Journal of Image and Signal Processing 图像与信号处理, 2023, 12(2), 128-135
Published Online April 2023 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2023.122013

Hans汉斯

基于ViT的中欧班列集装箱Logo图像分类识别研究

刘彦鹏[1,2]

[1]青岛科技大学信息科学技术学院，山东 青岛
[2]苏州市港航投资发展集团有限公司，江苏 苏州

收稿日期：2023年3月24日；录用日期：2023年4月14日；发布日期：2023年4月26日

摘 要

随着“一带一路”、长江经济带等国家战略的叠加实施，中国与西方的贸易也呈直线上升的趋势。伴随着中欧班列高质量运行，中欧班列集装箱的分类统计也成为中国对外贸易的一大重点。集装箱上代表着各供应商的logo图像在运输途中会产生褪色破损情况从而给识别增加了很多的难度。因此，公司为了有效识别各个标识，提出了基于ViT的logo图像分类模型，使用激活函数GELU代替传统的RELU。实验表明，改进后的模型可以很好地分类识别，准确率达到98%，且优于其他的分类模型。

关键词

logo分类，ViT，中欧班列，GELU

Research on Classification and Recognition of Container Logo Image of China-Europe Train Based on ViT

Yanpeng Liu[1,2]

[1]School of Information Science &Technology, Qingdao University of Science and Technology, Qingdao Shandong
[2]Suzhou Port & Shipping Investment and Development Group Co., Ltd., Suzhou Jiangsu

Received: Mar. 24th, 2023; accepted: Apr. 14th, 2023; published: Apr. 26th, 2023

Abstract

With the overlapping implementation of national strategies such as the “the Belt and Road” and

文章引用: 刘彦鹏. 基于 ViT 的中欧班列集装箱 logo 图像分类识别研究[J]. 图像与信号处理, 2023, 12(2): 128-135.
DOI: 10.12677/jisp.2023.122013

the Yangtze River Economic Belt, China's trade with the West is also on the rise. With the high-quality operation of the China-Europe train, the classified statistic of the containers of the China-Europe train has also become a major focus of China's foreign trade. The logo image representing each supplier on the container will be faded and damaged during transportation, which adds a lot of difficulties to the identification. Therefore, in order to effectively identify each logo, the company proposed a logo image classification model based on ViT, using the activation function GELU instead of the traditional RELU. The experiment shows that the improved model can classify and recognize well, with an accuracy of 98%, and is superior to other classification models.

Keywords

Logo Classification, ViT, China-Europe Train, GELU

1. 引言

集装箱 logo 图像的分类难点有很多，首先就在于其所处的特殊环境，集装箱在运输过程中一直风吹日晒，logo 图像时常会产生褪色和破损的情况，从而造成一些主要细节的丢失，且集装箱的堆放是层级堆放，在拍摄识别的时候会出现拍摄图像目标高低不一，远近有别的情况，就会时常在识别的时候会出现差错。并且很多 logo 图像都是字母组成，区别不大，比如“Olvi”和“Ovl”是两个不同 logo 图像标志的代表。其实传统的神经网络模型在图像粗分类问题上已经产生了较好的效果，例如 AlexNet [1]、VGG [2]、ResNet [3]等著名的神经网络模型。不过传统的分类模型是对整张图片进行特征提取，进而忽略了对全局的把控，使得这些分类网络模型在这种破损的数据分类上表现得不尽如人意。

传统模型的特征提取存在很多不完善的情况，其中 Sift [4]算法的实时性不高，对边缘光滑的目标无法进行准确地检测，Part R-CNN [5]需要将不同特征进行连接，接着用 SVM 进行分类。传统的方法，要么在识别的时候会出现偏差，要么会产生人工标签标注这种费时费力的情况。

2017 年，Google 提出 Transformer [6]模型用于解决自然语言类处理的问题，因为在并行计算上的优势使得其取代了 RNN 网络。2020 年 Google 再次提出 Version Transformer [7]模型，用于视觉领域。本次实验即采用该模型。实验证明，ViT 在图像全局结构的把握上产生了很好效果。

本文从公司的信息系统以及现场拍摄采集数据集，然后对于数据集进行预处理和图像增强操作，以此来提高图像的识别难度从而获得实验用的数据集，最后将数据集输入 ViT 模型中进行实验，以此来验证在不同激活函数和不同分类模型中的性能。

2. 相关技术

2.1. Vision Transformer (ViT)

Transformer 是一种用于自然语言处理的网络模型，模型包含两个部分，分别是 Encoder (编码器)和 Decoder (解码器)。Encoder 将输入的特征序列编码转为中间特征，而 Decoder 随机生成一个序列，再和中间序列进行结合最终输出结果。

随着 Transformer 在自然语言处理领域取得了成功并且广泛应用，大家逐渐将目光投向计算机视觉任务。在不改变 Transformer 结构的条件下，将语句的分类应用于图片中。本文的 ViT 网络结构模型因为是

分类模型，所以只用到了 Encoder 而没有用到 Decoder。结构如图 1 所示，

首先将图片分成规格相同的图片块，用 Encoder 对于图片块进行特征提取，不同的图片块只是整个图片的一部分，无法代表全部，因此需要将所有图片块结合成一个 class token。通过添加位置编码区分 token，再用多头注意力机制获得彼此之间的权重关系[8]。Class token 具有整张图片的信息，最后用于分类。

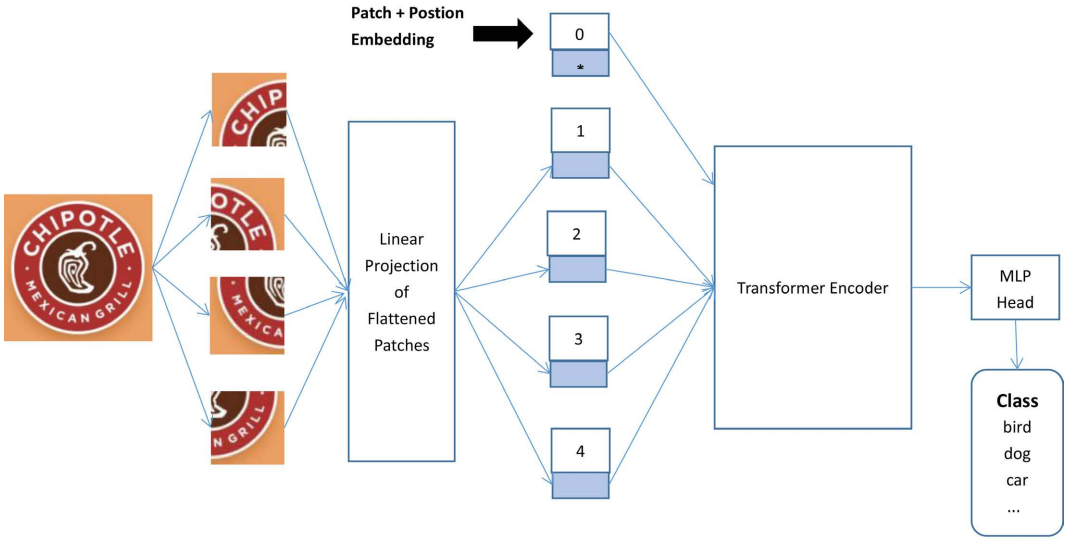

Figure 1. Structure diagram of vision transformer classification model
图 1. Vision Transformer 分类模型结构图

2.2. GELU 激活函数

在深度学习中，卷积层之间的过渡需要非线性激活函数来处理，从而使得网络模型可以解决比较复杂的问题。如果不使用激活函数，则网络模型无论多少层，也就好比一个线性的模型，因此激活函数对卷积神经网络至关重要。

目前大多数模型采用的激活函数以 RELU 为主，RELU 计算快且可以使得梯度快速收敛。但是 RELU 在输入为负数的情况下输出为 0，从而会存在大量神经元失效的情况。

高斯误差线性单元(GELU [9])是一种高性能的神经网络激活函数。GELU 函数在输出为负值的情况下仍然保留了一部分输出，避免出现了 RELU 全部归零的问题。在计算机视觉等任务上，使用 GELU 作为激活函数的模型性能常常超过 RELU。GELU 的公式以及导数曲线图如图 2 所示。

$$\mathrm{GELU}(x)=xP(X\le x)=x\Phi(x) \tag{1}$$

其中，$\Phi(x)$ 是正态分布的概率函数

GELU 函数的求导结果如下：

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{GELU}(x)=\Phi(x)+xP(X=x) \tag{2}$$

3. 算法实现

本文利用 ViT 模型来提取图像特征，最后进行图像分类，核心的模块主要分为五块：图像增强，特征提取策略，位置编码，图像编码，图像分类。

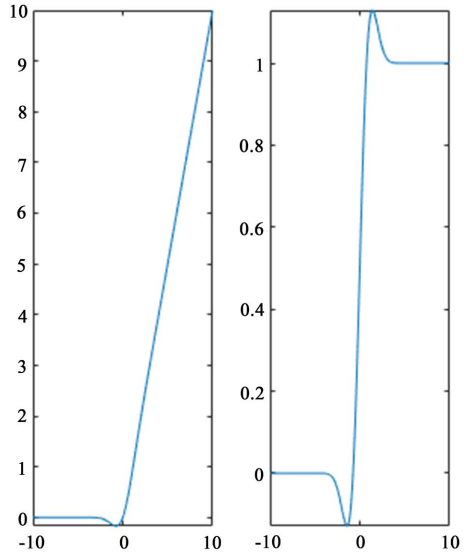

Figure 2. Curves and derivative graphs of the GELU activation function
图 2. GELU 激活函数的曲线和导数曲线图

3.1. 图像增强

图像增强是为了给模型增加难度，让模型适应各种各样的训练数据，从而达到一个很好的验证效果，本次图像增强中采用的是自动增强法，图像有翻转，倒置，对比度，饱和度等 16 种操作，每种操作数值不一，产生的概率不一，从而极大的提升了图像的多样性，进而提升模型质量。效果如图 3 所示。左边为原始图像，右边为自动增强后的图像。

(a) (b)

Figure 3. Image enhancement effect
图 3. 图像增强效果图

3.2. 特征提取策略

首先将图像设置成统一的 H*W，输入为 B*C*H*W，B 为 batch，C 为通道数，H 为高度，W 为宽度，输入的时候，会将每个图像分成一个个大小相同的图片块，图片块的大小为 P*P，从而个数则为 N = H*W/(P*P)，然后将 N 拉平拉长，而 C 也会卷积成一个高维向量 D，从而得到输出为 B*D*N。因为得到的 N 个 token 都是代表原图像中其中一块的特征，不具有全局性，无法去做分类，于是需要将得到的 N

个 token 融合起来产生一个新的 token，称为 class token，最终得到 N + 1 个 token。由此得到维度 B*D*(N + 1)。

3.3. 位置编码

位置编码是用于区分每一个 token 的必要方法，每个 token 都来自原来图像的不同位置，且相邻的 token 往往存在一定的联系，所以给每个 token 添加一个位置编码，保留 token 的空间位置信息，class token 具有全局性，不需要位置编码。位置编码往往采用一维向量。如果本环节的各个 token 不用位置编码进行区分，则各个位置的 token 就无法区别，从而降低识别准确率。

3.4. 图像编码

Encoder 图像编码中会产生自注意力机制，为了能够从多方面来考虑权重，所以采用了多头，多头的数量可以自己选择，每个 token 产生多个 Query，Key，Value，简称 q，k，v。各个 token 的 q 去寻找每个 token 对应的 k 从而得到彼此间的内积，再乘上对应的 v，彼此相加得到最终输出的特征[10]。如此往复，进行多层编码，q，k，v 会随时更新，得到最后的输出。图 4 是图像编码的过程图。

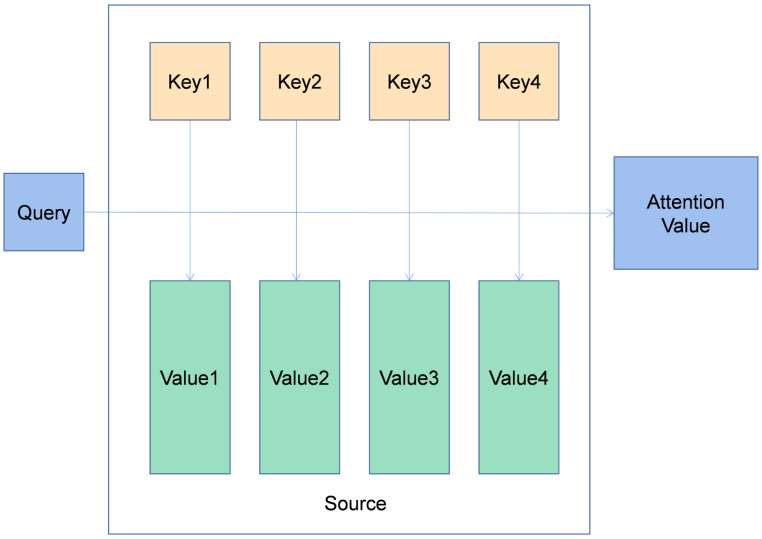

Figure 4. Process diagram of image encoding
图 4. 图像编码的过程图

3.5. 图像分类

在经历多层编码之后，模型用 class token 作为输出进入全连接层，将维度放大到 3072 维再缩小进行分类，分类器会计算出在每个类别中的概率值，最终将概率值最高的类别作为标签赋予。

4. 实验结果与分析

4.1. 数据集

本文的数据集是由信息系统存储的图像数据以及现场拍摄所得，拍摄角度也是由远及近，由左到右等多角度放大和缩小拍摄，由此想确保模型的说服力。同时通过自动数据增强方法提高图像的多样性。本次实验数据共有 2936 张，分为 59 个类，其中训练集有 2202 张，验证集 367 张，测试集 367 张，比例是 8:1:1。图 5 展示了部分的数据集图片。

Figure 5. Partial dataset image
图 5. 部分数据集图片

4.2. 实验设置与评价指标

实验将图片的尺寸统一为 224*224，图片块 patch 的大小为 16*16，224/16 = 14，因此 token 数量为 196。通过卷积将每个 token 的通道数 3 维转为 768 维。接着合成 class token，变成 197 个 token，class token 用于输出。每个 token 加入一维的位置编码，class token 不用位置编码。注意力机制采用 12 头，每个 token 产生 12 组 qkv。标签平滑为 0.1，训练 100 个 epoch，采用学习率预热，优化器为 Adam，batch 为 16，本实验通过得分作为分类指标，及通过概率转为分数作为图片的分类判断，最终输出准确率和损失值大小。

4.3. 实验结果

本次实验在训练 100 个 epoch 之后最终得到准确率为 98.0000%。损失值也从 4.0309 降到了 1.3879，准确率在中途曾经有过回流，不过在 20 个 epoch 之后，又开始呈现上升的趋势。12 头注意力机制的使用对于抓准目标特征起到了很大的作用，避开了无用的背景信息，对 logo 图像的轮廓有清晰的定位，找准目标里的数据特征注意力机制可视化如图 6 所示，亮色为模型所着重关注的地方。

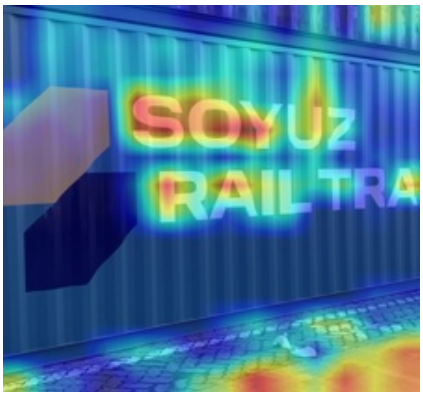

Figure 6. Visual image of attention mechanism
图 6. 注意力机制可视化图片

4.4. 不同激活函数的对比

实验首先对于不同激活函数之间的性能做出了对比，分别采用 GELU 和 RELU。实验表明，两者的

准确率和损失值的曲线走向基本上差不多，但是 RELU 的准确率是 95.6284%，损失函数为 1.5836，相较于 GELU 还是略逊一筹。由此可见，本实验采用 GELU 作为激活函数可以得到更好的效果。图 7 和图 8 分别是两者的准确率和损失值的曲线图。

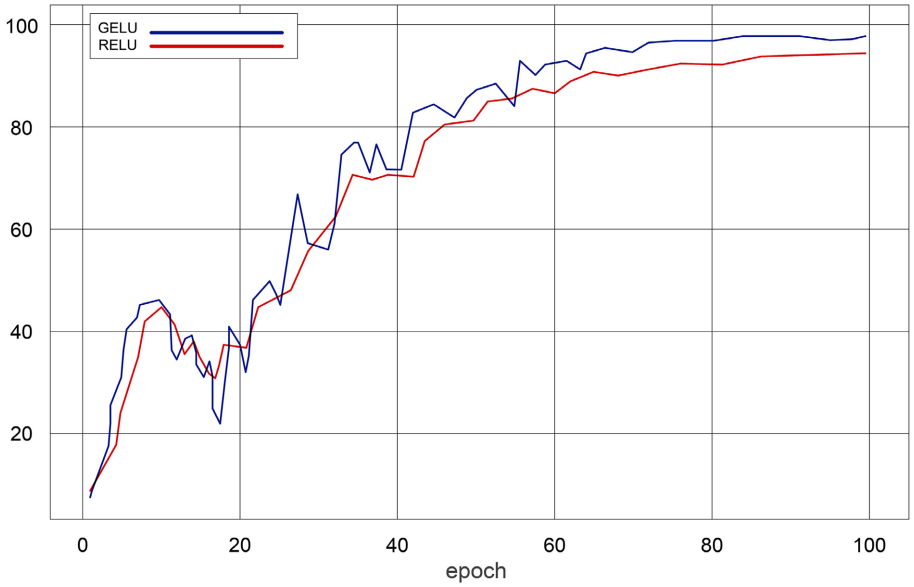

Figure 7. Accuracy comparison curve
图 7. 准确率对比曲线图

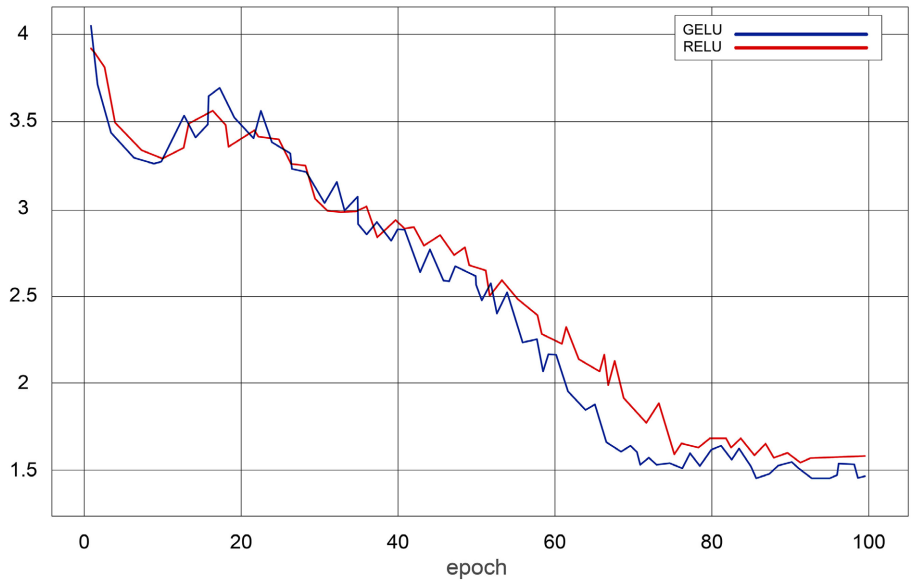

Figure 8. Loss value comparison graph
图 8. 损失值对比曲线图

4.5. 不同模型的对比

实验同样对比了不同网络模型之间的性能，本文的对比实验选取了常用了的 ResNet18，VGG16，AlexNet 模型对于实验数据集进行测试。表 1 是模型的对比情况。

Table 1. Comparison table of different models
表 1. 不同模型对比表格

模型	准确率 Acc	损失值 Loss
AlexNet	92.7345	1.9891
VGG16	93.9823	1.8056
ResNet18	94.6412	1.7153
Our	98.0000	1.3879

实验表明，本文的模型在识别破损的集装箱 logo 图像方面的性能优于其他的模型，准确率更高，损失值更小。改进后的 ViT 模型在图像识别方面能够更好的从全局出发，把握好 logo 图像的整体框架结构，即使图像出现破损的情况，也能在结构轮廓上对于图像有清晰的定位和识别，本文的 12 头注意力机制可以很好的抓住主要特征，采用 GELU 激活函数也能够在输入为负值的情况下进行激活，从而产生一个缓冲的效果。最终在性能上达到了预期。

5. 结语

集装箱上的 logo 图像因为其所处的特殊环境，所以与其他的图像存在一定的差异，因此识别起来会有不小的难度。本次实验采用 ViT 模型对集装箱 logo 图像进行识别，多头注意力机制对于数据的特征把握得十分精确，预测了 logo 的轮廓，成功克服了对于残缺和变色的 logo 图像无法准确判断的问题，采用 GELU 激活函数代替 RELU，使得模型在收敛的时候达到更好的效果，经过对比实验，改进后的模型的准确率达到了 98.0000%，性能优于其他模型。由此验证了本文的模型能够对集装箱 logo 图像进行较为准确地识别与分类。也为该公司在图像识别上提供了一定的参考价值。

参考文献

[1] Guo, Y. and Yang, L.D. (2022) Radar Moving Target Detection Method Based on SET2 and AlexNet. *Mathematical Problems in Engineering*, **2022**, Article ID: 3359871. https://doi.org/10.1155/2022/3359871

[2] Teng, B.W., Zhao, H.J., Jia, P., Yuan, J.F. and Tian, C.H. (2020) Research on Ceramic Sanitary Ware Defect Detection Method Based on Improved VGG Network. *Journal of Physics*: *Conference Series*, **1650**, Article ID: 022084. https://doi.org/10.1088/1742-6596/1650/2/022084

[3] Xiao, Y.T., Yin, H.S., Wang, S.H. and Zhang, Y.D. (2021) TReC: Transferred ResNet and CBAM for Detecting Brain Diseases. *Frontiers in Neuroinformatics*, **15**, Article ID: 781551. https://doi.org/10.3389/fninf.2021.781551

[4] 彭得阳, 邓安健. 基于 SIFT 的特征均匀提取改进研究[J]. 测绘与空间地理信息, 2021, 44(12): 46-50+56.

[5] 万子伦. 基于改进 Faster-RCNN 的多尺度人脸口罩检测算法研究[D]: [硕士学位论文]. 开封: 河南大学, 2022. https://doi.org/10.27114/d.cnki.ghnau.2022.001060

[6] 高金金, 李潞洋. 一种改进的点云 Transformer 深度学习模型[J]. 中北大学学报(自然科学版), 2021, 42(6): 515-523.

[7] 袁媛, 陈明惠, 柯舒婷, 王腾, 何龙喜, 吕林杰, 孙好, 刘健南. 基于集成卷积神经网络和 Vit 的眼底图像分类研究[J]. 中国激光, 2022, 49(20): 108-116.

[8] 杜显君. 基于 ViT 的高速公路车辆细分类研究[J]. 现代计算机, 2022, 28(12): 51-55.

[9] 郭文龙, 刘芳华, 吴万毅, 李冲, 肖鹏, 刘朝. 融合 ViT 卷积神经网络 的木板表面缺陷识别[J]. 计算机科学, 2022, 49(S2): 609-614.

[10] 顾昕, 叶海良, 杨冰, 曹飞龙. 结合信息保留的多头注意力图池化模型[J]. 中国计量大学学报, 2022, 33(2): 288-296.