Masked Auto-Encoders for Efficient End-to-End Particle Reconstruction and Identification for the CMS Experiment

Abstract

In high-energy particle physics experiments, such as Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC), effective classification of collision events is crucial. Traditional methods like the particle flow algorithm have been augmented by modern deep learning techniques, enabling improved accuracy in detecting and classifying particle events. We investigate the use of self-supervised learning and vision transformers (ViTs) for the classification of boosted top quark jets using simulated data from the CMS Open Data collection. Specifically, we employ masked autoencoders to learn representations through image reconstruction and leverage these representations for classification tasks. We find that a masked autoencoder with vision transformer architecture as base achieves an AUC score of 0.9830 for the task of classifying boosted top quark jets which outperforms baseline ResNet model.

1 Introduction

In high-energy physics experiments, such as those conducted at the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider (LHC), reconstructing particle collisions is crucial for investigating fundamental forces and particles, including searches for phenomena beyond the Standard Model (BSM) of particle physics. One key step of the reconstruction process is to properly classify the interaction process in the collision event. As an example, in this paper we will look at classification of jets which originated from either top quark or non-top quark jets.

Historically, event classification has relied on rules-based methods, such as the particle flow algorithm, [1] which clusters particle tracks and calorimeter deposits based on kinematic properties. However, machine learning (ML) approaches have recently gained prominence for their ability to extract more complex patterns from high-dimensional detector data. One recently introduced ML methodology is end-to-end reconstruction and classification in which one attempts to fully recover information about the event by looking at only minimally processed detector hit information. [2, 3, 4, 5, 6, 7]

The form of data used in these frameworks is typically constructed as either an image representation or graph representation. In the case of image representations, convolutional neural networks (CNNs) have traditionally been used. In recent years, vision transformers (ViTs) [8] have demonstrated significant potential in other image classification tasks by employing self-attention mechanisms to capture global and local features effectively.

ViT models are typically trained using supervised learning algorithms, which require large labeled datasets. However, for many tasks, there is a far greater proportion of unlabelled data as compared with the labeled data needed for supervised learning. This is also the case with many real particle physics datasets. One method for leveraging unlabelled data is Masked Image Modelling [9] which focuses on reconstruction of masked pixels. In our work, we utilize the Masked Image Modeling(MIM) approach by training masked autoencoders [10] to learn representations of particle collision images

Preprint. Under review.

through image reconstruction. These learned representations are then used to classify boosted quark jet images. By leveraging self-supervised learning, we reduce the need for extensive labeled datasets, making the classification process more efficient and scalable in the context of high-energy physics experiments.

2 Related Work

Several studies have applied end-to-end deep learning techniques to high-energy physics problems, focusing on the classification of collision events using low-level detector data. Andrews et al. [2] demonstrated the power of image-based classifiers on CMS Open Data to distinguish signal from background processes in LHC collisions, outperforming traditional kinematic approaches by leveraging the angular and energy distribution of photons. Chaudhari et al. [7] extended this work by introducing an end-to-end inference pipeline within the CMS framework, integrating ONNX and Docker for efficient event classification in real-time using GPUs.

In the domain of jet classification, Andrews et al. [3] achieved an AUC score of 0.982 for boosted top quark jet classification by processing raw detector data using end-to-end deep learning. A similar approach was taken by Andrews et al. [4] for quark versus gluon jet discrimination, demonstrating the potential of deep learning for capturing subtle differences in jet composition. Tumasyan et al. [5] introduced an innovative method for reconstructing particle decays using minimally processed data, bypassing rule-based techniques and further showcasing deep learning's effectiveness in high-energy physics.

In the broader field of computer vision, vision transformers (ViTs) introduced by Dosovitskiy et al. [8] have shown great promise in image classification tasks by leveraging self-attention mechanisms to capture both global and local features. Their work demonstrated that transformers can be scaled to large image datasets with excellent performance across various domains. Building on this, CrossViT [11] improved feature representation through cross-attention mechanisms, processing multi-scale patches for better performance.

Self-supervised learning, particularly in the form of Masked Image Modeling (MIM), has emerged as a powerful tool for unsupervised visual representation learning. BEiT [9], inspired by BERT [12], applied masked language modeling concepts to images, refining the ability to handle missing data. He et al. [10] extended this idea with Masked Autoencoders (MAEs), using an asymmetric encoder-decoder structure to efficiently learn from masked images, which has proven to be highly scalable for large-scale visual tasks.

3 Background

Classification of collision events is critically important in experiments at LHC. Traditionally, this was done by using methods such as the particle flow algorithm [1] which would cluster tracks and then attempt to classify them based on properties like spatial proximity and kinematics. Over time more modern methods have been introduced using machine learning algorithms to approach this task as laid out in section 2.

The task we explore in this paper involves using such approaches for the task of classifying boosted (high pT) top quarks. Work in [13], [14], and [15] established boosted top quarks as strong candidates for the study of Bell inequalities in the entanglement of top quarks. Should the inequalities not match theoretical expectations, this could provide support for theories such as supersymmetry which extend beyond the standard model (BSM) of particle physics. Improving classification of top quark events could further improve data resolution of these BSM searches. This provides strong motivation to explore classification on boosted top quark datasets.

4 Dataset and Pre-processing

The dataset was obtained from the CMS Open Data collection [16] contains simulated particle collision detector data generated using Madgraph 2.6.6 with Pythia6 used for parton showering with the Z2Star tune. Full info about the dataset can be found at [3]. The target signal in this case is top quark jets with a transverse momentum (pT) greater than 400GeV with a background consisting of

non-top quark jets (also with pT > 400GeV) generated through quantum chromodynamics processes. The sample contains approximately 4 million images (2mil. signal, 2mil. background). The images are obtained by cropping a window along ieta and iphi (indexed positions in the detector) around a candidate jet. Each image contains 8 distinct channels: the track transverse momentum, the longitudinal impact parameter, the transverse impact parameter, electron calorimeter energy deposition, hadronic calorimeter energy deposition, and barrel pixel layer energy deposition.

The data is further pre-processed by the following: 1) pixel values less than 1e-3 are suppressed to 0; 2) each channel is separate z-score normalized; 3) each channel is clipped outside of 500 times the standard deviation of the pixel value; 4) each channel is min-max scaled.

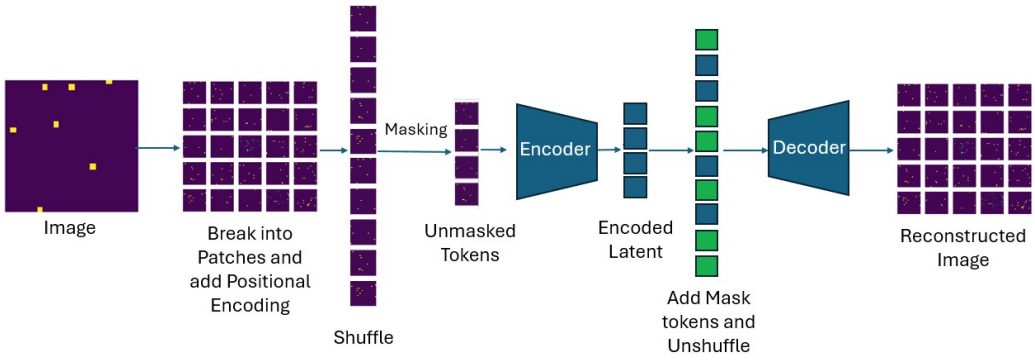

Figure 1: Flow Chart for Masked AutoEncoder

5 Model and Training

We trained different Masked AutoEncoder models and compared their performance against ResNet-15. We utilized a total of 3.8 million images, with 3 million for training and 800k split equally between validation and test sets. The 3 million training data for pre-training was split in a 2:1 ratio, where two-thirds were used for pre-training and the remaining third for fine-tuning. For fine-tuning, 1 million images were used for training, 400k for validation, and 400k for testing. We used standard evaluation metrics such as accuracy and AUC-ROC score for evaluation of our models. The BASE MAE model, with 4.9 million parameters, demonstrated the best overall performance in both linear probing and fine-tuning. This suggests that the representations learned during pre-training effectively generalized across both evaluation methods, offering high accuracy and AUC-ROC scores. Meanwhile, the Depth-Conv MAE, which uses depthwise convolutions, also performed competitively during fine-tuning, achieving a higher accuracy but slightly lower AUC-ROC score compared to the Base MAE.

Pre-training was conducted using two A100 GPUs over 47 hours, running between 80 to 100 epochs. For linear probing and fine-tuning, the models were trained for 10 epochs with a learning rate of 1.5e-4, using the AdamW optimizer. Each of the models had embedding dimensions of 128 and masking ratio set to 0.75. The detailed performance of each model is presented in Table 1.

6 Results and discussion

The experimental results demonstrate that the Base MAE model outperformed ResNet-15 when fine-tuned. Specifically, the Base MAE achieved a 0.9306 accuracy and 0.9830 AUC-ROC score,

Table 1: Model Performance Comparison

No.	Model	Accuracy	AUC-ROC Score	Trainable Params
Linear Probing				
1	Base MAE	**0.9035**	**0.9747**	385
2	Depth-Conv MAE	0.8885	0.9719	385
Finetuning				
1	Base MAE	0.9306	**0.9830**	4.9M
2	Depth-Conv MAE	**0.9376**	0.9816	4.9M
5	ResNet-15	-	0.9824	90K

compared to the ResNet-15 model's AUC-ROC of 0.9824. The Depthwise Convolution MAE model also exhibited competitive performance during fine-tuning, achieving an accuracy of 0.9376, which was higher than that of the Base MAE. However, its AUC-ROC score of 0.9816 was slightly lower.

Both the Base MAE and Depthwise Convolution models demonstrated high parameter efficiency during linear probing, with only 385 trainable parameters, achieving accuracy and AUC-ROC scores close to those of ResNet-15. This efficiency is notable as it indicates that the MAE models captured high-quality representations during pre-training, requiring minimal retraining to perform well in classification tasks. This result highlights the potential of such models to reduce computational costs while maintaining performance, making them especially useful in data-constrained environments.

Overall, these findings emphasize the parameter efficiency and strong performance of attention-based models compared to traditional architectures like ResNet-15. The scalability and generalization capabilities of MAEs make them highly promising for a wide range of classification tasks, with the potential to outperform convolutional networks, especially when pre-training on large datasets is possible.

7 Conclusion

The results of our experiments provide important insights into the potential of self-supervised learning to develop models that generalize well, even in the context of linear probing, thereby reducing both inference time and computational costs. The strong performance of these MAE models (0.9816 and 0.9830 AUC), particularly when compared to baseline architectures like ResNet-15 (0.9824 AUC), demonstrates that transformers can be effectively utilized as a base architecture for tasks such as particle identification. By employing attention mechanisms and efficient parameterization, these MAE models outperform traditional convolutional architectures during fine-tuning, and only slightly lag behind during linear probing (0.9719 and 0.9747 AUC) in which almost all of the model parameters are frozen during training, making them suitable for resource-constrained environments.

This research underscores the potential of transformer-based models in scientific applications, where accuracy, efficiency, and scalability are critical. The findings suggest that with further tuning and parameter optimization, these models could become a powerful tool for particle identification and other similar tasks.

References

[1] Florian Beaudette. The cms particle flow algorithm, 2014.

[2] M. Andrews, M. Paulini, S. Gleyzer, and B. Poczos. End-to-end physics event classification with cms open data: Applying image-based deep learning to detector data for the direct classification of collision events at the lhc. *Computing and Software for Big Science*, 4(1), March 2020.

[3] M. Andrews, B. Burkle, Y. Chen, D. DiCroce, S. Gleyzer, U. Heintz, M. Narain, M. Paulini, N. Pervan, Y. Shafi, W. Sun, E. Usai, and K. Yang. End-to-end jet classification of boosted top quarks with the cms open data. *Physical Review D*, 105(5), March 2022.

[4] M. Andrews, J. Alison, S. An, B. Burkle, S. Gleyzer, M. Narain, M. Paulini, B. Poczos, and E. Usai. End-to-end jet classification of quarks and gluons with the cms open data. *Nuclear*

Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 977:164304, 2020.

[5] CMS Collaboration. Reconstruction of decays to merged photons using end-to-end deep learning with domain continuation in the cms detector. *Phys. Rev. D*, 108:052002, Sep 2023.

[6] Andrews, Michael, Burkle, Bjorn, Chaudhari, Shravan, Di Croce, Davide, Gleyzer, Sergei, Heintz, Ulrich, Narain, Meenakshi, Paulini, Manfred, and Usai, Emanuele. Accelerating end-to-end deep learning for particle reconstruction using cms open data. *EPJ Web Conf.*, 251:03057, 2021.

[7] Purva Chaudhari, Shravan Chaudhari, Ruchi Chudasama, and Sergei Gleyzer. End-to-end deep learning inference with cmssw via onnx using docker, 2023.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[9] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[11] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[13] Yoav Afik and Juan Ramón Muñoz de Nova. Entanglement and quantum tomography with top quarks at the lhc. *The European Physical Journal Plus*, 136(9), September 2021.

[14] Zhongtian Dong, Dorival Gonçalves, Kyoungchul Kong, and Alberto Navarro. Entanglement and bell inequalities with boosted $t\bar{t}$. *Phys. Rev. D*, 109:115023, Jun 2024.

[15] CMS Collaboration. Observation of quantum entanglement in top quark pair production in proton-proton collisions at $\sqrt{s}$ = 13 tev, 2024.

[16] CMS Collaboration. Simulated dataset ttjets_hadronicmgdecays_8tev-madgraph enriched with tracker hits in aodsim format for 2012 collision data. CERN Open Data Portal, 2019.