Machine Learning Aboard the ADAPT Gamma-Ray Telescope

Ye Htet
Marion Sudvarg
Andrew Butzel
Jeremy D. Buhler
Roger D. Chamberlain
James H. Buckley

Ye Htet, Marion Sudvarg, Andrew Butzel, Jeremy D. Buhler, Roger D. Chamberlain, and James H. Buckley, “Machine Learning Aboard the ADAPT Gamma-Ray Telescope,” in *Proc. of Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis* (SC-W), November 2024.

Dept. of Computer Science and Engineering
Washington University in St. Louis

Dept. of Physics
Washington University in St. Louis

Machine Learning Aboard the ADAPT Gamma-Ray Telescope

Ye Htet*, Marion Sudvarg†, Andrew Butzel*, Jeremy D. Buhler*, Roger D. Chamberlain*, James H. Buckley†
*Department of Computer Science and Engineering, †Department of Physics
Washington University in St. Louis
St. Louis, Missouri, USA
{htet.ye, msudvarg, a.butzel, jbuhler, roger, buckley}@wustl.edu

***Abstract*—The Advanced Particle-astrophysics Telescope (APT) is an orbital mission concept designed to contribute to multi-messenger observations of transient phenomena in deep space. APT will be uniquely able to detect and accurately localize short-duration gamma-ray bursts (GRBs) in the sky in real time. Current detection and analysis systems require resource-intensive ground-based computations; in contrast, APT will perform on-board analysis of GRBs, demanding analytical tools that deliver accurate results under severe size, weight, and power constraints.**

In this work, we describe a neural network approach in our computation pipeline for GRB localization, demonstrating the capabilities of two neural networks: one to discard signals from background radiation, and one to estimate the uncertainty of GRB source direction constraints associated with individual gamma-ray photons. We validate the accuracy and computational efficiency of our networks using a physical simulation of GRB detection in the Antarctic Demonstrator for APT (ADAPT), a high-altitude balloon-borne prototype for APT.

***Index Terms*—machine learning, neural networks, multi-messenger astrophysics**

I. Introduction

The Advanced Particle-astrophysics Telescope (APT) is a mission concept for a space-based observatory designed to observe high-energy gamma rays and cosmic rays in support of multi-wavelength and multi-messenger astrophysics [1]–[5]. APT will promptly detect energetic transient events in the distant universe, especially gamma-ray bursts (GRBs) in the MeV energy range, and will rapidly communicate these events to narrow-band instruments for follow-up observation. APT will be deployed in a Sun-Earth Lagrange L_2 orbit, achieving nearly full-sky field of view and order-of-magnitude improvement in GRB detection sensitivity compared with current instruments such as the Fermi Gamma-ray Space Telescope [2], [6], [7]. The Antarctic Demonstrator for APT (ADAPT), a small-scale technology demonstration mission for APT's hardware design and computational capabilities, will launch using a high-altitude balloon in late 2025 [4], [8]–[10].

APT will infer a GRB's location in the sky by observing how the GRB's gamma rays scatter in its detector. To inform rapid follow-up observations by other telescopes with narrow fields of view, GRBs must be localized with high accuracy, ideally to within a degree or less. Moreover, the dim, short-duration GRBs of particular interest to APT might be visible for only a few seconds – less than the 5-second light-speed delay expected for communication from the instrument at L_2 to a terrestrial computing facility. Hence, to avoid communication delays, and to support a possible co-located optical telescope at L_2, the computations to detect and localize GRBs must occur on the APT platform in space (with all its implied limitations on computational power) in real time.

The latency and resource constraints imposed by APT's operation preclude traditional, compute-intensive offline analysis to correct for uncertainties caused by factors such as instrument noise and background radiation. We have instead designed a real-time analysis pipeline for APT that runs on a low-power heterogeneous computing system utilizing ASICS, FPGAs, and multicores, which we previously described in [4], [10]. That prior work demonstrated both rapid localization and high accuracy; however, as we refine our physical models of both the detector hardware [9] and the sensing environment [8], maintaining these desirable properties becomes ever more challenging, requiring new computational approaches.

In this work, we introduce a machine-learning component to our pipeline to improve localization accuracy in the presence of instrument noise and background radiation. We design neural network models to efficiently address two challenges: first, the need to reject signals caused by naturally occurring background radiation particles unrelated to any gamma-ray burst; and second, the inability of existing analytical methods based on propagation of error to correctly estimate uncertainty in the information (the *Compton ring*) inferred from each incident gamma-ray photon. We consider not only the accuracy of our networks but also their computational suitability for our low-latency, low-resource application.

We validate our networks using a detailed simulation of the detector hardware and sensing environment for the ADAPT demonstrator. ADAPT has a much smaller detector than APT (and so sees fewer gamma rays per GRB) and will be subject to diffuse background radiation from Earth's atmosphere, so it actually creates a *more* challenging detection task than for APT at a given GRB brightness. We demonstrate that, with our novel machine-learning components, ADAPT can localize GRBs of brightness one to a few MeV/cm^2 with significantly less error versus the prior pipeline without ML. Moreover, adding ML to the pipeline incurs reasonably low additional

Supported by NASA award 80NSSC21K1741. Author A. B. was supported by an NSF REU site award. Author M. S. was supported by Washington U. seed grant CC0001285 (PJ000030737).

latency. For background rejection, this cost can be reduced by quantization, especially for FPGA-based deployment, while preserving much of the original model's accuracy.

II. Background and Motivation

A. ML for Multi-Messenger Astrophysics

Multi-messenger astrophysics [11], [12] aligns observations of transient cosmological phenomena from multiple instruments over several modalities (gravitational waves, neutrinos, electromagnetic) to provide insights into their evolution. Many instruments perform omnidirectional event detection but rely on compute-intensive pattern-matching searches through raw data [13]–[15] and so must run on large clusters [16].

Deep learning is a promising approach to accelerate and improve the accuracy of pattern-matching search. In gravitational-wave astronomy, neural networks detect transient events such as binary neutron star mergers [17] and can process months of data from the LIGO detector within 50 seconds [18]. In the electromagnetic realm, the AGILE X-ray/gamma-ray satellite uses an anomaly detection autoencoder convolutional neural network (CNN) to detect GRBs in real time using data from its anticoincidence system [19]. COSI, a NASA Small Explorer satellite currently in development, may use multi-layer perceptrons (MLPs) to reconstruct the trajectories of individual gamma-ray photons in its detector [20]. Such approaches are often deployed on ground-based computers, but the limited communication bandwidth of space-based instruments can incur substantial delay before analysis.

To minimize the volume of data sent to the ground, some future space-based missions plan to perform data reduction and analysis *in orbit*. For example, the POLAR-2 GRB polarimeter will be installed on the Tiangong space station in 2025 or '26, where it will have access to a GPU. This will allow it to provide GRB alerts with degree-scale localization within two minutes of detection [21]. However, we address the more challenging problem of efficient and accurate localization under tight size, weight, and power constraints – e.g., aboard a standalone satellite such as APT [2].

B. The ADAPT Detector

The Antarctic Demonstrator for the Advanced Particle-astrophysics Telescope (ADAPT) will serve as a technology demonstrator for APT, launching on a high-altitude balloon at the end of 2025. It will detect and localize GRBs using on-board computational hardware, showing the feasibility of doing so on the future APT mission.

ADAPT's gamma-ray detector has four layers of scintillating tiles that emit light when incoming gamma-ray photons scatter within them. This light is captured by perpendicular arrays of wavelength shifting (WLS) optical fibers that line the top and bottom surfaces of each tile, then measured by silicon photomultipliers (SiPMs) placed at the end of each fiber, as shown in Figure 1. This overlay of 1-dimensional fiber arrays into a 2-dimensional mesh, along with the relative position of the tile, resolves the 3-dimensional position of each gamma-ray/scintillator interaction. Each gamma-ray photon

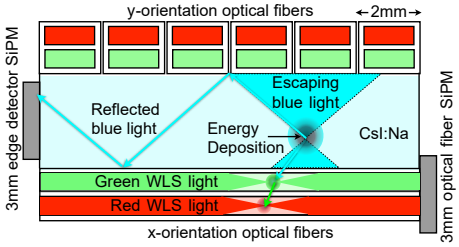

Fig. 1. Light collection in detector (from [2]).

may Compton-scatter multiple times in one or more layers before it is photoabsorbed.

Hereafter, we refer to the measurements of a single gamma-ray photon as an *event*, which consists of a list of its interactions, or *hits*, within the detector. Each hit i has an associated 3-dimensional vector of spatial coordinates $\mathbf{r}_i$ and an amount of energy E_i that it deposits within the detector. Using the information from each hit, we can reconstruct the photon's likely trajectory through the detector as described in [3], [4], [22]. The result of this reconstruction is a *Compton ring* (see Figure 2) that constrains the angle (equivalently, its cosine η) between the vector $\mathbf{c}$ through the spatial positions of the photon's first and second hits in the detector, and the direction vector $\mathbf{s}$ pointing to its source (the GRB) in space. Given Compton rings reconstructed from multiple events, their common source direction (i.e., the location of the GRB) can be derived by intersecting the rings in a process known as *localization*.

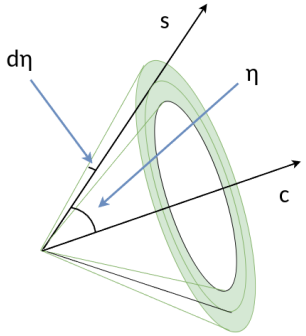

Fig. 2. A reconstructed Compton ring, for the true GRB source vector s, is defined by the vector c through the first two hits and the cosine η and uncertainty $d\eta$ of the scattering angle.

Computational Pipeline: Our GRB analysis pipeline performs both reconstruction of individual event trajectories and localization of the burst from the resulting set of Compton rings. Hereafter, we focus on the localization stage.

For each Compton ring that enters localization, we have not only its parameters $\mathbf{c}$ and η but also an estimate of the uncertainty $d\eta$ in the ring's opening angle. The GRB's source direction $\mathbf{s}$ is likely to lie within the area defined by the ring and its uncertainty (the green area in Figure 2)[1]. A "thicker" ring (larger $d\eta$) corresponds to greater uncertainty in the photon's source. Following prior work [22], we estimate $d\eta$ by propagation of error from the detector's known uncertainties in the measured position and energy of each hit.

The localization computation [3] intersects the Compton rings from all collected events to infer a single common source direction $\mathbf{s}$ for their photons. Because of the uncertainties $d\eta$, the rings rarely intersect at a single point, so we must instead derive a maximum likelihood estimate of $\mathbf{s}$ using a probabilistic model of the actual source direction given the

[1]Specifically, $d\eta$ parameterizes the width of a radially symmetric Gaussian probability density for the source direction centered at radius $\cos^{-1}\eta$.

rings. Moreover, not every ring entering localization passes near the true source direction. Some rings may be incorrectly reconstructed by earlier computations [23], yielding incorrect $\mathbf{c}$ and η values, while others may arise not from GRB photons but rather from events caused by *background* radiation in Earth's upper atmosphere, as shown in Figure 3. Localization must therefore be robust enough to discard likely background events while accounting for the uncertainty inherent in each non-background Compton ring.

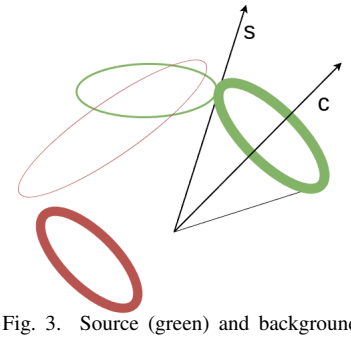

Fig. 3. Source (green) and background (red) Compton rings.

Our localization algorithm consists of two stages, *approximation* and *iterative refinement*. Approximation picks a small random sample of incoming Compton rings and considers the set of candidate source directions that lie close to at least one of these rings, choosing the direction $\mathbf{s_0}$ that maximizes the joint likelihood of the sample. Refinement uses all the rings to refine $\mathbf{s_0}$. Maximizing the joint likelihood of the source direction $\mathbf{s}$ given a set of rings R is equivalent to solving an almost-linear least-squares problem [4]. To ensure robustness against background particles, we iteratively choose the set of rings R_i with high enough likelihood given the current source estimate $\mathbf{s}_i$ and apply least-squares to R_i to infer an updated estimate $\mathbf{s}_{i+1}$, until the estimates converge to our final source prediction.

Limitations of the Existing Pipeline: The motivation for the new approaches described in the next section is twofold. First, we have observed that the uncertainty estimates $d\eta$ obtained by propagation of error are frequently incorrect. In particular, many rings have much larger actual errors in η than our estimates predict, either because our detector error model is incomplete or because the photon's trajectory was incorrectly reconstructed. False certainty about $d\eta$ can lead our likelihood model astray and result in incorrect source reconstruction. Second, although our pipeline is designed to resist the influence of incorrect Compton rings, rings caused by background particles still present a major challenge. Within the time window of a short GRB, localization typically receives 2-3$\times$ as many Compton rings from background particles as from the GRB itself. Unlike other gamma-ray missions such as COSI [24], ADAPT cannot afford a heavy anticoincidence shield, so we must suppress the background computationally.

To quantify these concerns, Figure 4 shows the accuracy of ADAPT's analysis pipeline on a simulated GRB of fluence (i.e., time-integrated brightness) 1 MeV/cm^2 occurring normally incident to the detector. The simulation used Geant4 [25] to simulate particle and gamma-ray interactions with the detector, together with a detailed model of ADAPT's electronics [9]. Localization error is reported as the angle in degrees between the true source direction and the direction inferred by our pipeline. We report two values, *68% and 95% containment*,

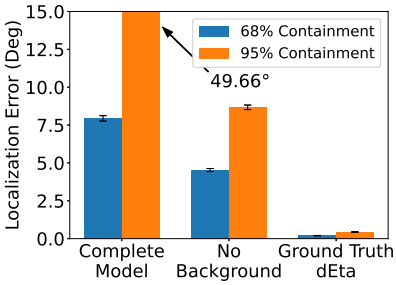

Fig. 4. Impact of background particles and $d\eta$ error on localization accuracy. Error bars are over ten meta-trials.

reflecting the largest error observed in at most 68% and 95% of 1000 trials with randomly-generated simulated particles. The leftmost results include the influence of both background radiation and inaccurate $d\eta$ estimation; the middle remove all background Compton rings, while the rightmost instead replace the pipeline's estimates of $d\eta$ with the true errors in the inferred η for each simulated ring. Fully correcting for background and $d\eta$ inaccuracy would substantially improve our pipeline's localization accuracy. While we cannot hope to achieve these best-case results in practice, we can mitigate the two sources of error using the methods of the next section.

III. Neural Networks

We now describe two neural network models that respectively attempt to correct for $d\eta$- and background-associated errors in localization. The **background network** classifies a Compton ring as originating from either a GRB photon or a background particle, while the **dEta network** performs regression on the observed properties of a Compton ring to estimate its $d\eta$ value. The models share a common set of input features and a common multilayer feed-forward structure.

Input Features: From each Compton ring reconstructed by our pipeline, we use twelve features of the associated detection event as input to our predictive models. They are: the total energy deposited by the event that produced the Compton ring; the four parameters (three spatial coordinates plus deposited energy) associated with each of the event's first and second hits, and the measurement uncertainties associated with each of the three energy measurements (total plus two deposited). These parameters, plus the uncertainty in the hits' spatial coordinates, are all the information used to reconstruct an event's Compton ring. However, ADAPT's reported events exhibit much greater uncertainty in energy than in spatial position [23]; hence, our model considers only energy uncertainty.

In addition to these twelve features, the model input includes a guess at the source direction's *polar angle*, that is, whether the particle entered the detector from above (angle 0°), from the side (angle 90°), or in between. (Earth obscures ADAPT's field of view, blocking any GRBs that originate below the detector.) Empirically, we found that prediction performance at the lowest and highest angles improves given a roughly correct estimate (to within about 10°) of this angle; a comparison is shown in Figure 7. Of course, a GRB's source direction is not known *a priori*, since that is what the pipeline ultimately computes; however, we provide the angle inferred without the use of machine learning as our initial guess.

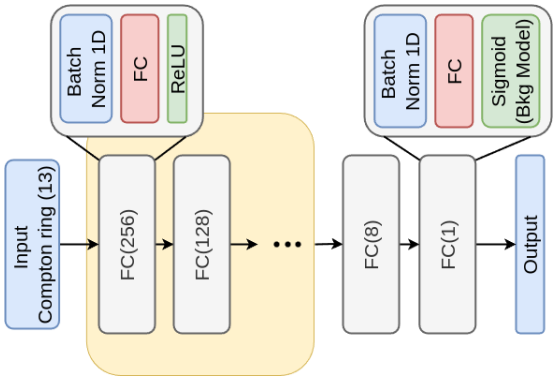

Fig. 5. Model architecture.

Model Architecture: Our models use a feed-forward architecture, inspired by work of Takashima [20], which is shown in Figure 5. The network consists of a series of *blocks*, each with a one-dimensional batch normalization (Batch Norm 1D) followed by a fully-connected layer (FC) and finally a rectified linear unit (ReLU) activation. The architecture's depth and the width of each FC layer in the yellow region of Figure 5 may be tuned as hyperparameters. The background network outputs a probability that the Compton ring arose from a background particle, which is thresholded to yield an output label. The dEta network's output is a prediction of the *natural log* of $d\eta$, as this value can range over several orders of magnitude.

Model Usage: Removal of background Compton rings and assignment of $d\eta$ values to the rest may in principle be performed immediately upon entry to localization. However, our choice to utilize polar angle as an input to the networks instead requires an iterative approach, illustrated in Figure 6, which applies the models in the *middle* of localization. We first iterate between determining a source direction $\hat{\mathbf{s}}$ (with its implied polar angle) and applying the background model using this angle to identify and remove background rings. This iteration is more effective at removing background Compton rings than a single application of the model using the first estimate of $\hat{\mathbf{s}}$. Once $\hat{\mathbf{s}}$ has converged, or after a predetermined maximum number of iterations (currently five), we update the estimated $d\eta$ of all surviving rings according to the dEta network's output, then re-run localization using the last $\hat{\mathbf{s}}$ as an initial estimate to produce the final source direction $\mathbf{s}$.

Our iterative approach allows us to trade predictive accuracy for efficiency. If the system is heavily loaded, or if our models suggest that further iteration is not needed to achieve a given level of accuracy in the source direction, we may at any point halt and report the current source direction guess $\hat{\mathbf{s}}$.

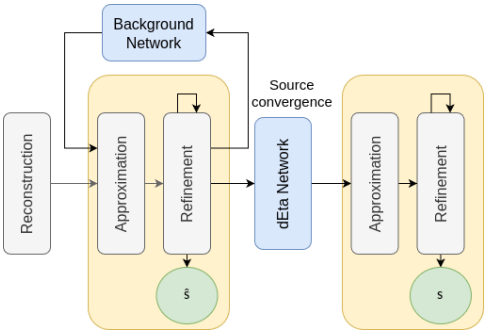

Fig. 6. Model usage.

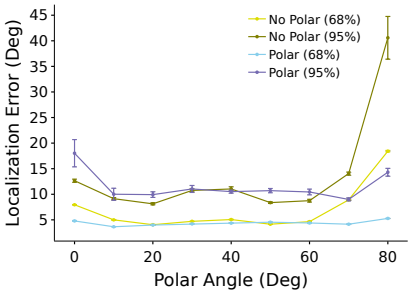

Fig. 7. Impact of including polar angle as an input.

Model Training: We train our models using simulated data from the Geant4-based detector simulations described in Section II. These simulations use the most realistic available models of the ADAPT instrument design and are the only way to obtain large numbers of incident particles for which the true source (GRB or background) and error in η are known.

Our full data set consists of 270 million GRB photons, evenly divided across nine source polar angles from $0°$ to $80°$ in ten-degree increments. For background particles, we used the models in [8] to estimate the number of events expected to occur within a 1-second exposure and generated 1350 times as many simulated events. After filtering these events through our detector model and reconstruction, we retained only rings that the pre-localization stages of the pipeline deemed correctly reconstructed. There were around one million such rings, split approximately 60%/40% between GRB photons and background particles. These rings were slightly biased toward lower polar angles, which are less likely to be rejected by earlier filters. From this reduced dataset, we used an 80/20 training/testing split, with the training set further split 80/20 for training/validation. For the dEta network, we also remove the background rings from the training set.

We trained the models using hyperparameter tuning via the Weights and Biases (WandB) platform [26] to search over different combinations of batch size, learning rate, and architectural variables including the number of FC layers, the maximum width of any layer, and the width of each layer relative to the maximum. Networks were trained using the SGD optimizer; the background network was trained using binary cross-entropy loss and the dEta network using ℓ_2 loss. Training ran for up to 120 epochs with early stopping if validation loss ceased to improve. For the background network, we divided the range of input polar angles into ten-degree bins and chose an output threshold for each bin that minimized training loss; the threshold is then selected dynamically at inference time based on the input polar angle.

For experiments in subsequent sections, we used a background network model trained using batch size 4096 and learning rate 5.204e-4, and a dEta network trained using batch size 256 and learning rate 4.375e-3. Both networks used four FC layers in total. The background network had a maximum width of 256 in its first FC layer, with subsequent layers gradually decreasing in width, with the dEta nework had a maximum width of 16 in the middle and shorter widths at the beginning and end.

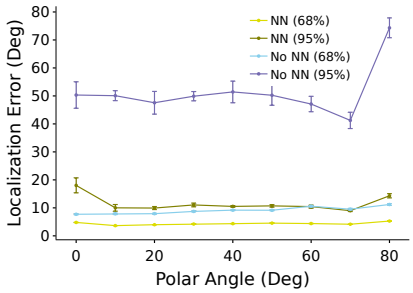

Fig. 8. Localization accuracy vs. polar angle.

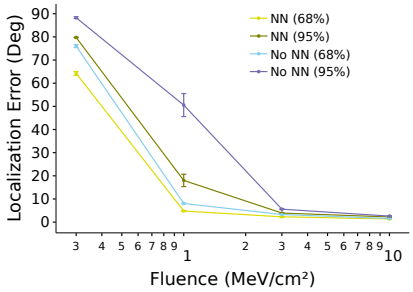

Fig. 9. Localization accuracy vs. fluence.

IV. Localization Results

To measure the impact of our neural network models on GRB localization accuracy, we test our improved pipeline using the detailed physical simulation model described in Section II. All experiments use simulated 1-second GRBs with equivalent background exposure[2] – short GRBs, characteristic of binary neutron star mergers, typically last 10ms to 2s [27]–[31]. We measure localization accuracy as a function of GRB brightness (fluence) for normally-incident bursts (polar angle 0°), as well as accuracy as a function of polar angle for a fixed brightness of 1 Mev/cm^2.

Figure 8 quantifies the impact of our neural network models on the pipeline's localization accuracy for varying incident polar angles, while Figure 9 shows the impact for varying fluences of a normally-incident GRB. Incorporating our models consistently approves accuracy, especially for the tail of the error distribution (95% containment) and for dimmer GRBs. We predict that across all polar angles, ADAPT can localize GRBs with fluence at least 1 MeV/cm^2 to within 6° of error at least 68% of the time.

Robustness: Despite our efforts to capture all known sources of noise in our device simulation model, unforeseen properties of the physical instrument might incur additional measurement errors during flight. To characterize how robustly our pipeline handles additional uncertainty, we add Gaussian noise to the spatial and energy values of each hit prior to reconstruction. For an input with value x, we perturb its value to $x' \sim \mathcal{N}\left(x, (x \cdot \epsilon/100)^2\right)$, testing $\epsilon \in \{0, 1, 5, 10\}$. In other words, noise with standard deviation $\epsilon\%$ of the input's value is added.

Figure 10 shows that, even under increasing perturbation noise, our models continue to improve localization accuracy versus our prior work. Moreover, 68% containment error increases more slowly with noise when incorporating our networks than without them.

[2]Spectral energy parameters and light curves match those in [4], [9], except that we use a fixed $\beta = -2.35$ and a minimum simulated energy of 30 keV.

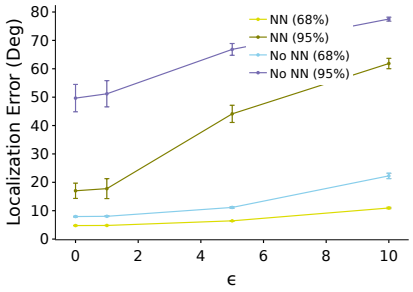

Fig. 10. Localization accuracy with perturbed inputs.

TABLE I
Timing results on **RPi 3B+**.

Stage	Mean Time (ms)	Range (ms)
Reconstruction	36.9	35-44
Localization Setup	35.4	34-99
DEta NN Inference	31.0	17-41
Bkg NN Inference	36.1	22-58
Approx + Refine	91.7	89-107
Total (Max 5 iter)	834.0	730-1116

TABLE II
Timing results on **Atom**.

Stage	Mean Time (ms)	Range (ms)
Reconstruction	18.6	15-26
Localization Setup	12.1	12-13
DEta NN Inference	5.5	5-6
Bkg NN Inference	14.7	14-15
Approx + Refine	18.5	17-21
Total (Max 5 iter)	220.7	204-246

Timing: To assess our algorithms' efficiency, we ran the pipeline on two computational platforms. One platform, a Raspberry Pi 3B+, uses a 1.4 GHz quad-core Cortex-A53 (ARMv8) CPU and 1 GB of LPDDR2 DRAM. It serves as a proxy for the capabilities of space-qualified processors suitable for a satellite mission [32]–[34]. The second platform, a WINSYSTEMS EBC-C413 industrial single-board computer, uses a quad-core, 1.92 GHz Intel Atom E3845 CPU and 8 GB of DDR3L DRAM. It is a likely candidate to fly on ADAPT and has been used with other high-altitude balloon-based telescopes.

We measured elapsed times in milliseconds for event reconstruction, initial approximation of source direction, iterative refinement (both before and after the neural network stage), and network inference. All stages of the pipeline, where possible, were parallelized with OpenMP to use all four cores on each processor. We repeated the experiment 300 times using a 1 MeV/cm^2, normally-incident burst. Tables I and II show that even running all 5 iterations of our complete pipeline typically only takes around 220 ms on the Atom and 830 ms on the RPi 3B+.

V. Quantization

Neural Networks can be optimized post-training using a number of techniques [35], such as pruning, knowledge distillation, and quantization. Such optimizations are especially

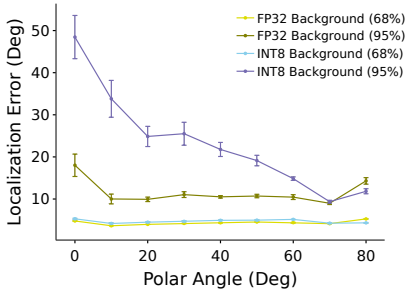

Fig. 11. Localization accuracy with quantized background model.

important in latency- and resource-constrained environments such as space-based embedded platforms. Here, we focus on quantization, as it not only significantly reduces the model size but may also improve inference latency and power usage, especially when accelerated with GPUs or FPGAs [36]. Since ADAPT will already fly with dozens of FPGAs to process streaming sensor data [10], such an architecture may be suitable for offloading our proposed ML models.

In this section, we perform a preliminary investigation using PyTorch's quantization framework to convert our 32-bit floating point (FP32) background model into an 8-bit integer (INT8) representation. The resulting model maintains localization accuracy most of the time yet provides about $1.75\times$ the throughput of its FP32 counterpart on an FPGA and about $3.6\times$ speedup (even at a very conservative 100 MHz clock) vs. the worst-case time on the Atom.

Methodology: We perform quantization-aware training (QAT) using PyTorch's Eager Mode, which requires marking the points in the code where tensor precision switches and fusion of batch-norm layers with another layer. We enable fusion of the fully-connected linear, batch normalization and ReLU layers by retraining the background model with an updated architecture that reverses the order of these two layers within a block compared to that shown in Figure 5. We then use PyTorch's default 'x86' configuration to perform the quantization-aware training.

Accuracy: We compare localization accuracy using the INT8-quantized versus the FP32 (regular) version of our background network (still in conjunction with the FP32 version of the dEta model). Figure 11 shows results for a 1 MeV/cm^2 burst across polar angles. The INT8 model performs almost as well as FP32 68% of the time. However, 95% containment values become less accurate.

FPGA Deployment: To evaluate performance gains realized by quantization when deploying to an FPGA, we implement the model as an FPGA kernel using high-level synthesis (HLS). HLS expresses hardware designs using procedural and generic programming languages; Vitis HLS allows kernels to be authored in C++ with additional pragmas providing hardware-specific optimization guidance to the compiler. HLS has proven effective for FPGA synthesis of other operations in our pipeline [10].

Our FPGA kernel matches the updated (layer-swapped) model architecture, except that the final sigmoid is not implemented – because a sigmoid is a bijective function, a prior threshold can instead be applied, eliminating the cost of sigmoid evaluation. The HLS code uses several C++ preprocessor definitions and templates to allow easy duplication of layers and type switching between INT8 and FP32. We optimize the kernel to achieve the best speed, parallelizing computational logic to the extent possible and enabling deep dataflow pipelining, which allows multiple inputs to occupy separate layers of the network concurrently.

TABLE III
QUANTIZATION RESULTS ON FPGA.

Statistic	INT8	FP32
Latency (cycles)	881	1891
Initiation Interval (cycles)	692	1209
BRAM Blocks	15	144
DSP Slices	4,304	7,467
Flip-Flops	366,545	651,014
Lookup Tables	775,986	817,041
Latency (ms) for 597 rings	4.13	7.22

We synthesize the kernel in Vitis HLS 2021.1, then perform C/RTL co-simulation with a conservative 10 ns clock cycle to account for possibly constrained power budgets. We use a C++ testbench that passes feature inputs and receives the network's output over AXI-based memory interfaces. Speeds and logic resources utilized are listed in Table III. Because of the pipelined implementation, the initiation interval (II) is shorter than the latency (L), so for n inputs, total latency can be computed as $n \cdot II + (L - II)$ [37]. For the timing results in Section IV, the first iteration of the background network processed 597 rings on average. For the same number of rings, the INT8-quantized network on the FPGA takes only 4.13 ms (compared to 22-58 ms on the RPi 3B+ and 14-15 ms on the Atom), and achieves $1.75\times$ the throughput of FP32 while using significantly fewer logic resources.

VI. CONCLUSION

ADAPT and the future space-based APT mission seek to provide timely, accurate localization of gamma-ray transients on a resource-constrained on-board hardware platform. This work demonstrates that neural network models improve the resilience of our analysis pipeline to background radiation and measurement uncertainty, all within the constraints of low-resource, real-time computation. We also provide preliminary evidence to support further accelerating these models in FPGA logic. ADAPT, with the help of our networks, is expected to localize a moderately bright, short-duration GRB in under a second with a typical localization accuracy of six degrees.

Future work will include consideration of additional sources of error, such as multiple events that arrive simultaneously to within the detection latency of the instrument. We will also investigate a broader range of quantization strategies for our models, including different configurations of quantization and other libraries outside of PyTorch. Finally, we will study the impact of our models on the full APT instrument, whose much larger detector demands event processing at a higher rate yet could allow localization of even dim (< 0.1 MeV/cm^2) GRBs to within a degree or less.

REFERENCES

[1] W. Chen *et al.*, "The Advanced Particle-astrophysics Telescope: simulation of the instrument performance for gamma-ray detection," in *Proc. 37th Int'l Cosmic Ray Conf.*, vol. 395, 2021, pp. 590:1–590:9.

[2] J. Buckley *et al.*, "The Advanced Particle-astrophysics Telescope (APT) project status," in *Proc. 37th Int'l Cosmic Ray Conf.*, vol. 395, Jul. 2021, pp. 655:1–655:9.

[3] M. Sudvarg *et al.*, "A fast GRB source localization pipeline for the Advanced Particle-astrophysics Telescope," in *Proc. of 37th Int'l Cosmic Ray Conf.*, vol. 395, Jul. 2021, pp. 588:1–588:9.

[4] Y. Htet *et al.*, "Prompt and accurate GRB source localization aboard the Advanced Particle Astrophysics Telescope (APT) and its Antarctic Demonstrator (ADAPT)," in *Proc. 38th Int'l Cosmic Ray Conf.*, vol. 444, Jul. 2023, pp. 956:1–956:9.

[5] J. H. Buckley, J. D. Buhler, and R. D. Chamberlain, "The Advanced Particle-astrophysics Telescope (APT): computation in space," in *Proc. 21st ACM Int'l Conf. Computing Frontiers Workshops and Special Sessions*, May 2024, pp. 122–127.

[6] C. Meegan, G. Lichti, P. N. Bhat *et al.*, "The Fermi gamma-ray burst monitor," *Astrophysical J.*, vol. 702, no. 1, pp. 791–804, Aug. 2009.

[7] J. D. Meyers, Curator, "Overview of the Fermi GBM," *https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_Introduction/GBM_overview.html*, National Aeronautics and Space Administration Goddard Space Flight Center, Jan. 2020, accessed: 26 Oct, 2022.

[8] W. Chen *et al.*, "Simulation of the instrument performance of the Antarctic Demonstrator for the Advanced Particle-astrophysics Telescope in the presence of the MeV background," in *Proc. 38th Int'l Cosmic Ray Conf.*, vol. 444, Jul. 2023, pp. 841:1–841:9.

[9] M. Sudvarg *et al.*, "Front-end computational modeling and design for the Antarctic Demonstrator for the Advanced Particle-astrophysics Telescope," in *Proc. 38th Int'l Cosmic Ray Conf.*, vol. 444, Jul. 2023, pp. 764:1–764:9.

[10] M. Sudvarg, C. Zhao, Y. Htet, M. Konst *et al.*, "Hls taking flight: Toward using high-level synthesis techniques in a space-borne instrument," in *Proc. 21st ACM Int'l Conf. Computing Frontiers*, 2024, pp. 115–125.

[11] P. Mészáros, D. B. Fox, C. Hanna, and K. Murase, "Multi-messenger astrophysics," *Nature Reviews Physics*, vol. 1, no. 10, pp. 585–599, 2019.

[12] A. Neronov, "Introduction to multi-messenger astronomy," *J. Physics: Conf. Series*, vol. 1263, no. 1, p. 012001, 2019.

[13] D. George and E. A. Huerta, "Deep neural networks to enable real-time multimessenger astrophysics," *Physical Review D*, vol. 97, no. 4, p. 044039, 2018.

[14] E. A. Huerta *et al.*, "Enabling real-time multi-messenger astrophysics discoveries with deep learning," *Nature Reviews Physics*, vol. 1, no. 10, pp. 600–608, 2019.

[15] IceCube Collaboration, "Towards a more robust reconstruction method for IceCube's real-time program," in *Proc. 38th Int'l Cosmic Ray Conf.*, vol. 444, Jul. 2023, pp. 1186:1–1186:9.

[16] M. Aartsen *et al.*, "The IceProd framework: Distributed data processing for the IceCube neutrino observatory," *Journal of Parallel and Distributed Computing*, vol. 75, pp. 198–211, 2015.

[17] R. Qiu, P. G. Krastev, K. Gill, and E. Berger, "Deep learning detection and classification of gravitational waves from neutron star-black hole mergers," *Physics Letters B*, vol. 840, p. 137850, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370269323001843

[18] P. Chaturvedi, A. Khan, M. Tian, E. A. Huerta, and H. Zheng, "Inference-optimized AI and high performance computing for gravitational wave detection at scale," *Frontiers in Artificial Intelligence*, vol. 5, p. 828672, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2022.828672

[19] N. Parmiggiani, A. Bulgarelli, A. Ursi, A. Macaluso *et al.*, "A deep-learning anomaly-detection method to identify gamma-ray bursts in the ratemeters of the agile anticoincidence system," *Astrophysical J.*, vol. 945, no. 2, p. 106, Mar. 2023.

[20] S. Takashima, H. Odaka, H. Yoneda, Y. Ichinohe *et al.*, "Event reconstruction of compton telescopes using a multi-task neural network," *Nucl. Instrum. Methods Phys. Res. A*, vol. 1038, p. 166897, 2022.

[21] M. Kole, G. Koziol, and D. Droz, "HAGRID - High Accuracy GRB Rapid Inference with Deep learning," in *Proc. 38th Int'l Cosmic Ray Conf.*, vol. 444, 2023, pp. 724:1–724:8.

[22] S. E. Boggs and P. Jean, "Event reconstruction in high resolution Compton telescopes," *Astronomy and Astrophysics Suppl. Series*, vol. 145, no. 2, pp. 311–321, 2000.

[23] Y. Htet, M. Sudvarg, J. D. Buhler, R. D. Chamberlain, and J. H. Buckley, "Localization of gamma-ray bursts in a balloon-borne telescope," in *Proc. Wkshps. Int'l Conf. High Performance Computing, Network, Storage, and Analysis (SC-W)*, Nov. 2023, pp. 395–398.

[24] J. Beechert, H. Lazar, S. E. Boggs, T. J. Brandt *et al.*, "Calibrations of the compton spectrometer and imager," *Nucl. Instrum. Methods Phys. Res. A*, vol. 1031, p. 166510, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900222001243

[25] S. Agostinelli, J. Allison, K. Amako *et al.*, "Geant4 — a simulation toolkit," *Nucl. Instrum. Methods Phys. Res. A*, vol. 506, no. 3, pp. 250–303, 2003.

[26] L. Biewald, "Experiment tracking with weights and biases," 2020. [Online]. Available: https://www.wandb.com/

[27] D. Gruber, A. Goldstein, V. W. von Ahlefeld *et al.*, "The Fermi GBM gamma-ray burst spectral catalog: four years of data," *Astrophysical J. Suppl. Series*, vol. 211, no. 1, p. 12, Feb. 2014.

[28] A. von Kienlin, C. A. Meegan, W. S. Paciesas *et al.*, "The second Fermi GBM gamma-ray burst catalog: The first four years," *Astrophysical J. Suppl. Series*, vol. 211, no. 1, p. 13, Feb. 2014.

[29] P. N. Bhat, C. A. Meegan, A. von Kienlin *et al.*, "The third Fermi GBM gamma-ray burst catalog: the first six years," *Astrophysical J. Suppl. Series*, vol. 223, no. 2, p. 28, Apr. 2016.

[30] A. von Kienlin, C. A. Meegan, W. S. Paciesas *et al.*, "The fourth Fermi-GBM gamma-ray burst catalog: a decade of data," *Astrophysical J.*, vol. 893, no. 1, p. 46, Apr. 2020.

[31] E. Berger, P. A. Price, S. Cenko, A. Gal-Yam *et al.*, "A merger origin for short gamma-ray bursts inferred from the afterglow and host galaxy of grb 050724," *Nature*, vol. 238, pp. 988–990, 2005.

[32] X. Iturbe, B. Venu, E. Ozer, and S. Das, "A triple core lock-step (tcls) arm® cortex®-r5 processor for safety-critical and ultra-reliable applications," in *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop (DSN-W)*, 2016, pp. 246–249.

[33] J. Keller, "Boeing to develop next-generation radiation-hardened space processor based on the ARM architecture," *Military Aerospace Electronics*, vol. 28, 2017.

[34] W. A. Powell, "High-performance spaceflight computing (HPSC) project overview," in *Radiation Hardened Electronics Technology Conference (RHET)*, 2018.

[35] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: a survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221010894

[36] U. Kulkarni, A. S. Hosamani, A. S. Masur, S. Hegde, G. R. Vernekar, and K. Siri Chandana, "A survey on quantization methods for optimization of deep neural networks," in *2022 Int'l Conf. Automation, Computing and Renewable Systems*, 2022, pp. 827–834.

[37] C. Zhao, C. J. Faber, R. D. Chamberlain, and X. Zhang, "HLPerf: demystifying the performance of HLS-based graph neural networks with dataflow architectures," *ACM Transactions on Reconfigurable Technology and Systems*, 2024.