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 10 月 26 日

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 [1]。之后又出现了 ResNet [2] 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 [3]。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 [4]。在此基础上 Facebook 开发出 LeViT 模型，是其进一步分演进和发展 [5]

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 AlexNet 对数据的识别分析

3.1 对 AlexNet 的介绍

AlexNet 是一个深度卷积神经网络，它的结构如下：[1]

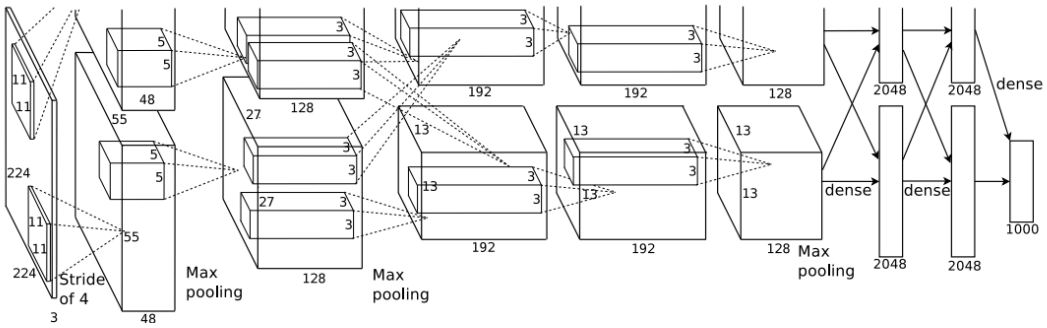

图 2: AlexNet 结构

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。

3.2 使用 AlexNet 对图像进行识别的效果分析

训练结果如下：

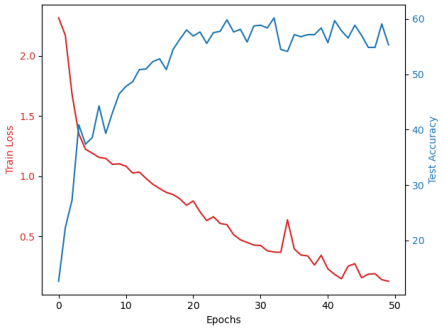

图 3: AlexNet 训练结果

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。

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.