Imperial College London

HEPH-Zenz-1

Developing Machine Learning techniques for Simplified Template Cross Sections in $H \rightarrow \gamma\gamma$

Dominika Vasilkova

Supervisor: Dr Seth Zenz

Assessor: Prof. Gavin Davies

Word count: 2487

Abstract

The simplified template cross section method for collecting data about the Higgs is considered as a potential target for machine learning methods, for identification and classification of events from the diphoton decay. This combination is designed to be applied to upcoming LHC data, to measure the properties of the Higgs more accurately and look for deviation from the Standard Model. Deep learning, a new machine learning technique, is analysed as a potential upgrade to the methods being used currently.

Introduction

In 2012, the ATLAS [1] and CMS [2] collaborations at the Large Hadron Collider (LHC) announced the discovery of a new particle believed to be the Higgs boson. This was the culmination of over 50 years of theoretical and experimental work to complete the Standard Model of Particle Physics (SM), the mathematical theory which predicted the existence and properties of many new particles [3]. However, whether this particle discovered was the SM Higgs needed testing, leading to an ongoing study of its properties and couplings to other particles. Any discrepancy found would give evidence for new physics, and possibly help to explain some of the gaps in current theories.

The Higgs boson is a scalar (spin 0) boson with even parity. At the LHC, the primary production mode is gluon fusion, followed by vector boson fusion (VBF) [4].

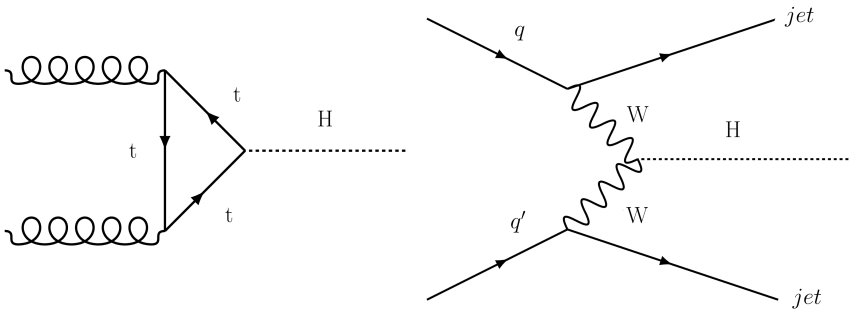

Figure 1: Feynman diagrams for gluon fusion (left) and vector boson fusion (right). The forward jets from VBF can be used to identify it compared to other modes [5].

Alternative production modes include associated production with a W boson and the recently observed associated production with top quarks [6]. These modes are all individually useful as they have distinct signatures and can be used to probe the couplings of the Higgs with the different particles involved.

The decay modes include decays into vector bosons, two tau leptons, two b quarks and two photons [7], which is the decay mode considered in this project. Despite having a relatively small cross-section [2], the diphoton mode has the advantage of being a very clean mode, since the photons do not produce jets. It also has very good mass resolution, as the invariant mass of the photons gives a sharp peak [8]. Additionally, the diphoton decay is interesting since it proceeds via a loop process, due to the photon being massless and therefore having no direct coupling to the Higgs. This makes the decay mode more sensitive to interference effects from potential new particles [9].

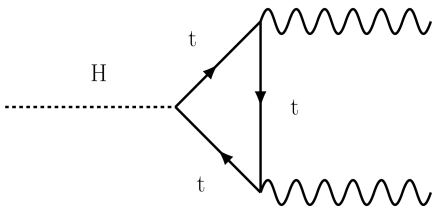

Figure 2: A possible Feynman diagram for the diphoton decay via a top quark loop. Another possibiliy is a W boson loop [5].

So far, all results seen are consistent with the SM, and the Higgs's spin-0 and even parity properties have been confirmed [10] [11]. For each of the production and decay modes, a variable σ/σ_{SM} is defined to test compatibility with the SM, referred to as μ. This is expected to be close to 1, and indeed all measurements of it so far have been consistent with expectations [1] [10].

Simplified template cross sections are a method for interpreting data first used in LHC Run 1, where the cross sections of each production mode are separated and split into multiple 'bins', mutually exclusive areas of phase space, in order to systematically reduce the theory dependence of results. This can help reduce the theoretical uncertainties associated with measured values, as well as removing the dependencies on the underlying physics model being used, something that is particularly helpful when looking for BSM physics.

With the large volumes of data the LHC produces, finding a way to make this data more manageable is important. In recent years, machine learning has increased in popularity as a tool to do this. These algorithms can be trained to identify event signatures from backgrounds based on a set of criteria.

The aim of this project is to analyse and improve the current machine learning techniques being used at CMS in H$\rightarrow \gamma\gamma$, in order to increase its usage for event selection and categorisation into the 'bins' defined by the simplified template cross section method, as well as potentially looking at how the binning choices could be expanded for future data analyses. This will allow better measurement of the Higgs boson's properties, allowing for more detailed tests of the SM Higgs and further probing for BSM physics in the diphoton decay.

Background material

Theory

Spontaneous symmetry breaking is where a system in the ground state does not have the same symmetry as the Lagrangian that describes it. This is observed in nature, but the physical origin of the breaking remains a mystery. Including this broken symmetry and quantising the fields predicts the existence of 'gauge' bosons such as the W and Z. However, this requires all bosons to be massless due to Goldstone's theorem [12], which presents a problem as the weak force is known to have a very short range so any mediator particles are required to be massive.

A solution was provided by Schwinger [13], Englert and Brout [14], Higgs [12] and Guralnik, Hagen and Kibble [15]. Their insights were that introducing a new scalar field to break the symmetry allowed the W and Z bosons to acquire mass by 'absorbing' three

of the four degrees of freedom in the field [16]. The Higgs boson is the fourth degree of freedom, and the particle seen is the excitation of this new scalar field. This theory does not predict the mass of the Higgs boson, rather it is a parameter that if known allows all the expected coupling strengths to be predicted.

It was later observed by Weinburg that via a Yukawa interaction (an interaction between a scalar field like the Higgs field and a Dirac field, which describes the behaviour of fermions) this would also give mass to fermions [17]. Combining the unified electroweak model with this new information produces what is now referred to as the Standard Model of Particle Physics.

Methodology of the diphoton channel

The expected signature of a $H \rightarrow \gamma\gamma$ decay is a narrow peak above the background found in the mass spectrum of the two photon system [18]. The observation of this has two main stages - the identification of the photons, and the measurement of their properties in order to reconstruct the decay. In the first stage, energy deposits in the electromagnetic calorimeter (ECAL) consistent with a photon are looked for. If an event has two of these, they will be analysed further to make sure that the particles are photons, since electrons and fragments from hadron jets such as π^0 will leave similar deposits [19]. To do this, a set of criteria called 'Photon identification' are developed. To find the invariant mass of the photon pair, angles are needed so the point of decay needs to be reconstructed (as the photons don't leave any deposits in the inner part of the detector). This is difficult, since it can only be done statistically, and any errors in the vertex position will increase the width of the mass peak [20]. Another uncertainty which could increase the width are imperfect energy measurements in the ECAL, but these can be corrected for [21]. Finally, additional criteria designed to differentiate between a Higgs decay and a background decay are applied to the data. The main background in this decay channel is QCD production of two photons, and is irreducible [22].

Thus, as well as working on upgrading the LHC and the detectors themselves, the methods of photon identification, reconstruction and signal/background identification have needed improvements with each data run [23]. With improvements to the LHC itself, an additional challenge is presented by the increasing data volumes, which has led to greater need for better data processing methods such as machine learning.

Machine learning methods

Generally, the aim for any machine learning algorithm being used on LHC data is to cut away as much uninteresting data as possible. This could be immediately discarding bad quality events, or identifying particles based on detector deposits and separating this from expected backgrounds. The pattern recognition abilities of machine learning algorithms appear to be an efficient way to do this. These can be trained in two different ways; aided, where the training data is labelled to show what the features mean, and unaided, where the training data is unlabelled and the algorithm is left to make its own decisions. In particle physics, aided data sets are usually used, created by running simulations of the expected interactions in the detector [24].

Boosted decision trees (BDTs) are the most common method used currently [25]. Here, each input variable is a 'node', and the algorithm makes multiple cuts until some predefined end point is reached. The cuts are across multiple variables at once, and are chosen to maximise the change in some classifying variable, for example the signal-to-background

ratio of a sample. Each data point is then classified as 'signal' or 'background'. One problem with this method is that small statistical fluctuations in the training data can have a large impact. The solution to this is to use 'boosting', where multiple decision trees are combined to reduce the errors and increase the overall stability [26]. BDTs have been used successfully in experiments such as MiniBooNE [27], and are good at handling large data inputs, but even with the boosting are still not massively robust and can have issues with overfitting, where the tree created is specialised to the training data and doesn't work with real data as well [28].

Both CMS and ATLAS have used BDTs in their analyses in a variety of different ways. For example, using a BDT to identify tau leptons in decays compared to electrons and jets has been implemented in ATLAS, with tests confirming the expected score distributions when run on different sets of simulated data [29]. Similarly, ATLAS has also tested b tagging using BDTs [30]. CMS made extensive use of BDTs during Run 1, using them for photon identification, uncertainty estimation and photon energy correction [31]. Using a BDT allowed the photons to be classified by signal purity, allowing an additional analysis to be done on a smaller sample of data. The impact of this was estimated to be equal to collecting 50% more data [24]. These successes show machine learning in general is a useful tool in high energy physics.

In recent years, the improvements in BDTs for LHC data have started to slow down, leading to a search for alternatives [32]. One class of algorithms that has received much interest recently is Deep Learning, which differs from 'shallow' methods by teaching itself how to spot the features in the data rather than being taught. This makes them much more versatile than shallow methods and can lead to the algorithm itself finding better distinguishing features than the current set being used. It can also work with a larger set of criteria compared to current methods, however the training does take longer and is still prone to overfitting [33].

A study testing the possible effectiveness attempted to identify the signatures of exotic particles using a deep neural network, and found that it out-performed both BDT and neural network shallow methods even when the shallow methods were explicitly given criteria based on expected physics [32]. Despite a few cases where the data quality became the limiting factor rather than the algorithm [34], deep learning has had many successes. ATLAS tested one for b-tagging, which reduced the misidentification by a factor of 4. The method combined with a BDT for faster training still reduced this by a factor of 3 [35]. It has also been used successfully to identify the tau Higgs decay against the background [36]. These results show that deep learning has promise for event classification, making it a very useful tool should one need to categorise events according to specific criteria.

Simplified template cross sections

Simplified template cross sections involve measuring the cross section of each production mode divided into regions defined by their kinematic variables. This is done directly rather than measuring the signal strength, removing the theory dependence on the efficiency of the Higgs decay and acceptance limits, which are usually estimated computationally [37]. Since the signal strength is calculated from a combination of the cross-section and these theoretical estimates, measuring just the cross section gives more accurate data to work with while still making it possible to calculate the signal strength in the old way.

The bins are chosen differently for each production mode, allowing them to be designed for the expected signatures and topologies. The boundaries of the phase space regions are chosen to minimise extrapolation and maximise the sensitivity of the measurements.

They are also chosen to minimise the number of bins required, and to allow potential BSM effects to be identified clearly [38]. Splitting the data in this way gives more information about the Higgs while having the benefit of combined decay mode information, allowing the testing of theoretical models in more detail than before.

There are three 'stages' of subdivision, designed to be measured in succession and updated based on measurements made before it. At stage 0, each production mode has one bin, with some modes like VBF and associated Higgs production grouped together due to similarities in signature. Stage 1 separates these bins further by kinematic properties. The aim is for all analyses to reach the specific subdivisions defined. Stage 2 divides the stage 1 bins, but before these can be defined stage 1 information is needed [39].

While the stage 0 bins are similar to Run 1 categorisations, so far only ATLAS has managed to measure stage 1 bins at $\sqrt{13}$ TeV. The diphoton section is split into 29 categories, using kinematic variables such as transverse momentum, number of jets, missing energy and invariant mass. Due to not having enough data sensitivity, the regions were merged into 9 phase space regions. All of the regions were found to be compatible with the Standard Model. The precision in measurement of the diphoton cross section was improved, but still limited by the data uncertainties. Thus, it was concluded that better quality data is needed to allow greater stage 1 splitting [40].

Conclusion

In conclusion, while the discovery of the Higgs and subsequent confirmation of its properties have been a great triumph for modern physics, there are still many things in and beyond the Standard Model that are not understood. The LHC is aiming to answer these questions in the future, but with upgrades come new challenges to deal with the increases in data volume. A promising way to do this is with machine learning, which can be used to identify events and other interesting patterns in data from the background, but the currently used methods in Particle Physics are starting to struggle so new approaches are needed. Simultaneously, the methods used for data analysis can be improved, and simplified template cross sections look like a good method for increasing the quality and usefulness of experimental measurements, as well as flagging up any BSM physics that is being seen at the LHC.

References

[1] The ATLAS collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Phys.Let.B*, 716:1, 2012.

[2] The CMS collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". *Phys. Lett. B*, 716:30–61, 2013.

[3] J. Woithe, G.J. Wiener and F. F Van der Veken. "Let's have a coffee with the Standard Model of particle physics!". *Physics Education*, 52, 2017.

[4] Michael Kramer. "Higgs Production at the LHC", 2005. Theorie-Seminar Universitat Bielefeld.

[5] Alec Aivazis. Draw Feynman diagrams online. https://feynman.aivazis.com/.

[6] The CMS collaboration. "Observation of ttH production". arXiv:1804.02610 [hep-ex].

[7] A. Djouadi. "Decays of the Higgs Bosons", 1997. Talk given at the International Workshop on Quantum Eects in the MSSM Barcelona, Spain, September 913.

[8] C. J. Seez et al. "Photon decay modes of the intermediate mass Higgs". 1990.

[9] The CMS collaboration. "Combined measurements of the Higgs bosons couplings at sqrt(s) = 13 TeV". CMS Analysis summary (PAS HIG-17-031).

[10] The CMS collaboration. "Observation of the diphoton decay of the 125 GeV Higgs boson and measurement of its properties". *Eur. Phys. J. C*, 74:3076, 2014.

[11] The CMS collaboration. "Constraints on anomalous Higgs boson couplings using production and decay information in the four-lepton final state". *Phys. Lett. B.*, page 1, 2017.

[12] P.W. Higgs. "Broken symmetries, massless particles and gauge fields". *Phys. Lett. B*, 12:132–133, 1964.

[13] J. Schwinger. "Gauge Invariance and Mass". *Phys. Rev.*, 125:397–398, 1962.

[14] F. Englert and R. Brout. "Broken symmetry and the mass of gauge vector mesons". *Phys. Rev. Lett.*, 13:321, 1964.

[15] G. S. Guralnik, C. R. Hagen, and T.W. B. Kibble. "Global conservation laws and massless particles". *Phys. Rev. Lett.*, 13:585, 1964.

[16] G. S. Guralnik, C. R. Hagen, and T.W. B. Kibble. "Broken symmetries and the Goldstone theorem", 1967. Advances in Physics 2.

[17] S. Weinberg. "A Model of Leptons". *Phys. Rev. Lett*, 19:1264, 1967.

[18] S. Zenz. "Understanding the Higgs Boson: Where We Are, Where We're Going, and How To Get There". UCL HEP seminar series.

[19] The CMS collaboration. Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at sqrt(s) = 8 TeV. *JINST*, page P08010, 2015.

[20] S. Bhattacharya and S. Jain. "A review of the discovery of SM-like Higgs boson in $H \to \gamma\gamma$ decay channel with the CMS detector at the LHC". *J. Phys.*, page 35, 2016.

[21] The CMS collaboration. "Combined results of searches for the standard model Higgs boson in pp collisions at sqrt(s) = 7 TeV". *Phys. Lett. B*, 710:26, 2012.

[22] The ATLAS collaboration. "Combined search for the Standard Model Higgs boson in pp collisions at sqrt(s)= 7 TeV with the ATLAS detector". *Phys. Lett. D*, 86:032003, 2012.

[23] The CMS collaboration. "Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at sqrt(s) = 13 TeV". arXiv:1804.02716 [hep-ex], 2018.

[24] A. Radovic et al. "Machine learning at the energy and intensity frontiers of particle physics". *Nature*, 560:41, 2018.

[25] Adrian Bevan. "Machine Learning in High Energy Physics". Talk given as part of HORSE2017, 2017.

[26] Y. Freund and R.E. Schapire. "A short introduction to boosting". *Japanese Society for Articial Intelligence*, 14:771, 1999.

[27] B. P. Roe et al. "Boosted decision trees as an alternative to artifical neural networks for particle identification". *NIMA*, page 577, 2005.

[28] G. James, D. Witten, T. Hastie, R, Tibshirani. *"An Introduction to Statistical Learning: with Applications in R"*. New York: Springer, 2017.

[29] The ATLAS collaboration. "The ATLAS Tau Trigger in Run 2". 2017. ATLAS-CONF-2017-061.

[30] The ATLAS collaboration. "Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run". ATL-PHYS-PUB-2016-012, 2016.

[31] The CMS collaboration. "Updated measurements of the Higgs boson at 125 GeV in the two photon decay channel". 2013. CMS-PAS-HIG-13-001.

[32] P. Baldi, P. Sadowski, and D. Whiteson. "Searching for Exotic Particles in High-Energy Physics with Deep Learning". *Nature Commun.*, 5:4308, 2014.

[33] J. Schmidhuber. "Deep learning in neural networks: An overview". *Neural Networks*, 61:85, 2015.

[34] A. Morningstar, P. Savard, and R. Teuscher. "A BDT optimization study and assessment of deep learning in selecting VBF events in the H $\rightarrow$ ZZ* $\rightarrow$ 4l channel". 2015.

[35] The ATLAS collaboration. "Identication of Jets Containing b-hadrons with Recurrent Neural Networks at the ATLAS Experiment". ATL-PHYS-PUB-2017-003, 2017.

[36] E. Barberio et al. "Deep learning approach to the Higgs boson CP measurement in H to tau tau decay and associated systematics". *Physical Review D*, 96, 2017.

[37] M. Duehrssen-Debling, P. Francavilla, F. Tackmann, K. Tackmann. "Simplied Template Cross Sections". Talk given at QED@LHC, 2016.

[38] LHC Higgs Cross Section Working Group. "Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector". *CERN yellow reports*, 2, 2017.

[39] F. Tackmann. "Simplified template cross sections". Talk given at Higgs/B9 meeting, DESY, 2017.

[40] "Measurements of Higgs boson properties in the diphoton decay channel using $80 fb^{-1}$ of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector". ATLAS-CONF-2018-028, 2018.