基于机器学习的

验证码识别

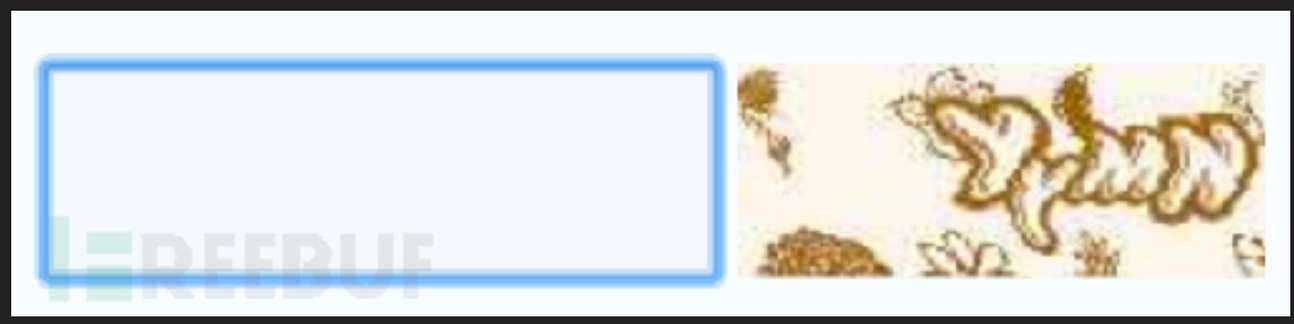
COMPLETELY AUTOMATED PUBLIC TURING TEST TO TELL COMPUTERS AND HUMANS APART

全自动区分计算机和人类的公开图灵测试

CAPTCHA

首先，先介绍下验证码程序的提出者，路易斯·冯·安(Luis von Ahn)。2002年，路易斯和他的小伙伴在卡内基梅隆第一次提出了CAPTCHA(验证码)这样一个程序概念。该程序是指，向请求的发起方提出问题，能正确回答的即是人类，反之则为机器。这个程序基于这样一个重要假设：提出的问题要容易被人类解答，并且让机器无法解答。

在当时的条件下，识别扭曲的图形，对于机器来说还是一个很艰难的任务，而对于人来说，则相对可以接受。yahoo在当时第一个应用了图形化验证码这个产品，很快解决了yahoo邮箱上的垃圾邮件问题，因此图形类验证码开始了大发展时期。




验证
>>
请按住滑块，拖动到
边
下一步


Select all images with a cat.


验证码：
请点击下图中所有的 洋葱
刷新

如何生成复杂验证码

- 随机字体
- 字母+数字
- 噪点
- 交互
- 复杂图片（眼力）
- 智力问题（脑力）：比如选出4条腿的动物

今天的主角

- 大写字母+数字随机组合
- 数字没有1
- 不规则的曲线干扰

- 图案有重叠（颜色）
- 图案有一定的随机位移

准备工作：采集样品

- 通过程序生成
- 通过用户收集
- 人肉采集

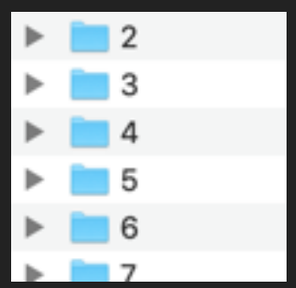
第一步：生成训练集

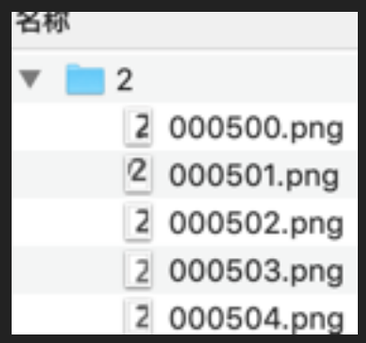
- 原图

- 降噪

- 分割

- 灰度化
- 生成训练集

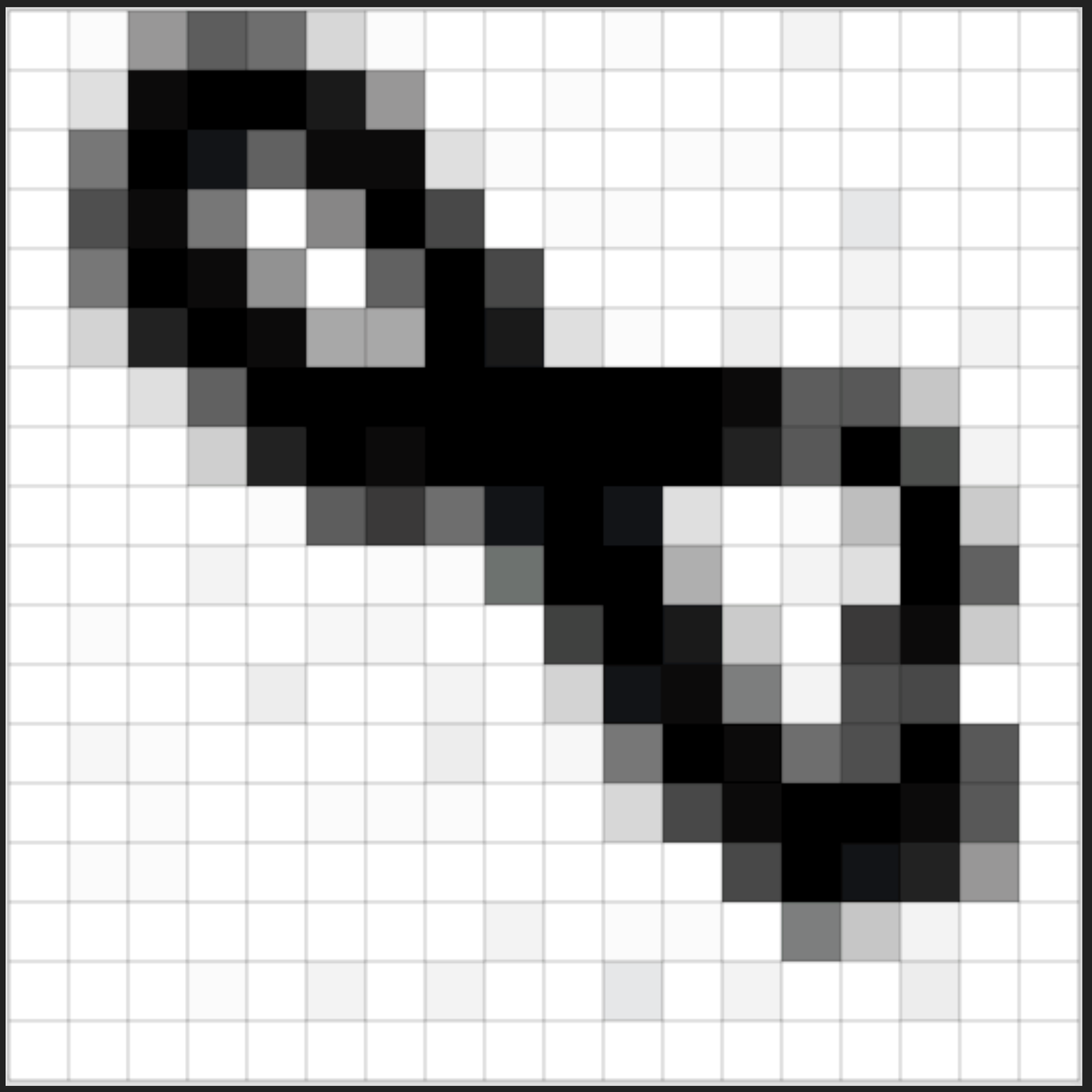




常用降噪方法

- 平滑
- 模糊
- 图像运算，平均值降噪
- 线性滤波
- 高斯模糊
- 。。。

对于解决我们的问题，不是很理想



降噪

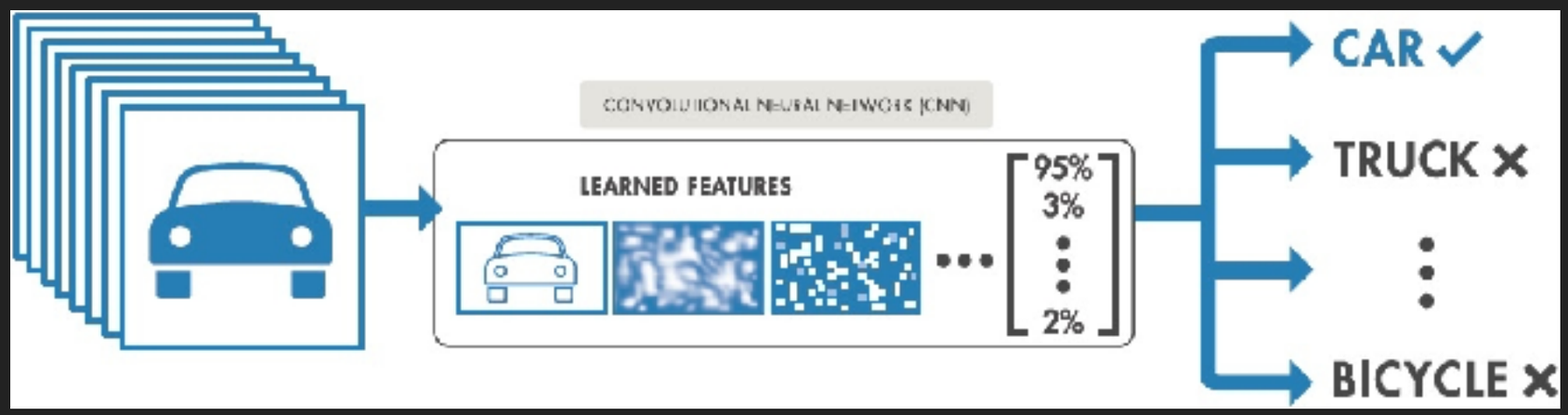
- 1. 底色变白
- 2.去杂线
- 3.去杂块（横向）
- 4.去杂块（纵向）
- 5.底色变白
- 6. 去杂线
- 7. 回填颜色?

分离

- 等分
- 通过颜色摘取

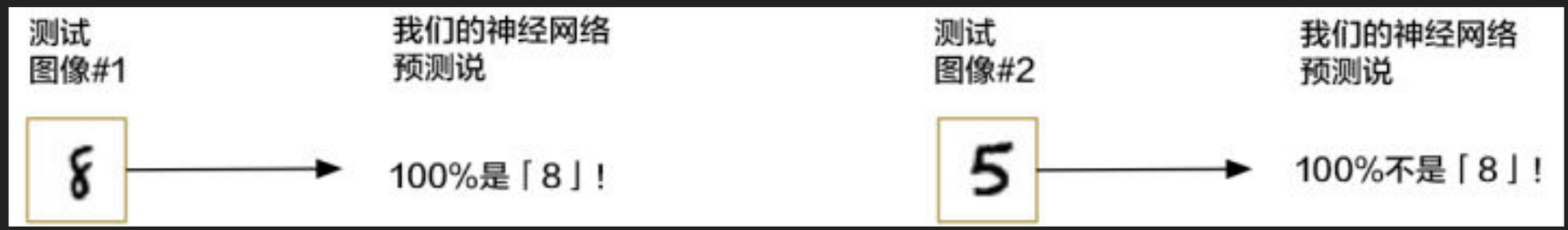
使用CNN进行训练

- 消除了手动提取特征的需要。-直接学习
- 可以产生先进的识别结果
- 可以重新训练完成新的识别任务

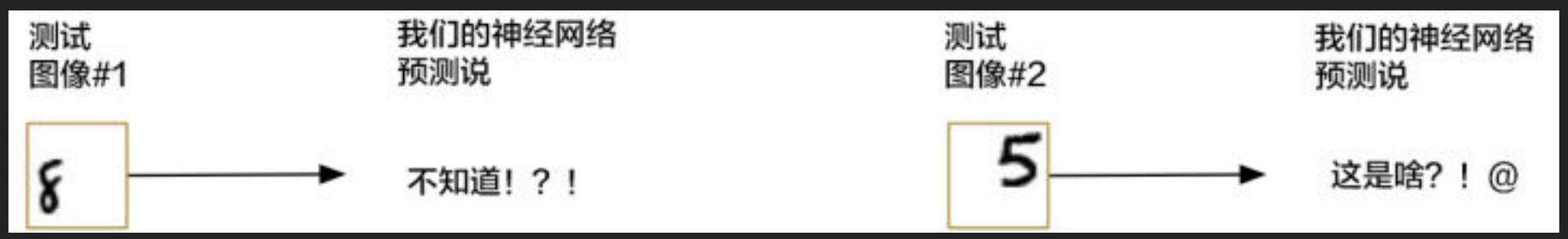


使用CNN进行训练

当文本在中央的时候，很好识别

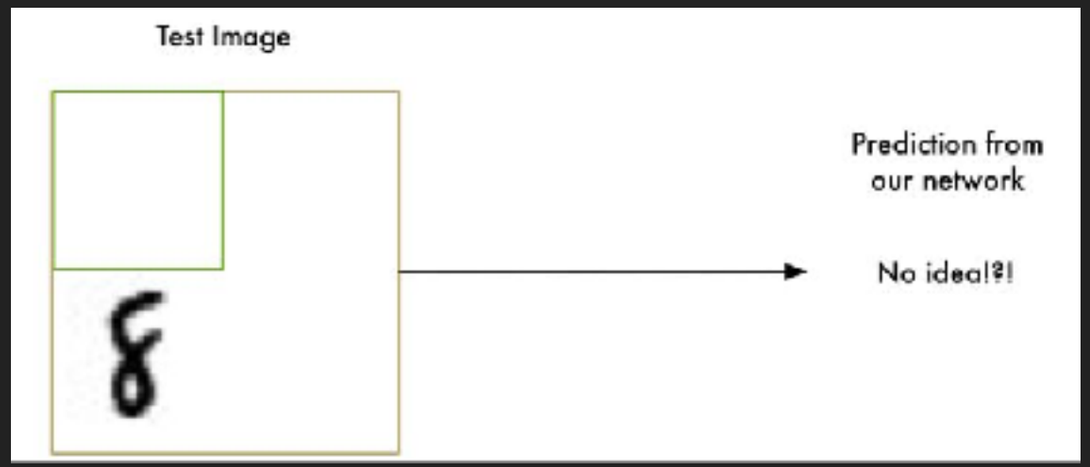


当文本发生位移后，识别不出来

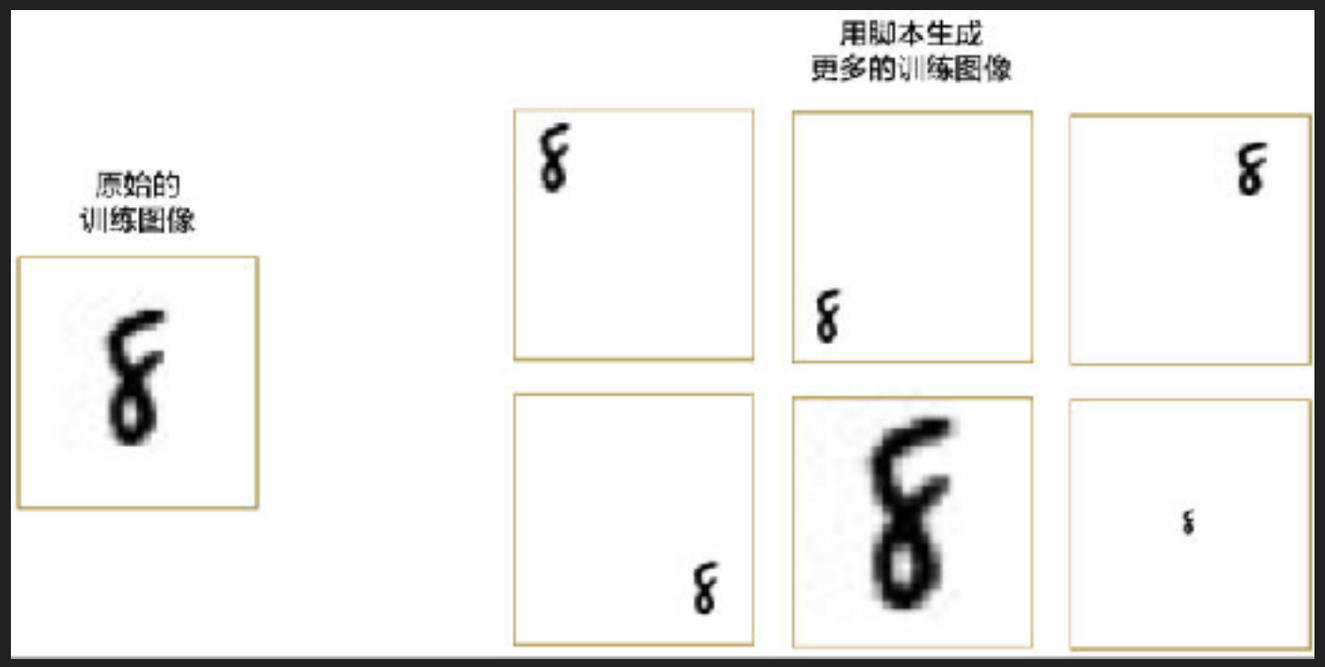


使用CNN进行训练

滑框搜索

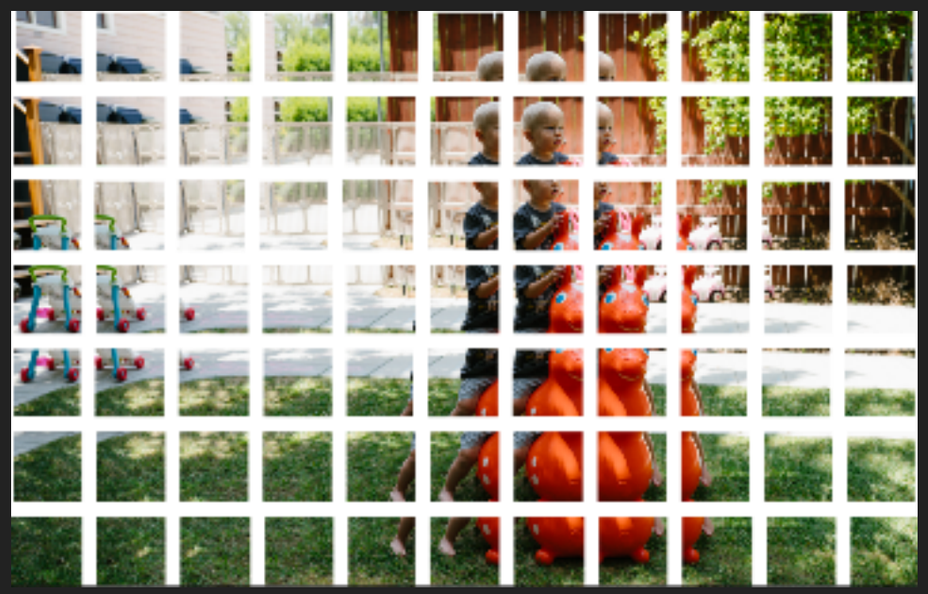


暴力方法

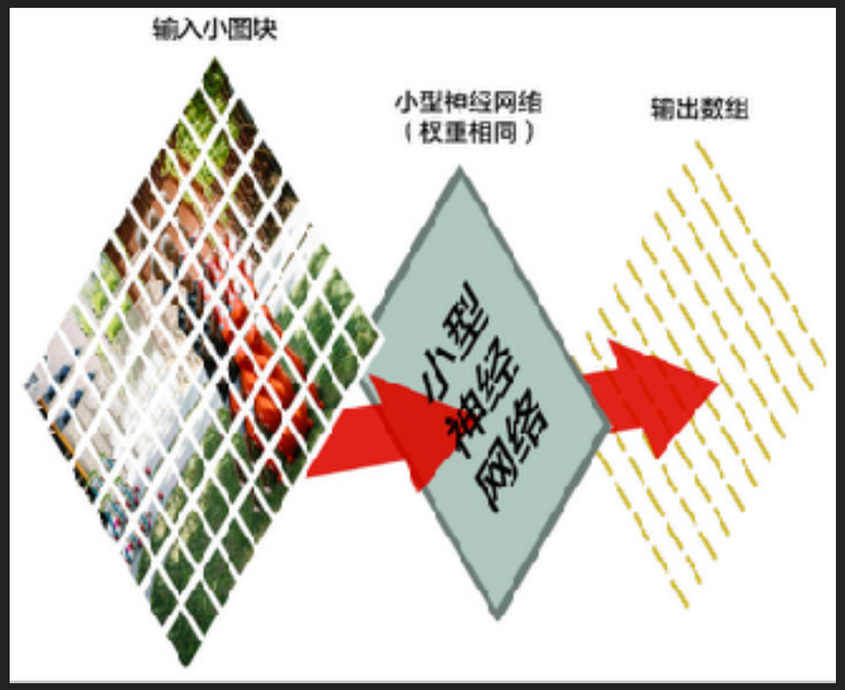


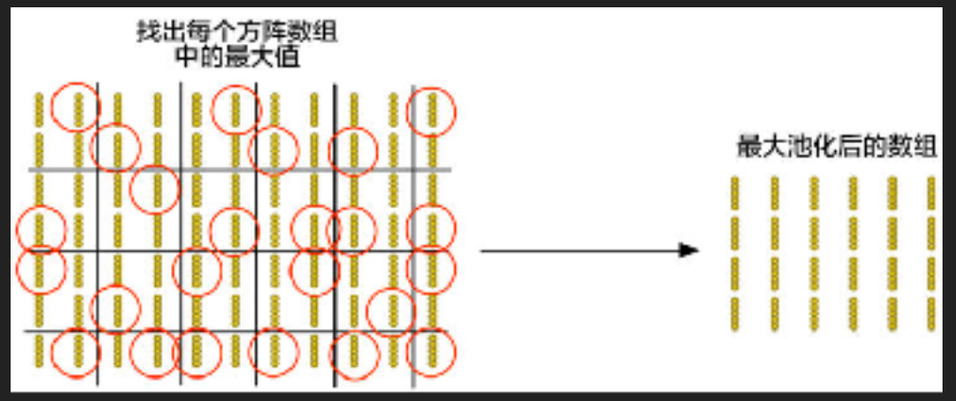
使用CNN进行训练

卷积



使用CNN进行训练



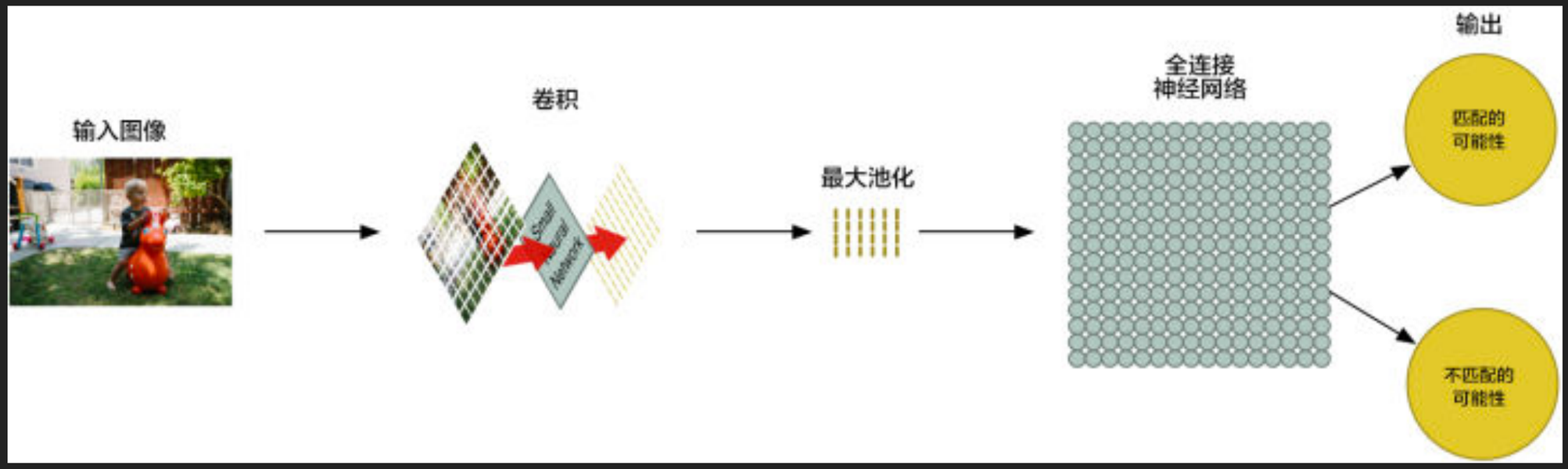


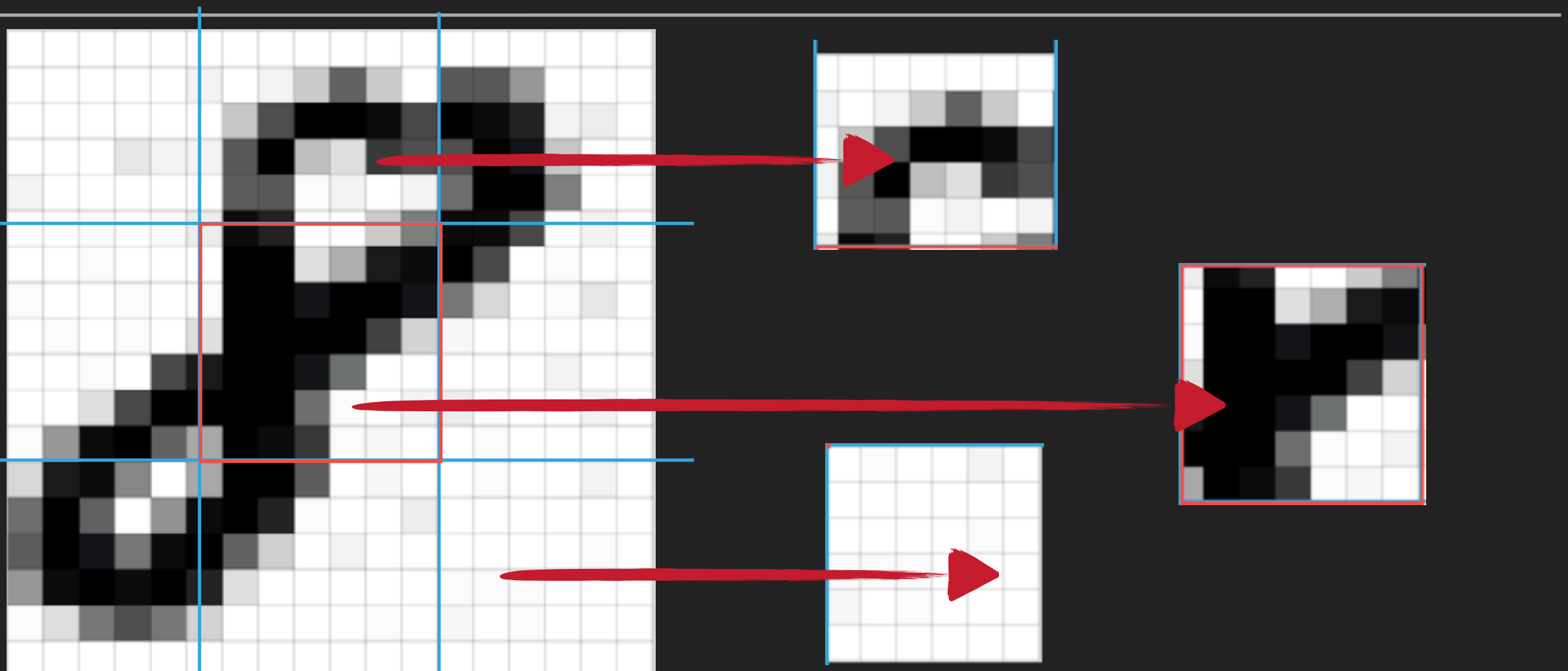
通过多层卷积之后，一步一步识别复杂的图案：

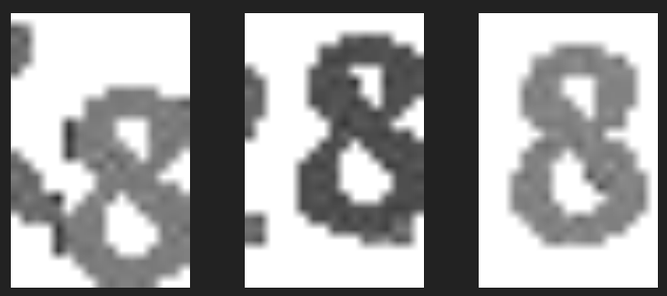
假如要识别一只鸟：第一层识别尖锐的东西，第二个卷积在尖锐图案中寻找鸟类的喙，

最后识别整只鸟

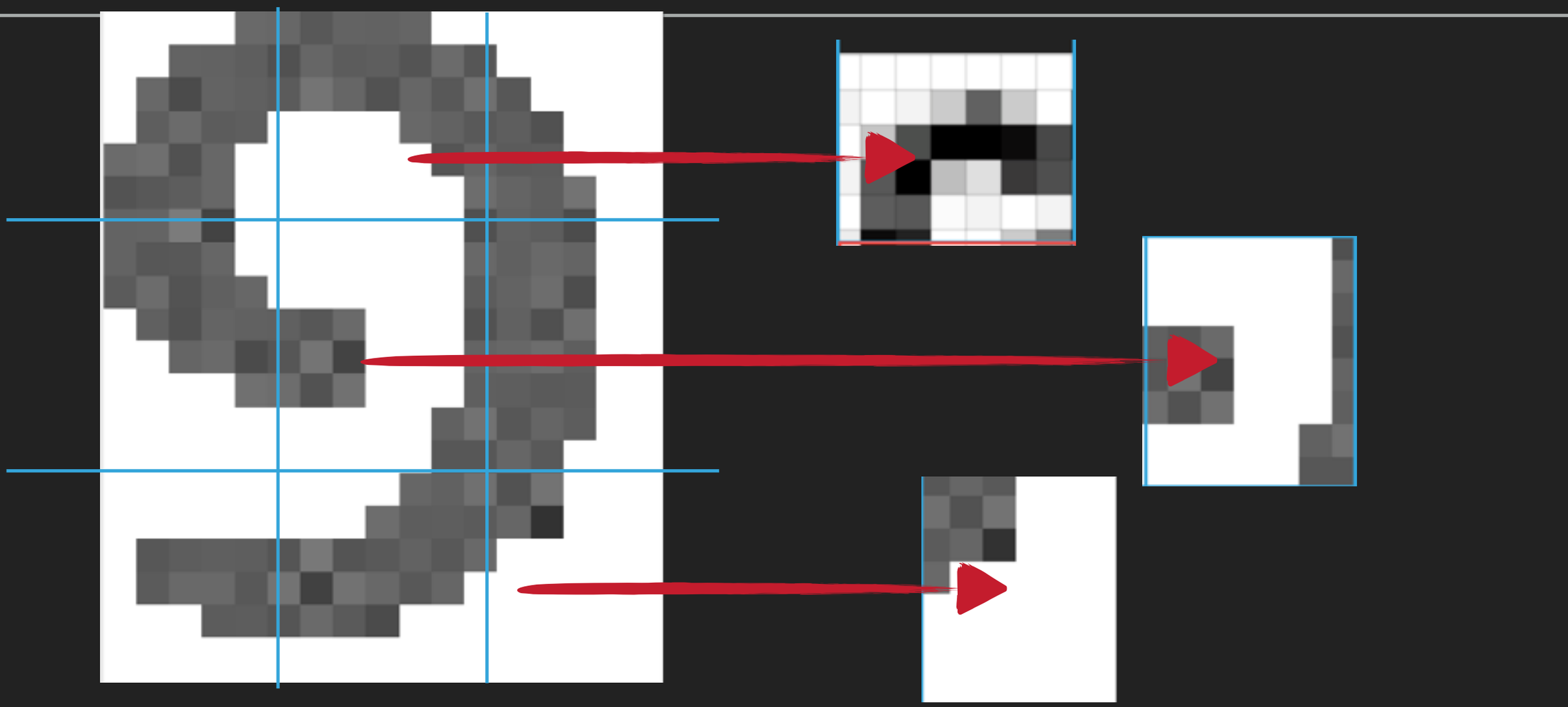
使用CNN进行训练


输入图像
卷积
Small Neural Network
最大池化
全连接
神经网络
输出
匹配的
可能性
不匹配的
可能性

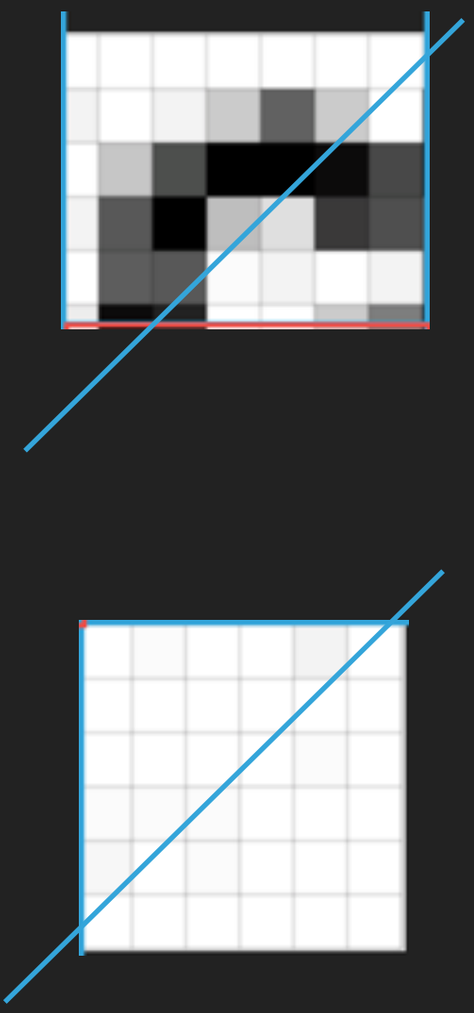


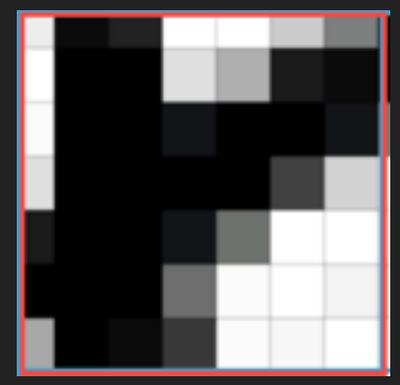


卷积后生成的8可能的特征

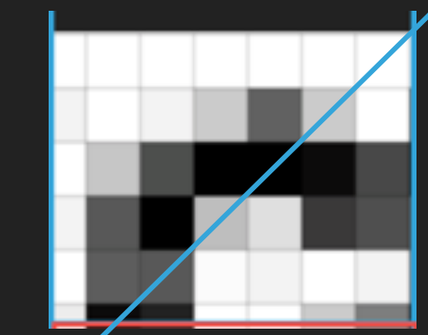


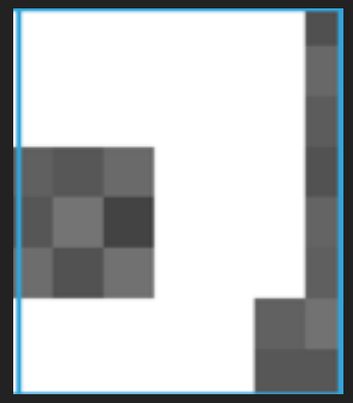
卷积后生成的9可能的特征





8





9

思考