

高能物理喷注标记的深度学习模型综述

高六龙^{1,2}, 黄正坤^{1,2}, 姜晓巍^{1,2}, 孙功星¹, 李佳枫^{1,2}

(1. 中国科学院高能物理研究所, 北京 100049; 2. 中国科学院大学, 北京 100049)

摘要: 近年来,深度学习在计算机视觉、自然语言处理等应用领域取得了巨大的成功,致使高能物理研究者也开始关注深度学习技术,并探索其在强子喷注标记任务中的应用。最初研究者将喷注数据转化成图像和序列数据,采用卷积神经网络(CNN)和循环神经网络(RNN)对喷注进行标记,但存在计算效率低和可解释性差的问题。为了解决这些问题,研究者对网络结构进行了多方面的改进,并在构建的多种喷注标记数据集上进行训练,提升了模型分类的性能。本文对新型网络模型的重要模块进行深入分析综述,包括基于集合表示喷注的方法、等变性神经网络的应用以及喷注基础模型的探索。同时,对各种标记分类器进行了分析和比较,评估各种网络结构的性能,并对相关模型现状进行了分析与总结,探讨了深度学习模型在喷注标记任务中的应用前景。

关键词: 卷积神经网络; Transformer 模型; 喷注标记; 等变性神经网络; 基础模型

源代码链接: <https://github.com/sysu17363020/Models-For-Jet-Tagging/tree/main>

中图分类号: TP39

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0252689

Survey of Deep Learning Models in High Energy Physics Jet Tagging

GAO Liulong^{1,2}, HUANG Zhengkun^{1,2}, JIANG Xiaowei^{1,2}, SUN Gongxing¹, LI Jiafeng^{1,2}

(1. Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】 In recent years, deep learning has achieved tremendous success in application fields such as computer vision and natural language processing. This has led researchers in high energy physics to also turn their attention to deep learning technologies and explore their application in hadronic Jet Tagging tasks. Initially, researchers converted Jet data into image and sequence data, and used Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to tag Jets. However, these approaches suffered from problems such as low computational efficiency and poor interpretability. To address these issues, researchers have made improvements to network architectures from multiple perspectives and conducted training on various constructed Jet Tagging datasets, thereby enhancing the classification performance of the models. This paper provides an in-depth review of the key modules of novel network models, including methods for representing Jets based on sets, the application of equivariant neural networks, and the exploration of Jet foundation models. Meanwhile, this paper analyzes and compares various tagging classifiers, evaluates the performance of different network architectures, summarizes the current status of relevant models, and discusses the application prospects of deep learning models in Jet Tagging tasks.

【Key words】 Convolutional Neural Network (CNN); Transformer model; Jet Tagging; equivariant neural network; Foundation Model (FM)

0 引言

高能物理(HEP)主要研究微观尺度上的基本粒子以及支配它们相互作用的基本力量。这一领域通过极高能量和极端条件下的实验探索宇宙的基本构造。为了研究在传统方法下无法实现的能量水平,科研人员运用了许多复杂的仪器和设备,其中包括欧洲核子中心(CERN)的大型强子对撞机

(LHC)。在 LHC 中,两束质子被加速到接近光速,并以每秒 4 000 万次的频率碰撞。碰撞产生的不稳定粒子在量子色动力学(QCD)的理论支撑下,经过部分子级联和强子化过程,会衰变并产生大量的射出粒子。该过程所生成的准直粒子射流被称为喷注(Jet)。物理学家根据探测器捕获的射出粒子的位置、能量、动量、类型等信息为每次碰撞重建一个物理事件,再通过分析碰撞数据识别涉及新物理过程

基金项目: 国家自然科学基金(11775249)。

作者简介: 高六龙,男,博士研究生,主研方向为神经网络;黄正坤,博士研究生;姜晓巍,高级工程师、博士;孙功星(通信作者),研究员、博士;李佳枫,硕士研究生。

收稿日期: 2025-07-01

修回日期: 2025-11-09

E-mail: sungx@ihep.ac.cn

的事件,例如希格斯玻色子的发现。碰撞数据分析中一个关键步骤是喷注标记(Jet Tagging),旨在对碰撞产生的喷注进行分类,以识别引发喷注的粒子类型,例如顶夸克标记(Top Tagging)、夸克/胶子标记(Quark-gluon Tagging)和 W、Z、H 玻色子标记等^[1]。级联效应引发喷注的粒子会辐射,而辐射产生的粒子又会进一步辐射,最终导致大约产生 100 个末态粒子。这使得喷注标记成为一项具有挑战性的任务。传统的喷注标记方法依赖于由 QCD 驱动的手工观测量特征,例如基于喷注质量、衰变产物角分布和次级顶点等特征的多变量分析^[2-4]。这些方法需要大量的人工干预,并且受限于物理学家在设计观测量时的创造性,这限制了喷注标记效率的提高,而更高的标记效率意味着可以从单位时间、单位碰撞次数中提取更多有价值的信息,从而优化数据分析的效率,并减少处理和存储无意义数据所需的资源。

近年来,深度学习的兴起为喷注标记带来了新的发展。在深度学习的应用中,首要的关键步骤是确定数据的表征方式。喷注具有复杂的内部结构,其辐射模式信息没有唯一的编码范式^[5]。受到深度学习在计算机视觉^[6]、自然语言处理^[7]和三维点云^[8]等领域应用的启发,研究者以不同的视角(图像、序列、集合)表示喷注,并设计了针对不同表示的神经网络模型。

此外,喷注结构满足物理学中的一些基本对称性,例如洛伦兹对称性^[9]。为了将洛伦兹对称性引入神经网络,研究者对几何深度学习模型和等变性神经网络开展了大量研究。其核心思想是通过数学中的对称性约束,使神经网络能够自动适应输入数据的几何结构,从而在保持模型表达能力的同时显著提升对结构化数据的泛化能力。

基础模型(FM)是一类新型的深度学习模型,它们通过广泛的数据集和多样化的问题进行训练,从而能够泛化到多种下游任务与数据集。典型代表包括大语言模型(LLM),如 GPT-3^[10]和 BERT^[11]等。近年来,为了更好地推动喷注物理的全面发展,研究者对构建喷注基础模型进行了大量尝试。

在此背景下,撰写深度学习模型在高能物理喷注标记领域的综述具有迫切必要性与重要学术价值:一方面,需整合近十年的关键研究成果,厘清不同深度学习模型的结构与性能边界,为领域研究者提供清晰的技术图谱;另一方面,需重点分析当前研究所面临的挑战,如模型存在的不足,神经网络与物理理论的适配性,进而为未来研究方向提供指引,推

动深度学习技术更高效地服务于高能物理的前沿探索。

本文首先介绍喷注标记领域使用的数据集,接着阐述不同喷注表示、等变性神经网络和基础模型在喷注标记发展中的历程,并对相关模型进行分析与总结。根据喷注标记发展的现状,分析相关模型的优势与局限性,并展望未来深度学习在喷注标记领域的研究方向。

1 基准数据集分析

本章对喷注标记领域的三大基准数据集进行总结与分析。当前国际上最有代表性的高能物理数据集可以通过欧洲核子研究中心(CERN)开放数据门户进行访问。这些数据集包括有关碰撞、粒子以及喷注的信息。在喷注标记领域,主要的数据集来源于仿真工具,如 PYTHIA8^[12]、DELPHES^[13]和 MAD-GGRAPH^[14]。这些数据集包括 Top Tagging(简称为 TOP)数据集^[15]、Quark-gluon Tagging(简称为 QG)数据集^[16]和 JETCLASS 数据集^[17]。

TOP 数据集包括 120 万个训练样本、40 万个验证样本、40 万个测试样本。数据集中的每个样本对应于一个单独的喷注。该数据集中的喷注来源于高能顶夸克(信号喷注)或轻夸克和胶子(背景喷注)。事例由 PYTHIA8 蒙特卡罗模拟过程产生,并经过 DELPHES 软件包进行了 ATLAS 探测器模拟。该数据集中喷注横动量 $p_T \in [550 \text{ GeV}, 650 \text{ GeV}]$,伪快度 $\eta \in (-2, 2)$ 。单个喷注最多含有 200 个粒子。

QG 数据集包括 160 万个训练样本、20 万个验证样本、20 万个测试样本。该数据集是通过 PYTHIA8 蒙特卡罗模拟生成轻夸克(信号喷注)和胶子(背景喷注)创建的。数据集中喷注横动量 $p_T \in [500 \text{ GeV}, 550 \text{ GeV}]$,伪快度 $\eta \in (-2, 2)$ 。事例没有经过探测器模拟。

JETCLASS 数据集包括 1 亿个训练样本、500 万个验证样本、200 万个测试样本。数据集一共有 10 种喷注(包含 9 种信号喷注和背景胶子喷注),均分上述 3 个样本集。首先该数据集中的喷注是用 LHC 实验中使用的标准蒙特卡罗事件发生器模拟的,顶夸克和 W、Z、H 玻色子的产生和衰变用 MAD-GGRAPH 生成。然后使用 PYTHIA8 对产生的粒子进行演化,即进行部分子淋浴和强子化,并产生最终的末态粒子。最后使用 DELPHES 模拟了探测器效应。该数据集中的喷注横动量 $p_T \in [500 \text{ GeV}, 1\,000 \text{ GeV}]$,伪快度 $\eta \in (-2, 2)$ 。

上述 3 个基准数据集包含的物理属性如表 1 所示,其中,“√”表示包含该属性,“×”表示不包含该属性。

表 1 基准数据集包含的物理属性对比
Table 1 Comparison of physical properties included in benchmark datasets

物理属性	JETCLASS	TOP	QG
粒子 p_x 动量	√	√	√
粒子 p_y 动量	√	√	√
粒子 p_z 动量	√	√	√
粒子能量	√	√	√
粒子方位角	√	√	√
粒子伪快度	√	√	√
粒子标识符	×	×	√
粒子为中性强子	√	×	√
粒子为光子	√	×	√
粒子为带电强子	√	×	√
粒子为电子	√	×	√
粒子为缪子	√	×	√
粒子电荷数	√	×	√
喷注横动量	√	√	√
喷注方位角	√	√	√
喷注伪快度	√	√	√
喷注能量	√	√	√
喷注质量	√	√	√
喷注含有粒子数	√	√	√
横向冲击参数的双曲正切	√	×	×
横向冲击参数误差	√	×	×
纵向冲击参数的双曲正切	√	×	×
纵向冲击参数误差	√	×	×

表 2 展示了数据集中包含的喷注类型以及数量,其中“—”表示原文献无该指标值,下同。除了上述数据集以外,还包括 CMS (Compact Muon Solenoid) 公开数据集^[18]、ATLAS (A Toroidal LHC Apparatus) 顶夸克标记数据集^[19]、JetNet 数据集^[20]、OmniFold 数据集^[21]以及 LHC R&D 数据集^[22]等。由于这些数据集在格式、喷注类型和物理量的统一性方面存在问题,因此主流模型未采用它们。因此,构建更为标准、完善的喷注标记数据集是值得进一步研究的方向。

2 喷注标记模型的不同喷注表示方法

机器学习技术在喷注标记中的效率和有效性很大程度上取决于喷注的表示方式。本章将回顾 3 种主流的喷注表示方法:基于图像的表示,基于序列的表示和基于集合的表示,并分析与不同表示方法相关的模型。

表 2 基准数据集包含的喷注类型和数量对比
Table 2 Comparison of Jet types and quantities included in benchmark datasets

喷注类别	JETCLASS	TOP	QG
$H \rightarrow b\bar{b}$	10	—	—
$H \rightarrow c\bar{c}$	10	—	—
$H \rightarrow gg$	10	—	—
$H \rightarrow 4q$	10	—	—
$H \rightarrow l\nu qq'$	10	—	—
$Z \rightarrow qq$	10	—	—
$W \rightarrow qq'$	10	—	—
$t \rightarrow bqq'$	10	1	—
$t \rightarrow bl\nu$	10	—	—
uds/g	10	1	2

2.1 基于图像的表示

如果将探测器看成相机,那么喷注可以看成像素化后的二维灰度图像,如图 1 所示。其中像素点位置使用探测器单元标准位置参数 $\eta-\phi$ (伪快度与方位角之差)作为参考系,像素强度则表示在特定位置沉积能量的所有粒子的能量(或横向动量)。

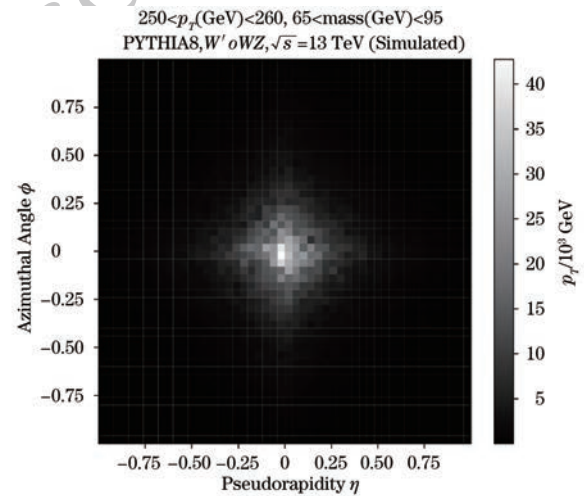


图 1 喷注图像表示示意图

Fig.1 Schematic diagram of Jet image representation

基于这一表示方法,DE OLIVEIRA 等^[23]于 2017 年首次采用卷积神经网络(CNN)对喷注图像进行分类,以区分 W 玻色子喷注与背景胶子喷注。实验结果表明,卷积神经网络的分类性能优于 3 种基于 QCD 理论的传统方法。受到 RGB 图像的启发,KOMISKE 等^[24]将喷注图像拓展为包含带电粒子的横向动量、中性粒子的横向动量和带电粒子多重数 3 个关键信息通道,从而丰富了喷注图像的特征信息。实验是在基于 PYTHIA 和 HERWIG 软件模拟生成的夸克/胶子喷注数据上进行的。实验结果显示,卷积神经网络的分类效率优于传统的物理驱动变量方法和增强决策树。在顶夸克标记任务

上, KASIECZKA 等^[25]使用 Delphes3 快速探测器模拟生成的喷注图像作为输入, 构建了卷积神经网络进行训练, 预测结果达到了与基于 QCD 的顶夸克标记器相当的性能。

综上所述, 在喷注标记任务上, 卷积神经网络的性能相比于基于 QCD 理论的传统多变量方法取得了显著的改进。然而, 将喷注表示为图像有 2 个主要的缺点: 1) 喷注作为图像会导致表示稀疏性问题, 因为典型的喷注具有 $O(10)$ 到 $O(100)$ 数量级的粒子, 而一个喷注图像通常需要 32×32 个像素点才能完整表示喷注, 这可能导致大量像素点为空白, 从而影响卷积神经网络的计算效率; 2) 将喷注作为图像的前提是要求探测器的探测单元均匀分布且大小相同, 从而避免喷注像素化时信息丢失, 然而实际情况无法满足这一假设。尽管存在上述缺点, 但将喷注视为图像并使用当时最先进的卷积神经网络进行处理无疑是一次伟大的尝试, 这将激励更多研究者将喷注标记与深度学习方法相结合。

2.2 基于序列的表示

受到自然语言处理的启发, 研究者开始将喷注表示为其组成粒子的有序序列。这种表示方式不仅允许将更多粒子特征作为输入, 还符合不同喷注通常包含不同数量粒子的特点, 有效解决了图像表示稀疏的问题。如图 2 所示, 喷注粒子按照距离碰撞原点的最近距离进行排序, 并且使用循环神经网络 (RNN) 进行特征提取。具体而言, 基于序列表示的喷注标记任务通常有以下两种解决方案。

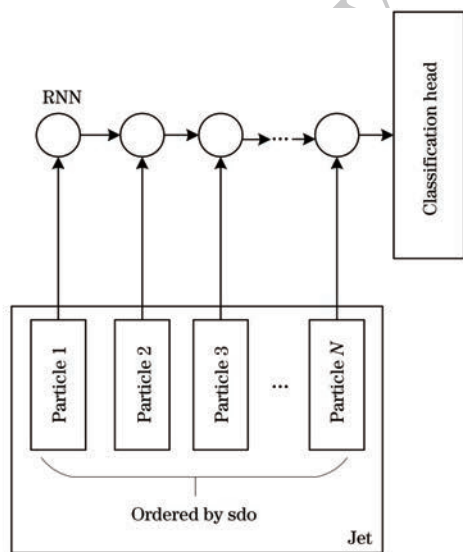


图 2 喷注序列表示示意图

Fig.2 Schematic diagram of Jet sequence representation

1) 将构成喷注的粒子按特定方式排序, 然后利用 RNN 及其变体长短时记忆 (LSTM) 网络和门控循环单元 (GRU) 进行处理。2016 年, GUEST 等^[26]

通过对粒子在径向方向上距离碰撞原点的最近距离进行排序, 首次使用 LSTM 来分类轻味夸克喷注和重味夸克喷注。实验结果表明, 仅使用粒子轨迹和顶点信息的网络在分辨准确率上优于使用专家特征的网络。2020 年, BOLS 等^[27]设计了 DeepJet 模型, 该模型同时考虑了所有喷注组成粒子的完整特征信息, 包括次级顶点 (按飞行距离排序)、喷注物理属性、带电粒子 (按冲击参数排序) 和中性粒子 (按到次级顶点的最短角距离排序)。RNN 层是 DeepJet 模型的重要组成部分, 因为它允许模型捕捉带电粒子轨迹中的顺序信息, 从而利用此信息提高分类性能。DeepJet 模型发布时在喷注标记和夸克/胶子标记中取得了当年的最佳效果 (SOTA)。

2) 将喷注表示为二叉树。由于喷注是由组成粒子根据顺序喷注算法重建而成的, 聚类历史自然形成了一棵二叉树。在这棵树中, 叶子节点由粒子的四维动量表示, 根节点则使用多维向量表示喷注。LOUPPE 等^[28]认为粒子通过喷注重建算法组合成了喷注, 类似于单词通过语法组成句子, 从而创建了上述喷注表示方法。由于根节点可以通过叶子节点递归计算获得, 这种网络被称为递归神经网络 (RecNN), 其中可训练的参数是二叉树的边。

然而, 基于序列的表示需要人为地对粒子进行排序, 这种排序可以是显式的 (第一种表示方法) 或者隐式的 (第二种表示方法)。然而, 由于量子力学的基本原理, 无法确定与喷注内部最终状态强子相关联的唯一顺序或历史。此外, 序列化模型难以进行并行计算, 从而极大地影响了模型的训练效率, 这促使研究者寻求新的表示方式。

2.3 基于集合的表示

一个比粒子序列或树结构更自然的表示方式是无序、置换不变的粒子集合。该表示方式具备其他喷注表示的所有优点, 特别是能够灵活地为每个粒子包含任意特征。研究者将这种喷注表示称为粒子云, 类似于计算机视觉中的三维点云。粒子云能够有效地表达构成粒子之间的相互关联信息以及喷注内部丰富的几何结构。图 3 展示了一个喷注中所含粒子的 $\eta-\phi$ 空间分布示意图 (彩色效果见《计算机工程》官网 HTML 版, 下同), 每个“点”都具备高维的物理属性, 研究者通常将这些“点”作为矩阵整体表示喷注的输入。

基于粒子云的喷注表示方法, KOMISKE 等^[16]于 2019 年提出了基于粒子身份识别的粒子流网络 (PFN-ID) 模型。PFN-ID 模型基于 Deep Sets 框架, 包含完整的粒子类别信息。Deep Sets

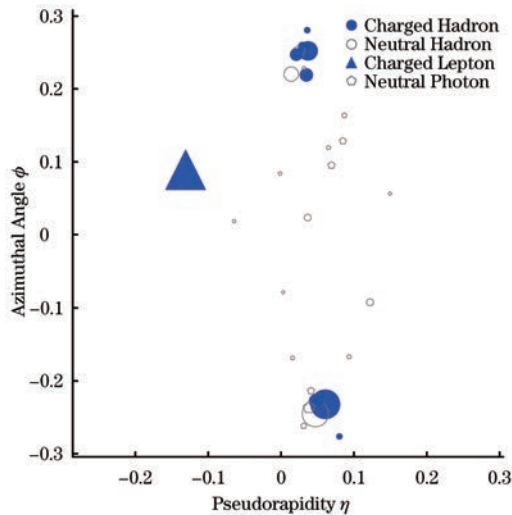


图 3 喷注集合表示示意图

Fig.3 Schematic diagram of Jet set representation

框架是一种机器学习框架,允许直接从集合中进行学习^[29]。Deep Sets 框架首先使用共享函数将集合中的每个元素映射到隐空间;然后通过对称函数聚合这些隐空间的表示;最后使用共享函数将聚合后的隐空间表示映射回输出空间。对于喷注标记任务,该过程首先将每个粒子映射到隐空间并进行收集;随后可观测量被表示为该隐空间上的函数;最后将可观测量输入分类网络以对喷注进行分类。

2020 年,QU 等^[30]提出了 ParticleNet 模型。该模型使用动态图卷积神经网络(DGCNN)^[31]处理构成喷注的无序粒子集合。其架构由 3 个边卷积块(EdgeConv)组成,第一个边卷积块使用粒子在伪快度-方位角空间中的二维坐标来计算距离,而后续的边卷积块则使用学习到的特征向量作为坐标。每个边卷积块具有不同数量的通道。EdgeConv 块结构如图 4 所示。边卷积块首先将点云表示为图,其中顶点为粒子本身,边则是每个顶点与其 k 个最近邻(k -NN)点之间的连接;随后边卷积块根据“坐标(coordinates)”输入计算距离,找出每个粒子的 k -NN 粒子(k -NN-index);然后根据 k -NN 粒子的索引,将中心点特征和邻居特征拼接成“边特征(Edge-features)”;接着使用具有可学习参数的“边函数”将每个粒子投影到潜在特征空间中,从而为每个粒子有效地分配新的“坐标”。“边函数”通过一个 3 层的多层感知器(MLP)实现,每一层包括线性变换(Linear)、批量归一化(BatchNorm)和修正线性单元(ReLU)激活。此外,受 ResNet 启发,每个块都集成了与 EdgeConv 操作并行的跳跃连接,以便于输入特征的直接传递。EdgeConv 块由两个超参数

定义:邻居数量(K)和通道数量(C),分别表示考虑的邻居数量和线性变换中的隐层神经元数量。这些 EdgeConv 块在学习粒子云的局部特征并将其聚合成喷注的全局特征向量中发挥着关键作用。经过 EdgeConv 块后,采用全局平均池化(Pooling)对粒子特征进行汇总,然后通过一个包含 256 个神经元的全连接层,经过 ReLU 激活和 Dropout 层处理,最终输出 Softmax 层用于分类。ParticleNet 模型结构如图 5 所示。该模型在 2020 年的顶夸克标记和夸克-胶子标记基准数据集测试中取得了 SOTA。

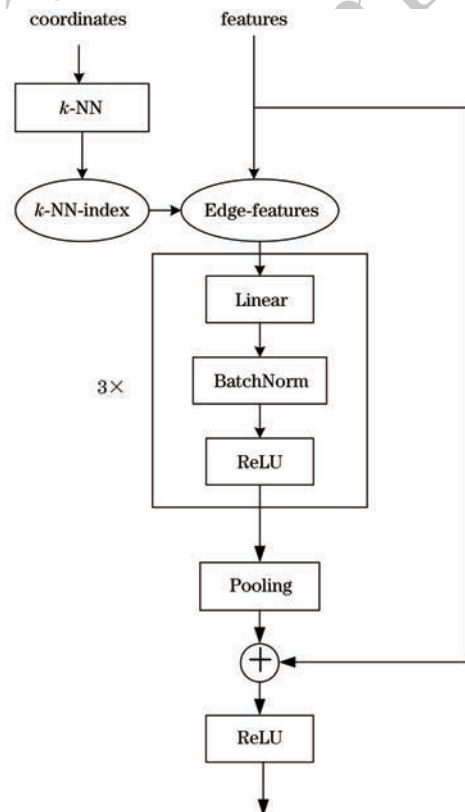


图 4 EdgeConv 块结构

Fig.4 Architecture of EdgeConv block

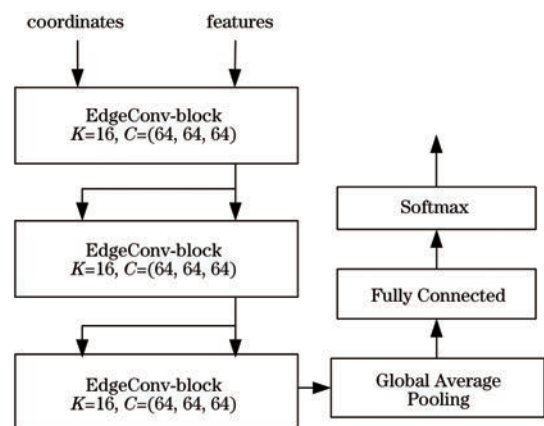


图 5 ParticleNet 模型结构

Fig.5 Architecture of ParticleNet model

2022 年, QU 等^[17] 引入了注意力机制和 Transformer 基础架构, 提出了 ParT (Particle Transformer) 模型, 并发布了亿级喷注分类数据集 JETCLASS。ParT 的创新之处在于使用了两组输入: 一是粒子特征输入, 包括与喷注中所有粒子相关的特征; 二是粒子相互作用输入 U , 编码喷注中每一对粒子的成对作用特征。粒子相互作用输入计算如下:

$$\begin{aligned}\Delta &= \sqrt{(\eta_a - \eta_b)^2 + (\phi_a - \phi_b)^2} \\ k_T &= \min(p_{T,a}, p_{T,b}) \Delta \\ z &= \min(p_{T,a}, p_{T,b}) / (p_{T,a} + p_{T,b}) \\ m^2 &= (E_a + E_b)^2 - \|\mathbf{p}_a + \mathbf{p}_b\|^2\end{aligned}\quad (1)$$

ParT 模型结构如图 6 所示。粒子特征输入 (Particles) 和粒子相互作用输入 (Interactions) 分别通过一个嵌入层 (Embedding) 投影到潜在嵌入空间中。粒子特征嵌入随后经过一系列多头注意力机制层 (Particle Attention Block)。在每个注意力层中, 使用相同的粒子相互作用嵌入作为权重来修改注意力分数, 计算方法如下:

$$S_{\text{PMHA}} = \text{Softmax}(\mathbf{QK}^T / \sqrt{d_k} + U) \mathbf{V} \quad (2)$$

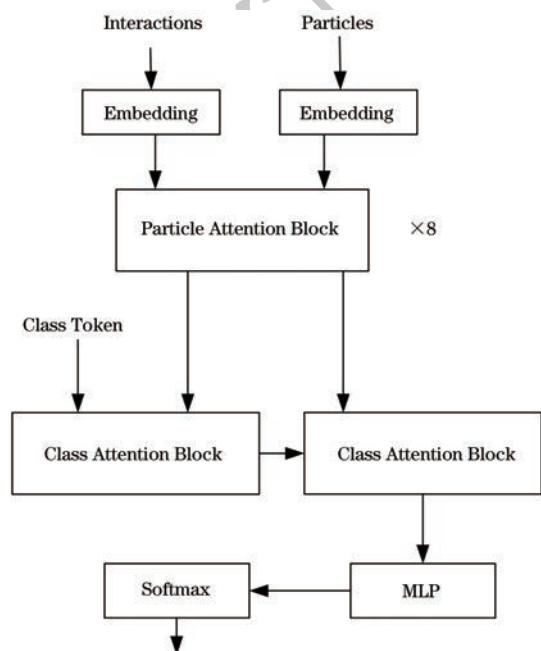


图 6 Particle Transformer 模型结构

Fig.6 Architecture of Particle Transformer model

这个机制对于将物理过程先验信息融入粒子表达中至关重要, 能够使网络利用喷注中粒子相互作用下的 QCD 原理。最后一个注意力层的输出连同全局类别标记 (Class Token) 一起输入到两个堆叠的类别注意力模块 (Class Attention Block), 该模块计算类别标记与所有粒子之间的注意力得分。这使得模型能够评估不同粒子相对于类别标记的重

要性, 并在为整个喷注做出分类决策时, 有选择地考虑来自各个粒子的信息。第二个类别注意力模块的输出随后传递给一个含有 10 个输出神经元的多层感知器, 以生成最终的分分类得分。在最终的分分类性能上, ParT 模型在 JETCLASS 上达到了 86.1%, 顶夸克标记数据集上达到了 94.4%, 夸克/胶子标记数据集上达到了 85.2%。

2025 年, WU 等^[32] 提出了增强交互粒子 Transformer (MIParT) 模型。该模型创新性地引入了增强交互注意力 (MIA) 机制, 通过嵌入更加精细的粒子相互作用, 提升了喷注标记的性能。该模型在处理 JETCLASS 数据集时, 将粒子相互作用嵌入维度扩大了一倍, 同时显著降低了模型复杂度。在顶夸克标记数据集和夸克/胶子标记数据集上的测试表明, MIParT 在保持与领先模型近乎相同的准确率和 AUC 指标的同时, 背景拒绝率分别提升了 25 和 3 个百分点。通过在 JETCLASS 数据集上进行微调后, 模型性能进一步提高了 39 和 6 个百分点。

基于深度学习的喷注标记算法已在 LHC 的实际数据分析中得到广泛应用。例如, CMS 合作组开发了 DeepAK8 模型, 用于标记真实探测器中由顶夸克、希格斯玻色子、W 玻色子或 Z 玻色子产生的喷注^[33]。此外, CMS 合作组利用 ParticleNet 模型, 在强子对撞机上首次观测到 Z 玻色子衰变为一对粲夸克的过程^[34]。显然, 喷注标记技术的进步对于真实物理数据的分析起着至关重要的作用。

虽然基于集合的表示方法在标记精度和扩展性方面优于基于图像和基于序列的表示方式, 但当前主流模型将粒子特征直接通过嵌入层进行映射。粒子特征包含离散变量 (粒子类型、电荷等) 和连续变量 (能量、动量以及粒子在空间中的位置信息)。对粒子进行更有效的 Token 化方式仍然值得进一步研究。

3 喷注标记领域的等变性神经网络应用

本章主要对喷注标记领域中应用的等变性神经网络进行分析综述。在数学中, 等变性是一种性质, 指的是一个函数在某种对称操作下的行为方式。具体来说, 定义群 G 及两个集合 X 和 Y , 定义群作用 $\Phi: G \times X \rightarrow X$ 和 $\Psi: G \times Y \rightarrow Y$, 定义函数 $f: X \rightarrow Y$ 。对于任意的 $g \in G$ 和 $x \in X$, $f(\Phi(g, x)) = \Psi(g, f(x))$ 成立, 则称函数 f 在群 G 上是等变的。换句话说, 进行群作用操作后再应用函数 f 的效果等同于先应用函数 f 再进行群作用操作。若进行群作用操作后再应用函数 f 的效果等同于直接进行函

数 f 操作, 即 $f(\Phi(g, x)) = f(x)$, 则函数 f 在群 G 上是不变的。近年来, 研究者通过引入洛伦兹对称性: 确保粒子四动量在隐空间中始终保持四维, 且其闵可夫斯基度量保持不变, 这使得输入数据的四动量进行任意洛伦兹变换都不会对喷注标记分类结果造成影响。因此, 洛伦兹等变性神经网络结构减少了对特定参考系模拟数据的依赖, 提升了模型从模拟到真实探测器数据的泛化性。

2020 年, BOGATSKIY 等^[35] 提出了洛伦兹群神经网络 (LGN)。LGN 模型包含多个层级结构, 即线性输入层、迭代的 CG (Clebsch-Gordan) 张量积层以及多层感知器层。初始线性层处理来自喷注中的粒子的四动量以及标签、电荷、自旋等特征。迭代的 CG 张量积层通过洛伦兹群表示的张量积进行 CG 分解, 实现等变的非线性变换。CG 层与仅作用于洛伦兹不变量的 MLP 层交替排列。输出层计算激活值的算术和, 最终通过全连接线性层生成用于二元分类的两个标量权重。实验表明, 在参数量减少 10~1 000 倍的情况下, LGN 模型仍能保持与当前最优模型相当的分类性能。然而, 由于在隐藏层上的连续张量积操作, LGN 模型对高阶张量进行显式计算时, 计算速度缓慢。

为了解决上述问题, GONG 等^[36] 于 2022 年提出了新的保持洛伦兹对称性的深度学习模型——LorentzNet。该模型由洛伦兹群等变模块 (LGEB) 堆叠而成。LGEB 的结构包含多个关键层级, 即闵可夫斯基范数与内积层、求和池化层以及多层感知器。该模块以一组四动量向量 (E, p_x, p_y, p_z) 和洛伦兹不变标量 (质量, 电荷, 粒子 ID) 作为输入, 通过闵可夫斯基范数与内积层的变换, 提取出洛伦兹不变的几何特征量, 计算公式如下:

$$m_{ij}^l = f_e(h_i^l, h_j^l, \psi(\|x_i^l - x_j^l\|^2), \psi(\langle x_i^l, x_j^l \rangle)) \quad (3)$$

式中: $\psi(\|x_i^l - x_j^l\|^2)$ 为闵可夫斯基距离; $\psi(\langle x_i^l, x_j^l \rangle)$ 为闵可夫斯基内积。

求和池化层对这些几何特征进行聚合, 生成输入的标量表征。随后, 多层感知器层学习从标量表征到新特征空间的非线性映射实现四动量和标量更新。

$$\begin{aligned} x_i^{l+1} &= x_i^l + c \sum_{j \in [N]} f_x(m_{ij}^l) x_j^l \\ h_i^{l+1} &= h_i^l + f_h\left(h_i^l, \sum_{j \in [N]} w_{ij} m_{ij}^l\right) \\ w_{ij} &= f_m(m_{ij}^l) \end{aligned} \quad (4)$$

式中: f_e, f_x, f_h, f_m 为需要参数学习的多层感知器。

LGEB 块结构如图 7 所示。实验在顶夸克标记和夸克/胶子标记基准数据集上测试, 取得了较优的性能, 同时验证了当输入的四动量进行洛伦兹变换后, 分类结果保持不变的特性。

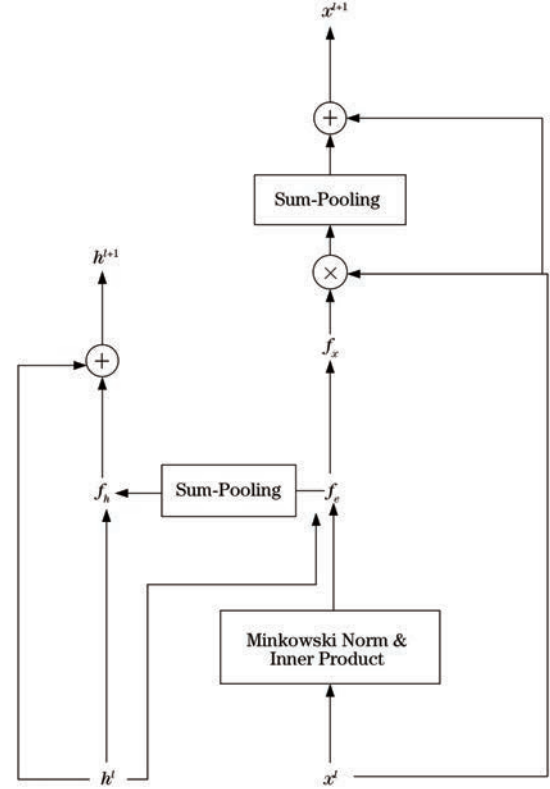


图 7 LGEB 块结构

Fig. 7 Architecture of LGEB block

2023 年, BRANDSTETTER 等^[37] 提出了基于克利福德代数的等变图神经网络 (CGENN)。这一创新框架通过克利福德代数构建在 $O(n)$ 群和 $E(n)$ 群上具备等变性的网络。CGENN 充分利用克利福德代数的几何特性 (如几何积) 来参数化等变神经网络层。这些网络层操作于多向量——一种包含标量、向量及高维几何特征的结构, 从而实现对称性感知计算。在喷注标记任务中, 该模型将粒子数据嵌入多向量空间, 以多向量的形式表示喷注。

2024 年, BOGATSKIY 等^[38] 提出了 PELICAN 模型。该模型首先将喷注粒子的四动量输入作为点云进行处理, 随后采用迹运算和求和聚合函数进行特征降维, 接着将二阶输入张量 (喷注内粒子的四动量矢量的两两点积) 规约为排列不变的标量, 最后通过全连接层和交叉熵损失函数进行优化。

同样受几何代数的启发, BREHMER 等^[39] 于 2024 年提出了洛伦兹几何代数 Transformer (L-GATr)。该模型基于几何代数 Transformer 结构^[40], 并创新性地将洛伦兹对称性与注意力机制相

结合,能够处理四维时空粒子物理数据。该模型内置专为闵可夫斯基空间设计的洛伦兹等变注意力算法,以及可处理由粒子类型和四动量向量参数化的粒子数据的归一化层。实验表明,L-GATr 的性能达到或超越同类的洛伦兹等变网络。

4 喷注标记领域基础模型的相关研究

本章综述了喷注标记领域的基础模型并总结了基础模型的预训练策略。LLM 展现了通过海量文本预训练学习通用语言表征的卓越能力。类似地,高能物理中的基础模型能够学习编码粒子相互作用、探测器响应和物理定律的“语言”,为实验数据分析和解释提供通用工具。LLM 的成功表明,基础模型可能彻底变革高能物理中数据驱动的发现方式,为探索新物理提供可扩展的方法。LHC 已开始收集海量数据以探测新物理过程。随着数据量的增加,对撞机上部署模型的训练时间将显著延长。由于基础模型通过海量数据预训练,并且在下游任务中通过少量微调表现优异,因此它们有望为高能物理领域节省大规模计算资源。此外,计算资源有限的研究者也能够从高能物理专用基础模型中受益。

2024 年,BIRK 等^[41]提出了 OmniJet- α 模型,首次实现了无监督任务(喷注生成)与经典监督任务(喷注标记)之间的迁移学习。该模型首先使用 JETCLASS 数据集单独训练一个矢量量化变分自编码器^[42](VQ-VAE),将喷注编码为 N 个标记(Token)的集合;然后模型以无监督方式在 JETCLASS 数据集中利用 VQ-VAE 解码器训练自回归 Jet 生成模块(该模块由带有下一标记预测头的 Transformer 主干网络实现);最后模型通过 VQ-VAE 编码器以及将“下一标记预测头”替换为“分类头”的 Transformer 主干网络进行监督训练,验证了生成喷注模型中所编码的信息也能迁移到喷注分类任务中。

2024 年,GOLLING 等^[43]同样提出利用 VQ-VAE 对喷注进行表征学习。不同的是,该团队引入了“掩码机制”,并构建了掩码粒子模型(MPM)。在该模型中,部分粒子会被替换为可学习向量,随后输入 Transformer 主干网络。训练目标是预测被遮蔽粒子的离散标记身份(该身份由预训练 VQ-VAE 的编码器定义)。实验利用 JETCLASS 数据集对 Transformer 主干网络进行有监督微调,并进行分类准确率测试。实验结果表明,与完全从头训练的监督式模型相比,采用预训练的方法能带来显著的

性能提升。

受 OmniJet- α 的启发,MIKUNI 等^[44]于 2025 年提出了一种名为 OmniLearn 的基础模型,旨在同时推进喷注物理领域的多项任务。OmniLearn 的核心在于学习喷注的共享表示(shared representation),该表示随后被输入至特定任务的神经网络。这种架构设计不仅提升了模型的灵活性和效率,还通过仅加载共享表示和相关任务网络,显著减少了下游任务应用的整体模型规模。OmniLearn 的主要构建模块包括点边变换器(PET)、分类器头(Classifier Head)和生成器头(Generator Head)。PET 作为网络的共享表示部分,接收喷注内的粒子信息以及扩散时间参数作为输入。PET 通过时间嵌入层将时间信息编码至更高维度空间,并与输入粒子信息相结合。此外,PET 利用动态图卷积层创建局部编码,以增强粒子间的几何信息。分类器头用于不同喷注类型的分类任务,其输入为 PET 的输出以及喷注的运动学信息。生成器头则用于生成任务,输入包括粒子嵌入、喷注运动学信息、时间信息以及类别标签。OmniLearn 的损失函数综合了分类和生成任务的多个损失项,涵盖交叉熵损失以及生成任务的均方误差损失。OmniLearn 在跨喷注类型和跨探测器的泛化能力方面表现出色。在顶夸克数据集和夸克/胶子数据集上,OmniLearn 的性能显著优于从头开始训练的模型,同时与经过微调的 PartT 模型相当,甚至有所超越。此外,在 CMS 公开数据集上,OmniLearn 在分类性能上优于从头开始训练的分类器,并且训练速度提高了 2 倍。

近年来,LLM 在通用任务中展现出强大能力,但在大规模数值数据分析驱动的科学领域(如高能物理实验)中面临显著瓶颈。主要原因是 LLM 处理数值数据的固有缺陷以及任务特定模型的局限性。为了解决上述问题,WU 等^[45]提出了面向粒子碰撞数据分析的 LLM 架构 BBT-Neutron。其核心创新是二进制分词技术,可将文本、数值、图像等多模态数据统一转换为字节序列,解决了主流 LLM 采用的 BPE(Byte Pair Encoding)分词处理数值数据时存在的碎片化、上下文依赖及定量关系失真问题。该架构基于 Transformer 解码器框架,通过 Patch 嵌入、旋转位置编码、Patch-Attention 等组件适配科学数据特性,在喷注标记任务中,BBT-Neutron 与 ParticleNet、Particle Transformer 等任务特定模型性能相当,且展现出独特的数据驱动“涌现能力”,为高能物理等大科学领域提供基础

模型范式。

为了探究模拟数据到探测器真实数据的迁移问题,BAALOUCH 等^[46]提出采用领域对抗神经网络(DANN)实现模拟到真实的领域自适应模型。研究以 LHCb 实验的公开真实探测器数据集(含 67 533 个样本,62%为信号事件)为训练集,测试结果显示 DANN 在模拟数据上准确率为 84.0%,并且模拟与真实数据的柯尔莫哥洛夫-斯米尔诺夫(K-S)距离为 0.06(低于 0.09,满足性能守恒标准),这证明了 DANN 模型成功消除了模拟/真实数

据分布偏差。2025 年,AMRAM 等^[47]推出了包含约 1.8 亿个喷注的 AOJ(AspenOpenJets)数据集,其源自 CMS 2016 年公开的真实探测器数据。研究使用 AOJ 对 OmniJet- α 基础模型进行预训练,并在 JETCLASS 数据集上微调以完成生成顶夸克喷注和 QCD 喷注,结果显示预训练模型在训练数据量少于 100 万喷注时,生成性能(通过 KL 散度和 Wasserstein-1 距离衡量)显著优于从零训练的模型,如图 8 所示,且性能指标在非饱和阶段遵循 $D^{-\gamma}$ ($0.2 < \gamma < 0.45$) 幂律缩放。

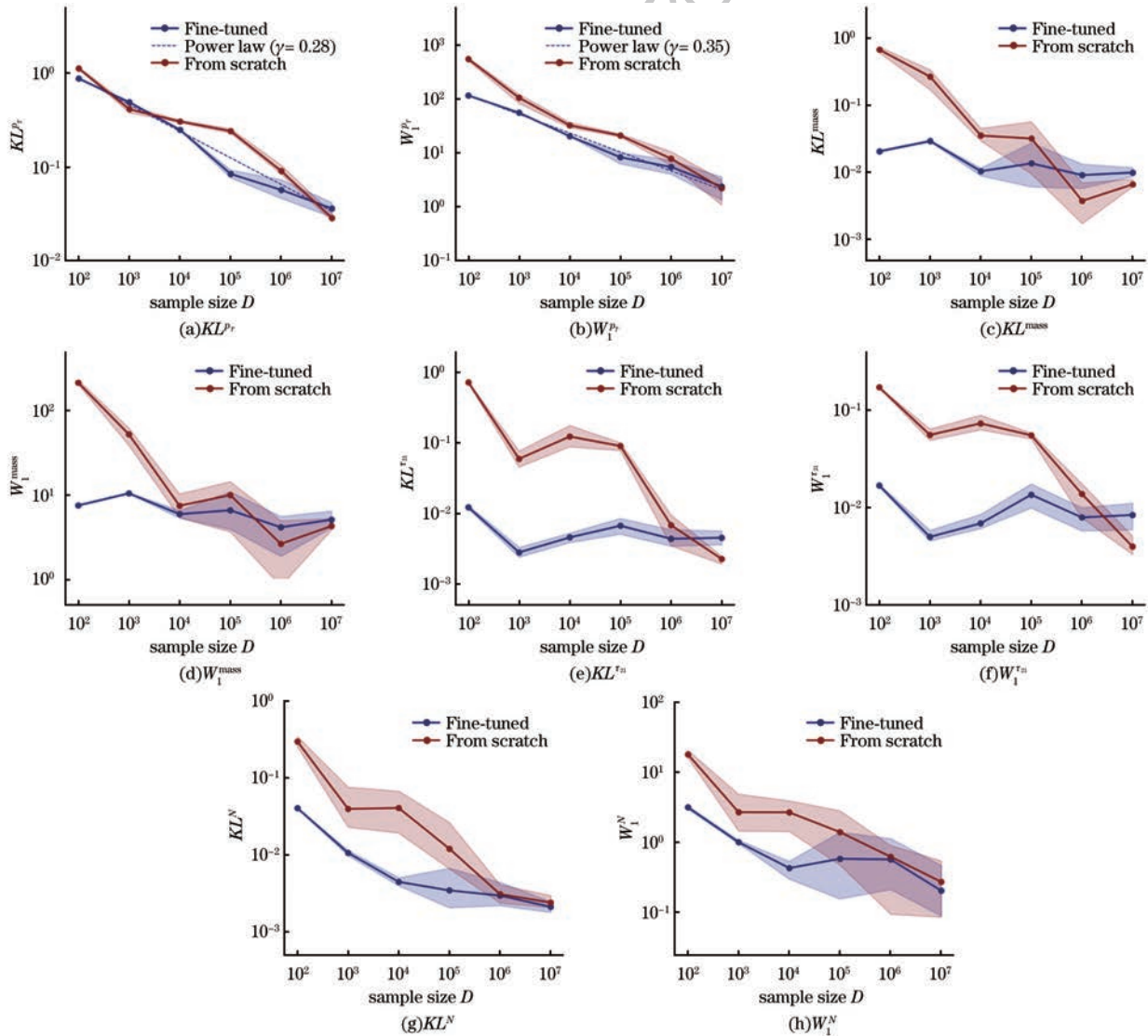


图 8 基于 JETCLASS 数据集 QCD 喷注训练的模型在不同训练样本量 D 下的生成质量对比^[47]

Fig.8 Comparison of the generation quality of models trained on QCD Jets from JETCLASS dataset for different training sample sizes D ^[47]

表 3 总结了基础模型的预训练策略比较以及优缺点分析。但是,与基线模型相比,基础模型的预训练成本显著提高。以在 JETCLASS 数据集上训练 5 个 Epoch 为例,在单张 H100 GPU 上训练 ParticleNet、Particle Transformer 和 L-GATr 需要

的时间分别为 27、38 和 180 h。OmniLearn 的预训练则需要在 128 个 GPU 上训练 200 个 Epoch^[44]。因此,训练成本低、精度高的轻量化模型更加适合单一的喷注标记领域。表 4 总结了喷注标记领域轻量化模型设计的相关思路。

表 3 基础模型预训练策略比较

Table 3 Comparison of pre-trained strategies for FMs

策略类型	核心思想	代表模型	优点	缺点
自监督学习	利用无标签数据通过掩码预测等任务学习通用特征表示,减少对人工标注数据的依赖	MPM	有效利用海量无标签数据,降低标注成本;学习到的特征具有通用性,可迁移至多种下游任务	对掩码策略敏感,需设计适合喷注物理特性的掩码方法
迁移学习	将在源任务(如通用喷注标记或者喷注生成任务)上预训练的模型参数迁移至目标任务(如特定喷注标记),通过微调适应新任务	OmniJet- α	显著减少目标任务的标注需求和训练时间;适用于跨实验的模型迁移	源任务与目标任务需具有相似的数据分布,否则迁移效果有限
多任务学习	同时训练模型完成多个相关任务(如喷注标记、喷注生成),共享底层特征提取层,增强特征的泛化能力	OmniLearn	通过任务间互补性提升模型鲁棒性;共享参数减少计算资源消耗	任务间可能存在负迁移,需设计合理的任务权重与损失函数
多模态预训练	融合数值数据(粒子四动量)、文本(物理公式)和图像(量能器信号)进行预训练,提升模型对复杂数据的理解能力	BBT-Neutron	充分利用多模态数据,提升模型泛化能力;适用于跨探测器数据分析	数据对齐和融合难度大,需设计高效的多模态特征融合方法

表 4 轻量化模型设计思路

Table 4 Lightweight model design ideas

设计策略	核心思想	代表模型
注意力优化	改进注意力计算方式,更有效地捕捉喷注中的关键信息,提升喷注标记任务的性能,同时兼顾计算效率和模型复杂度	ESMAIL 等 ^[48] 提出一种高效 Transformer 架构,通过动态稀疏注意力和可学习交互矩阵,在参数仅为 ParT 的 1/10 的同时顶夸克喷注标记任务 AUC 达到 SOTA 水平; MURNANE ^[49] 提出 GravNetNorm 架构,通过在学习嵌入空间中构建几何约束注意力机制,仅依据距离定义注意力权重,并动态学习每个节点的邻域大小,在顶夸克喷注标记任务推理测试中,内存占用仅为 ParticleNet 的 7.4%,耗时为 25%
物理对称融入	利用洛伦兹对称性提升模型物理一致性与泛化性	FAVARO 等 ^[50] 提出洛伦兹局部规范化(LLoCa)框架,可在计算开销最小的情况下为任意神经网络实现精确的洛伦兹等变性,核心是为每个粒子预测局部参考系,并在系间传递任意阶量信息,模型参数匹配的情况下,训练速度比 L-GATr 快 6 倍
硬件适配与推理加速	适配高能物理实验的边缘计算设备(如 FPGA),降低实时推理延迟	ODAGIU 等 ^[51] 实现并优化了 3 种机器学习模型——MLP、Deep Sets 和图神经网络(GNN);通过 8 bit 量化感知训练和 32 个粒子场景下的剪枝,3 种模型均可在 FPGA 上实现约 100 ns 量级的延迟,且精度保留率大于 0.9
知识蒸馏	以高精度大模型为“教师”,向轻量化“学生”模型传递喷注分类知识	LIU 等 ^[52] 用 LorentzNet 模型(教师)蒸馏轻量化 MLP 模型(学生),学生模型计算量降低 60%,同时背景抑制率提升约 2 倍

总体而言,目前喷注物理领域的基础模型仍处于探索阶段。有关模型的参数量、训练成本、使用的数据集规模、喷注输入特征的选择,以及主干网络设计等关键问题,仍亟需进一步研究和验证。

5 模型性能对比

分析不同网络模型在高能物理喷注标记基准数据集上的分类效率及其网络规模,对于模型的改进至关重要。本章分别分析了相关模型在顶夸克标记数据集、夸克/胶子标记数据集和 JETCLASS 数据集上的性能表现以及模型复杂度、优势和局限性。

表 5 列出了相关模型在顶夸克标记数据集上的性能。其中:准确率是模型正确预测的比例,包括真正例和真负例的识别;AUC 是 ROC 曲线下的面积;Rej x 是在特定信号效率下的背景拒绝率,即

特定召回率(TPR)下误识别率(FPR)的倒数(以上指标在表 6、表 7、表 8 中同理)。分析得出,PELICAN 模型在准确率上达到了最优。在 AUC 指标上,PELICAN 和 L-GATr 模型都达到了最优。在顶夸克信号效率为 50%时,L-GATr 模型误识别率最低。在顶夸克信号效率为 30%时,PELICAN 模型误识别率最低。上述结论表明,等变性网络结构在提高顶夸克标记效率方面发挥了重要作用。

表 6 列出了相关模型在夸克/胶子标记数据集上的性能。PELICAN 模型在准确率、AUC、Rej50%、Rej30%上都取得了最优效果。但是,与在顶夸克标记数据集上的性能对比,准确率落后约 9 个百分点,AUC 落后约 6 个百分点,Rej30%落后约 1 个数量级。这说明当前模型在区分轻夸克和胶子上仍存在巨大进步空间。

表 5 相关模型在顶夸克标记数据集上的性能比较

Table 5 Performance comparison of relevant models on the TOP dataset

模型	准确率	AUC	Rej50%	Rej30%
PFN-ID	0.932	0.981 9	247±3	888±17
ParticleNet	0.940	0.985 8	397±7	1 615±93
ParT	0.940	0.985 8	413±16	1 602±81
MIParT	0.942	0.986 8	505±8	2 010±97
LGN	0.929	0.964 0	—	435±95
LorentzNet	0.942	0.986 8	498±18	2 195±173
PELICAN	0.943	0.987 0	—	2 250±75
CGENN	0.942	0.986 9	500	2 172
L-GATr	0.942	0.987 0	540±20	2 240±70

表 7 列出了相关模型在 JETCLASS 数据集上的性能。L-GATr 模型在准确率、AUC、9 种信号喷

表 7 相关模型在 JETCLASS 数据集上的性能比较

Table 7 Performance comparison of relevant models on the JETCLASS dataset

模型	准确率	AUC	$H \rightarrow b\bar{b}$	$H \rightarrow c\bar{c}$	$H \rightarrow g\bar{g}$	$H \rightarrow 4q$	$H \rightarrow l\nu qq'$	$t \rightarrow bqq'$	$t \rightarrow bl\nu$	$W \rightarrow qq'$	$Z \rightarrow qq$
			Rej50%	Rej50%	Rej50%	Rej50%	Rej99%	Rej50%	Rej99.5%	Rej50%	Rej50%
PFN-ID	0.772	0.971 4	292 4	841	75	198	265	797	721	189	159
ParticleNet	0.844	0.984 9	763 4	2 475	104	954	3 339	10 526	11 173	347	283
ParT	0.861	0.987 7	106 38	4 149	123	1 864	5 479	32 787	15 873	543	402
MIParT	0.861	0.987 8	107 53	4 202	123	1 927	5 450	31 250	16 807	542	402
L-GATr	0.865	0.988 4	121 95	4 819	128	2 304	5 764	37 736	19 231	580	427

表 8 给出了相关模型的模型复杂度。其中：Params 是模型中所有可学习参数总量；MFLOPs 表示每秒 100 万次浮点运算次数。结合表 5、表 6 可以看出，具有等变性网络结构的 PELICAN 模型同时具备在顶夸克标记数据集和夸克/胶子标记数据集上分类效率高、模型参数少的优点。结合表 7 可以看出，在维持和前代模型 ParT 相近的性能下，MIParT 模型参数量减少了 66%，计算需求降低了 47%。

根据模型具备的结构特点和在不同数据集上的性能表现，表 9 总结了相关模型的优势和局限性。

表 9 相关模型的优势和局限性比较

Table 9 Comparison of advantages and limitations of relevant models

模型	优势	局限性
PFN-ID	首次以集合的视角表示喷注	在大规模数据集上性能表现不佳
ParticleNet	使用了较为先进的 EdgeConv 提取特征	粒子之间的交互半径仅由超参数 k 决定，模型可解释性较差
ParT	使用了全局自注意力机制，扩大了粒子的感受野	参数量巨幅增加，但模型分类效率提升不明显
MIParT	大幅缩小了模型复杂度	相较于 ParT，粒子相互作用输入维数没有扩大
LGN	将洛伦兹对称性引入神经网络	张量积层计算开销大
LorentzNet	更加高效地引入了洛伦兹对称性	模型没有考虑喷注层面的特征输入，网络中的消息传递仅包含闵氏空间的内积，粒子间的交互信息远低于注意力机制
PELICAN	引入了排列等变性和洛伦兹不变性的同时降低了模型参数量	粒子间的相互作用仅包含四动量上的内积
CGENN	利用克利福德代数系统表示和提取喷注特征	几何积计算开销大，仅在顶夸克数据集上进行了实验
L-GATr	结合了几何代数、洛伦兹对称性和 Transformer 结构	计算开销较大，未在夸克/胶子标记数据集上进行验证

注 Rejx 上均取得了最优性能。但是，从 Rejx 指标可以分析出表中 5 个模型均存在分类效率不均衡的问题，即不能很好地从背景胶子喷注中分辨出特定信号。例如， $H \rightarrow gg$ 和 $t \rightarrow bqq'$ 相比，误识别率相差约 295 倍。

表 6 相关模型在夸克/胶子标记数据集上的性能比较

Table 6 Performance comparison of relevant models on the QG dataset

模型	准确率	AUC	Rej50%	Rej30%
PFN-ID	—	0.905 2	37.4±0.7	—
ParticleNet	0.840	0.911 6	39.8±0.2	98.6±1.3
ParT	0.849	0.920 3	47.9±0.5	129.5±0.9
MIParT	0.851	0.921 5	49.3±0.4	133.9±1.4
LorentzNet	0.844	0.915 6	42.4±0.4	110.2±1.3
PELICAN	0.855	0.925 2	52.3±0.3	149.8±2.4

表 8 相关模型的模型复杂度

Table 8 Model complexity of relevant models

模型	Params	MFLOPs
PFN-ID	86.1×10^3	4.62
ParticleNet	370×10^3	540.00
ParT	2.14×10^6	340.00
MIParT	720.9×10^3	180.00
LGN	4.5×10^3	—
LorentzNet	224×10^3	—
PELICAN	209×10^3	—

6 未来展望

深度学习模型在高能物理喷注标记基准测试中展现了优异的结果。然而,还有一些方面可以进一步改进。

1)更加完备的喷注数据集:分类器的性能取决于其是否能够有效学习真实喷注的数据分布。然而,现有模型仍无法很好地分类夸克/胶子数据集和 JETCLASS 数据集的部分喷注,这很可能是由于训练集规模不足所致。目前,CERN 开放数据门户提供了 PB 级的原始数据。未来的工作可以利用这些数据,选择性地提取出适合人工智能使用的喷注数据集。

2)改进等变性神经网络:等变性网络为喷注标记模型提供了良好的可解释性,并显著降低了模型的参数量。然而,目前的等变性模型仍存在一定的局限性,大部分模型仅引入了洛伦兹对称性,并面临计算开销较大的问题。未来的研究可以探讨引入更多对称性,例如动量和能量守恒,并着重降低模型的计算复杂度。

3)改进基础模型:基础模型的研究可能在未来彻底改变喷注物理的研究范式,包括喷注标记、喷注生成、四动量重建以及希格斯玻色子质量回归等任务。然而,目前喷注物理领域的基础模型仍处于探索阶段。未来的喷注物理基础模型应重点关注喷注物理特征选择、喷注的 Token 化方法、主干网络的设计、模型的训练方式等方面。

7 结束语

深度学习为实现高效且精准的高能物理喷注标记提供了可能性。本文回顾了从图像和序列的视角出发且利用相应网络模型处理喷注的相关研究,特别关注了集合视角和等变性神经网络在喷注处理中的最新进展,并介绍了相关基础模型的探索性研究。此外,本文还比较了相关分类模型在性能和设计方面的优势与局限性。下一步研究将系统分析现有喷注数据集的特性与偏差,针对性构建高质量基准数据集,并采用当前主流模型开展全面的性能验证。

参考文献

- [1] SALAM G P. Towards jetography [J]. The European Physical Journal C, 2010, 67(3): 637-686.
- [2] GALLICCHIO J, HUTH J, KAGAN M, et al. Multivariate discrimination and the Higgs + W/Z search [J]. Journal of High Energy Physics, 2011(4): 69.
- [3] CHATRCHYAN S, KHACHATRYAN V, SIRUNYAN A M, et al. Search for supersymmetry in events with opposite-sign dileptons and missing transverse energy using an artificial neural network [J]. Physical Review D, 2013, 87(7): 072001.
- [4] BALDI P, SADOWSKI P, WHITESON D. Enhanced Higgs boson to $\tau^+ \tau^-$ search with deep learning [J]. Physical Review Letters, 2015, 114(11): 111801.
- [5] LARKOSKI A J, MOULT I, NACHMAN B P. Jet substructure at the large Hadron Collider: a review of recent advances in theory and machine learning [J]. Physics Reports, 2020, 841: 1-63.
- [6] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用 [J]. 计算机学报, 2019, 42(3): 453-482.
- [7] 郑诚, 李鹏飞. 基于双超图神经网络特征融合的文本分类 [J]. 计算机工程, 2025, 51(6): 127-135.
- [8] 陈思帆, 杨家志, 黄琳, 等. 融合可变形核和自注意力的点云分类分割边卷积网络 [J]. 计算机工程, 2025, 51(6): 146-154.
- [9] CHEN S F, YANG J Z, HUANG L, et al. Edge convolutional network for point cloud classification and segmentation incorporated deformable kernel and self-attention [J]. Computer Engineering, 2025, 51(6): 146-154.
- [10] GRIPAIO B, HADDADIN W, LESTER C G. Lorentz- and permutation-invariants of particles [J]. Journal of Physics A: Mathematical and Theoretical, 2021, 54(15): 155201.
- [11] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2005.14165>.
- [12] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional Transformers for language understanding [EB/OL]. [2025-06-07]. <https://arxiv.org/abs/1810.04805>.
- [13] CIESIELSKI R, GOULIANOS K. MBR Monte Carlo simulation in PYTHIA8 [EB/OL]. [2025-06-07]. <https://arxiv.org/abs/1205.1446>.
- [14] DE FAVEREAU J, DELAERE C, DEMIN P, et al. DELPHES 3: a modular framework for fast simulation of a generic collider experiment [J]. Journal of High Energy Physics, 2014, 2014(2): 57.
- [15] ALWALL J, FREDERIX R, FRIXIONE S, et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations [J]. Journal of High Energy Physics, 2014(7): 79.
- [16] KASIECZKA G, PLEHN T, BUTTER A, et al. The machine learning landscape of top taggers [J]. SciPost Physics, 2019, 7(1): 014.
- [17] KOMISKE P T, METODIEV E M, THALER J. Energy flow networks: deep sets for particle jets [J]. Journal of High Energy Physics, 2019(1): 121.
- [18] QU H, LI C, QIAN S. Particle transformer for Jet Tagging [C]// Proceedings of International Conference on Machine Learning. New York, USA: ACM Press, 2022: 18281-18292.
- [19] KOMISKE P T, MASTANDREA R, METODIEV E M, et al. Exploring the space of jets with CMS open data [J]. Physical Review D, 2020, 101(3): 034009.
- [20] ATLAS Collaboration. ATLAS Top Tagging open data set [EB/OL]. [2025-06-07]. <https://opendata.cern.ch/record/15013>.
- [21] KANSAL R, DUARTE J, SU H, et al. Particle cloud generation with message passing generative adversarial

- networks[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2106.11535>.
- [21] ANDREASSEN A, KOMISKE P T, METODIEV E M, et al. OmniFold: a method to simultaneously unfold all observables[J]. *Physical Review Letters*, 2020, 124(18): 182001.
- [22] KASIECZKA G, NACHMAN B, SHIH D, et al. The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics[J]. *Reports on Progress in Physics*, 2021, 84(12): 124201.
- [23] DE OLIVEIRA L, KAGAN M, MACKEY L, et al. Jet-images—deep learning edition[J]. *Journal of High Energy Physics*, 2017(7): 69.
- [24] KOMISKE P T, METODIEV E M, SCHWARTZ M D. Deep learning in color: towards automated quark/gluon jet discrimination[J]. *Journal of High Energy Physics*, 2017(1): 110.
- [25] KASIECZKA G, PLEHN T, RUSSELL M, et al. Deep-learning top taggers or the end of QCD?[J]. *Journal of High Energy Physics*, 2017(5): 6.
- [26] GUEST D, COLLADO J, BALDI P, et al. Jet flavor classification in high-energy physics with deep neural networks[J]. *Physical Review D*, 2016, 94(11): 112002.
- [27] BOLS E, KIESELER J, VERZETTI M, et al. Jet flavour classification using DeepJet[J]. *Journal of Instrumentation*, 2020, 15(12): 12012.
- [28] LOUPPE G, CHO K, BECOT C, et al. QCD-aware recursive neural networks for jet physics[J]. *Journal of High Energy Physics*, 2019(1): 57.
- [29] ZAHEER M, KOTTUR S, RAVANBAKHS S, et al. Deep sets[C]//*Proceedings of Advances in Neural Information Processing Systems (NIPS)*. Long Beach, USA: Curran Associates Inc. Press, 2017: 3394-3404.
- [30] QU H L, GOUSKOS L. Jet Tagging via particle clouds[J]. *Physical Review D*, 2020, 101(5): 056019.
- [31] WANG Y, SUN Y B, LIU Z W, et al. Dynamic graph CNN for learning on point clouds[J]. *ACM Transactions on Graphics (TOG)*, 2018, 38: 1-12.
- [32] WU Y, WANG K, LI C, et al. Jet Tagging with more-interaction particle Transformer[J]. *Chinese Physics C*, 2025, 49(1): 013110.
- [33] SIRUNYAN A M, TUMASYAN A, ADAM W, et al. Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques[J]. *Journal of Instrumentation*, 2020, 15(6): 06005.
- [34] TUMASYAN A, ADAM W, ANDREJKOVIC J W, et al. Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions $\sqrt{s}=13$ TeV[J]. *Physical Review Letters*, 2023, 131(6): 061801.
- [35] BOGATSKIY A, ANDERSON B, OFFERMANN J T, et al. Lorentz group equivariant neural network for particle physics[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2006.04780>.
- [36] GONG S Q, MENG Q, ZHANG J, et al. An efficient Lorentz equivariant graph neural network for Jet Tagging[J]. *Journal of High Energy Physics*, 2022(7): 30.
- [37] BRANDSTETTER J, FORRÉ P, RUHE D. Clifford group equivariant neural networks[C]//*Proceedings of the Advances in Neural Information Processing Systems 36*. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 62922-62990.
- [38] BOGATSKIY A, HOFFMAN T, MILLER D W, et al. Explainable equivariant neural networks for particle physics: PELICAN[J]. *Journal of High Energy Physics*, 2024(3): 113.
- [39] BREHMER J, BRESÓ V, DE HAAN P, et al. Lorentz-equivariant geometric algebra Transformers for high-energy physics[C]//*Proceedings of the Advances in Neural Information Processing Systems 37*. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 22178-22205.
- [40] BREHMER J, DE H P, BEHREND S, et al. Geometric algebra Transformer[C]//*Proceedings of the Advances in Neural Information Processing Systems 36*. New Orleans, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 35472-35496.
- [41] BIRK J, HALLIN A, KASIECZKA G. OmniJet- α : the first cross-task foundation model for particle physics[J]. *Machine Learning: Science and Technology*, 2024, 5(3): 035031.
- [42] VAN D O A, VINYALS O. Neural discrete representation learning[C]//*Proceedings of the Advances in Neural Information Processing Systems(NIPS)*. Long Beach, USA: Curran Associates Inc. Press, 2017: 6309-6318.
- [43] GOLLING T, HEINRICH L, KAGAN M, et al. Masked particle modeling on sets: towards self-supervised high energy physics foundation models[J]. *Machine Learning: Science and Technology*, 2024, 5(3): 035074.
- [44] MIKUNI V, NACHMAN B. Method to simultaneously facilitate all jet physics tasks[J]. *Physical Review D*, 2025, 111(5): 054015.
- [45] WU H K, CHI P P, ZHU Y F, et al. Scaling particle collision data analysis[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2412.00129>.
- [46] BAALOUCH M, DEFURNE M, POLI J P, et al. Sim-to-real domain adaptation for high energy physics[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/1912.08001>.
- [47] AMRAM O, ANZALONE L, BIRK J, et al. Aspen Open Jets: unlocking LHC data for foundation models in particle physics[J]. *Machine Learning: Science and Technology*, 2025, 6(3): 030601.
- [48] ESMAIL W, HAMMAD A, NOJIRI M. IAFormer: interaction-aware Transformer network for collider data analysis[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2505.03258>.
- [49] MURNANE D. Graph structure from point clouds: geometric attention is all you need[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2307.16662>.
- [50] FAVARO L, GERHARTZ G, HAMPRECHT F A, et al. Lorentz-equivariance without limitations[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2508.14898>.
- [51] ODAGIU P, QUE Z Q, DUARTE J, et al. Ultrafast jet classification at the HL-LHC[J]. *Machine Learning: Science and Technology*, 2024, 5(3): 035017.
- [52] LIU R, GANDRAKOTA A, NGADIUBA J, et al. Efficient and robust Jet Tagging at the LHC with knowledge distillation[EB/OL]. [2025-06-07]. <https://arxiv.org/abs/2311.14160>.