Об использовании RNN с LSTM для определения ускоренных top-кварков в адронных струях в экспериментах Большого адронного колайдера

В статье описан подход, с помощью которого осуществляется достаточно важная задача определения потоков частиц с адронных и глюонных распадов, физики элементарных частиц где-то за рамками Стандартной модели. Как отмечено в статье, для определения типа частиц используются нейронные сети. Упоминается подход одного из использований рекурсивных сетей для генерации потоков W бозонов – переносчиков слабого взаимодействия, там входными данными служили 4-импульсы частиц; далее данные передавались на вход fully-connected DNN. Основными характеристиками сигнала служили ортогональная составляющая момента импульса (p_T), псевдобыстрота (угловое отклонение от потока распространения, η) и азимутальный угол (ϕ). Также применено Лоренцево преобразование поворота, которое сыграло важную роль в перфомансе сети. В данной работе предлагается заменить списков импульсов частиц фиксированной длины предлагается использовать триплеты (p_T, η, ϕ), вариативные по длине, последовательно даваемые на вход LSTM-ячейкам.

Авторы уделили очень большое внимание препроцессингу и моделированию исследуемого сигнала. Отметим самые важные моменты:

- ✗ данные сгенерированы методом Монте-Карло, они состоят из $Z^{|}$-бозонов, распадающихся на топ-кварки (разрешены только определённые распады), и фоновых струй глюонов, с одинаковыми распределениями параметра η. Также проведена симуляция детектора;
- ✗ Проведена подготовка потоков частиц с помощью FastJet: часть частиц собрана алгоритмами кластеризации, разложены на подкластеры, чтобы сохранить исходную структуру, часть – подверглась триммингу с дефолтным порогом по модели, часть была подвержена сэмплированию, чтобы алгоритм не смог быстро приспособиться к законам распределения импульсов в модели – этакая регуляризация. Итого получили 7 миллионов потоков, разбитых на сигнал и фон, которые были разбиты на обучение, валидацию и тест как 80:10:10 и подготовлен набор из независимых 11 миллионов потоков (половина сигнал, половина фон) для финального тестирования. Порядок следования потоков при каждой эпохе обучения предполагался перемешанным.
- ✗ Произведён препроцессинг, а именно:
  - преобразование Лоренц-инвариантым поворотом – этакая аугментация;
  - центрирование по всему триплету параметров;

- сдвиг системы координат, чтобы среднее значение оставалось бы в большей степени положительным, поворот системы координат;

✗ Произведено выравнивание последовательности потоков частиц:
- Рекурсивно + помощь анти-k_T алгоритма (декластеризация) установлен порядок построения подпотоков исходного потока (зачем это сделано – см. препроцессинг) – это позволило определить лучший порядок следования потоков, как заявляют авторы статьи

Архитектура является эвристикой, найденной авторами. Оптимизатор – Адам (неудивительно). Сеть состоит из LSTM на 128 выходов, выведенных на 64 узла Dense layer. Как видно, достаточно простая архитектура.

Авторы статьи по данным ROC-кривым утверждают, что добились в 2 раза большей эффективности отвержения не тех потоков на уровне 50% (достаточно высокий уровень, не факт, что это есть хорошо…), также продемонстрирована эффективность проведённого командой препроцессинга.

На самом деле, авторы данной статьёй показали очень важную вещь: они очень много трудились над предобработкой данных: аугментировали, настраивали регуляризацию потоков их разбиением, строили подпорядки разбиения потоков алгоритмами кластеризации на специальной метрике, чтобы показать, что если дать LSTM подходящей структры последовательность, которая не хаотична, а можно построить её иерархию, устнановить порядок, то эта модель окажется чуть ли не на порядок лучше обычной DNN, хотя устроена проще – то есть, очень важна природа тех признаков, которые пытается «извлечь» каждый слой сети. Но, хотелось бы отметить и минусы такого подхода – это жестокий препроцессинг, делавшийся очень умными людьми. Возможно, стоило бы поискать компромисс между подготовкой данных и сложности сети для максимизации выгоды при минимуме затрат.