Доменное состязательное обучение для понижения смещения прогноза при поиске бозона Хиггса в детекторе ATLAS

Фатхуллин Ильяс Фаизович[1]⋆ ilyas.fn979@gmail.com
Стрижов Вадим Викторович[1] strijov@phystech.edu
[1]Долгопрудный, Московский физико-технический институт

Детальное изучение бозона Хиггса, включая его редкие режимы рождения, является целью экспериментов на Большом адронном коллайдере [1]. Анализируется обнаружение бозона Хиггса, образующегося в ассоциации с топ анти-топ кварковой парой (известной как режим рождения ttH). Измерение этого процесса проверяет взаимодействие Юкавы между бозоном Хиггса и топ-кварком. Наиболее вероятным распадом бозона Хиггса является распад на два нижних кварка (измеренные как b-джеты). Данный сигнал требуется отличать от большого фона tt+b-джетов. Это задача бинарной классификации с 41-им признаком, которые имитируют отклик детектора ATLAS на рождение бозона Хиггса. Соответствующим типом события, меткой класса, является сигнал или фон. Симулированные размеченные наборы данных Монте-Карло используются для обучения модели классификации. Однако смещение при обучении в сторону конкретного Монте-Карло генератора снижает обобщающую способность модели. Текущие результаты [1] поиска ttH ($H \to bb$) ограничиваются моделированием неопределенности симуляции фона, рассчитываемой как расхождение отклика классификатора (смещение при обучении) на различных генераторах Монте-Карло. Цель данной работы состоит в том, чтобы уменьшить это расхождение при обучении.

Два набора данных S_1 и S_2 состоят из 100000 ttH(bb) сигнальных событий в результате моделирования MadGraph/Herwig6. Каждый из них также содержит 100000 фоновых событий. Но S_1 заполнен фоновыми событиями из MadGraph/Pythia6, а S_2 из Powheg Pythia8. Каждое событие описывается 41-м признаком такими как *количество джетов*, *потери поперечной энергии*, *масса ближайших b-джетов* и т.д.

Распределение признаков фоновых событий немного отличается для S_1 и S_2. Целью данного исследования является постро-

ение модели классификации, которая показывает адекватное качество на S_2 при обучении на S_1. Качество классификации оценивается по трем критериям: точность, площадь под ROC-кривой и значимость. Последнее оценивается как

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}}, \tag{1}$$

где s, b соответствуют числу сигнальных и фоновых событий, превышающих фиксированное значение на графике распределения ответов классификатора, а σ_b^2 есть разница между уровнями двух фонов, превышающих фиксированное значения на графике распределения ответов классификатора.

Изучено применение нейросетей с градиентным реверсивным слоем к поиску ttH(bb) в ATLAS. Это полносвязная трехслойная нейронная сеть, в которую встроена дополнительная нейронной сетью под названием *классификатор доменов*. Классификатор доменов используется только во время обучения и предназначен для сближения распределений признаков в скрытом пространстве для двух заданных наборов данных.

Эксперименты показали, что состязательная доменная адаптация снижает смещение прогноза к заданной симуляции, сохраняя при этом адекватное качество классификации.

[1] *Collaboration ATLAS.* Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\overline{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, 2019. `arXiv:1712.08895`.

Domain Adversarial Learning to Reduce Training Bias in ttH(bb) Search at ATLAS

Fatkhullin Ilyas[1]⋆ ilyas.fn979@gmail.com
Strijov Vadim[1] strijov@phystech.edu
[1]Dolgoprudny, Moscow Institute of Physics and Technology

The detailed study of the Higgs boson including its rare production modes is a primary activity of the Large Hadron Collider (LHC) experiments [1]. The detection of the Higgs boson produced in the association with a top anti-top quark pair (known as ttH production mode) is analysied. A measurement of this process tests the Yukawa coupling between the Higgs boson and the top quark. The most probable decay of the Higgs boson is to two bottom quarks (measured as b-jets) and the signal needs to be distinguished from a large background of tt+b-jets. Formally, this is a binary classification problem with 41 input variables (features) that simulate ATLAS detector response on the processes. Its corresponding event type (class label) is either signal or background. The simulated labelled Monte Carlo datasets are utilised to train the classification model. However, the training bias towards a specific Monte Carlo generator reduces the generalisation capabilities of the model. The current results [1] for the ttH (H $\to$ bb) search was limited to model the uncertainty of the background simulation, calculated as the discrepancy of the classifier response (training bias) to different Monte Carlo generators. The aim is to reduce this training bias.

Two datasets S_1 and S_2 consist of 100000 ttH(bb) signal events from MadGraph/Herwig6 simulation. Each of them also have 100000 background events. But S_1 is filled with background events from MadGraph/Pythia6, while S_2 from Powheg Pythia8. Each event is described by 41 features such as *number of jets*, *missing transverse energy*, *mass of closest b-jets* etc.

The feature distributions of the background events appear to be slightly different for S_1 and S_2. The goal of this study is to build a classification model, which shows an adequate performance on S_2 while being trained on S_1. The classification performance is mainly assessed based on accuracy, area under ROC-curve, and significance. The latter is approximated as follows:

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}}, \tag{1}$$

where s, b correspond to the number of signal and background events, above the fixed cut in the classifier response plot, and σ_b^2 is the difference between two backgrounds above the cut.

The application of Neural Networks with a gradient reversal layer to the ttH(bb) search at ATLAS is studied. This is a simple Feed Forward NN with an additional NN called *domain classifier*. The domain classifier is used during training only and aimed to match the feature distributions in the latent space for the two given datasets.

Extensive experiments have demonstrated that adversarial domain adaptation reduces the training bias towards a given simulation while preserving adequate classification performance.

[1] *Collaboration ATLAS.* Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, 2019. `arXiv:1712.08895`.