不同图像识别技术的对比与探究

孟悦琦　朱宸慷　杨钧博　李昌玟　尹容乎

2023 年 11 月 11 日

摘要

车辆识别是计算机视觉领域的重要问题。目前有很多识别方法可以完成车辆识别。这些方法大抵可以分为传统方法和神经网络方法两种。本文通过使用 Adaboost、AlexNet、YOLO 等多种识别方法，对比分析不同方法之间的原理与差别，得到较为合适的模型。最后提出可能的改进方案。

1　背景介绍

随着城市化进程的加速，车辆数量激增，交通管理难度加大。为了改善城市交通状况，保障出行安全，开发自动车辆识别系统势在必行。开发一个车辆识别系统可以提高交通管理效率、提升交通安全性。同时，车辆识别模块可以应用在诸多方面的问题上。

有很多 CV 的方法可以进行车辆识别，这里主要可以分为传统方法和神经网络方法。传统主要包括支持向量机、马尔可夫网络等。其特点是图像识别使用的模型相对简单、包含的参数数量有限。同时，传统方法可能很依靠数据集的选取，在不同的数据集上可能有截然不同的表现。

在 CV 领域，2010 年出现了基于深度神经网络的模型。例如 AlexNet，便是基于卷积神经网络的模型。AlexNet 具有高于传统 CV 方法的识别准确度 [**?**]。之后又出现了 ResNet [**?**] 等方法，进一步提升其识别的准确度。近些年，又出现了 YOLO 模型。YOLO 是一种划时代的单阶段目标检测算法。YOLO 使用单次前馈网络即可完成检测，检测速度极快；整图预测充分利用全局信息，检测精度高，因此被广泛使用 [**?**]。另外还有，Transformer 模型，其是一种采用自注意力机制的深度学习模型，这一机制可以按输入数据各部分重要性的不同而分配不同的权重。通过借鉴 Transformer 的设计思想，Google 设计出 ViT 模型，也是一种识别准确度颇高的模型，且具有很强开创性的模型 [**?**]。在此基础上 Facebook 开发出 LeViT 模型，是其进一步分演进和发展 [**?**]

本文将通过对比传统模型和神经网络模型，来比较分析不同模型的优劣，并分析其中的原理。

2　数据集介绍

现在有很多开源的数据集。考虑到本文要使用一些传统方法进行识别，我们选择的数据集不宜太大。最终我们选择了 kaggle 上的 Multilabel car and color dataset[1]作为数据集。在数据集中，共包含三个品牌各三种颜色的车辆图片数据。数据集的部分图片如下：

[1]可以在网站 https://www.kaggle.com/datasets/julichitai/multilabel-small-car-and-color-dataset 中获取。

(a) matiz blue (b) tiggo black (c) rio blue (d) matiz red

图 1: 数据集样例

此数据集共有 9 个类，同时样本数量较少，不同品牌间视觉特征可能相近，如何提升模型泛化能力和防过拟合是关键。此数据集很适合用来考察传统模型和神经网络之间的差别。

3 层次聚类方法对图像的识别分析

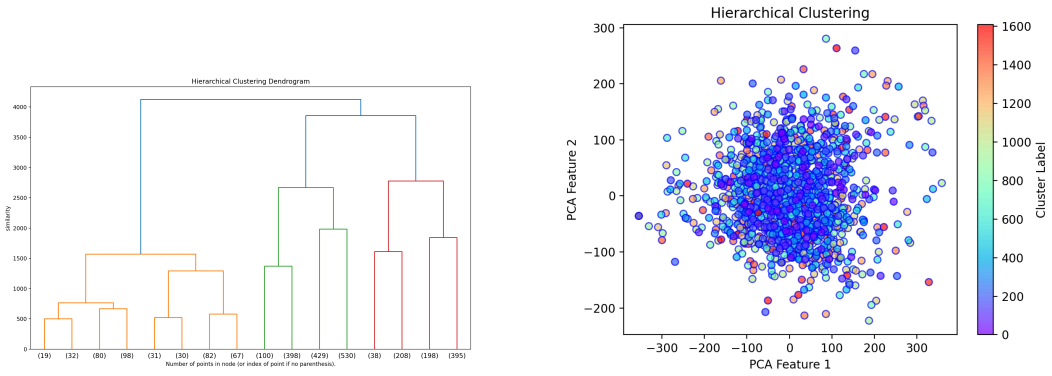
3.1 层次聚类介绍

在最开始，每个数据点作为单独的群集开始，然后合并相邻的群集以形成树状结构。简单来说，这可以被看作是自下而上的聚类方法。层次聚类是一种强大的无监督学习算法，它通过创建数据的层次结构来对相似的数据进行分组。该算法的优点在于能够帮助可视化和理解数据间的相似性。此外，通过层次结构可以创建更详细的分组，并理解群组间的关系和层次结构。

它主要应用于社交网络分析、客户细分、地理信息系统等领域。例如，在生物学中，它可以用于理解基因分析、物种分类和进化关系等。总之，由于这种算法提供了基于数据相似性的分析，因此它对于识别数据的结构和模式、提取信息非常有用。

通过测量数据之间的距离和聚类合并过程，可以获得有价值的信息。

3.2 使用层次聚类对图像进行识别的效果分析



(a) 训练结果 1 (b) 训练结果 2

数据集中的图像之间的差异在于车型和颜色。层次聚类是将彼此相似的图像聚集在一起形成群集的过程，形成得越快的群集其相似性就越大。为了便于理解，我们来看看树状图的黄色条形图，首先，X 轴上的节点

数，如 19、32、80、98 等，表示最初聚集在一起的群集中的图像数量。这是最初聚集在一起的最相似的图像群集。接下来，由 19 个图像组成的群集和由 32 个图像组成的群集在相似度数值 500 处形成了一个群集。这意味着与其他群集相比，由 19 个图像组成的群集和由 32 个图像组成的群集更为相似（相似度数值为 500），因此它们聚集在一起形成了下一个群集。相似度数值越接近 0，就越相似。通过这种层次化的方式，最终形成一个群集。在最开始群集的形成标准是车型和颜色中的哪一个，需要自主判断。

4 AdaBoost 对图像的识别分析

4.1 AdaBoost 简介

AdaBoost 是 Adaptive+Boosting 的组合词，其算法的定义如下：

弱分类器 (weak classifier) 通过顺序 (sequential) 学习相互补充，并将它们组合在一起以最终提高强分类器 (strong classifier) 的性能。

工作原理如下：

弱分类器 (weak classifier) 一次一个地顺序进行学习。首先学习的分类器会产生正确分类的数据和错误分类的数据。首先学习的分类器将正确分类的结果信息和错误分类的结果信息传递给下一分类器。下一分类器利用从前一个分类器接收到的信息来提高分类不佳的数据的权重 (weight)。也就是说，通过不断调整前一个分类器错误分类样本的权重，使其更集中于错误分类的数据，从而使学习效果更好。因此，名称中带有“adaptive”。最终分类器 (strong classifier) 通过对先前学习的弱分类器分别应用权重并进行组合来进行学习。

总结一下，就是将预测性能较低的弱分类器组合在一起，最终形成一个性能稍好一些的强分类器。弱分类器通过相互补充 (adaptive) 的方式进行学习，并通过组合这些弱分类器来形成一个分类器，因此称为 boosting。

用公式表示如下：

$$H(x) = \alpha_1 h_1(x) + \alpha_2 h_2(x) + \cdots + \alpha_t h_t(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{1}$$

其中：

$H(x)$：最终强分类器，也称为加权多数投票分类器。

h：弱分类器，也称为基分类器。

α：弱分类器的权重，用于衡量弱分类器对最终分类器的重要性。

t：迭代次数，表示弱分类器的数量。

AdaBoost 算法的工作原理如下：

1. 初始化训练数据集的每个样本的权重为 1/N，其中 N 是训练数据集的样本数。
2. 训练一个弱分类器。
3. 计算弱分类器的错误分类率。
4. 将错误分类率高的样本的权重增加，将错误分类率低的样本的权重减少。
5. 重复步骤 2-4，直到满足某个终止条件。
6. 将所有弱分类器的输出通过加权求和得到最终的强分类器的输出。

AdaBoost 算法具有以下优点：

1. 可以有效地提高弱分类器的性能。
2. 可以处理异常值。
3. 可以处理不平衡数据集。

4. AdaBoost 算法在分类、回归、异常检测等领域都有广泛应用。

5 AlexNet 对图像的识别分析

5.1 AlexNet 介绍

AlexNet 是一个深度卷积神经网络，它的结构如下：[?]

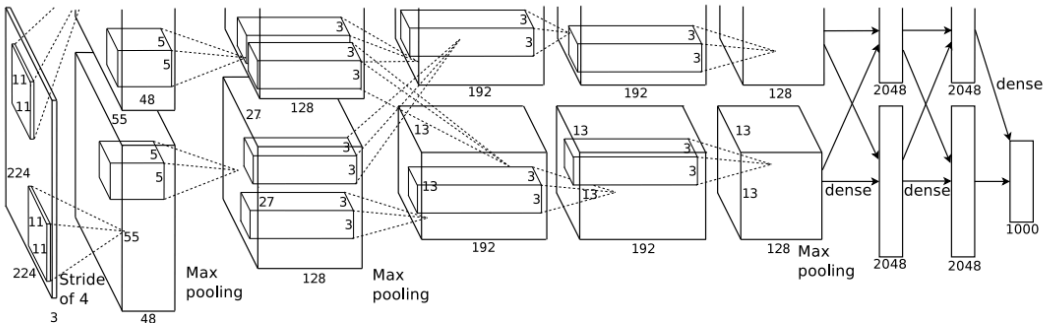


图 2: AlexNet 结构

AlexNet 主要特点：

1. 更深的网络结构：AlexNet 由 5 个卷积层、3 个全连接层和最后的 Softmax 分类层组成。
2. ReLU 激活函数：AlexNet 是第一个大规模使用 ReLU 作为激活函数的网络，这加速了训练过程。
3. Dropout：为了减少过拟合，AlexNet 在全连接层中使用了 Dropout。
4. 局部响应归一化（LRN）：在某些卷积层后使用了局部响应归一化。
5. 数据增强：为了进一步减少过拟合，AlexNet 使用了图像平移、翻转和颜色变化等数据增强技术。

5.2 使用 AlexNet 对图像进行识别的效果分析

训练结果如下：

Epoch	Loss	Test Accuracy
10	1.10	46.48
20	0.75	58.00
30	0.42	58.74
40	0.34	57.37
50	0.13	55.33

表 1: 随训练轮数增加 Loss 和 Test Accuracy 的变化

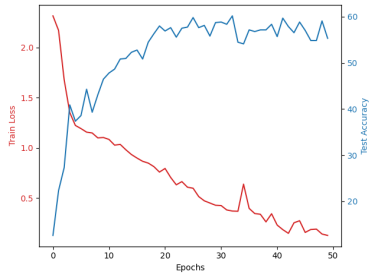


图 3: AlexNet 训练结果

在训练过程中，我们将数据集按 7：3 的比例划分为训练集和测试集。采用 AlexNet 进行训练，训练 50 轮后，发现模型对于测试集的预测准确率不断上升，最后趋于稳定，稳定在 55% 左右；损失函数值也在不断下降，最终达到了 0.13。考虑到样本集中的数据量只有 2700 左右，数据量偏少，因此训练效果不是很理想。如果增大样本集规模，模型预测的准确率应该会有进一步提升。

在训练过程中发现，虽然损失函数值在下降，但是预测准确率却并未上升，有时甚至下降，这说明模型可能存在过拟合现象，这也可能是导致模型预测准确率不高的原因之一。

6 ResNet 对图像的识别分析

6.1 ResNet 介绍

ResNet 是一个划时代的深度神经网络架构，它是由 Kaiming He，Xiangyu Zhang，Shaoqing Ren，Jian Sun 于 2015 年提出的。[?] 残差学习是利用残差块的设计，实现前层特征的直接传递。这样的结构成功地解决了随着网络加深，梯度会逐渐消失的问题。这使得 ResNet 可以使用很深的神经网络进行训练。ResNet-152 层的网络在 ImageNet 图像分类任务上取得了 3.57% 的 top-5 的错误率，对当时视觉领域造成了巨大的影响。

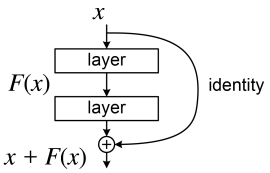


图 4: ResNet 中残差块的结构

总的来说，ResNet 开创了深度残差网络的新范式。它简单而高效的架构设计，成功地训练了百层级甚至千层级的超深网络，使得超深的神经网络成为现实。同时，ResNet 在识别的过程中也有着很高的成功率，是图像识别神经网络的一个重要选择。

6.2 使用 ResNet 对图像进行识别的效果分析

训练结果如下：

Epoch	Loss	Test Accuracy
10	0.92	67.15
20	0.70	67.97
30	0.57	79.67
40	0.47	78.44
50	0.40	86.24

表 2: 随训练轮数增加 Loss 和 Test Accuracy 的变化

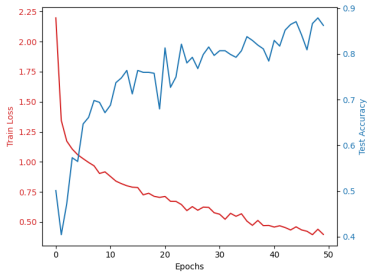


图 5: ResNet 训练结果

通过分析识别的准确率可以发现，使用 ResNet 进行训练，随着训练的轮数增加，Loss 逐渐下降，而准确率不断上升。在训练 45 轮左右时，准确率基本稳定在 85% 左右，且仍有上升势头。发现在 50 轮训练的过程中并未出现过拟合的现象，原因可能是 ResNet 的网络深度比较深，较少轮数的训练不足以达到过拟合的情况。

总的来说，ResNet 识别的准确率相较于 AlexNet 高出近 20%，具有长足提升。这可能得益于其具有的更深的神经网络和残差块的结构。

7 YOLO 对图像的识别分析

7.1 YOLO 模型简介

YOLO 模型有很多版本，本文章使用的是较新的 YOLOv8[2]模型。

通过架构图，对比之前的 YOLO 模型，YOLOv8 主要进行了一下改动：

1. 提供了一个全新的 SOTA 模型，包括 P5 640 和 P6 1280 分辨率的目标检测网络和基于 YOLACT 的实例分割模型。和 YOLOv5 一样，基于缩放系数也提供了 N/S/M/L/X 尺度的不同大小模型，用于满足不同场景需求。

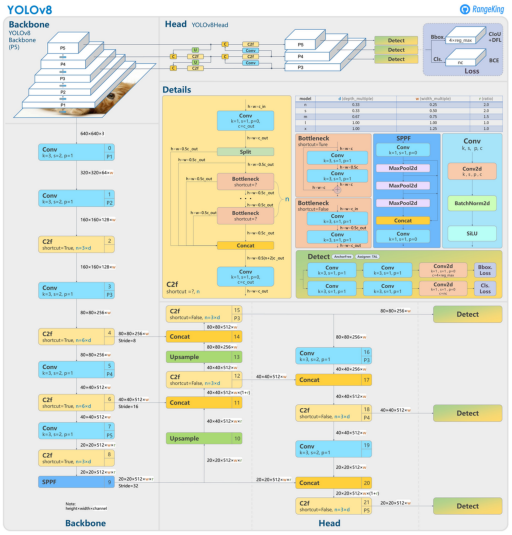


图 6: YOLOv8 架构

2. 骨干网络和 Neck 部分可能参考了 YOLOv7 ELAN 设计思想，将 YOLOv5 的 C3 结构换成了梯度流更丰富的 C2f 结构，并对不同尺度模型调整了不同的通道数，属于对模型结构精心微调，不再是无脑一套参数应用所有模型，大幅提升了模型性能。不过这个 C2f 模块中存在 Split 等操作对特定硬件部署没有之前那么友好了。

3. Head 部分相比 YOLOv5 改动较大，换成了目前主流的解耦头结构，将分类和检测头分离，同时也从 Anchor-Based 换成了 Anchor-Free。

4. Loss 计算方面采用了 TaskAlignedAssigner 正样本分配策略，并引入了 Distribution Focal Loss。

5. 训练的数据增强部分引入了 YOLOX 中的最后 10 epoch 关闭 Mosiac 增强的操作，可以有效地提升精度。

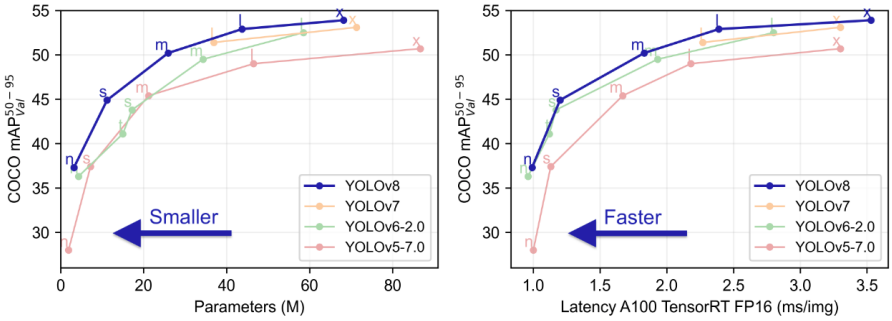


图 7: 历代 YOLO 对比

7.2 使用 YOLO 模型对图像进行识别的效果分析

[2]YOLOv8 的详细信息可以从 https://github.com/ultralytics/ultralytics 处获取。

参考文献

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1106–1114. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[3] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 779–788. [Online]. Available: https://doi.org/10.1109/CVPR.2016.91

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[5] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jegou, and M. Douze, "Levit: A vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 12 259–12 269.