IL NUOVO CIMENTO **47 C** (2024) 250
DOI 10.1393/ncc/i2024-24250-5

Communications: SIF Congress 2023

Particle Transformer for τ lepton pair invariant mass reconstruction for the $HH \to b\bar{b}\tau^+\tau^-$ CMS analysis

V. Camagni(*) on behalf of the CMS Collaboration

Dipartimento di Fisica G.Occhialini, Università di Milano Bicocca and INFN, Sezione di Milano Bicocca - Milan, Italy

received 30 January 2024

Summary. — One of the most interesting channels to probe theories beyond the Standard Model at the Large Hadron Collider is the production of a new massive particle, that decays into pairs of Higgs bosons which, in turn, decay into a pair of b-quarks and a pair of τ leptons. A fundamental discriminant variable to separate HH signal from the backgrounds is the invariant mass of the di-τ system. In order to reconstruct it special techniques are needed, as the presence of neutrinos from τ decay does not allow for a complete reconstruction of the event. To this end, a transformer-based architecture (Particle Transformer) has been implemented, showing better results with respect to the most common used algorithm in CMS, which is, in addition, extremely CPU consuming.

1. – Introduction

The fundamental interactions of nature are investigated at the Large Hadron Collider (LHC) at CERN, where the Compact Muon Solenoid (CMS) experiment collects signals generated by proton-proton (p-p) collisions occurring every 25 ns when two high-energy proton bunches cross. The posterior analysis of these data allows to probe the Standard Model (SM) of particle physics. In particular, the precise characterization of the properties and couplings of the Higgs boson (H) is now of utmost importance, since deviations from SM predictions may point to physics beyond the SM (BSM). In this regard, a central property of H is its self-coupling, which is proportional to its mass and whose experimental evidence is yet to be found. The most promising method to directly probe the H self-coupling is via the study of Higgs boson pair production (HH). Observing this process at the LHC is particularly difficult because it has a small cross-section, which is roughly 1500 times smaller than the single H production cross-section. Nevertheless, this small cross-section is highly sensitive to the presence of BSM contributions, that could manifest directly as new states X of mass $m_X > 2m_H$ decaying into a HH system [1].

(*) E-mail: `v.camagni1@campus.unimib.it`

Among the decay channels of HH, the one involving two bottom quarks and two τ leptons ($b\bar{b}\tau^+\tau^-$ or, for simplicity, $bb\tau\tau$) represents one of the best options for HH searches. The $bb\tau\tau$ final state benefits from a sizeable branching fraction of 7.3% and concurrently profits from the high-selection purity of the τ leptons that keeps background contamination contained. However, both $H \to b\bar{b}$ and $H \to \tau^+\tau^-$ signals are challenging to be extracted since there are several background processes with similar final state or that result in jets from quarks, or gluons which are misidentified as the signal τ or b jets. This discrimination between signal and background events can be performed based on the invariant mass of the di-τ system. Also this $m_{\tau\tau}$ estimation is arduous: τ leptons are not stable particles as they in turn decay into electrons or muons (*i.e.,*, leptonically) or into quarks (*i.e.,*, hadronically), both accompanied by neutrinos which weakly interact with matter and escape detection. As a consequence only a partial reconstruction of the di-τ is possible, in the form of an invariant visible mass ($m_{\tau\tau}^{vis}$), which has a resolution that does not allow to efficiently discriminate the HH signal from the background. To exploit at the most the information collected by the detector and improve the precision on $m_{\tau\tau}^{vis}$, the invariant mass of the $\tau\tau$ pair is reconstructed using the CMS algorithm Secondary Vertex Fit (SVFit) [2]: taking as inputs the visible τ decay products and the Missing Transverse Momentum (MET, *i.e.,* the total imbalance in the transverse momentum representative of all escaped neutrinos in the event), SVFit reconstructs, through a maximum likelihood approach, the most probable kinematics of the missing neutrinos in the final state. The significant computational time combined with a reduced resolution of the mass reconstructed by this algorithm opens opportunities for new ML-based strategies. In this work, a deep learning model —Particle Transformer (ParT)— has been implemented to estimate the four-momentum of the neutrinos involved in the τ decay for a high-resolution reconstruction of the corresponding invariant mass.

2. – Data sets and methods

2˙1. *Data.* – The HH signal and backgrounds samples used to train, validate and test the ParT model have been produced with a full detailed simulation of the CMS detector, the *Geant4* package [3] has been used to simulate the interaction of particles through the detector:

- The signal sample is represented by the gluon-gluon fusion (ggF) production of a massive resonance X of spin 0 decaying in a couple of Higgs bosons. Independent samples are generated for different values of m_X ranging from 250 GeV to 550 GeV every 50 GeV. The choice to focus on this mass interval range, instead of considering the entire range up to 1 TeV, is due to the fact that it constitutes the most difficult part to be reconstructed by SVFit. Indeed, in the absence of a Lorentz boost, the two τ are often produced back to back and the missing momentum associated with their neutrinos partially cancels out. As a result, the invariant mass of a resonance cannot be directly reconstructed using the MET and the visible decay products of the τ leptons. At this level, pre-selection criteria on τ and b-jets are applied, leading to the survival of τ pair objects that can decay into one of the $\tau_h\tau_h$, $\tau_\mu\tau_h$, $\tau_e\tau_h$ final states with a total of 2 or 3 neutrinos (those coming from the same τ decay are reconstructed as a single system).

- The considered background processes are:

  1) Drell-Yan: production of a Z boson decaying into a pair of τ leptons.

2) $t\bar{t}$: production of a top anti-top pair with each one decaying into a W boson and a b-quark. One W boson is then decaying into a lepton (e, μ or τ) and a neutrino, while the other into a pair of quarks (generating light jets).

2·2. *Representation of the input.* – Each collider event obtained in the data generation phase is converted to an event graph as the input for the ParT, reason why it can be seen as a Graph Neural Network. A node represents a final state object while an edge represents a set of pair-wise features between two nodes. This object can be a τ lepton, a b-jet or MET. Each node has a twelve-dimensional feature vector:

$$x_i = (p_x, p_y, p_z, E, p_T, \eta, \phi, dm, bjet^{dF}, bjet^{b}_{pNet}, bjet^{c}_{pNet}, bjet^{uds}_{pNet}), \tag{1}$$

which contains the most relevant properties of the corresponding final state. For the elements of a feature vector, (p_x, p_y, p_z, E) represent the the four-momenta of the object, p_T, η, ϕ and dm, respectively, the transverse momentum, the pseudo-rapidity, the azimuthal angle and the decay mode. The remaining ones $(bjet^{dF}, bjet^{b}_{pNet}, bjet^{c}_{pNet}, bjet^{uds}_{pNet})$, instead, represent variables characterising jets, indicating a score about the probability that the jet comes from a b-quark or c-quark. Each pair of nodes is linked by an edge which is weighted to a four-dimensional feature vector:

$$u_i = (\Delta, k_T, z, m^2). \tag{2}$$

These features are derived from the energy-momentum 4-vector $p = (p_x, p_y, p_z, E)$ of each particle, *i.e.*, node. Specifically, for a pair of particles a, b with 4-vector p_a, p_b, u_i are calculated as:

$$\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2}, \tag{3}$$

$$k_T = \min(p_{T,a}, p_{T,b})\Delta, \tag{4}$$

$$z = \min(p_{T,a}, p_{T,b})/(p_{T,a}, p_{T,b}), \tag{5}$$

$$m^2 = (E_a + E_b)^2 - ||\vec{p_a} + \vec{p_b}||^2, \tag{6}$$

where y_i is the rapidity, ϕ_i is the azimuthal angle, $p_{T,i} = \sqrt{p^2_{x,i} + p^2_{y,i}}$ is the transverse momentum (p_T), $\vec{p_i} = (p_{x,i}, p_{y,i}, p_{z,i})$ is the momentum 3-vector and $||.||$ is the norm, for $i = a, b$. Since these variables tipically have a long-tail distribution, for each particle pair (lnΔ, lnk_T, ln$z$, lnm^2) has been used as edge. The choice of this set of pair-wise features is motivated in [4].

2·3. *ParT.* – The transformer model, as presented in [5], stands as a prominent deep learning architecture that has found widespread adoption across diverse domains, including Natural Language Processing (NLP), computer vision (CV), and speech processing. Beyond its prolific use in language-related applications, the transformer has found applications in other disciplines, such as chemistry and life sciences. A recent work [6] demonstrated how this class of models can achieve good results also for high-energy physics applications, in particular for the jet tagging task outperforming the previous state-of-the-art network. This led to the choice of investigating such architecture to address the di-τ mass regression problem.

The ParT is composed of only the encoder part with respect to the original published version, but on the other side, it combines both the task of regression of the neutrinos

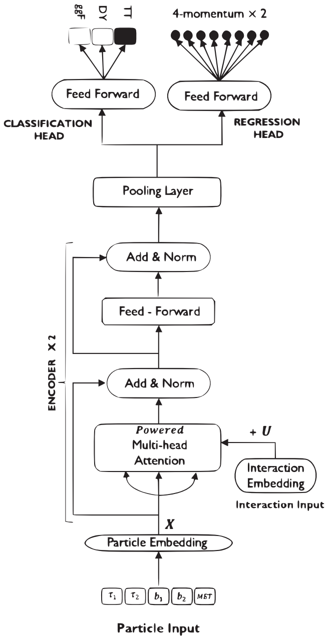
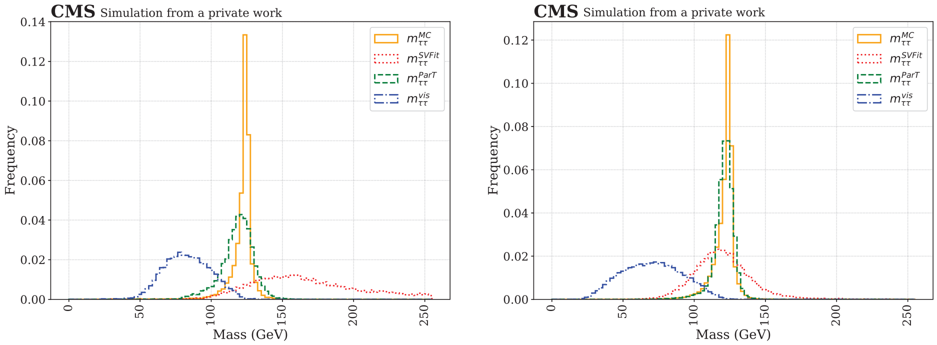


Fig. 1. – ParT architecture.

four-momentum, and the task of classification of the event (*i.e.,* the probability of belonging to the ggF, DY or $t\bar{t}$ class). The detailed architecture of the ParT is in fig. 1 and, as first step, particle and interaction inputs are each one followed by a Multi-Layer Perceptron (MLP) in order to project them to fixed-size vectors. Unlike transformers for NLP and vision, no *ad hoc* positional encodings has been added, since the particles in an event are permutation invariant. Then, the particle embedding X is fed into a stack of two encoder blocks to produce new embeddings via multi-head self-attention, the core of these kind of models. In this context the interaction embedded matrix U is used to augment the scaled dot-product attention (powered multi-head self-attention: P-MHA) by adding it as a bias to the pre-softmax attention weights. This allows P-MHA to incorporate particle interaction features designed from physics principles and modify the dot-product attention weights, thus increasing the expressiveness of the attention mechanism. After that, the last particle embedding is fed into the pooling layer that has the role to summarize the outputs of the encoder layers into a single fixed-size vector, used for the downstream tasks. The usage of an overall loss that takes into account the regression of the neutrinos four-momentum (MAE loss), the mass constraint ($E^2 < p^2$) (Huber loss) but also the classification (CrossEntropy loss) led to a better output, since the model optimizes the reconstruction of the mass not only based on the physics constraints, but also to efficiently separate signal and background.

3. – Results

To compare the performance of ParT and SVFit algorithms in reconstructing the SM Higgs boson mass from a resonant decay, three different indicators are studied: the mean of the reconstructed mass distribution, the resolution of the mass peak calculated as the standard deviation and the computational inference time. The first two indicators



(a) Invariant mass $\tau\tau$ final state for $M_X = 250$ GeV sample.

(b) Invariant mass $\tau\tau$ final state for $M_X = 550$ GeV sample.

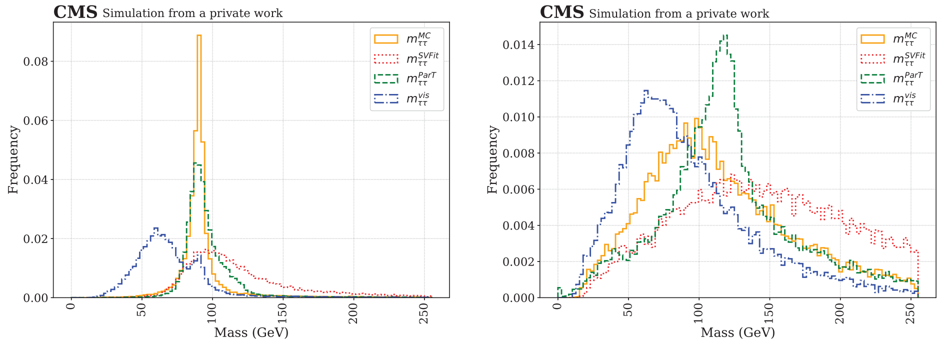
Fig. 2. – Invariant mass distributions for signal samples.

are calculated after a Gaussian fit. Regarding the multi-class classification task, the confusion matrix on the test set is discussed.

Figures 2 and 3 show the $m_{\tau\tau}$ distribution predicted by the ParT ($m_{\tau\tau}^{ParT}$), by the SVFit algorithm ($m_{\tau\tau}^{SVFit}$), computed combining Monte Carlo neutrinos four-momentum ($m_{\tau\tau}^{MC}$) and the pair of reconstructed τ, and the visible one ($m_{\tau\tau}^{vis}$), without any neutrino contribution. Table I summarizes the μ and σ results on signal and background samples.

The obtained results, compared with those estimated using the SVFit algorithm, show an invariant mass distribution of the di-τ system more centered in the target value and narrower (therefore better resolved) since it provides a smaller bias. In addition to these advantages, the ParT algorithm requires a much shorter computation time in the inference phase, of the order of milli-seconds compared to seconds of SVFit.

From the classification results, the confusion matrix on the test set showed that the 90% and 91% of ggF and DY events, respectively, are correctly identified. The $t\bar{t}$ sample is, instead, the one that is recognized with a lower precision, with a considerable 18% of events incorrectly classified as signal.



(a) Invariant mass $\tau\tau$ final state for Drell-Yan sample.

(b) Invariant mass $\tau\tau$ final state for $t\bar{t}$ sample.

Fig. 3. – Invariant mass distributions for background samples.

TABLE I. – *μ and σ after a Gaussian fit on the $m_{\tau\tau}$ computed by the analyzed algorithms.*

Sample	μ, σ ($m_{\tau\tau}^{ParT}$) [GeV]	μ, σ ($m_{\tau\tau}^{SVFit}$) [GeV]	μ, σ $m_{\tau\tau}^{MC}$ [GeV]
$M_X = 250$ GeV	122.65, 5.10	159.23, 36.63	124.02, 1.92
$M_X = 550$ GeV	123.01, 4.19	120.24, 18.16	123.85, 1.89
DY	88.98, 4.92	101.64, 21.05	89.92, 2.19
$t\bar{t}$	113.20, 52.84	150.19, 87.62	101.74, 47.49

4. – Conclusion

The mass reconstruction of the di-τ system resulting from a Higgs boson decay carries an important role in searches of new massive particles decaying into two Higgs bosons. Currently, the so-called SVFit algorithm in the CMS Collaboration is used based on a likelihood approach. One disadvantage of this approach is its high CPU time of $O(1\text{ s})$ per event, which can be a limiting factor for very big data sets. Because of this, new and computationally less expensive methods for the reconstruction are under study. In this context, the transformer architecture has demonstrated to be a competitive approach. The presented work only shows the performance of ParT on a very specific, Monte Carlo generated event sample, which is restricted to the gluon-gluon fusion production of a new massive particle. In order to generalize these results it would be necessary to train the ParT algorithm also for other production processes and the fully leptonic decay channels of the pair of τ.

* * *

The author is grateful to S. Gennai, P. Govoni, M. Paganoni and A. Biancini for their contribution to the results presented in this work.

REFERENCES

[1] CMS Collaboration, *Phys. Lett. B*, **842** (2023) 137531.

[2] Bianchini Lorenzo *et al.*, *J. Phys.: Conf. Ser.*, **513** (2014) 022035.

[3] Agostinelli, Sea *et al.*, *Nucl. Instrum. Methods Phys. Res. Sect. A*, **506** (203) 250.

[4] Dreyer Frédéric A. and Qu Huilin, *J. High Energy Phys.*, **2021** (2021) 23.

[5] Vaswani Ashish *et al.*, *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (Neural Information Processing Systems Foundation, Inc. (NeurIPS)) 2017.

[6] Qu Huilin *et al.*, *Particle transformer for jet tagging*; in *Proceedings of the 39th International Conference on Machine Learning* (PMLR) 2022, p. 18281.