Title: Edge-Optimized Deep Learning: Harnessing Generative AI and Computer Vision with Open-Source Libraries.

Proposers' names, titles, and affiliations: (See the bio sketches into the attached document**)**

Samet Akcay, Paula Ramos, Ria Cheruvu, Alexander Kozlov, Zhen Zhao, Zhuo Wu, Raymond Lo, and Yury Gorbachev

Preference half of full day: Full-day, Three speakers will deliver this tutorial in-person and two online.

Course description (Topics, brief outline, and important details):

This tutorial addresses the challenge of navigating the increasingly complex deep learning (DL) landscape, characterized by many frameworks with specialized functionalities. It aims to equip researchers and practitioners with the necessary skills to develop efficient and accessible DL models for diverse applications. This tutorial encompasses critical aspects of the DL pipeline, including robust data management, diverse training methodologies, optimization strategies, and efficient deployment techniques. Emphasis is placed on the utility of open-source libraries, such as, OpenVINO toolkit, OpenVINO Training eXtensions (OTX), and Neural Network Compression Frameworks (NNCF), in streamlining the DL development process. Through hands-on experiences with OpenVINO, OTX, and NNCF, participants will gain proficiency in managing data effectively, utilizing various training methods, and implementing optimizations across the AI lifecycle including computer vision pipelines and Generative AI (Gen AI). Furthermore, the tutorial dives into the concept of fine-tuning generative AI models, specifically Stable Diffusion SD with LoRA, adaptors for edge computing environments. This section highlights the advantages of customized models in reducing latency and enhancing efficiency. Ultimately, this comprehensive tutorial provides a valuable learning experience, equipping participants with the knowledge and skills necessary to navigate the complexities of modern DL and achieve success in their respective fields.

Tutorial attendees will evidence how OTX 2.0 [1] streamlines the complex deep learning ecosystem by providing a unified API and powerful CLI that integrate various frameworks, simplifying the process for researchers and developers. It ensures a consistent experience across different platforms (MMLab [2], Lightning [3], or Anomalib [4]) focusing on training, inference, and end-to-end optimization for edge deployment. Initially, the session will focus on fine-tuning downstream Computer Vision tasks such as detection and segmentation, covering techniques in supervised and semi/self-supervised learning. Additionally, the tutorial will explore the fine-tuning visual prompting tasks, including Segment Anything Model (SAM), showcasing the versatility of OTX 2.0 in addressing a wide range of computer vision challenges.

Building on the foundation of Computer Vision tasks, the tutorial will transition to the fine-tuning of Gen AI models. This tutorial will explain how to fine-tune a SD model with custom data using multiple acceleration methods [5, 6], and how to deploy the fine-tuned model by inserting subgraph files with multiple LoRA weights into a single SD model. For the last one, we will use only the SD model graph once [7], instead of looping the SD model graph multiple times for multiple LoRA weights as in [8], so there are no extra execution costs in the forwarding operation of LoRA weights and shortened model compiling time could then be achieved, through using OpenVINO Transformation Passes API [9]. As a result, we can get an image with multiple features that represented in LoRA in just one inference.

Finally, the tutorial will focus on the model optimization capabilities for the inference phase of the AI lifecycle. OpenVINO toolkit and OTX library enhance model optimization by integrating with the Neural Network Compression Framework (NNCF) [10], allowing users to refine neural networks during and after training. It facilitates accuracy-aware optimizations to maintain performance while compressing models and offers post-training techniques like quantization to decrease model size and improve inference speed, particularly on edge devices with limited resources.

In this tutorial, attendees will learn as well how to optimize DL models using NNCF, as we showcase computer vision pipelines, such as Object Detection and Generative AI pipelines including Stable diffusion, LLMs, and multi-modal models. We will have demos for evidencing how OpenVINO runtime API is enabling real-time inference on laptops, edge devices, and resource-constrained hardware by more than 10x[1] in latency reduction for Stable Diffusion workloads.

[1]Performance varies by use, configuration and other factors. Learn more at intel.com/performanceindex

Outline

1. OpenVINO, OpenVINO Training eXtensions (OTX) and NNCF. Fundamentals
2. Module 1: Data management, training, and fine-tuning downstream Computer Vision tasks.
3. Module 2: Optimize and run Gen AI pipelines on your laptop. SD with LoRA weights.
4. Module 3: Optimization with NNCF for Computer Vision and Gen AI (Multimodal).
5. Module 4: Evaluate and deploy your solution as an edge-computing system. Multiple Computer Vision tasks and Gen AI pipelines on a wide range of HW.

Important details: Participants should have access to an Intel-based laptop or server and will be provided with access to the Intel Dev Cloud for the Edge. A fundamental understanding of deep learning and downstream computer vision tasks, such as classification, detection, segmentation, anomaly detection and generative AI such as LLMs and Stable Diffusion, is required for the full benefit of the tutorial. This tutorial expects 100+ in-person attendees and 5.000+ on-line.

Key Benefits and audience: Participants will: i) Master the integration of various deep learning frameworks using a single OTX 2.0 API, ii) Import and export data in various formats using OTX's support for over 35 public vision data formats. iii) Train models using diverse methods including supervised, semi-supervised and self-supervised learning, iv) Optimize computer vision models for deployment on edge devices through techniques like quantization and OpenVINO, ensuring efficient performance in resource-constrained environments, v) how to optimize popular LLM model having CPU only and get real-time performance in chatbot application as well as generating images with OpenVINO locally within a few seconds, vi) Address Generative AI deployment efficiently for customizes SD models. vii) Gain hands-on experience in model optimization and deployment on edge devices**. Composition and number:** Designed for those with a foundational grasp of deep learning concepts, this tutorial is ideal for: i) Professionals seeking to use a unified API across multiple deep learning frameworks, ii) Individuals interested in leveraging Intel GPU capabilities for model training, iii) Practitioners aiming to train and deploy models on edge devices.

A description of how this proposal relates to tutorials/short courses appearing at CVPR, ICCV, and ECCV within the last three years: The current tutorial is related to previous CVPR tutorials “OpenMMLab: A foundational Platform for computer Vision Research and Production” [11], and Boosting Computer Vision Research with OpenMMLab and OpenDataLab [12], our tutorial will focus on model training and efficiency deployment compared to these prior tutorials focus on benchmarking. Another previous related tutorial: "How to get quick and performant model for your edge application. From data to application" [13], where we showed the workflow of porting to Intel DL deployment tools. In this submission, we will be focusing on the advancement in various acceleration and optimization techniques that are more recently published and deployed with the OpenVINO toolkits.

Links to a few previous recorded talks

[KubeCon][2023][GenAI – Stable Diffusion with Optimum-Intel and OpenVINO][R. Cheruvu][P. Ramos]
[ComputerVisionMeetup][2023][Breaking the Bottleneck of AI Deployment at the Edge with OpenVINO][P. Ramos]
[OpenCV Webinar Series][2023][Anomalib: A Deep Learning Library for Anomaly Detection][P. Ramos]
[OpenVINO DevCon Series][2023][Generative AI with OpenVINO][P. Ramos][R. Cheruvu]
[CVPR][2022] [How to get quick and performant model for your edge application. ...][P. Ramos]
[OpenCV Webinar Series][2023][Anomalib: A Deep Learning Library for Anomaly Detection][S. Akcay][P. Ramos]
[Technovation Podcast][2023][Building Trustworthy and Explainable AI Models][R. Cheruvu]
[CVPR][2022] [How to get quick and performant model for your edge application. ...][P. Ramos]
[Edge AI Reference Kit][2023][Defect Detection with Anomalib Edge AI Reference Kit][P. Ramos]
[Edge AI Reference Kit][2023][Code Demo | Defect Detection with Anomalib Edge AI Reference Kit][P. Ramos]
[IoT North meetup][2022][IoT North Meetup - Intel GETi & OpenVINO updates][P. Ramos]
[OpenCV Webinar Series][2021][Smarter Agriculture with OpenCV AI Kit feat. Team Benchbotics][P. Ramos]
[OpenCV Webinar Series][2022][From OpenCV Competition to Intel][P. Ramos]
[India Electronics & Semiconductor Association (IESA) Vision Summit][2021][Ethical AI and Data Use][R. Cheruvu]

Description (links) to planned materials to be distributed to attendees

Repositories with technical material for OTX, OpenVINO Model API, OpenVINO Runtime, and Datumaro. We will provide access and examples to different libraries, such as Anomalib, OTX, OpenVINO Model API, OTX Notebooks, OpenVINO Documentation. Attendees can also access the Intel Developer Cloud for the edge for running the tutorial demos and hands-on experience, Intel® Developer Cloud for the Edge Home page. Attendees will have full access to the deck, presentations and material during and after the tutorial.

List of citations/URLs/products by organizer

- **Intel Corporation,** "OpenVINO™ Training Extensions", [Online]. Available: https://github.com/openvinotoolkit/training_extensions. Intel Corporation, 2023. [Accessed 27 November 2023].
- **S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja,** "Anomalib: A Deep Learning Library for Anomaly Detection," in Proceedings of the 2022 IEEE International Conference on Image, 2022.
- **Intel Corporation,** "Defect Detection with Anomalib", [Online], Available: https://www.intel.com/content/www/us/en/developer/articles/training/defect-detection-with-anomalib.html. Intel Corporation, 2023. [Accessed 27 November 2023].
- **P. Ramos,** "Hands on lab how to perform automated defect detection using Anomalib". [Online], Available: https://medium.com/openvino-toolkit/hands-on-lab-how-to-perform-automated-defect-detection-using-anomalib-5c1cfed666b4. Medium, 2023
- **Intel Corporation,** "CAM-Visualizer: Class Activation Map Visualization Toolkit", [Online]. Available: https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/class-activation-map-visualizer.html Intel Corporation, 2023. [Accessed 27 November 2023].
- **Intel Corporation,** "OpenVINO™ Automatic Model Manifest Add-On", [Online]. Available: https://www.intel.com/content/www/us/en/developer/articles/reference-implementation/openvino-automatic-model-manifest-add-on.html. Intel Corporation, 2023. [Accessed 27 11 2023]

References

[1] **Intel Corporation,** "OpenVINO™ Training Extensions", [Online]. Available: https://github.com/openvinotoolkit/training_extensions. Intel Corporation, 2023. [Accessed 27 November 2023].

[2] **OpenMMLab**, [Online]. Available: https://github.com/open-mmlab. [Accessed 27 November 2023].

[3] **Lightning.ai**, [Online]. Available: https://lightning.ai/. [Accessed 27 November 2023].

[4] **Intel Corporation,** "Anomalib", [Online]. Available: https://github.com/openvinotoolkit/anomalib/tree/main. Intel Corporation, 2023. [Accessed 27 November 2023].

[5] **Q. Fu, et al.,** "Deep Learning Models on CPUs: A Methodology for Efficient Training," arXiv preprint arXiv:2206.10034, 2022.

[6] **Intel Corporation,** "Remote Tensor API of GPU Plugin," Intel corporation, 2023. Available online: OpenVINO Documentation [Accessed 5 November 2023].

[7] **Intel Corporation,** "OpenVINO Stable Diffuison (with LoRA) C++ pipeline," Intel Corporation, 2023. Available online: OpenVINO Documentation. [Accessed 5 November 2023].

[8] **S.Luo, et al.,** " LCM-LoRA: A Universal Stable-Diffusion Acceleration Module," arXiv:2311.05556 Available online: *arXiv* [Accessed 4 December 2023].

[9] **Zhen Zhao and Kunda Xu,** "Enable LoRA Weights with Stable Diffusion Controlnet Pipeline," Intel Corporation, 7 Aug. 2023. Available online: Intel Community Blog. [Accessed 5 November 2023].

[10] **Intel Corporation,** "Neural Network Compression Framework (NNCF)", [Online]. Available: https://github.com/openvinotoolkit/nncf. Intel Corporation, 2023. [Accessed 27 November 2023].

[11] **K. Chen, C. Loy, H. Hu, H. Zhao, and H. Duan,** "OpenMMLab: A Foundational Platform for Computer Vision Research and Production", [Online]. Available: https://openmmlab.com/community/cvpr2022-tutorial. OpenMMLab, 20 June 2022. [Accessed 27 November 2023].

[12] **K. Chen, C. He, Y. Zeng, S. Zhang, and W. Zhang,** "Boosting Computer Vision Research with OpenMMLab and OpenDataLab", [Online]. Available: https://openmmlab.com/community/cvpr2023-tutorial. OpenMMLab, 18 June 2023. [Accessed 27 November 2023].

[13] **P. Ramos, Z. Wu, Y. Gorvachev, and R. Lo,** "How to get quick and performant model for your edge application. From data to application", [Online]. Available: https://paularamo.github.io/cvpr-2022. Intel Corporation, 19 June 2022. [Accessed 27 November 2023].