KAN 网络是否超过了 MLP 和 CNN 网络？

冯欣

1 我们的工作

今年 4 月，来自麻省理工学院（MIT）的博士生提出了 KAN（Kernel Attention Network）[1]网络，并将其与现有的 MLP（多层感知机）网络进行了深入对比。这项研究的初衷在于探索 KAN 网络是否具备替代并超越 MLP 网络的潜力。在这篇文章中，阐述 KAN 网络作为一种创新的可学习架构，其核心思想在于将传统神经网络中的权重参数替换为可学习的单变量函数。这一变革不仅提升了神经网络的性能，还增强了模型的可解释性，为深度学习领域带来了新的可能性。KAN 网络的出现无疑对现有的 MLP 网络构成了强有力的挑战。通过其卓越的性能和可解释性，KAN 网络展现了巨大的潜力和价值。随着研究的深入和技术的成熟，我们有理由相信 KAN 网络将在未来神经网络领域占据重要的地位。为了全面验证 KAN（Kernel Attention Network）神经网络模型在分类任务中的性能，我们特别在 Fashion-MNIST 数据集上进行了详尽的实验。除了与经典的 MLP（多层感知机）网络进行对比外，为了满足实验要求，我们还进一步将 KAN 网络拓展至卷积神经网络（CNN）的架构，并在相同的 Fashion-MNIST 数据集上进行了测试，以探究其是否具备替代 CNN 的潜力。本报告主要聚焦于 MLP、KAN、卷积 KAN 以及 CNN 网络在 Fashion-MNIST 数据集上的表现。为了确保对比的公正性和有效性，我们严格控制了每种模型的层数以及其他关键参数，以确保实验结果的可靠性。通过这一系列的实验，我们期望能够深入了解 KAN 网络及其卷积变体在图像分类任务中的性能，并评估它们是否能够在保持高效性能的同时，提供比传统 MLP 和 CNN 网络更高的可解释性和灵活性。这些发现不仅有助于推动神经网络架构的创新，还可能为未来的研究提供有价值的参考。

2 实验方法

在本节中，我们主要给出每种使用的神经网络部分原理和数据集的处理操作，我们分别介绍多层感知机、KAN 网络、一般卷积神经网络和 KAN 卷积神经网络等相关原理。

2.1 多层感知机和 KAN 网络结构

2.1.1 多层感知机。

MLP，即多层感知机（Multilayer Perceptron），也称为人工神经网络（Artificial Neural Network, ANN），它以其卓越的能力著称，能够通过一个或多个隐藏层将输入层的输入向量值精确地映射到输出层的输出向量值。这种映射过程模拟了生物神经网络中的信息处理和传递机制，使得 MLP 在模式识别、分类、回归等任务中展现出强大的性能。其中在 MLP 中单个神经元如图 1 所示。

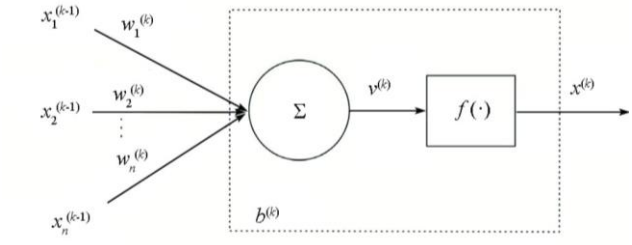

图 1 单个神经元模型

在图 1 中，$x = (x_1^{k-1}, x_2^{k-1}, \ldots x_n^{k-1},)$表示神经元的 n 个输入；$w_1^{(k)}$、$w_2^{(k)}$、…、$w_n^{(k)}$为输入权重；$b^{(k)}$表示神经元的偏置；f(·)为激活函数；v(k)为神经元权值向量和输入向量的线性加权；x(k)为神经元的输出，则单个的神经元输出可表示由$v^{(k)} = \sum_{i=1}^{n} w_i^{(k)} x_i^{(k-1)}$。当没有激活函数时，输入和输出展现的是一种线性组合形式。加上激活函数后，非线性因素被引入神经元中，使得神经网络可以任意逼近任何非线性函数，这样神经网络就可以解决更复杂的非线性问题了。然而仅仅单层神经网络无法解决更为复杂的线性不可分问题，由此引入了多个神经元并列分层的方式，在输入层和输出层之间增加 1 个隐藏层，即多层感知机。多层感知机的简化模型如图 2 所示。

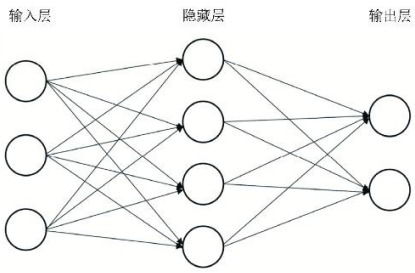

图 2 多层感知机简化模型

2.1.2 KAN 网络结构

KAN 的架构设计来自一个数学问题：对一个由输入输出对 {xi, yi} 组成的有监督学习任务，寻找函数 f 使得所有数据点的 yi ≈ f (xi)。其核心在于找到合适的单变量函数Φq, p（内部函数）和 Φq（外部函数）。在 KAN 中，使用 B-spline（B 样条）来构建。对于 B-spline，函数在其定义域内、在结点（Knot）都具有相同的连续性。其多项式表达可由 Cox-de Boor 递推公式表达：

$$B_{i,0}(x) := \begin{cases} 1 & \text{if } t_i \le x < t_{i+1}, \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{i,k}(x) := \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x).$$

例如 KAN 定理的内部函数可以定义为带有参数的矩阵计算。矩阵中的每个元素事实上是一个函数或算子。其中 KAN 层可以定义为下式:

$$\mathbf{\Phi} = \{\phi_{q,p}\}, \qquad p = 1, 2, \cdots, n_{\text{in}}, \qquad q = 1, 2 \cdots, n_{\text{out}},$$

那么根据 KAN 定理，理论上只要 2 个 KAN 层就可以充分表征实数域的各类有监督学习任务。2 层的 KAN 中，激活函数放置在边缘而不是节点上（在节点上进行简单求和），并且 2 层中间有 2n+1 个变量。一般的 KAN 表征形式是：

$$\text{KAN}(\mathbf{x}) = (\mathbf{\Phi}_{L-1} \circ \mathbf{\Phi}_{L-2} \circ \cdots \circ \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0)\mathbf{x}.$$

其中Φ_l是第 l 个 KAN 层所对应的函数矩阵（B-spline 函数矩阵），x 为输入矩阵。图 3 为 MIT 团队给出了 KAN 的基本网络流程。

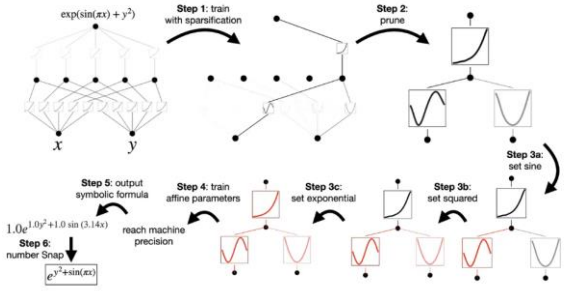

图 3 KAN 网络流程图

2.2 . 卷积神经网络和卷积 KAN 神经网络结构

2.2.1 卷积神经网络

卷积层（Convolutional Layer）是 CNN 的核心部分之一，该层利用卷积核对输入 张量进行卷积运算，并通过反向传播算法对卷积核参数进行优化，实现对输入张量的 高维特征提取。相较于传统的全连接层，卷积层能够更有效提取输入信息特征，同时 可以显著降低模型训练参数的数量。卷积的计算过程的本质是卷积核与输入张

量进行互相关运算（cross-correlation），通过卷积核在输入张量上的逐行滑动，计算出输出张量。卷积核可视为提取输入张量局部特征的算子，同一个卷积核可以提取输入张量不同位置的局部信息，从而降低了模型的参数数量。假设[**X**]为卷积层的输入张量，[**V**]为形状为$k_{\hbar} * k_w$的卷积核，卷积层的操作过程表达式如公式所示。

$$[H]_{i,j} = u + \sum_{k_h}\sum_{k_w}[V]_{k_h k_w}[X]_{i+k_h j, j+k_w}$$

其中uu是一个常数，表示卷积层的偏置，[**HH**]为卷积层的输出张量，$[H]_{i,j}$和$[H]_{i,j}$分别表示[**XX**]和[**HH**]中位置(ii, jj)处的值。为灵活设计 CNN 的输入输出形状，通常采用填充（padding）和步幅（stride）来控制卷积层的控制输出张量的形状。步幅指的是在卷积运算过程中卷积核滑动的步长，以张量中的元素个数作为单位。卷积层在进行计算时，卷积核从输入张量的左上角开始，向下、向右滑动，逐步计算出输出张量。为了使 CNN 可以处理 RGB 彩色图像，通常会引入通道（channel）的设计，卷积层输入张量的维度称之为输入通道，输出张量的维度称之为输出通道。并且，为让卷积运算能过够提取或整合到更多高维度的特征信息，卷积层的输入通道与输出通道并不相等。综上所述，卷积层是可通过卷积核实现从输入张量中提取高维特征目的的神经网络运算层，可以通过设置卷积层的卷积核大小、步幅和通道，来实现对输出张量大小的控制，若对上述卷积层的超参数进行巧妙设置，卷积层的输入与输出则会呈现出一定的数学美。

2.2.2 KAN 神经网络架构

通过将 KAN 的理论应用到卷积神经网络当中，相当于将卷积中的参数按照 KAN 层的方式进行修改，由于 KAN 是根据$Kolmogorov - Arnold$定理对拟合曲线进行逼近，KAN 卷积与卷积非常相似，但不是在内核和图像中相应像素之间应用点积，而是对每个元素应用可学习的非线性激活函数，然后将它们相加。KAN 卷积的内核相当于 4 个输入和 1 个输出神经元的 KAN 线性层。对于每个输入i，应用ϕ_i可学习函数，该卷积步骤的结果像素是$\phi_i(x_i)$的总和。对于 KAN 卷积中的参数，假设一个 KxK 内核，对于该矩阵的每个元素，都有一个ϕ，其参数计数为：$gridsize + 1$，其中ϕ定义为：

$$\phi = w_1 * spline(x) + w_2 * b(x)$$

v 这为激活函数b提供了更多的可表达性，线性层的参数计数为$gridsize + 2$。因此，KAN 卷积总共有 K^{2*}(gridsize + 2)个参数，而普通卷积只有K^2。

3 实验评价

在本节中，我们首先介绍我们使用的相关数据集和预处理的方法，然后是每种模型所挑选的参数等相关评价指标。

3.1 数据集与初始化设置

Fahion-mnist 数据集。Fashion-MNIST 数据集，由德国 Zalando 公司精心打造，是一个专为衣物图像设计的基准测试数据集。它包含 60,000 个样本的训练集和 10,000 个样本的测试集，每个样本均为 28x28 像素的灰度图像，并与 10 个不同的衣物类别标签紧密关联。如图 4。Fashion-MNIST 数据集旨在作为原始 MNIST 手写数字数据集的直接替代品，为机器学习算法提供更具挑战性和实用性的基准测试平台。

图 4 Fashion-MNIST 数据集简介

数据预处理。在利用 torchvision 库中的 Fashion-MNIST 数据集进行神经网络模型的训练和测试时，我们采用了特定的预处理步骤。其中，主要的数据预处理步骤是应用 transforms.ToTensor()方法。这一方法首先将图像转换为 PyTorch 张量，随后自动将图像在每个通道中的像素值从原始范围[0, 255]归一化至[0, 1]的范围内。处理后的张量遵循(C, H, W)的维度顺序，其中 C 代表通道数，H 表示图像高度，W 代表图像宽度。这样的预处理不仅简化了后续的网络训练过程，还有助于提升模型的收敛速度和性能。

3.2 神经网络结构模型参数设置

表格 1 神经网络参数设计

Hyperparameters	MLP	KAN	CNN	CNN_KAN
Mini-batch Size	[64,128,256]	[64,128,256]	[64,128,256]	[64,128,256]
Optimizer	Adam	Adam	Adam	Adam
Loss function	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss	CrossEntropyLoss
Learning Rate	[0.01,0.001]	[0.01,0.001]	[0.01,0.001]	[0.01,0.001]
Training Epochs	100	100	1000	100
Activation Functions	Tanh	Tanh	Tanh	Tanh
Pooling Layers	MAXPOOL	MAXPOOL	MAXPOOL	MAXPOOL
Conv/KAN-Layer	0	0	2	2
Linear-Layer	4	0	2	2
KAN	0	4	0	0

3.3 评估指标

在我们的实验中主要对不同分类下的数据集进行评估，我们使用分类器输出的预测概率来对目标追踪进行分类。如果一张图片被正确的分类，我们将记录为真正例（TP），否则记录为假负例（FN）。在 fashionmnist 数据集下，使用准确率、精确率、召回率和 F1 度量来评估不同模型在相同数据集下的表现情况，为了尽可能地保证数据训练和检测过程中的公平性和准确性，实验过程使用加权平局的召回率和 F1 度量作为指标。其中$Accuracy$等于$(TP+TN)/(TP+FN+TP+TN)$，精确率等于$TP/(FP+TP)$，回归率等于$TP/(FN+TP)$，F1 度量等于$2TP/(FN+FP+2TP)$，在多分类任务中我们分别使用每种分类的加权均值作为最后的结果。

4 实验结果

在这一节，我们展示了每个模型在一定程度相同的结构、参数和神经网络下，采用 MLP、卷积神经网络和 KAN 卷积 KAN 下的区别和对比情况。

4.1 MLP 和 KAN 对比

在 Fashion-MNIST 数据集上，我们进行了详尽的模型训练和测试，对比了不同参数设置下的 MLP（多层感知机）和 KAN（Kernel Attention Network）模型的性能。这些实验专注于十分类任务，并探索了学习率和批次大小对模型性能的影响。根据表 2 中的结果，我们发现较小的学习率（0.001）对两个模型的性能和稳定性有着显著的提升。具体而言，MLP 模型在正确率、精确率、召回率和 F1 度量等关键指标上均显示出明显的性能改善。尤其是在采用较大批量大小时，MLP 模型展现出了最优的性能。同样，KAN 模型在较小的学习率下也表现出了更高的性能，尽管其提升幅度相较于 MLP 模型略为逊色。然而，在相同条件下，KAN 模型的性能指标普遍高于 MLP 模型，特别是在学习率为 0.01 和中等批量大小（128）时，KAN 模型展现出了突出的性能。综上所述，MLP 和 KAN 模型在采用小学习率和大批量大小时均能达到最佳性能。然而，值得注意的是，KAN 模型在多个测试条件下均显示出了比 MLP 更高的性能指标，这体现了 KAN 模型在处理图像分类任务时的优势和潜力。这些发现为我们进一步优化神经网络模型提供了有价值的参考。

表格 2 MLP 与 KAN 对比结果图

模型	学习率	批次	正确率	精确率	召回率	F1 度量
MLP	Ir=0.01	64	0.70010	0.69665	0.70010	0.66570
		128	0.75610	0.74812	0.75610	0.72553
		256	0.81720	0.81853	0.81720	0.81274
	Ir=0.001	64	0.87700	0.87647	0.87700	0.87627
		128	**0.88390**	**0.88464**	**0.88390**	**0.88369**
		256	0.89080	0.89093	0.89080	0.89076
KAN	Ir=0.01	64	0.71720	0.73516	0.71720	0.69257
		128	0.85030	0.85012	0.85030	0.84897
		256	0.87030	0.86970	0.87030	0.86785
	Ir=0.001	64	0.89320	0.89380	0.89320	0.89302
		128	**0.89030**	**0.89214**	**0.89030**	**0.89042**
		256	0.88780	0.88899	0.88780	0.88811

为了更全面地比较 MLP（多层感知机）和 KAN（Kernel Attention Network）两类模型在 Fashion-MNIST 数据集上的训练和测试表现，我们选取了相同参数设置下的训练和测试曲线进行绘制。从图 5 中可以看到，尽管每种网络在图像识别上都展现出自身的优势，但二者之间的区别度并不十分明显。这表明 MLP 和 KAN 在处理该数据集时都具有一定的有效性。然而，从图 6 的收敛曲线可以看出，MLP 相较于 KAN 网络具有更好和更快的收敛性质。这意味着在训练过程中，MLP 能够更快地找到问题的最优解，从而节省训练时间。同时，对于模型的训练时间而言，MLP 大大优于 KAN 网络。但值得注意的是，当前 KAN 网络仍处于起步阶段，我们期待随着未来对模型的进一步优化，KAN 的训练时间将会大大减少。此外，我们还绘制了两模型在最后一刻的混淆矩阵，如图 5 所示。从混淆矩阵中，我们可以观察到两模型都对于十分类任务有着很好的检测效果。这进一步验证了 MLP 和 KAN 在处理 Fashion-MNIST 数据集时的有效性。通过比较混淆矩阵，我们可以更直观地了解模型在各类别上的分类性能，并找出可能存在的误分类情况，为后续的模型优化提供指导。

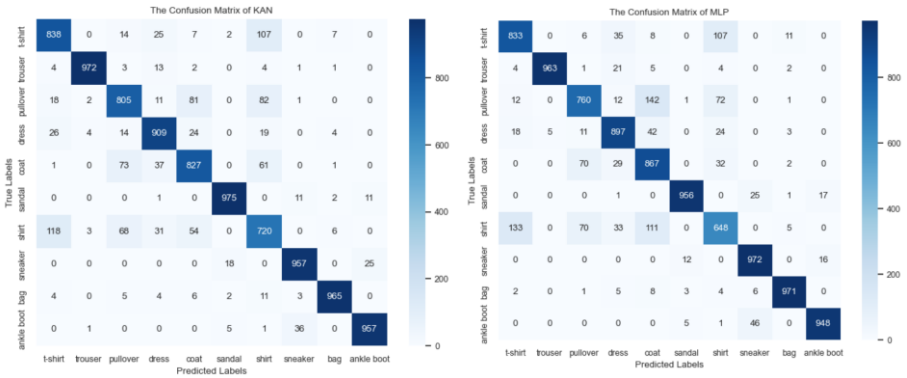

图 5 MLP 和 KAN 混淆矩阵分类图

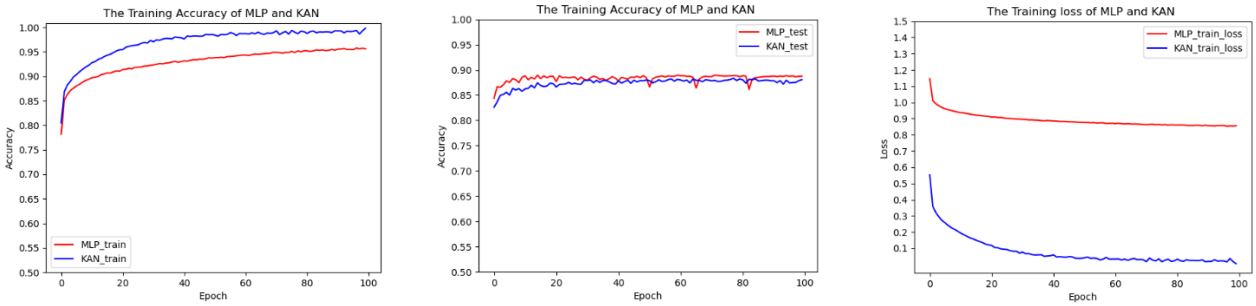

图 6 MLP 和 KAN 训练测试图

4.2. CNN 和卷积 KAN 网络比较

在本次 CNN（卷积神经网络）与 KAN（Kernel Attention Network）卷积模型的对比实验中，我们深入探究了两种模型在不同学习率（Ir）和批量大小（Batch Size）下的性能表现。实验结果表明，较低的学习率（0.001）普遍增强了模型的性能和稳定性。具体来讲，CNN 模型在学习率为 0.001 且批量大小为 256 的配置下，取得了最优越的性能，其各项指标均超过 0.908，展现出了出色的分类能力。相比之下，KAN 模型在相同学习率下，于批量大小为 128 时表现最佳，其正确率为 0.88920，F1 度量为 0.88755，尽管性能也相当可观，但相较于 CNN 模型仍有一定差距。值得注意的是，KAN 模型在学习率为 0.01 时虽表现出相对较好的性能，然而其整体性能依旧未能超越 CNN 模型。这一结果暗示，在处理类似 Fashion-MNIST 这样的图像分类任务时，CNN 模型凭借其强大的特征提取和分类能力，展现出了更优越的性能和稳定性。综上所述，CNN 模型在各项测试条件下均表现出更高的性能指标，尤其在较低学习率和较大批量大小的配置下表现尤为突出。而 KAN 模型虽有其独特的优势，但在本次对比实验中，其在处理相同任务时的性能稍逊于 CNN 模型。

表格 3 CNN 与 KAN 卷积对比结果图

模型	参数	批次	正确率	精确率	召回率	F1 度量
CNN	Ir=0.01	64	0.90210	0.90209	0.90210	0.90191
		128	0.89590	0.89675	0.89590	0.89523
		256	0.89800	0.89767	0.89800	0.89737
	Ir=0.001	64	0.90050	0.90097	0.90050	0.90017
		128	**0.90680**	**0.90652**	**0.90680**	**0.90656**
		256	0.90840	0.90844	0.90840	0.90834
KAN 卷积	Ir=0.01	64	0.87360	0.87600	0.87360	0.87415
		128	0.88020	0.88341	0.88020	0.88129
		256	0.88010	0.88019	0.88010	0.87978
	Ir=0.001	64	0.88710	0.88835	0.88710	0.88758
		128	**0.88920**	**0.88845**	**0.88920**	**0.88755**
		256	0.88640	0.88602	0.88640	0.88539

为了更细致地对比 CNN（卷积神经网络）和 KAN（Kernel Attention Network）两类模型在训练和测试过程中的性能，我们选取了相同的参数设置，并绘制了它们的训练和测试曲线。如图 7 所示，可以明显观察到 CNN 相较于 KAN 网络具有更好和更快的收敛性质。在训练过程中，CNN 能够更迅速地找到问题的最优解，其训练曲线显示出更平滑且更快的下降趋势，这表明 CNN 在优化过程中更加高效。同时，从训练时间上来看，CNN 也大大优于 KAN 网络。然而，需要指出的是，这主要是因为 KAN 网络目前还处于发展的初期阶段。我们有理由相信，

随着对 KAN 模型进行更多的研究和优化，其训练时间将会得到显著的减少。此外，我们还展示了两个模型在训练结束时的混淆矩阵（如图 7 示）。从混淆矩阵中，我们可以观察到 CNN 和 KAN 模型在十分类任务上都取得了很好的检测效果。这表明，无论是 CNN 还是 KAN，在处理类似 Fashion-MNIST 这样的数据集时，都展现出了较高的分类准确率和可靠性。综上所述，虽然 CNN 在收敛速度和训练时间上具有明显优势，但 KAN 网络作为一种新兴的神经网络结构，也展现出了其独特的潜力和价值。我们期待在未来通过不断的优化和改进，KAN 网络能够在图像分类等任务上取得更好的性能于发展的初期阶段。

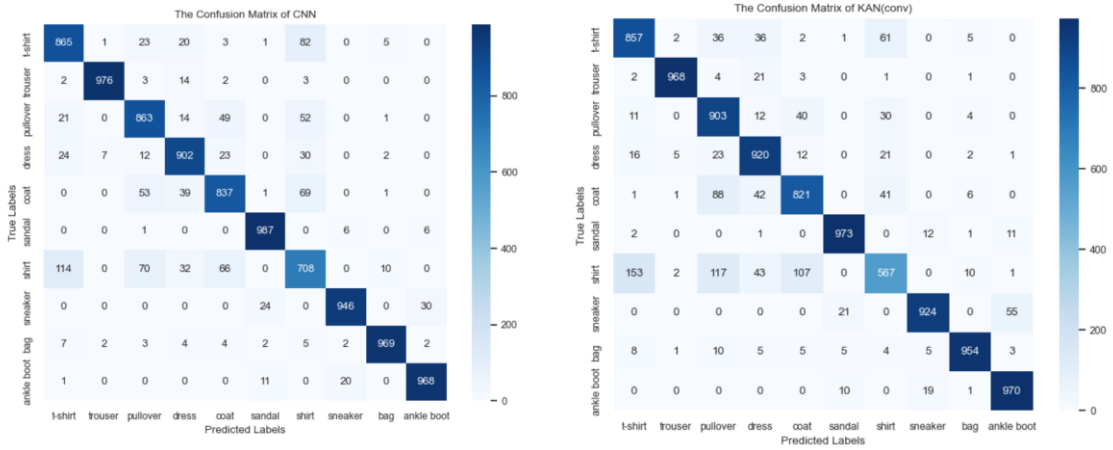

图 7 CNN 和 KAN 混淆矩阵分类图

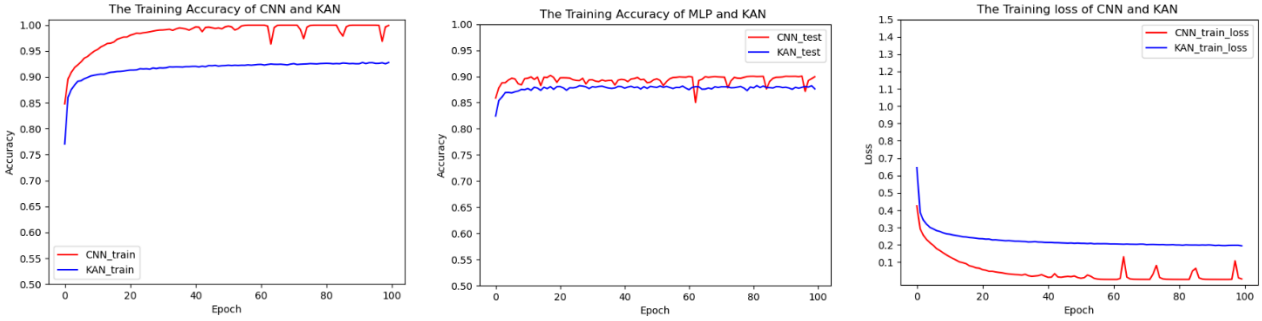

图 8 CNN 和 KAN 训练和测试图

5. 总结

经过近期对 KAN（Kernel Attention Network）模型与经典 MLP（多层感知机）和 CNN（卷积神经网络）的对比测试，我们在设置相同层数、学习率和批次大小的条件下，发现 KAN 模型并未如网络所描述的那样展现出卓越的收敛性和准确度。在相同的时间范围内，MLP 和 CNN 模型在性能上显著优于 KAN 及其卷积版本。然而，值得注意的是，KAN 作为一个新兴的研究方向，目前还处于发展的初期阶段。尽管在当前的测试条件下表现不如成熟的 MLP 和 CNN 模型，但 KAN 模型的参数量在同层次模型中显示出优势，这一点为其未来的发展提供了潜力。回顾 MLP 和 CNN 的发展历程，它们的优秀检测效果和收敛性并非一蹴而就，而是通过持续的研究和改进逐渐获得的。因此，我们有理由相信，随着对 KAN 模型研究的深入和技术的不断优化，它有望在未来成为神经网络领域的一个新的发展趋势，并找到属于自己的应用天地。

6. 参考文献

[1].Liu, Ziming, et al. "Kan: Kolmogorov-arnold networks." *arXiv preprint arXiv:2404.19756* (2024).