FERMILAB-CONF-22-561-AD-SCD
5th North American Particle Accel. Conf. NAPAC2022, Albuquerque, NM, USA JACoW Publishing
ISBN: 978-3-95450-232-5 ISSN: 2673-7000 doi:10.18429/JACoW-NAPAC2022-MOPA28

SEMANTIC REGRESSION FOR DISENTANGLING BEAM LOSSES IN THE FERMILAB MAIN INJECTOR AND RECYCLER

M. Thieme*†, J. Arnold, M. R. Austin, M. A. Ibrahim, K. J. Hazelwood*
V. P. Nagaslaev, A. Narayanan‡, D. J. Nicklaus, G. Pradhan, A. L. Saewert
B. A. Schupbach, K. Seiya, R. M. Thurman-Keup, N. V. Tran, D. Ulusel
Fermi National Accelerator Laboratory§, Batavia, IL USA
H. Liu, S. Memik, R. Shi
Northwestern University, Evanston, IL USA

Abstract

Fermilab's Main Injector enclosure houses two accelerators: the Main Injector (MI) and the Recycler (RR). In periods of joint operation, when both machines contain high intensity beam, radiative beam losses from MI and RR overlap on the enclosure's beam loss monitoring (BLM) system, making it difficult to attribute those losses to a single machine. Incorrect diagnoses result in unnecessary downtime that incurs both financial and experimental cost. In this work, we introduce a novel neural approach for automatically disentangling each machine's contributions to those measured losses. Using a continuous adaptation of the popular UNet architecture in conjunction with a novel data augmentation scheme, our model accurately infers the machine of origin on a per-BLM basis in periods of joint and independent operation. Crucially, by extracting beam loss information at varying receptive fields, the method is capable of learning both local and global machine signatures and producing high quality inferences using only raw BLM loss measurements.

READS OVERVIEW

The Real-time Edge AI for Distributed Systems (READS) project is a collaboration between the Fermilab Accelerator Division and Northwestern University. The project has two main goals: 1) to create a Machine Learning (ML) system for real-time beam loss de-blending in the Main Injector (MI) accelerator enclosure [2], and 2) to create a separate ML system for slow spill regulation in the Delivery Ring [3] used in the Mu2e experiment [4, 5]. In this paper, we extend our previous work [6] and introduce a novel approach to beam loss de-blending inspired by semantic segmentation models originally developed for biomedical imaging [7].

Beam Loss De-blending

The MI and RR accelerators share a tunnel and one beam loss monitoring (BLM) system. When originally constructed, the 8 GeV permanent magnet Recycler was used as an anti-proton storage ring for the Tevatron collider [8]. As the the 8 GeV anti-proton losses from RR were relatively insignificant compared with the 120 GeV proton losses from MI, there was little need to monitor ionization beam losses from RR. However, when the Tevatron was decommissioned, RR was re-purposed as a proton stacker for MI 120 GeV NuMI beam operation [9] as well as for 8 GeV Muon g-2 experiment beam delivery [10]. As a consequence, normal operation of the accelerator complex sees high intensity beams in both Main Injector and RR simultaneously, and beam losses from both machines are now a significant concern. However, while the origin of radiative losses measured on any of the 259 operational BLMs can be difficult to attribute to a single machine, experts can often attribute losses to either MI or RR based on timing, machine state, and physical location within the ring.

Using streamed, distributed BLM readings and real-time ML inference hardware, this project aims to replicate and then improve upon the machine expert's ability to de-blend, or disentangle, each machines' contribution to the measured losses.

PRELIMINARIES

BLMs are spaced (approximately) evenly within the tunnel and report the incident flux in mR/s. Because this flux is generated when beam is lost from the accelerators, i.e. when beam scrapes the edges of the beampipe and generates a spray of particles that then exit the accelerator, we often refer to it as 'loss'. When we discuss the 'BLM loss profile' we are referring to the pattern of flux measurements over the BLMs at a given time. This is not to be confused with 'loss' in machine learning, which refers to the penalty incurred for prediction errors. In this paper, when we refer to 'loss' or 'BLM loss', it is these flux measurements that we refer to.

TRAINING ON BLM LOSS PROFILES

Following recent progress in Pirate Card development [11], which now allows for the collection of high frequency (333 Hz) data in real-time directly from the BLMs, we have constructed a training dataset using actual accelerator operations data.

A single training example is composed of a single BLM loss profile, also called a 'tick', collected at some time i, and is represented as a 1D vector $x_i \in \mathbb{R}^{259}$ in which each of the 259 element represents the flux over each of the 259

* Equal contribution.
† Performed at Northwestern University with support from the Departments of Computer Science and Electrical and Computer Engineering.
‡ Also at Northern Illinois University, DeKalb, IL USA.
§ Operated by Fermi Research Alliance, LLC under Contract No.De-AC02-07CH11359 with the United States Department of Energy. Additional funding provided by Grant Award No. LAB 20-2261 [1].

ISBN: 978-3-95450-232-5 ISSN: 2673-7000 doi:10.18429/JACoW-NAPAC2022-MOPA28

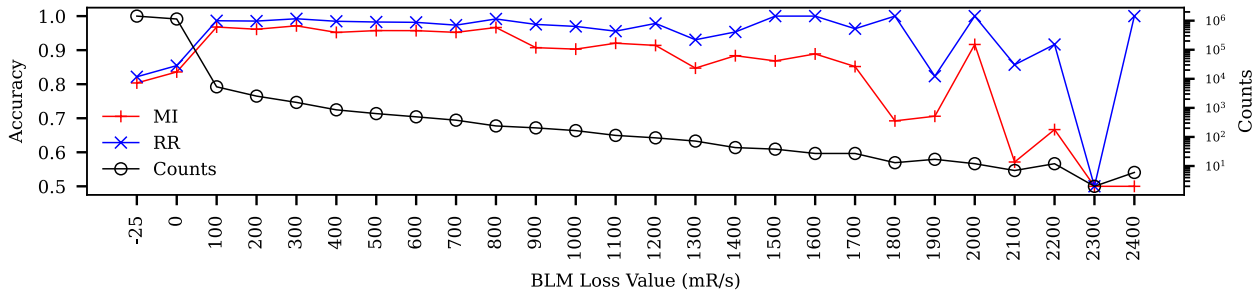

Figure 1: Accuracy is high in the primary region of interest, between 10 and 2500 mR/s. Counts reflects the number of observations with loss in that range. Note that noise in the accuracy is driven by the declining counts with a given loss value.

BLMs at time i. To date, we have collected and processed data from 12M such ticks at 33 Hz, 9M of which are used for training, 1M for validation and 2M for testing.

We take a supervised learning approach to attributing BLM losses to one machine or another so, in order to generate a training signal, we also need targets for each example. Targets for a given tick are defined by the operational states of MI and RR at that time. Specifically, whether or not each accelerator is carrying beam. For a given tick $x_i \in \mathbb{R}^{259}$ we construct a target vector $y_i \in [0,1]^{259\times k}$, where the number of classes $k = 2$ and the value of the label reflects whether MI or RR are carrying beam.

When neither machine is in operation, values measured on the BLMs could not have originated in either machine and must be background noise. Labels $y_{i,j}$ for each BLM_j in this setting would be assigned values $[0., 0.]$ as the probabilities that the loss originated in either machine are zero. It is important to include these during training to ensure the model is able to distinguish the signatures of each machine from background noise. Next, settings in which only a single machine is in operation at once, referred to here as 'single operation', the losses measured on all BLMs could only have originated in that machine. In this case, depending on whether MI or RR are the single machine in operation, the labels for each BLM_j at that tick will be either $[1., 0.]$ or $[0., 1.]$, for MI or RR, respectively.

In the next setting, when both machines are in operation simultaneously, referred to here as 'joint operation', we do not have a clean way to attribute the losses to one machine or the other. Indeed, this is the difficulty motivating the project. Data collected during periods of joint operation are handled in one of two ways: 1) We hold these data out during training and use them only for testing, relying on expert machine operators to assess the quality of the predictions. This is what we will discuss in our Results section. 2) We can construct synthesized data which approximate, to the best of our ability, the relative percentages of the loss originating in MI or RR. In the second case, synthetic data and labels can be created by summing loss profiles from periods of single-operation of each machine (controlled for machine states) and, for each label, normalizing the target probabilities for MI and RR to sum to 1. As training has not been completed on synthetic data, we confine our discussion to results obtained using method 1).

SEMANTIC REGRESSION

Here, we motivate the architectural choices made in designing our model as well as detail how our semantic *regression* model differs from the popular UNet [7] architecture for semantic *segmentation*.

Most generally, for each input example x_i, we are looking for an estimator $f(\cdot)$ parameterized by learnable parameters θ mapping 1D BLM loss profiles onto predicted class probabilities for each BLM:

$$f_\theta : x_i \in \mathbb{R}^{259} \rightarrow \hat{y}_i \in [0,1]^{259\times k} \quad (1)$$

where, for each index j corresponding to BLM_j at tick i, $\hat{y}_{i,j} \in [0,1]^{1\times 2}$ reflects the probability that the loss measured on BLM_j at tick i originated in either of the $k = 2$ machines (MI or RR). This task is similar to one known as semantic segmentation. The object of semantic segmentation - considered primarily in computer vision applications [12] - is to classify each pixel in the input into one of k categories. For example, in biomedical imaging, we may want to know whether some pixel contains normal or abnormal tissue or, in autonomous driving settings, we can imagine the utility of knowing whether some pixel represents road or sidewalk.

The UNet architecture has seen wide-ranging success in such settings. At a high level, it is composed of cascaded convolution operations between two distinct halves. In the first half, known as the contracting path, we apply repeated convolution operations, each time increasing the channel dimension while the original spatial dimensions are downsampled via pooling. In the second half, known as the expanding path, we invert the operations of the first, de-convolving at each layer and decreasing the channel dimension, until we arrive at an output with the same spatial dimension as the original input image but with k channels corresponding to k classes. To better incorporate information from varying receptive fields and avoid information loss, feature maps are passed layer-wise from the contracting path to their analog in the expanding path. Finally, softmax is applied over the channel dimension of the final layer and class labels are obtained.

ISBN: 978-3-95450-232-5 ISSN: 2673-7000 doi:10.18429/JACoW-NAPAC2022-MOPA28

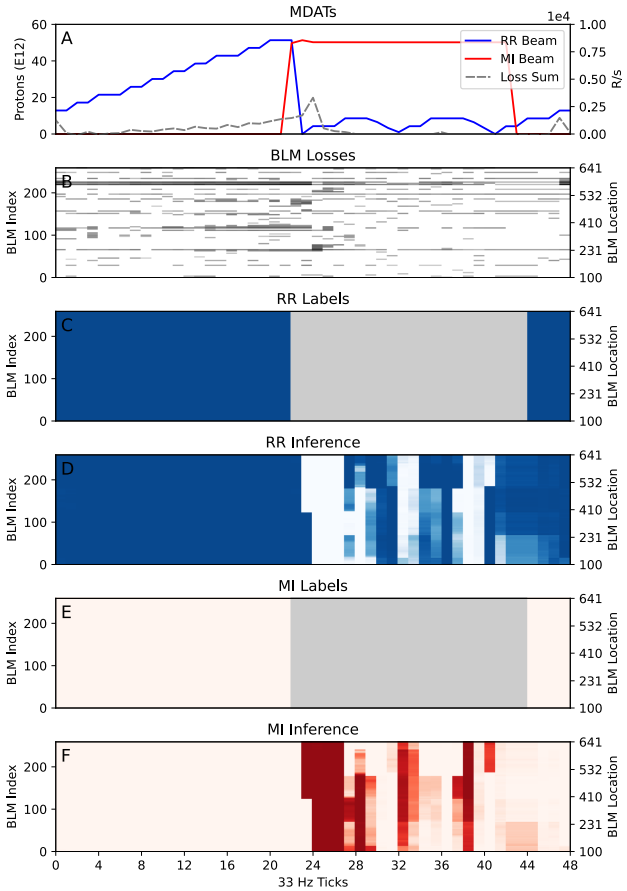

Figure 2: Inferences made on BLM losses during a period of joint operation. A: Beam intensities (R/s) in RR and MI over 48 ticks. B: BLM loss profiles - darker = more loss. C and E: Labels for RR and MI, where gray indicates that the machine of origin is unknown. D and F: UNet model inferences for RR and MI, respectively. Intensity corresponds to the inferred probability that the losses on a particular BLM at a particular loss originated in RR or MI.

While similar to the original UNet, our model differs in two important ways: 1) our input $x_i \in \mathbb{R}^{259}$ is 1D, so we replace 2D convolutions with 1D convolutions, and 2) we are not trying to predict a *single class* per BLM, but *probabilities for each class*, each of which may scale independently between 0 and 1, so we replace UNet's final softmax layer with a sigmoid. Losses are calculated using the MSE between predictions and labels: $\ell = \mathrm{MSE}(\hat{y}_i, y_i)$.

RESULTS

Figure 1 details prediction accuracy per-label in periods of single operation (MI *or* RR *or* neither, i.e. labels [1., 0.], [0., 1.] or [0., 0.]) vs. BLM loss value. We report accuracy here as it is a more interpretable quantity than MSE, using a 20% threshold for the accuracy calculation, i.e. if a prediction is further or closer that 0.2 away from the label, we consider the inference incorrect or correct, respectively. This figure shows that our model is able to learn robust representations of each machine's unique signature using only BLM loss profiles. It also shows that prediction performance is highest on BLM loss values of interest, namely those above the background noise (10 mR/s) and below the existing abort threshold (2500 mR/s). While accuracy appears to fall above 1800 mR/s, this is primarily due to a paucity of data in this range, of which the counts in log-scale are visible on the right axis.

Inference on Losses of Unknown Origin

Of central interest in Fig. 2 are panels D and F, which contain inferences made during periods of joint operation. Panel A shows the circulating beam intensity in each machine as a function of time (ticks), B shows the BLM loss profiles collected for each tick (these are the inputs $x_i \in \mathbb{R}^{259}$ that our model ingests), and C/D displays the labels, where the gray indicates regions in which we cannot assign a unique label. These inferences agree well with known behavior in MI and RR, and we call the readers attention to two features of these inferences that demonstrate the model's ability to generalize beyond the labeled training data.

First, in panel D, the heavy band of RR inferences starting around BLM index 200 are in agreement with known behaviors in RR. The two RR bumps between ticks 24 and 42 are in the beam to Muon campus, and reflect a process that involves coalescing 53 MHz RF bunched beams into larger 2.5 MHz RF bunches. During this process, some beam is lost, and ultimately the beam is extracted around BLM index 190. Since some portion of the beam is not captured during the coalescing process, this beam continues around in the machine until the end of cycle where it is finally lost at the RR collimators and abort line around BLM index 210. This region after BLM index 200 is where we would expect large RR contributions to the loss, and the model agrees across both RR bumps.

Second, in panel F, ticks 32 and 38 happen to lie in the few milliseconds between events where there is no beam in RR. Our model's high confidence that the losses in these ticks originated in MI is likely due to the line interpolation and lower frequency data rate. Our labeling system cannot label this type of event correctly every time on 33 Hz data, however, it is likely there are samples where it is labeled correctly because the sample happened to occur at an opportune time in which losses and beam events overlapped. What we can gather from F is that our model was able to learn these profiles and correctly impute that, given the loss profiles at ticks 32 and 38, the loss likely originated in MI.

Future Work

Towards the goal of real-time inference, this model is being adapted for implementation on FPGA. We are also exploring additional methods for improving the generalization and accuracy of the model in periods of joint operation, including the generation of representative synthetic data, the incorporation of temporal information, and further statistical analysis of synergistic effects observed during periods of joint operation.

ISBN: 978-3-95450-232-5 ISSN: 2673-7000 doi:10.18429/JACoW-NAPAC2022-MOPA28

REFERENCES

[1] Department of Energy, Office of Science, "Data, Artificial Intelligence, and Machine Learning at DOE Scientific User Facilities," Tech. Rep. DOE National Laboratory Program Announcement Number: LAB 20-2261, 2020. https://science.osti.gov/-/media/grants/pdf/lab-announcements/2020/LAB_20-2261.pdf

[2] K. Seiya *et al.*, "Accelerator Real-time Edge AI for Distributed Systems (READS) Proposal," *arXiv preprint*, 2021. doi:10.48550/arXiv.2103.03928

[3] A. Narayanan *et al.*, "Optimizing Mu2e Spill Regulation System Algorithms," in *Proc. IPAC'21*, Campinas, Brazil, May 2021, pp. 4281–4284. doi:10.18429/JACoW-IPAC2021-THPAB243

[4] L. Bartoszek *et al.*, "Mu2e Technical Design Report," Tech. Rep. FERMILAB-TM-2594, FERMILAB-DESIGN-2014-01, 2014. doi:10.2172/1172555

[5] V. Nagaslaev and others, "Third Integer Resonance Slow Extraction Using RFKO at High Space Charge," Tech. Rep. FERMILAB-CONF-11-475-AD, TRN: US1200088, 2011. https://www.osti.gov/biblio/1031169

[6] K. J. Hazelwood *et al.*, "Real-Time Edge AI for Distributed Systems (READS): Progress on Beam Loss De-Blending for the Fermilab Main Injector and Recycler," in *Proc. IPAC'21*, Campinas, Brazil, May 2021, pp. 912–915. doi:10.18429/JACoW-IPAC2021-MOPAB288

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *arXiv preprint*, 2015. doi:10.48550/arXiv.1505.04597

[8] G. Jackson, "The Fermilab Recycler Ring Technical Design Report. Revision 1.2," Tech. Rep. FNAL-TM-1991, ON: DE97051388; BR: KAHEP; TRN: US200506%%505, 1996. doi:10.2172/16029

[9] R. Ainsworth *et al.*, "High Intensity Proton Stacking at Fermilab: 700 kW Running," in *Proc. HB'18*, Daejeon, Korea, Jun. 2018, pp. 136–140. doi:10.18429/JACoW-HB2018-TUA1WD04

[10] J. Grange *et al.*, "Muon (g-2) Technical Design Report," *arXiv preprint*, 2018. doi:10.48550/arXiv.1501.06858

[11] J. Berlioz *et al.*, "Synchronous High-Frequency Distributed Readout for Edge Processing at the Fermilab Main Injector and Recycler," presented at NAPAC'22, Albuquerque, New Mexico, USA, Aug. 2022, paper MOPA15, this conference.

[12] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," en, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022. doi:10.1109/TPAMI.2021.3059968