WOTAN: Weakly-supervised Optimal Transport Attention-based Noise Mitigation

Nathan T. Suri
Department of Physics
Yale University
New Haven, CT 06511
nathan.suri@yale.edu

Vinicius Mikuni
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
vmikuni@lbl.gov

Benjamin Nachman
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
bpnachman@lbl.gov

Abstract

We improve upon the existing literature of denoising techniques studied at the Large Hadron Collider (LHC) for the task of disentangling proton-proton collisions. The primary technique that serves as the foundation for this work is known as Training Optimal Transport using Attention Learning (TOTAL). The TOTAL methodology relies on the use of a transformer architecture using a loss function inspired by optimal transport problems to learn full event descriptions. By comparing matched samples with and without noisy interactions present, the TOTAL network robustly learns an accurate description of noise as a transport function without any need for assumptions of the nature of noise derived from simulations. In this work, we develop an improved version of TOTAL known as Weakly-supervised Optimal Transport Attention-based Noise Mitigation (WOTAN) by reducing the degree of its self-supervision. The reduction of the self-supervision allows us to demonstrate the power of optimal transport-based denoising in being able to use data for particle classification instead of solely simulations. In spite of the reduced supervision, our work still outperforms existing conventional pileup mitigation approaches. Such an extension of the TOTAL methodology allows for more robust denoising, one that would truly be the first fully data-driven machine learning denoising strategy at the LHC.

1 Denoising at the Large Hadron Collider

A collision of proton bunches occurs every 25 nanoseconds at the Large Hadron Collider (LHC) at CERN; each possibly holding the key to expanding our current understanding of the world at the subatomic level either through precision measurements of key values of the Standard Model or searches for beyond the Standard Model processes. A data instance at the LHC is known as an event and corresponds to a bunch crossing of protons, yielding showers of energetic particles. On average, during the latest run of the LHC, approximately 50 simultaneous interactions, yielding charged and neutral showers collectively known as pileup vertices, were recorded per event. This number is expected to only increase at the High Luminosity LHC with predicted values as high as 200 (1). With such a high rate of data collection, the task of denoising naturally proves to be a formidable one. Not only is there systematic noise resulting from detector effects, but the aforementioned simultaneous interactions exist as physical processes of low enough data quality that are not usable for searches or

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

precision measurements. The vast majority of search analyses and precision measurements are reliant on accurate determinations of particles stemming from the primary interaction. As such, removing simultaneous interactions from data presents a salient problem that, if not checked, hinders the search for new physics as well as Standard Model precision measurements.

1.1 Existing Methodologies

As the noise sourced from noisy interactions continues to scale up, the existing data analysis pipeline struggles to keep pace in outputting results with the highest sensitivity possible. This swelling need for more effective and efficient rejection of particles from noisy interactions has directly contributed to the burgeoning sector of dedicated denoising literature within the context of the LHC. The current detectors at the LHC are well-designed to discriminate charged particles originating from noisy interactions due to superior spatial resolution of their respective tracking systems. Thus, the task of noise mitigation need only be restricted to that of neutral particle discrimination. Currently, the leading conventional denoising algorithms such as PUPPI (2) are focused on tackling the problem using a similar strategy to charged noise mitigation in that the rejection of noisy particles is done using sets of physics-motivated rules to generate per-particle probabilities on whether the particle originates from a noisy interaction or not. A key limitation of such rule-based approaches is the need to assume a consistent definition of noise either through kinematic selections or through MC simulations. In either case, deviations from the truth can cause significant mismodeling of the data and ineffective denoising.

2 Optimal Transport as a Solution

In order to avoid the limitations of rule-based denoising, significant work has been done to build new, more accurate models using machine learning, which can utilize more low-level information to output more realistic decisions. Such attempts include using image recognition techniques (3), graph neural networks (4), or transformers (5) to identify reweight particles likely to have arisen from noisy interactions and have proven to be more effective than conventional approaches. Despite their success, these existing ML implementations are limited by being fully-supervised models. Thus, they require accurate labels for each particle as to whether they originate from a noisy interaction or not. For studies with simple, limited detector simulations, this prerequisite is serviceable. However, the techniques do not translate well to more complex simulations, let alone real data as the physical granularity of detector systems would obfuscate any accurate definition of ground truth labels. For the aforementioned ML denoising techniques, the lightweight simulations are done in a software tool known as DELPHES (6), in which one identify whether a particle originated from a noisy interaction or not. However, in a full-scale simulation setup as with Geant4 (7), the realistic reconstruction makes the labeling of noisy particles ambiguous.

2.1 TOTAL

Despite the plethora of existing denoising strategies researched for the LHC, both the conventional and supervised ML denoising strategies have glaring shortcomings that hamper their effectiveness. To counter these limitations, a self-supervised approach using low-level information leveraged by an optimal transport setup was developed. Known as Training Optimal Transport using Attention Learning (TOTAL, (8)), the TOTAL methodology relies on the use of a transformer architecture (9) using a loss function inspired by optimal transport problems to learn a mapping between two samples, one containing only particles from the primary interaction (clean) and the other also containing noisy particles in addition to the particles from the primary interaction. The mechanics of the denoising done by TOTAL is described in the following section. Unlike other conventional or even ML alternatives, TOTAL does not require any assumptions of the nature of noise derived from simulations. TOTAL has already proven to outcompete existing conventional denoising techniques. Furthermore, by avoiding reliance on a ground truth or assumptions of noise derived from simulations, TOTAL proves to be far more successfully operable out of the box than its competitors in addition to its superior performance.

2.1.1 Implementation

The key to the success of the TOTAL methodology lies in the self-supervision provided to the transformer by the optimal transport-based loss function. Instead of relying on truth labels of each particle

as required by our supervised competitors, TOTAL utilizes a loss function that determines the optimal set of labels that transforms the whole noisy sample into the clean sample. Our implementation relies on the Wasserstein distance for this discrimination by leveraging geometric information present in the feature space of both samples to estimate their distance as seen in Equation 1. However, the Wasserstein distance only has a closed form solution in one dimension, leading to poor scaling in terms of computational complexity for higher dimensions. The sliced Wasserstein metric (SWD) allows us to approximate the full Wasserstein metric as an integral over one-dimensional transport problems.

$$L = \mathrm{SWD}(x'_n, x_c) + \lambda \mathrm{MSE}(p_T^{\mathrm{miss}}(x'_n), p_T^{\mathrm{miss}}(x_c)) \tag{1}$$

Here, $x_n = x_{noise} + x_c$ refers to the full sample inclusive of particles from both the primary interaction as well as noisy ones, while x_c refers to the sample only containing particles from the primary interaction. Additionally, p_T^{miss} denotes the missing traverse momentum of a sample, a key sample-level kinematic that accounts for particles unable to be directly reconstructed.

The model outputs a set of weights $\omega \in [0, 1]$ for each particle, such that particles from noisy interactions are given weights closer to zero and particles from the primary interaction occupy values closer to one. The weights are learned by minimizing the SWD as function of $x'_n = \omega x_n$ and x_c with the model being only given a limited set of kinematics for each particle $(p_T, \eta, \phi, \mathrm{charge})$. The second term in Equation 1 allows us to add additional physics constraints to the loss calculation via the mean square error (MSE) between the missing transverse momentum p_T^{miss} of each sample.

2.2 WOTAN

While TOTAL has demonstrated significant improvement over conventional rule-based techniques such as SoftKiller and PUPPI, it is not without its limitations. Chiefly, TOTAL requires that at each training instance, x_n and x_c correspond to the exact same sample. However, this is only achievable in simulations. To push TOTAL to be a fully data-driven denoising technique would require the ability to compare x_n and x_c from two independent samples. In the existing TOTAL methodology, no longer requiring that x_n and x_c are from the exact same physical event leads to a reduction in the available information given to the model as the primary interaction between compared samples is no longer guaranteed to be the same after the aforementioned shuffling. Thus, we introduce Weakly-supervised Optimal Transport Attention-based Noise Mitigation (WOTAN). WOTAN rewrites the input data into a batch as an ensemble of samples to mitigate the information loss from comparing independent samples.

$$L = \mathrm{SWD}(\psi'_n, \psi_c) + \lambda \mathrm{MSE}(p_T^{\mathrm{miss}}(\psi'_n), p_T^{\mathrm{miss}}(\psi_c)) \tag{2}$$

Unlike in Equation 1, we no longer compare single samples x_n and x_c, but instead ensembles of individual, independent samples ψ_n and ψ_c. The mathematical difference explicitly manifests when looking at the dimensions of the inputs into their corresponding loss function. In Equation 1, x_n and x_c are of size $N_{\mathrm{batch}} \times N_{\mathrm{particles}} \times N_{\mathrm{features}}$. In Equation 2, ψ_n and ψ_c are of size $(N_{\mathrm{batch}} \times N_{\mathrm{particles}}) \times N_{\mathrm{features}}$. Now transformed into ensembles of individual, independent samples, the inputs are flattened and the loss is calculated per ensemble or now equivalently per batch. In doing so, the model is able to see a single ensemble of all samples in the batch, mitigating the information loss resulting from TOTAL post-sample shuffling. By being able to compare and denoise independent samples, WOTAN supersedes TOTAL in applicability to the real world as a fully data-driven technique, not requiring any level of ground truth information and thus is able to be trained directly on data without any need for simulations.

3 Dataset

The results of this study were generated using the simulated PUMML dataset (10). The particular signal is q-qbar light-quark-initiated jets from the decay of a Higgs-like scalar particle. The noisy interactions were generated by overlaying soft QCD on top of the aforementioned signal. The dataset comes in two flavors, one with a set count of simultaneous interactions ($\mu = 140$) and a varied scalar mass and another with a varying count of simultaneous interactions and a set scalar mass (500 GeV).

The study was done using the latter dataset with simultaneous interaction counts between 140 and 145.

4 Results

For physics experiments such as the LHC, denoising exists as a means to an end. Cleaning the data of undesired noisy interactions is only a partial step towards any search for new physics or a precision measurement. Regressing any important physics observables for these types of analyses from the remaining data is a non-trivial task as desired signals still remain buried beneath high rates of known physics-processes. Without effective denoising, the physics program at the LHC would suffer greatly from poor sensitivity to desired signal processes. It is for this reason that the most effective method to assess the quality of a denoising technique is in its ability to recover unbiased regressions of key physics observables. One such observable is the average transverse momentum of the most energetic collimated collection of particles in a sample. Known as a jet, the shower of particles produced by the hadronization of quarks and gluons are key features in many interesting physics processes and thus an accurate regression of related jet observables are critical for the success of many studies at the LHC.

Figures 1a and 1c demonstrates how both the leading rule-based denoising techniques (SoftKiller (SK) and PUPPI) and WOTAN are able to reweight particle kinematics to recover the true leading jet p_T and mass distributions, albeit WOTAN showing superior regressive results. We also can see how WOTAN performs in comparison to TOTAL both with and without sample matching. As expected, shuffled TOTAL (without sample matching) performs significantly worse than all of the other denoising techniques. Even though WOTAN does not best TOTAL with sample matching, we must realize that the former is correcting the shuffled distribution to a degree better than PUPPI without requiring the nonphysical supervision of sample matching. Both the TOTAL and WOTAN results are averaged over five independent trainings to best represent its Rashomon set. Since both SK and PUPPI are functional and thus deterministic methodologies, this is not necessary. To better visualize WOTAN's performance, we look at the response of the leading jet p_T and mass resolutions. As shown in Figures 1b and 1d, WOTAN is able to recover unbiased responses in situ unlike SK or PUPPI, which requires significant manual tuning to achieve similar results.

5 Conclusions

WOTAN is a fully-data driven denoising technique for the LHC, allowing for the identification and subsequent rejection of particles from simultaneous interactions. By comparing ensembles of samples with and without noise present, WOTAN robustly learns a description of noise as a transport function, which can be used to reject particles from noisy interactions in a weakly-supervised manner. As shown, WOTAN proves itself above both its conventional and ML-based denoising competitors on three fronts. Firstly, WOTAN is able to show noticeable improvements in terms of regressing key observables by reweighting particles from noisy interactions. Secondly, WOTAN does not rely on any ground truth labels or assumptions of noise based on simulations. This provides WOTAN with the unique property of being able to be trained directly on data. Lastly, WOTAN generates per-particle labels for its noise mitigation in situ and thus does not require any manual tuning unlike conventional competitors like SK and PUPPI. While WOTAN has proven its success here for denoising tasks for particle collider experiments such as the LHC, the technique is theoretically equally powerful in other contexts, provided that a comparison between noisy and clean samples can be provided. Thus, we encourage our colleagues in other areas of the physical sciences to test WOTAN in their experiments.

References

[1] High-Luminosity Large Hadron Collider (HL-LHC): Technical design report. *CERN Yellow Reports: Monographs* **CERN-2020-010** (2020).

[2] Bertolini, D., Harris, P., Low, M. & Tran, N. Pileup Per Particle Identification. *Journal of High Energy Physics* **2014**, 59 (2014). URL `http://arxiv.org/abs/1407.6013`. ArXiv:1407.6013 [hep-ex, physics:hep-ph].

[3] The ATLAS Collaboration. Convolutional Neural Networks with Event Images for Pileup Mitigation with the ATLAS Detector (2019). URL `https://cds.cern.ch/record/2684070`.

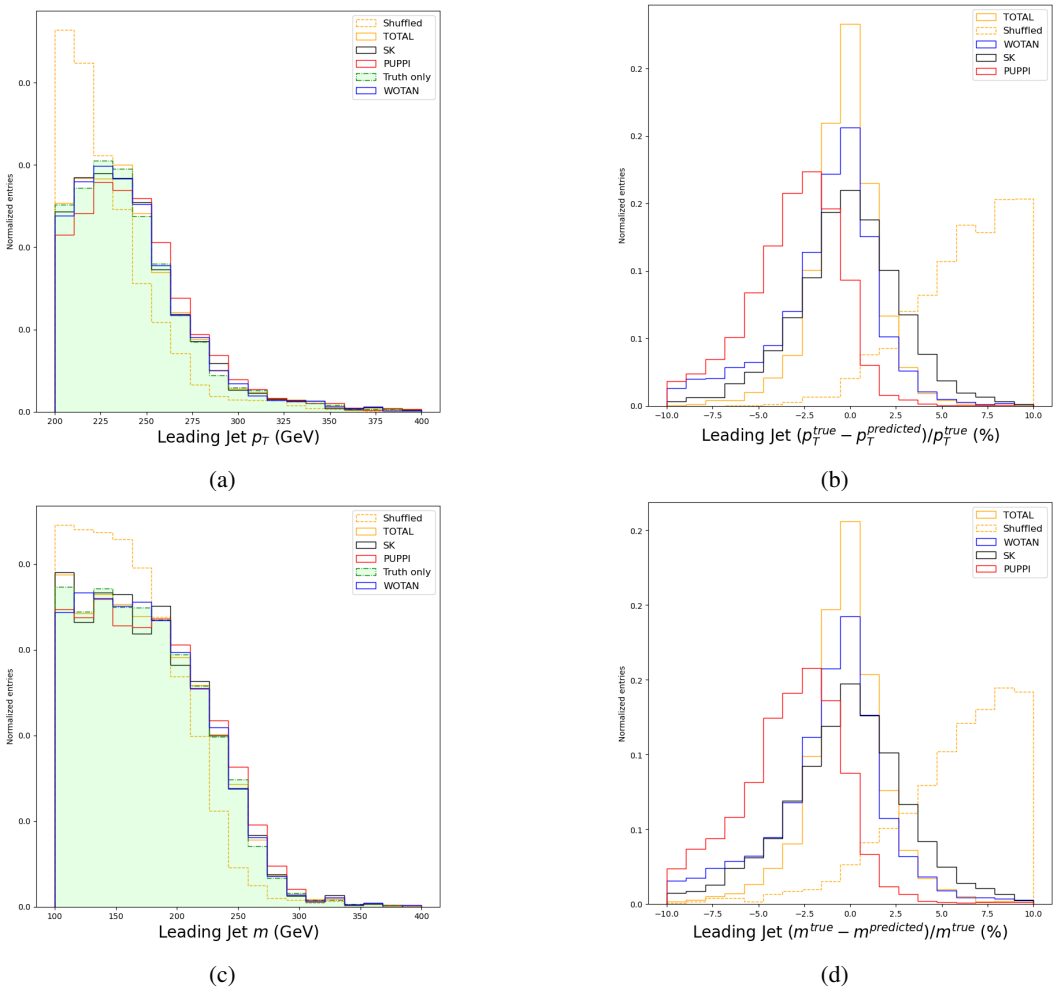

Figure 1: Results comparing the performance of WOTAN, SK, PUPPI, and TOTAL with and without shuffling with respect to leading jet p_T and mass in GeV

[4] Martinez, J. A., Cerri, O., Pierini, M., Spiropulu, M. & Vlimant, J.-R. Pileup mitigation at the large hadron collider with graph neural networks URL `https://arxiv.org/abs/1810.07988`.

[5] Maier, B. *et al.* Pile-Up Mitigation using Attention. *Machine Learning: Science and Technology* **3**, 025012 (2022). URL `http://arxiv.org/abs/2107.02779`. ArXiv:2107.02779 [hep-ex, physics:physics].

[6] de Favereau, J. *et al.* DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics* **2014**, 57 (2014). URL `http://arxiv.org/abs/1307.6346`. ArXiv:1307.6346 [hep-ex, physics:hep-ph].

[7] Agostinelli, S. *et al.* Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506**, 250–303 (2003). URL `https://www.sciencedirect.com/science/article/pii/S0168900203013688`.

[8] Gouskos, L. *et al.* Optimal transport for a novel event description at hadron colliders. *Physical Review D* **108**, 096003 (2023). URL `http://arxiv.org/abs/2211.02029`. ArXiv:2211.02029 [hep-ph].

[9] Mikuni, V. & Canelli, F. ABCNet: An attention-based method for particle tagging. *The European Physical Journal Plus* **135**, 463 (2020). URL http://arxiv.org/abs/2001.05311. ArXiv:2001.05311 [hep-ph, physics:physics].

[10] Komiske, P. T., Metodiev, E. M., Nachman, B. & Schwartz, M. D. Pileup Mitigation with Machine Learning (PUMML). *Journal of High Energy Physics* **2017**, 51 (2017). URL http://arxiv.org/abs/1707.08600. ArXiv:1707.08600 [hep-ex, physics:hep-ph, stat].