Short Research Report
Advanced Pile-Up Distribution Estimation using Graph Neural Networks in the CMS Experiment at the Large Hadron Collider

Mahtab Jalalvandi
Isfahan University of Technology
mahtab.jalal.vandi@cern.ch

October 8, 2024

Abstract

As luminosity levels increase with Large Hadron Collider (LHC) upgrades, high-energy physics experiments like CMS face growing challenges in estimating pile-up from simultaneous proton-proton collisions. This paper presents a novel Graph Neural Network (GNN) approach for pile-up estimation, leveraging the power of these networks to represent complex event data. Our method aims to estimate the number of pile-up interactions within each event, with potential applications in studying the linearity response of luminometers - a critical step for independent luminosity measurement. We implement two GNN architectures, PUSage and PUGAT, and evaluate their performance using metrics such as L1Loss and Poisson Negative Log-Likelihood Loss. Preliminary results are promising, demonstrating robust performance and effectiveness through visualizations including heatmap plots and kernel density estimates. This research explores the transformative potential of GNNs for pile-up estimation and luminosity measurement in high-energy physics experiments, representing a significant step forward in applying advanced machine learning techniques to these challenges.

1 Introduction

1.1 Background

As the luminosity of the LHC increases, the occurrence of multiple proton-proton interactions within a single bunch crossing, known as pile-up, becomes a more significant challenge for accurate data analysis and luminosity measurements. Accurate estimation of pile-up is crucial for precise luminosity measurements, which are essential for cross-section calculations and overall data quality [2]. Traditional methods for pile-up estimation often struggle with the complexity and high dimensionality of collision data, especially under high-luminosity conditions.

1.2 Motivation

Our research is driven by the need to improve simulation quality in particle collision experiments, a critical factor for precise analysis. As we prepare for the High-Luminosity LHC (HL-LHC), developing advanced pile-up estimation methods becomes increasingly important. Additionally, our work aims to establish an independent approach for luminosity estimation, potentially reducing systematic biases in current techniques. These objectives underscore the significance of applying Graph Neural Networks to pile-up estimation in high-energy physics, with implications for both current and future experiments.

1.3 Literature Review

Recent studies have highlighted the potential of GNNs in addressing pile-up challenges in high-energy physics. For instance, Qu and Gouskos [7] demonstrated a semi-supervised GNN for particle-level pile-up noise removal, achieving significant improvements over traditional methods. Pierini et al. [13] presented a GNN-based classifier designed as a refinement of the PUPPI algorithm, demonstrating improved pile-up rejection performance compared to state-of-the-art solutions. Mikuni et al. [12] introduced ABCNet, which excels in identifying pile-up particles in high-density environments. Additionally, foundational architectures like GraphSAGE [**?**] and Graph Attention Networks (GAT) [6] have inspired our approach to enhance pile-up estimation by effectively capturing particle interactions.

2 Methodology

2.1 Data Preparation and Feature Extraction

We utilized the CMSSW framework [1] to process n-tuple data from CMS collision events. Features extracted included particle momenta (p_T, p), energies, charges, vertex information, and spatial coordinates (η,

ϕ). Additionally, we employed a custom method for identifying the nearest vertex for each particle based on the dz distance metric, calculated as:

$$dz = |z_{\text{particle}} - z_{\text{vertex}}|$$

This method assigns each particle to its closest reconstructed vertex using the `dzAssociatePV` algorithm. The nearest vertex is identified by minimizing dz, providing a key feature for the graph representation of each event. The inclusion of the nearest vertex improves how we capture spatial relationships between particles and their associated vertices, which is crucial for accurate pile-up estimation.

2.2 Graph Construction

Each collision event was represented as a graph $G = (V, E)$, where nodes $v_i \in V$ correspond to particles, and edges $e_{ij} \in E$ represent proximity or interactions between particles. This graph-based approach is inspired by recent advancements in applying graph neural networks to particle physics problems [3]. Additionally, each particle node was associated with its nearest vertex, with the vertex identifier being added as a feature to the node. This representation of the event as a graph captures both the inter-particle relationships and the relationship between particles and their nearest vertices.

Edges were defined based on a distance threshold in the η-ϕ space, ensuring that particles within a certain angular distance were connected. The threshold was set to $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.2$, following common practices in jet reconstruction [10]. The nearest vertex feature was crucial for improving the model's ability to understand particle interactions with respect to collision vertices, especially in high pile-up conditions.

Figure 1 illustrates a sample event graph generated from the HDF5 data. The nodes are color-coded based on the Particle Data Group (PDG) codes of the particles. This visualization helps to conceptualize how the event data is structured into graphs and highlights the complexity and richness of the particle interactions that the GNN models aim to process.

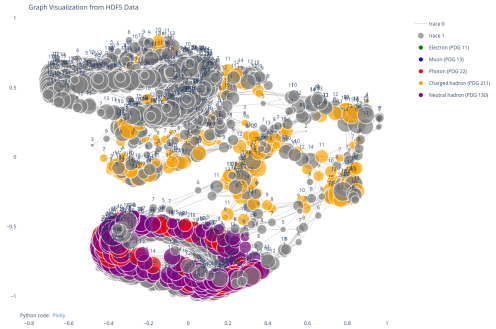

Figure 1: Event graph showing particle types and interactions based on PDG codes.

2.3 GNN Architecture

We explored two variations of Graph Neural Networks (GNNs):

- **PUSage**: This architecture is based on the GraphSAGE model [9], which aggregates information from a particle's neighborhood. In this approach, the nearest vertex information is leveraged as an additional feature during aggregation, allowing for more accurate pile-up estimation. The message passing function for PUSage is defined as:

$$h_v^{(k)} = \sigma\Bigg(W \cdot \text{MEAN}\Big(\{h_u^{(k-1)} : u \in \mathcal{N}(v)\} \cup \{h_v^{(k-1)}\}\Big)\Bigg) \tag{1}$$

  where $h_v^{(k)}$ is the feature vector of node v at layer k, $\mathcal{N}(v)$ is the neighborhood of v, and W is a learnable weight matrix.

- **PUGAT**: This model utilizes the Graph Attention Network (GAT) mechanism [6], where attention weights are computed to focus on the most relevant particle interactions. The nearest vertex feature helps the model assign higher importance to particles associated with the same vertex, improving the overall focus on interactions that are crucial for pile-up estimation. The attention mechanism in PUGAT is defined as:

$$\alpha_{ij} = \frac{\exp(LeakyReLU(a^T[Wh_i \| Wh_j]))}{\sum_{k \in \mathcal{N}(i)} \exp(LeakyReLU(a^T[Wh_i \| Wh_k]))}$$

  where α_{ij} is the attention coefficient between nodes i and j, a is a learnable attention vector, and $\|$ denotes concatenation.

Both models were trained using particle-vertex relationships as a key feature, enhancing their ability to capture complex inter-particle and vertex interactions in pile-up scenarios. The models were implemented using PyTorch Geometric [11], a library for deep learning on irregularly structured data.

2.4 Training and Optimization

The models were trained using the Adam optimizer with a learning rate of 0.001. We employed two loss functions:

1. **Mean Absolute Error (L1Loss)**:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^{N} |y_{\text{true}}^{(i)} - y_{\text{pred}}^{(i)}| \tag{2}$$

2. **Poisson Negative Log-Likelihood Loss (PoissonNLLLoss)**:

$$\mathcal{L}_{\text{Poisson}} = \sum_{i=1}^{N} \left(y_{\text{pred}}^{(i)} - y_{\text{true}}^{(i)} \log(y_{\text{pred}}^{(i)}) + \log(y_{\text{true}}^{(i)}!) \right) \tag{3}$$

The PoissonNLLLoss accounts for the Poisson nature of pile-up events, providing a more statistically robust evaluation.

3 Results

3.1 Linear Response

In high-energy physics, ensuring the linearity of detector response is crucial for accurate luminosity measurement and cross-section calculations. The linearity response tests how well the detector's response scales with the actual number of pile-up events. In this work, we evaluate the detector's linearity response under varying pile-up conditions, leveraging Graph Neural Network (GNN) models trained for pile-up estimation. As shown in Figure 2, the figure depicts a consistent linear response, with the slope close to 1 and minimal deviation from the ideal line, indicating that the detector's response maintains a high degree of linearity under varying pile-up conditions.

Model	PUSage	PUGAT	Baseline
Slope	1.01	1.01	1.00
Intercept	0.00	-0.65	0.00

Table 1: Comparison of results related to the slope and intercept for different models.

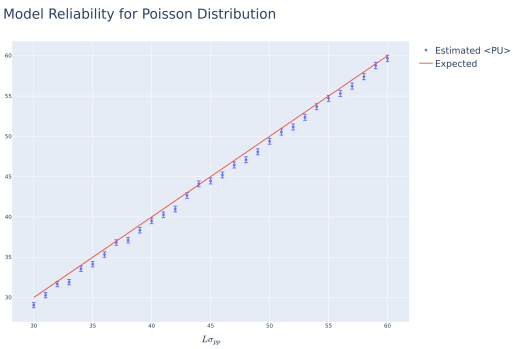

Figure 2: Model accuracy in estimating pile-up within a fixed luminosity.

3.2 Visualization

Figure 3 shows heatmap plots comparing predicted vs. true pile-up values for both GNN models and the baseline method.

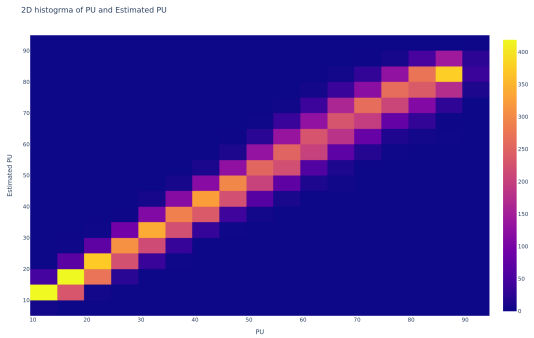

Figure 3: Heatmap plot of predicted vs. true pile-up values.

Kernel Density Estimation (KDE) plots in Figure 4 illustrate the distribution of predicted pile-up values compared to the actual distribution, demonstrating how closely the models' outputs align with the true pile-up distributions.

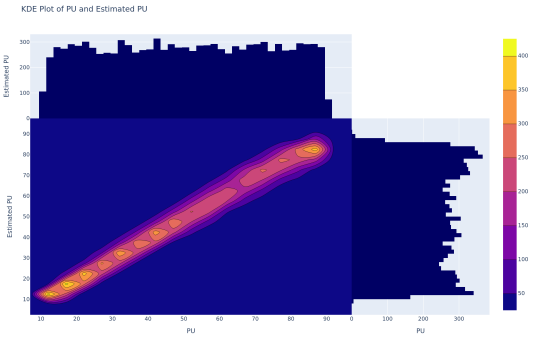

Figure 4: KDE plots of true vs. predicted pile-up distributions.

4 Discussion

The GNN models, particularly PUSage, demonstrated well performance in capturing complex particle interactions, leading to pile-up estimates and the attention mechanism in PUGAT allows the model to focus on the most relevant interactions, improving its predictive capabilities. Both models outperformed the baseline method. The use of PoissonNLLLoss provided insights into the models' performance concerning the Poisson distribution of pile-up events. Residual analysis indicated that errors were more significant in events with extremely high pile-up, suggesting areas for further improvement. Current limitations include increased computational complexity compared to traditional methods and the need for substantial computational resources. The models may also experience reduced performance in pile-up scenarios not well-represented in the training dataset. Our GNN-based approach outperforms recent machine learning methods for pile-up estimation

[5], particularly in scenarios with complex event topologies. The graph-based approach allows for better handling of varying numbers of particles per event compared to fixed-input neural networks.

5 Conclusion

This research demonstrates the effectiveness of Graph Neural Networks in estimating pile-up distributions for the CMS experiment. By leveraging the inherent graph structure of collision data, our models achieve significant improvements over traditional methods. The PUSage model, in particular, showed the best performance. The use of advanced loss functions like PoissonNLLLoss further validates the models' robustness. These results have important implications for improving data quality and luminosity measurements in high-energy physics experiments.

6 Future Work

Future research directions include:

- Investigating real-time implementation for online pile-up estimation during data acquisition.
- Extending the models to handle multi-task learning, simultaneously estimating pile-up and performing particle identification.
- Collaborating with the CMS software team to integrate the GNN models into the official data processing pipeline.
- **Integrating heterogeneous graph structures** that capture various particle types and interactions to enhance model robustness.
- Utilizing **foundation models** pre-trained on large datasets to improve training efficiency and performance.
- Employing **generative AI techniques** for synthetic data generation and data augmentation to address data scarcity and improve model generalization.
- Investigating **transfer learning** to leverage knowledge from related tasks for enhanced pile-up estimation accuracy.

References

[1] CMS Collaboration. (2008). The CMS experiment at the CERN LHC. JINST, 3, S08004.

[2] Karacheban, O. (2017). Luminosity Measurement at CMS. Brandenburgische Technische Universität Cottbus-Senftenberg.

[3] Shlomi, J., Battaglia, P., & Vlimant, J. R. (2021). Graph neural networks in particle physics. Machine Learning: Science and Technology, 2(2), 021001.

[4] CMS Collaboration. (2010). Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 7$ TeV. Physics Letters B, 722(1-3), 5-27.

[5] Sirunyan, A. M. et al. (2020). Pileup mitigation at CMS in 13 TeV data. Journal of Instrumentation, 15(09), P09018.

[6] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations (ICLR).

[7] Qu, H., & Gouskos, L. (2020). Jet tagging via particle clouds. Physical Review D, 101(5), 056019.

[8] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR).

[9] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems (pp. 1024-1034).

[10] Cacciari, M., Salam, G. P., & Soyez, G. (2008). The anti-kt jet clustering algorithm. JHEP, 2008(04), 063.

[11] Fey, M., & Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds.

[12] Mikuni, V., & Canelli, F. (2021). ABCNet: Attention-based graph neural network for particle tagging. Machine Learning: Science and Technology, 2(3), 035027.

[13] Pierini, M., Duarte, J. M., Tran, N., & Freytsis, M. (2019). Pileup mitigation at the Large Hadron Collider with graph neural networks. The European Physical Journal Plus, 134(11), 1–14.