利用神经网络CNN识别喷注味道

【高能对撞粒子分类挑战赛+第三名+CNS】

沈琢乔1 朱锐2 李甫鹏3

1. 中国海洋大学，山东青岛，266000
2. Yoho，江苏南京，210000
3. 浙江工业大学，浙江杭州，310000

【摘要】

重离子碰撞反应能将两束接近光速的高能粒子流对撞产生高温高密的粒子束。在如此极端条件下夸克胶子会出现短暂的共存，但是由于夸克的色禁闭原理导致其无法单独存在并且形成了新的高能粒子，由于相互作用的关系，高能粒子会形成高能粒子束，即喷注。喷注物理的主要目的就是通过探测到的末态喷注来判断初态的粒子性质，对初态粒子的研究有助于我们更好的了解早期宇宙以及其演化过程。在理论模型中我们可以很容易的识别出喷注的味道，但是在实验室中至今还无法准确的直接测量出真实的喷注类别。本文利用机器学习的方法，结合喷注物理中的可观测量来识别喷注的味道，为实验测量与基本粒子理论提供新的研究手段。

【关键词】重离子 高能粒子 喷注物理 深度学习 CNN

【正文】

1. 相关背景

宇宙中大多数物质由原子构成，原子又由原子核和电子组成。其中，电子是基本粒子，但原子核又可分为质子和中子，并可进一步分为夸克和胶子。这些夸克和胶子的相互作用非常强烈，以至于只有通过极高能量的质子对撞才能让它们摆脱束缚。在高能碰撞时可以产生包括夸克和中子在内的大量粒子，向某个方向射出，这些粒子团被称为喷注（jet）。  
  喷注可以分为1）胶体喷注，2）轻夸克喷注，3）魅夸克喷注，4）美夸克喷注。由于它们的不同内在特性（如质量和色量子数），不同种类的喷射经历不同的衰变过程，其内部结构也在实验中显示出不同的观测值。  
  尽管在理论物理模拟中可以很容易地识别出喷注的味道，但目前在实验中没有可靠的方法可以对所测量的真实喷注进行分类。因此，开发一种稳健的算法来识别喷注味道，将让我们可以更直接地比较实验测量和基本粒子理论。

现如今深度学习火热，尤其在视觉、自然语言的探索发挥了重要的作用。本次比赛我们使用了深度学习，并取得了第三的成绩。

1. **可观测量分析与特征量构建**

重离子碰撞产生大量的末态粒子在动量空间的信息，根据这些信息构建一些可观测量可以分析出其中的某种规律及其性质，如对粒子集体流的分析我们可以得到粒子集体运动的性质，从中看到末态粒子的集体效应，对于一般的流分析，主要计算直接流与椭圆流，对于高阶流来说通常影响不大。除此之外，针对喷注而言其可观测量有棱角（angularity），动量散布（momentum dispersion）以及喷注的宽度等，通过对末态粒子的这些观测量的分析，能大致的对不同种类的喷注在一定程度上进行分类。

除了这些可观测量外，基于对喷注产生的机制也进行了了解。喷注重建在喷注物理上也是非常常见的现象，由于在模型模拟的过程中我们可以获得一些相对应的标签，从而便于我们结合实验数据，验证结果的同时也能改进模型。喷注重建的方式也十分多样，常见的有通过对喷注半径的大小对喷注进行分类。喷注在动量空间的分布为圆锥形，这是由于在碰撞以后存在的夸克胶子等离子体种，会产生动量较大的单夸克，但是由于禁闭，但夸克是无法单独存在，迅速与其他的夸克相互作用产生新的粒子，由于相互作用的关系而产生了类似于流体力学中的粘滞作用，使得粒子呈现出圆锥状。而零头夸克的类别即是我们喷注的类别。因此喷注重建的算法也多种多样，在喷注物理中利用算法对喷注进行重建时比较有效的分类喷注的方法之一。除了以上的可观测量，还有比较常用的横动量，赝快度，方位角，以及能质比等等有一些物理可解释性的观测量。

综上，通过查阅相关文献以及数据探索我们构建了如下主要的特征量：

1.横动量

2.赝快度

3.喷注宽度

4.棱角

5.动量散布

6.直接流

7.椭圆流

8.方位角度

9.高阶流

10.能质比

11.横动量与能量比

以及其余的特征量（附件code中）等。

1. **主要思路及数据探索**

本次比赛数据有层级关系：一个event下有多个jet记录，一个jet记录下也有多个particle记录。如果使用树模型（lgb，xgb等等）训练数据的话，得构造不同层级（event,jet）的统计量（mean,sum,std,max,min等等），这样无疑舍弃了event->jet->particle的层级关系信息。我们使用了CNN 神经网络训练数据，类似textcnn（见图1），将particle 特征当成一个字的Embedding，句子程度为同event下粒子长度，不足maxlen补0。我们网络与textcnn有些区别。首先我们使用了Conv2D+kernel\_size(1,feat\_size),不同于textcnn使用了Conv1D+kernet\_size(2,feat\_size)。粒子之间好像没有顺序上的相关性，不需要捕获相邻粒子之间的信息，kernel\_size为1就行，捕获单独粒子特征之间的信息。相比Conv1D,Conv2D多了一个channel维度，我们使用了[128,256,512]渐进式的filters，再加上residual network，加深了网络深度切稳定的捕获了更多的非线性信息。最后在整个particle上做mean pooling，再经过几层MLP。

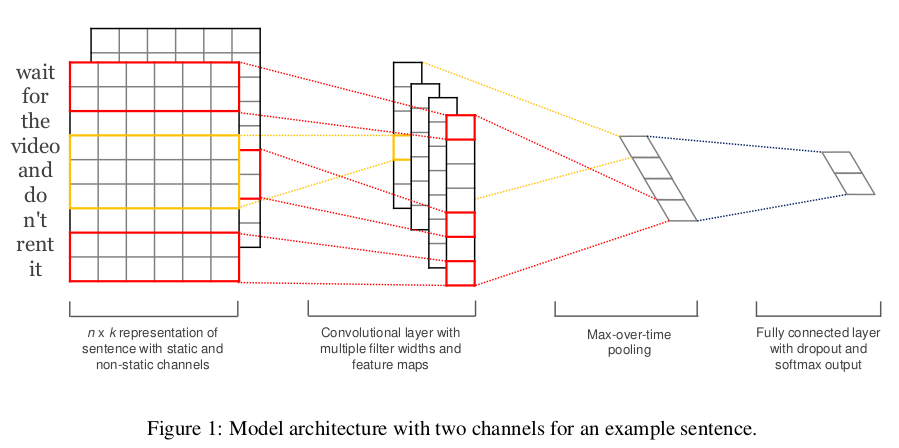


图1 textcnn 网络结构

训练集构造的话是以event为维度作为一条记录，类似句子。由于以event为维度，标签的分布也很均衡。训练的时候做了5折。测试集对5折预测概率结果进行了平均，线上分数达到了第二名。

1. **实验和试错**

复杂数据赛道开始的时候，我们沿用了简单数据赛道的方案，使用lstm对数据建模，但效果不佳。主要原因一是同event下particle序列过长，lstm对于过长的序列效果不佳以及耗时太长。二是同event下particle序列也没顺序上的依赖，按照energy或者mass排序反而降分。然后我们使用CNN对数据建模，训练速度以及效果提升了不少。后期我们尝试了构造3D数据进行训练（jet\_maxlen \* particle\_maxlen \* feats），但是jet，particle长度不同event有很大的区别，我们直接用0进行了padding,效果不佳就没继续尝试了。

1. **总结**

通过这次比赛，我们见识到了深度学习的强大力量，也学习到了不少。