一个平衡点来让模型的效果最佳。C 值较小时噪声干扰较大，C 值较大容易出现过拟合现象。

3.3. 多层感知机 (MLP)

多层感知机（MLP，Multilayer Perceptron）也叫人工神经网络（ANN，Artificial Neural Network），除了输入输出层，它中间可以有多个隐层，最简单的 MLP 只含一个隐层，即三层的结构.

隐藏层的神经元与输入层是全连接的，假设输入层用向量 x 表示，则隐藏层的输出就是 $f(W1X + b1)$，$W1$ 是权重（也叫连接系数），$b1$ 是偏置，函数 f 可以是常用的 sigmoid 函数或者 tanh 函数：

$$\sigma(x) = \frac{1}{1+e^{-x}}, tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

MLP 的基本模型图如图3所示。

图 3. 多层感知机网络结构图

4. 实验结论

此部分包括各种算法的实验设置、实验方法、参数调整过程以及运行实验后得到的结果，并根据结果对算法原理和数据集进行简单的分析和总结。

4.1. CNN

针对卷积神经网络分类方法，在 pytorch 环境下对代码进行测试，并根据已有的三个网络结构搭建新的神经网络，进行结果测试。

搭建网络遵循的基本公式为：

$$N = (W - F + 2P)/S + 1$$

其中 N 表示输出的图像 size，W 表示输入的图像 size，F 表示卷积核大小，P 表示填充值 (padding),S 表示步长。

Parameter	Accuracy
2×Conv-Pool-ReLu, 2×FC, Dropout	94.1
2×Conv-Pool-ReLu, 2×FC	93.3
1×Conv-Pool-ReLu, 2×FC	91.8
3×Conv-Pool-ReLu, 2×FC	91.1
1×Conv-Pool-ReLu, 2×FC, Dropout	91.1
3×Conv-Pool-ReLu, 2×FC, Dropout	91.1

表 1. 卷积神经网络结构与识别准确率

通过上表可以发现，对于卷积神经网络而言，适当提高网络层数可以提高算法的准确率，但提高层数的同时也需要考虑图像的维数，对于 28*28 的甲骨文图像，采用三层网络时最后一层的 MaxPool2d 参数不能设置为 2，最终的测试效果不如二层网络。Dropout 含义为适当丢弃一部分数据，让网络变“瘦”，有助于解决在训练过程中的过拟合问题，在网络层次搭建效果最佳的情况下可以适当提升性能。在不同的网络结构测试中，采用二层的神经网络并引入 Dropout 机制效果最佳。

4.2. SVC

对于支撑向量机分类算法，本文在实现基本代码后，在核函数分别为 rbf 和 poly 的情况下，对惩罚参数 C 进行修改，分别设置为 10,20,30...100，并进行测试。同时在每次的图片测试过程中，得到标签错误转移矩阵，对各类别的准确率进行输出整理和计算，得到两种核函数在不同的惩罚参数下的结果如图5所示。

Algorithm	Parameter	Test Accuracy
SVC	C=10, kernel=rbf	76.9
	C=20, kernel=rbf	77.0
	C=30, kernel=rbf	77.3
	C=40, kernel=rbf	76.6
	C=50, kernel=rbf	76.6
	C=60, kernel=rbf	76.3
	C=70, kernel=rbf	76.3
	C=80, kernel=rbf	76.3
	C=90, kernel=rbf	76.3
	C=100, kernel=rbf	76.1
	C=10, kernel=poly	75.0
	C=20, kernel=poly	75.6
	C=30, kernel=poly	75.1
	C=40, kernel=poly	74.7
	C=50, kernel=poly	74.5
	C=60, kernel=poly	74.3
	C=70, kernel=poly	74.4
	C=80, kernel=poly	74.3
	C=90, kernel=poly	74.4
	C=100, kernel=poly	74.3

图 4. SVC 算法测试结果

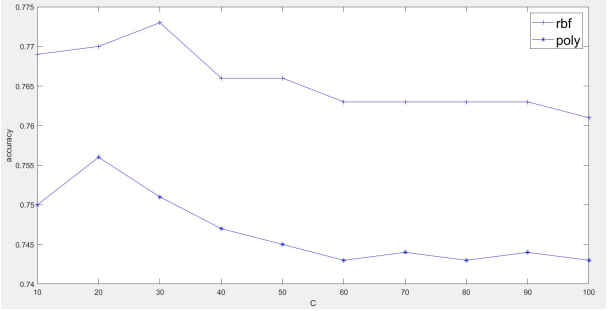

图 5. SVC 算法中不同核函数的性能比较

通过图5可以发现，在核函数选择上，rbf 比 ploy 效果更佳；同时，在 C=20 左右时性能最佳。C 越大代表这个分类器对在边界内的噪声点的容忍度越小，分类准确率高，但是容易过拟合，泛化能力差。所以一般情况下，应该适当减小 C，对在边界范围内的噪声有一定容忍。

同时，通过该算法的输出，可以分别得到 10 种标签测试集的准确率。本文采用 matlab 对不同标签的错误率进行求均值运算，得到 rbf 和 poly 核函数条件下 10 种标签的准确率分布情况如图6所示。

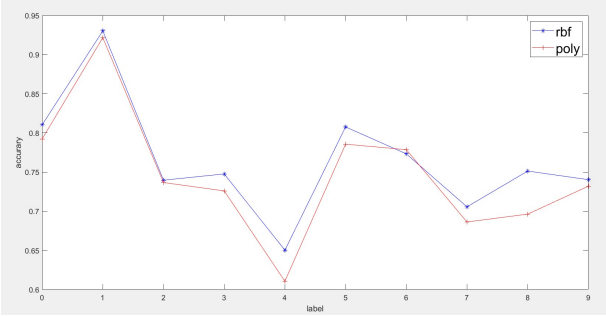

图 6. SVC 算法中针对两种核函数对于不同 label 的准确率

通过图6可以发现具体类别的甲骨文图像，rbf 的效果均优于 ploy，且类别间的正确率存在明显差异。对于正确率稍高的 rbf 核函数，通过 matlab 编程计算 C=10,20,...,100 的情况下的错误结果矩阵的均值，10 次结果的平均值存放在表2中，该表用矩阵表示为 A，其中的元素 $a_{i,j}$ 表示标签为 i 的图像被错误识别成标签 j 个数的平均值。

通过该表数据可以发现，部分的文字识别错误的分布情况存在一定的规律性，如标签为 1 的甲骨文出现错误的情况基本都对应标签为 5 的甲骨文，通过图6和表2可以了解到不同类型的甲骨文的错误率及错误分布情况，可以利用不同标签间结果的差异性侧面反映甲骨文的书写特点和相似性，针对性去完成识别任务，为更多方法的提出提供了思考的角度。

4.3. MLP

针对该算法，对于激活函数分别设定为 relu 和 tanh，同时对隐藏层的大小进行设置，运行代码进行效果测试，结果如表 3所示。

Parameter	Accuracy
ReLu, size=[256,128]	77.7
tanh, size=[256,128]	75.5
ReLu, size=[128,64]	75.4
ReLu, size=[100]	75.3
tanh, size=[128,64]	74.9
ReLu, size=[100,10]	71.6
tanh, size=[10,10]	58.7

表 3. 多层感知机识别准确率

表格中的参数代表多层感知机算法中采用的激活函数类型 (ReLu/tanh)，以及隐藏层的参数情况。通过该方法的测试，可以发现对于不同的参数测试结果，存在较为明显的差异; 根据有限结果显示，选择 ReLu 作为激活函数的效果更好一些；适当提高隐藏层的节点数，对于结果的准确率也有所帮助。

4.4. 结果分析

经过三种算法的测试，可以发现深度神经网络在图像数据处理上具有的强大优势，相比 SVC 和 MLP，对于更复杂的图像数据，CNN 的效果更佳。

CNN 的特征检测层通过训练数据进行学习，避免了显示的特征抽取，而隐式地从训练数据中进行学习；

label	0	1	2	3	4	5	6	7	8	9
0	0	0	4.6	7.4	2.5	0.2	5.6	19.3	11.6	5.7
1	0	0	0	0	0	19.9	0	0	1	0
2	6.3	2	0	6.2	16.6	3	22.3	7.5	6.1	8.1
3	3.1	2.1	4.8	0	10.8	6.9	8.7	5.1	19.6	14.6
4	4.7	0	24.6	20.1	0	2.3	10.5	9.8	28.8	4.1
5	0	29.7	7.2	6.7	4.1	0	1.2	7.2	0	1.6
6	9.1	1	13.9	6.8	17.2	3.7	0	8.3	8	0
7	23.8	0	20.3	4.8	15.8	1.9	6.5	0	9.7	5.8
8	3.8	1.7	4.6	12.1	12.6	10.1	10.9	8.1	0	10.7
9	4.5	2	10.8	24.3	7	3	2.9	8.5	14.9	0

表 2. average error number of each label

CNN 同一特征映射面上的神经元权值相同，所以网络可以并行学习，这也是卷积网络相对于神经元彼此相连网络的一大优势。卷积神经网络以其局部权值共享的特殊结构在语音识别和图像处理方面有着独特的优越性，其布局更接近于实际的生物神经网络，权值共享降低了网络的复杂性，特别是多维输入向量的图像可以直接输入网络这一特点避免了特征提取和分类过程中数据重建的复杂度。

SVM 算法的主要缺点为：观测样本较多时，效率不高；对非线性问题没有通用的解决方案，对于核函数的高维映射能力不强，对缺失数据较为敏感，在面对噪声影响较大的甲骨文数据集时性能有所影响。

MLP 网络有很多隐层组成，每个神经元都和上一层中的所有节点连接，而卷积核大小与每层输入大小相同会直接丢失非常多的输入空间信息，所以 MLP 这种运行模式不适合图像这种空间信息丰富的数据。

5. 总结延伸

在图像分类算法中，本文选取了几个较为经典的算法完成分类任务，此处对实验具体结果不再赘述。根据支持向量机算法中错误分布情况的输出，**提出一个新的思路**：对于构造的 10 分类数据集而言，想要提高其识别率，对于易错的文字进行研究和方法测试，让机器去模拟人的思路，先对当前的待分类图像给定一个可能性范围，并对于可能性较大的几个结果进行再次分类判决，可能对提高容易混淆的图像识别正确率有所帮助。

同时，在采用智能算法进行分类和测试过程中，对于其实际判别过程并没有进行深入的**可解释性分析**，对于甲骨文识别问题，算法是怎么实现结果判别的，对于该结果而言，其可靠度和置信度、结果判决的依据都成为值得深入研究的问题。若模型不可解释，当专家判断结果与机器检测结果不一致时，该如何决策以及维护人机间信任的问题都尚待解决。

Oracle-MINST 数据集在深度神经网络算法上实现了约 95% 的准确率，然而甲骨文数据集上的测试效果并不能完全代表其实际价值，真实的图像中存在的噪声干扰、图像残缺、同类文字的书写方式差异大等问题解决起来并不十分容易，在实际研究中不应仅仅停留在数据集的准确率提高上，如何在实际应用领域发挥其实际价值，切实做到减轻学者研究辅导的效果，是比较关键的问题。社会需求推动科研领域发展，构建的甲骨文数据集识别之路依旧任重道远。

课程收获. 通过本学期《模式识别与应用》课程的学习，对于模式识别的概念及常见的分类方法和思想有了较为清晰的认识，除理论知识外，课上还涉猎了很多知识应用场景和科研领域讲座，如脑机接口、人脸识别、甲骨文识别等，收获颇丰；在最后的大作业中通过对算法的实践和调参过程加深了对理论知识的理解。

新方法. 在本次实验中尝试了一种**签名转换方法**，文献链接为 https://arxiv.org/abs/2204.07953，其核心思想在于提取图像上的特征，最后将测试集的图像特征与之前提取的平均特征进行比较，选取差异最小的作为比对结果，在经典数据集 (MINST 等) 上实现了 100% 的准确率。在本文进行的实验中，在甲骨文数据集的测试结果为 100%，但考虑到部分经典数据集的标签准确性和可能在代码中出现的数据泄漏等问题，本文中没有展开讨论。该方法质疑声不断，其用于图像分类识别领域的可信度有待相关专家学者进行深入的研究，此处仅对该方法进行测试，对于该方法的代码讲解和质疑点分析已录制视频上传至哔哩哔哩。关于签名转换学习方法在经典数据集上 100% 准确率的尝试和研讨-哔哩哔哩 https://b23.tv/NYlESIW.

实验难点. 在本文完成过程中，较难的部分在于深度神经网络的搭建过程，如何调整网络参数使得网络的输入输出通道参数相匹配，通过 nn.Conv2d 中的数学计算理解了前三种网络结构的搭建过程，并按照计算规则搭建新的网络进行测试。在此过程中对于深度神经网络的结构和搭建过程有了更加深入的了解。

实验代码. 本次实验中用到的核心代码打包上传至百度网盘，共包括三个文件夹，其一为 SVC、签名转换方法，链接为：https://pan.baidu.com/s/19Q-i1P_xzfU-eOdQ7dsspw 提取码：i7sv. 其二为卷积神经网络算法，链接为：https://pan.baidu.com/s/10-vNKKvgEhVbn5jHNcRenA 提取码：jgiz，其三为多层感知机算法，链接为：https://pan.baidu.com/s/12KLBQzfZ78mepmIMcud1bg 提取码：uqez.

6. 参考文献

[1] 王玫. 麻省理工科技评论 APP.Stories Behind Science | 一眼千年——当 Al 遇见古文明的难解语言

[2]Mei Wang,Weihong Deng,Cheng-Lin Liu.Unsupervised Structure-Texture Separation Network for Oracle Character Recognition[J].arXiv:2205.06549 [cs.CV]

[3] 杨刚, 贺冬葛, 戴丽珍. 基于 CNN 和粒子群优化 SVM 的手写数字识别研究 [J]. 华东交通大学学报,2020,37(04):41-47.DOI:10.16749/j.cnki.jecjtu.2020.04.007.

[4]Oracle-MNIST: a Realistic Image Dataset for Benchmarking Machine Learning Algorithms. Mei Wang, Weihong Deng. arXiv:2205.09442

[5] 多层感知机实例代码 https://github.com/zhwwww/MLPImageClassification

[6] 张华美, 张皎洁. 基于人工智能的脱机手写数字识别研究综述 [J]. 南京邮电大学学报 (自然科学版),2021,41(05):83-91.DOI:10.14132/j.cnki.1673-5439.2021.05.012.

[7]MLP 多层感知机. 原文链接：http://deeplearning.net/tutorial/mlp.html-mlp

[8]SVC 中的参数说明与常用函数本文链接：https://blog.csdn.net/transformed/article/details/90437821

[9]J.de Curtò, I. de Zarzà, Hong Yan, Carlos T. Calafate.Learning with Signatures.[J]arXiv:2204.07953 [cs.CV]

[10] 简述 cnn 优点及 cnn 各个层作用，原文链接：https://blog.csdn.net/qq_42495866/article/details/88254911