A1. Reference Hierarchy

Model Development

Frameworks

[14] Chen Liu, Matthias Jobst, Liyuan Guo, Xinyue Shi, Johannes Partzsch, and Christian Mayr. 2024. Deploying Machine Learning Models to Ahead-of-Time Runtime on Edge Using MicroTVM. In Proceedings of the 2023 Workshop on Compilers, Deployment, and Tooling for Edge AI (CODAI '23). Association for Computing Machinery, New York, NY, USA, 37–40. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3615338.3618125

[78] "LiteRT overview," Google, March 4, 2025. [Online]. Available: https://ai.google.dev/edge/litert. [Accessed April 2025].

[79] "Caffe2," Facebook, [Online]. Available: https://caffe2.ai/docs/caffe-migration.html. [Accessed April 2025].

[80] "Apache MXNet," 2022. [Online]. Available: https://mxnet.apache.org/versions/1.9.1/. [Accessed April 2025].

Data Gathering/Preprocessing

[13] Oihane Gómez-Carmona, Diego Casado-Mansilla, Diego López-de-Ipiña, and Javier García-Zubia. 2019. Simplicity is Best: Addressing the Computational Cost of Machine Learning Classifiers in Constrained Edge Devices. In Proceedings of the 9th International Conference on the Internet of Things (IoT '19). Association for Computing Machinery, New York, NY, USA, Article 18, 1–8. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3365871.3365889

[17] Brendan Reidy, Sepehr Tabrizchi, Mohammadreza Mohammadi, Shaahin Angizi, Arman Roohi, and Ramtin Zand. 2024. HiRISE: High-Resolution Image Scaling for Edge ML via In-Sensor Compression and Selective ROI. In Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC '24). Association for Computing Machinery, New York, NY, USA, Article 275, 1–6. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3649329.3656539

Model Compression

Pre-Runtime

[19] Y. Gong, et al., “Compressing deep convolutional networks using vector quantization,” arXiv preprint arXiv:1412.6115, 2014.

[20] Denton, Remi, Zaremba, Wojciech, Bruna, Joan, LeCun, Yann, and Fergus, Rob. Exploiting linear structure within convolutional networks for efficient evaluation. In NIPS. 2014. Available: https://arxiv.org/pdf/1404.0736

Runtime Model Optimizations

[1] N. D. Lane et al., "DeepX: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices," 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Vienna, Austria, 2016, pp. 1-12, doi: 10.1109/IPSN.2016.7460664. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/stamp/stamp.jsp?tp=&arnumber=7460664

[9] Vidushi Goyal, Reetuparna Das, and Valeria Bertacco. 2022. Hardware-friendly User-specific Machine Learning for Edge Devices. ACM Trans. Embed. Comput. Syst. 21, 5, Article 62 (September 2022), 29 pages. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3524125

[81] F. Dong, D. Shen, Z. Huang, Q. He, J. Zhang, L. Wen and T. Zhang, "Multi-Exit DNN Inference Acceleration Based on Mult-Dimensional Optimization for Edge Intelligence," *IEEE Transactions on Mobile Computing,* vol. 22, no. 9, pp. 5389 - 5406, 2023.

Hardware Accelerators

[3] Kim, K.; Jang, S.-J.; Park, J.; Lee, E.; Lee, S.-S. Lightweight and Energy-Efficient Deep Learning Accelerator for Real-Time Object Detection on Edge Devices. *Sensors* **2023**, *23*, 1185. Available: https://doi.org/10.3390/s23031185

[10] Minkwan Kee and Gi-Ho Park. 2022. A Low-power Programmable Machine Learning Hardware Accelerator Design for Intelligent Edge Devices. ACM Trans. Des. Autom. Electron. Syst. 27, 5, Article 51 (September 2022), 13 pages. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3531479

[11] Amirali Boroumand, Saugata Ghose, Berkin Akin, Ravi Narayanaswami, Geraldo F. Oliveira, Xiaoyu Ma, Eric Shiu, and Onur Mutlu. 2024. Google Neural Network Models for Edge Devices: Analyzing and Mitigating Machine Learning Inference Bottlenecks. In Proceedings of the 30th International Conference on Parallel Architectures and Compilation Techniques (PACT '21). IEEE Press, 159–172. Available: https://doi-org.libweb.lib.utsa.edu/10.1109/PACT52795.2021.00019

[21] Gokhale, V., Jin, Jonghoon, Dundar, A., Martini, B., and Culurciello, E. A 240 g-ops/s mobile coprocessor for deep neural networks. In CVPR Workshops. 2013.

Resource Management

Inference

[7] Z. Safavifar, E. Gyamfi, E. Mangina and F. Golpayegani, "Multi-Objective Deep Reinforcement Learning for Efficient Workload Orchestration in Extreme Edge Computing," in IEEE Access, vol. 12, pp. 74558-74571, 2024, doi: 10.1109/ACCESS.2024.3405411. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/10538273

[8] S. Chouikhi, M. Esseghir and L. Merghem-Boulahia, "Energy-Efficient Computation Offloading Based on Multiagent Deep Reinforcement Learning for Industrial Internet of Things Systems," in

IEEE Internet of Things Journal, vol. 11, no. 7, pp. 12228-12239, 1 April1, 2024, doi: 10.1109/JIOT.2023.3333044. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/10318207

Training

[12] Bharath Sudharsan, John G. Breslin, and Muhammad Intizar Ali. 2020. Edge2Train: a framework to train machine learning models (SVMs) on resource-constrained IoT edge devices. In Proceedings of the 10th International Conference on the Internet of Things (IoT '20). Association for Computing Machinery, New York, NY, USA, Article 6, 1–8. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3410992.3411014

[16] Zeqian Dong, Qiang He, Feifei Chen, Hai Jin, Tao Gu, and Yun Yang. 2023. EdgeMove: Pipelining Device-Edge Model Training for Mobile Intelligence. In Proceedings of the ACM Web Conference 2023 (WWW '23). Association for Computing Machinery, New York, NY, USA, 3142–3153. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3543507.3583540

[18] Hayajneh, A.M., et al.: Tiny machine learning on the edge: a framework for transfer learning empowered unmanned aerial vehicle assisted smart farming. *IET Smart Cities*. 6(1), 10–26 (2024). Available: https://doi-org.libweb.lib.utsa.edu/10.1049/smc2.12072

Energy Consumption Analysis

[2] M. Kumar, X. Zhang, L. Liu, Y. Wang and W. Shi, "Energy-Efficient Machine Learning on the Edges," 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), New Orleans, LA, USA, 2020, pp. 912-921, doi: 10.1109/IPDPSW50202.2020.00153. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/stamp/stamp.jsp?tp=&arnumber=9150337

[4] Fanariotis, A.; Orphanoudakis, T.; Kotrotsios, K.; Fotopoulos, V.; Keramidas, G.; Karkazis, P. Power Efficient Machine Learning Models Deployment on Edge IoT Devices. *Sensors* **2023**, *23*, 1595. Available: https://doi.org/10.3390/s23031595

[5] Walid A. Hanafy, Tergel Molom-Ochir, and Rohan Shenoy. 2021. Design Considerations for Energy-efficient Inference on Edge Devices. In Proceedings of the Twelfth ACM International Conference on Future Energy Systems (e-Energy '21). Association for Computing Machinery, New York, NY, USA, 302–308. Available: https://doi.org/10.1145/3447555.3465326

[6] Xiaolong Tu, Anik Mallik, Dawei Chen, Kyungtae Han, Onur Altintas, Haoxin Wang, and Jiang Xie. 2024. Unveiling Energy Efficiency in Deep Learning: Measurement, Prediction, and Scoring across Edge Devices. In Proceedings of the Eighth ACM/IEEE Symposium on Edge Computing (SEC '23). Association for Computing Machinery, New York, NY, USA, 80–93. Available: https://doi.org/10.1145/3583740.3628442

[15] Giacomo Di Fabrizio, Lorenzo Calisti, Chiara Contoli, Nicholas Kania, and Emanuele Lattanzi. 2024. A Study on the energy-efficiency of the Object Tracking Algorithms in Edge Devices. In Proceedings of the IEEE/ACM 16th International Conference on Utility and Cloud Computing (UCC

'23). Association for Computing Machinery, New York, NY, USA, Article 29, 1–6. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3603166.3632541

Supplementary Works Not Included or Lightly Referenced

[22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv preprint arXiv:1602.07360*, 2016. Available: https://openreview.net/pdf?id=S1xh5sYgx

[23] S. Han, H. Mao, and W. Dally, "DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING, TRAINED QUANTIZATION AND HUFFMAN CODING," ICLR, 2016. Available: https://arxiv.org/pdf/1510.00149

Comment: A three-prong approach to compressing a model. Pruning weight connections below a certain threshold and retrain on the sparse weights. Trained quantization and weight sharing forces multiple connections to share the same weight and fine tune those weights to be represented in fewer bits. Use Huffman coding on quantized weights to further reduce memory overhead.

[24] A. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," Apr. 2017. Available: https://arxiv.org/pdf/1704.04861

Comment: A class of efficient models for mobile and embedded **computer vision applications**. Core layers of MobileNet architecture: Depthwise separable convolutions that split a conventional convolution into two simpler operations. Width multiplier is a parameter to reduce the number of channels per layer. Resolution multiplier is a parameter to reduce input image resolution.

[25] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017, Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/stamp/stamp.jsp?tp=&arnumber=7738524

Comment: A hardware accelerator designed for energy efficiency on the entire system running a DNN model. Main features include: A spatial architecture of 168 processing elements with a four-level memory hierarchy to minimize data access at high-cost levels. Row stationary – reconfigurable mapping to CNN shape. A reconfigurable communication architecture. Compression and data gating to reduce memory footprint

[26] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *proceedings.mlr.press*, Apr. 10, 2017. Available: https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

Comment: An introduction to federated learning. A model is trained on local data and weight updates are communicated to a central server that averages weight updates from multiple devices. Local data is never shared between devices. Introduces the FederatedAveraging Algorithm – local stochastic gradient descent and model averaging on a central server, eliminating communication overhead between devices.

[27] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge. In Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '17). Association for Computing Machinery, New York, NY, USA, 615–629. Available: https://dl.acm.org/doi/pdf/10.1145/3093337.3037698

Comment: Introduces a system to intelligently partition a DNN between the cloud and an edge device. Experiments were run on the AlexNet model. A performance analysis is done to determine data and computation characteristics of each layer and is combined with the communication latency between the cloud and the device. These metrics determine the ideal partitions for the model.

[28] J. Azar, A. Makhoul, M. Barhamgi, and R. Couturier, “An energy efficient IoT data compression approach for edge machine learning,” *Future Generation Computer Systems*, vol. 96, pp. 168–175, Jul. 2019, Available: https://doi.org/10.1016/j.future.2019.02.005

Comment: An introduction of a technique to reduce communication overhead of IoT devices in a ML context. The paradigm consists of fast data compression technique at the IoT device before transmission to an intermediate edge device, where the data is decompressed and analyzed before transmission to the cloud. The results show a 103x reduction in data transmission for the IoT device, significantly lowering power consumption.

[29] Z. Ali, L. Jiao, T. Baker, G. Abbas, Z. H. Abbas and S. Khaf, "A Deep Learning Approach for Energy Efficient Computational Offloading in Mobile Edge Computing," in IEEE Access, vol. 7, pp. 149623-149633, 2019, doi: 10.1109/ACCESS.2019.2947053. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8866714

Comment: Introduces a machine learning algorithm (EEDOS) for optimal computational offloading between a mobile edge device and the cloud. The algorithm considers a variety of conditions, including remaining battery life (possibly the first algorithm to do so), in picking the best set of partitions. The training dataset is exhaustive, which allows EEDOS to be 100% accurate. Thus, EEDOS outperforms previous approaches.

[30] W. Liu, B. Li, W. Xie, Y. Dai and Z. Fei, "Energy Efficient Computation Offloading in Aerial Edge Networks With Multi-Agent Cooperation," in IEEE Transactions on Wireless Communications, vol. 22, no. 9, pp. 5725-5739, Sept. 2023, doi: 10.1109/TWC.2023.3235997. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/10021296

Comment: This paper explores the problem of optimal computation offloading between mobile devices and a base station server with the support of UAVs that can aid in computation tasks or act as a communication relay. They leverage the concept of a digital twin edge network (DITEN) hosted on the base station to model the optimization process before deploying a solution to the devices on the network.

[31] Q. Wei, Z. Zhou and X. Chen, "DRL-Based Energy-Efficient Trajectory Planning, Computation Offloading, and Charging Scheduling in UAV-MEC Network," 2022 IEEE/CIC International Conference on Communications in China (ICCC), Sanshui, Foshan, China, 2022, pp. 1056-1061,

doi: 10.1109/ICCC55456.2022.9880711. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/9880711

Comment: An introduction of a deep reinforcement learning framework for optimizing multiple objectives in UAV operations for mobile edge computing networks. A priority mechanism is built into the learning framework to handle complex decision making. The objectives are energy efficiency, trajectory planning, communication scheduling, charge scheduling and task offloading.

[32] A. Cleary, K. Yoo, P. Samuel, S. George, F. Sun and S. A. Israel, "Machine Learning on Small UAVs," 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington DC, DC, USA, 2020, pp. 1-5, doi: 10.1109/AIPR50011.2020.9425090. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/9425090

Comment: A paper on the workflow design of incorporating machine learning models on small UAVs using consumer-off-the-shelf parts and free-open-source software. ML model focuses on target detection and interfaces with Draper UAV flight software. Compatible with CPU/GPU and CPU only systems.

[33] P. K. Barik, S. Shah, K. Shah, A. Modi and H. Devisha, "UAV-Assisted Surveillance Using Machine Learning," 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022, pp. 384-389, doi: 10.1109/PDGC56933.2022.10053282. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/10053282

Comment: There is some interesting related work cited in this paper. This paper showcases the ML model and related dataset used in suspicious object detection. It also contributes a resource allocation algorithm for creating communication links between surveillance drone, anchor drone and base station. The training data for object detection may be questionable as the example image is at ground level.

[34] I. Martinez-Alpiste, P. Casaseca-de-la-Higuera, J. Alcaraz-Calero, C. Grecos and Q. Wang, "Benchmarking Machine-Learning-Based Object Detection on a UAV and Mobile Platform," 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakesh, Morocco, 2019, pp. 1-6, doi: 10.1109/WCNC.2019.8885504. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/8885504

Comment: This paper tests various object detection CNNs run on mobile devices to detect objects in streaming video from commercial-off-the-shelf UAVS. They chose to focus on the scenario where the ML model is not run directly on the UAV. Metrics tested are accuracy, processing speed, and resource consumption.

[35] Y. M. Park, Y. K. Tun and C. S. Hong, "Optimized Deployment of Multi-UAV based on Machine Learning in UAV-HST Networking," 2020 21st Asia-Pacific Network Operations and Management Symposium (APNOMS), Daegu, Korea (South), 2020, pp. 102-107, doi: 10.23919/APNOMS50412.2020.9236987. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/9236987

Comment: In Korea, Rail-side units (RSUs) act as edge servers to maintain 5G cellular connectivity for passengers. Frequent handovers between RSUs represent a significant communication overhead. The authors present a reinforcement learning model to optimize the deployment of multiple UAVs to maintain longer, stable connections with criteria for UAV energy efficiency, trajectory, and altitude.

[36] M. A. Abd-Elmagid, A. Ferdowsi, H. S. Dhillon and W. Saad, "Deep Reinforcement Learning for Minimizing Age-of-Information in UAV-Assisted Networks," 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 2019, pp. 1-6, doi: 10.1109/GLOBECOM38437.2019.9013924. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/9013924

Comment: Another approach of optimizing UAV trajectory and scheduling of status updates for UAVs for UAV-assisted wireless networks. This one focuses on minimizing age-of-information for critical information updates along with the normal constraints of power consumption and optimal flight trajectory. Also uses a deep reinforcement learning model to optimize the state space similar to other papers. Results show improvement over distance-based and random walk models.

[37] Q. Dong, "Reinforcement Learning based Anti-UAV Three-dimensional Pursuit-evasion Game for Substation Security," 2024 5th International Conference on Mechatronics Technology and Intelligent Manufacturing (ICMTIM), Nanjing, China, 2024, pp. 224-227, doi: 10.1109/ICMTIM62047.2024.10629444. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/10629444

Comment: To enhance defensibility of power substations from UAV attack, this paper presents a deep reinforcement learning model to simulate a 3-dimensional pursuit evasion game, learning optimal interception strategies. They show improved interception success rates with the model.

[38] Y. Ding, Z. Yang, Q. -V. Pham, Y. Hu, Z. Zhang and M. Shikh-Bahaei, "Distributed Machine Learning for UAV Swarms: Computing, Sensing, and Semantics," in IEEE Internet of Things Journal, vol. 11, no. 5, pp. 7447-7473, 1 March1, 2024, doi: 10.1109/JIOT.2023.3341307. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/10353003

Comment: A paper investigating the complex task of deploying a UAV swarm to intelligently address multiple objectives. They present several deep learning frameworks such as Federated Learning, Multi-Agent Reinforcement Learning, Distributed Inference, and Split Learning. Multiple UAV swarm applications are investigated as well as areas for future research.

[39] M. Aljohani, R. Mukkamala and S. Olariu, "Autonomous Strike UAVs in Support of Homeland Security Missions: Challenges and Preliminary Solutions," in IEEE Access, vol. 12, pp. 90979-90996, 2024, doi: 10.1109/ACCESS.2024.3420235. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/document/10576029

Comment: An analysis of the challenges inherent in deploying autonomous UAVs for strike missions. The paper discusses the DRL model for training the UAV, smart contract and blockchain technology for ensuring data integrity, security, and auditability, and a simulation model to develop a synthetic data set for such missions.

[40] Ismi Abidi, Vireshwar Kumar, and Rijurekha Sen. 2021. Practical Attestation for Edge Devices Running Compute Heavy Machine Learning Applications. In Proceedings of the 37th Annual Computer Security Applications Conference (ACSAC '21). Association for Computing Machinery, New York, NY, USA, 323–336. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3485832.3485909

Comment: A software attestation tool to protect against malicious network attacks during runtime. Since computation overhead is a significant constraint in EdgeML, they design the tool to run integrity checks on small chunks of the software at random intervals to minimize runtime interruptions and combat attacks designed to run between attestation events. PracAttest achieves 50x-80x speedup over previous state-of-the-art baselines.

[41] Petros Amanatidis, George Iosifidis, and Dimitris Karampatzakis. 2022. Comparative Evaluation of Machine Learning Inference Machines on Edge-class Devices. In Proceedings of the 25th Pan-Hellenic Conference on Informatics (PCI '21). Association for Computing Machinery, New York, NY, USA, 102–106. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3503823.3503843

Comment: A performance analysis of running the inference step of an object classification ML model on three different edge device setups. Measurement metrics include throughput (fps), execution time, accuracy, energy efficiency, and total power consumption. There is discussion on cost considerations between hardware and the difficulty of implementing the ML model on each setup.

[42] Kyle Hoffpauir, Jacob Simmons, Nikolas Schmidt, Rachitha Pittala, Isaac Briggs, Shanmukha Makani, and Yaser Jararweh. 2023. A Survey on Edge Intelligence and Lightweight Machine Learning Support for Future Applications and Services. J. Data and Information Quality 15, 2, Article 20 (June 2023), 30 pages. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3581759

Comment: A survey paper that assesses current cloud computing limitations, domains for edge computing growth, lightweight ML algorithms, and potential future directions in edge intelligence.

[43] Arash Heidari, Nima Jafari Navimipour, Mehmet Unal, and Guodao Zhang. 2023. Machine Learning Applications in Internet-of-Drones: Systematic Review, Recent Deployments, and Open Issues. ACM Comput. Surv. 55, 12, Article 247 (December 2023), 45 pages. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3571728

Comment: A survey paper on drone swarm applications, an identification of the predominant use of CNNs in applications, and the common use of Python as the programming language of choice in drone applications.

[44] Mukhtar N, Mehrabi A, Kong Y, Anjum A. Edge enhanced deep learning system for IoT edge device security analytics. *Concurrency Computat Pract Exper*. 2023; 35(13):e6764. doi:10.1002/cpe.6764

Comment: This paper focuses on implementing ML models on IoT edge devices to identify side-channel attack vulnerabilities. Specifically, they target the Elliptic-Curve Cryptographic (ECC) algorithm used on these devices with a 96% accuracy in discovery of secret keys. Identified vulnerabilities can then be used to implement relevant mitigation efforts.

[45] Nimmagadda, Y. (2025). Training on Edge. In Model Optimization Methods for Efficient and Edge AI (eds P.R. Chelliah, A.M. Rahmani, R. Colby, G. Nagasubramanian and S. Ranganath). Available: https://doi-org.libweb.lib.utsa.edu/10.1002/9781394219230.ch11

Comment: A selected chapter from the book Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications that discusses the various considerations for training ML models on edge devices.

[46] Zhang, M., Zhang, F., Lane, N.D., Shu, Y., Zeng, X., Fang, B., Yan, S. and Xu, H. (2020). Deep Learning in the Era of Edge Computing: Challenges and Opportunities. In Fog Computing (eds A. Zomaya, A. Abbas and S. Khan). Available: https://doi-org.libweb.lib.utsa.edu/10.1002/9781119551713.ch3

Comment: A book on Fog Computing, including chapters on various applications, challenges, information security, energy harvesting, energy efficiency, etc.

[47] Peñaranda C, Reaño C, Silla F. Exploring the use of data compression for accelerating machine learning in the edge with remote virtual graphics processing units. *Concurrency Computat Pract Exper*. 2023; 35(20):e7328. doi:10.1002/cpe.7328

Comment: This paper investigates the utilization of on-the-fly data compression for remote GPU-edge device integration in machine learning. They study various compression libraries and an adaptive compression algorithm that dynamically decides compression level based on compression ratio, data size, and network speed.

[48] Wang Y, Meng W, Li W, Liu Z, Liu Y, Xue H. Adaptive machine learning-based alarm reduction via edge computing for distributed intrusion detection systems. *Concurrency Computat Pract Exper*. 2019; 31:e5101. Available: https://doi-org.libweb.lib.utsa.edu/10.1002/cpe.5101

Comment: This paper presents a framework for enhancing the efficiency of Distributed Intrusion Detection Systems (DIDS) with adaptive ML models and edge computing utilization. Key contributions include utilizing edge devices for computing to reduce latency and bandwidth usage as well as an adaptive machine learning model to reduce false alarms.

[49] Ali, Elmustafa Sayed, Hasan, Mohammad Kamrul, Hassan, Rosilah, Saeed, Rashid A., Hassan, Mona Bakri, Islam, Shayla, Nafi, Nazmus Shaker, Bevinakoppa, Savitri, Machine Learning Technologies for Secure Vehicular Communication in Internet of Vehicles: Recent Advances and Applications, *Security and Communication Networks*, 2021, 8868355, 23 pages, 2021. Available: https://doi-org.libweb.lib.utsa.edu/10.1155/2021/8868355

Comment: A review of applications of ML in the Internet of Vehicles. Applications range from computation offloading decisions, network resource management, quality of experience for the user, and security enhancement. Future work looks at integrating 6G, using ML models to forecast issues with self-driving cars, and further experience enhancements.

[50] Raad, H. (2020). Cloud and Edge. In Fundamentals of IoT and Wearable Technology Design, H. Raad (Ed.). Available: https://doi-org.libweb.lib.utsa.edu/10.1002/9781119617570.ch7

Comment: A selected chapter from Fundamentals of IoT and Wearable Technology Design, First Edition. Provides definition of Fog computing, cloud topologies and services, and several cloud platforms and considerations for choosing a platform for a specific application.

[51] Perumalla, M.M.R., Singh, S.K., Khamparia, A., Goyal, A. and Mishra, A. (2020). Machine Learning Frameworks and Algorithms for Fog and Edge Computing. In Fog, Edge, and Pervasive Computing in Intelligent IoT Driven Applications (eds D. Gupta and A. Khamparia). Available: https://doi-org.libweb.lib.utsa.edu/10.1002/9781119670087.ch4

Comment: A selected chapter from Fog, Edge, and Pervasive Computing in Intelligent IoT Driven Applications, First Edition. Gives definitions for fog, edge and pervasive computing. Provides an overview of machine learning frameworks designed for these applications. Defines ML techniques for edge computing such as Naïve Bayes, Support Vector Machines, K-nearest Neighbor and K-means.

[52] Nimmagadda, Y. (2025). Model Optimization Techniques for Edge Devices. In Model Optimization Methods for Efficient and Edge AI (eds P.R. Chelliah, A.M. Rahmani, R. Colby, G. Nagasubramanian and S. Ranganath). Available: https://doi-org.libweb.lib.utsa.edu/10.1002/9781394219230.ch4

Comment: A selected chapter from Model Optimization Methods for Efficient and Edge AI: Federated Learning Architectures, Frameworks and Applications, First Edition. An explanation of ML model optimization techniques categorized by predeployment, deployment-time, and postdeployment. Each category is given ample attention.

[53] Diego Méndez, Daniel Crovo, Diego Avellaneda. (2024). Chapter 15 - Machine learning techniques for indoor localization on edge devices: Integrating AI with embedded devices for indoor localization purposes, TinyML for Edge Intelligence in IoT and LPWAN Networks, Academic Press, Pages 355-376, ISBN 9780443222023, Available: https://doi.org/10.1016/B978-0-44-322202-3.00020-8.

Comment: A selected chapter from TinyML for Edge Intelligence in IoT and LPWAN Networks. This chapter present techniques for indoor localization of mobile devices, specifically two types of signal fingerprinting techniques. Indoor obstacles can impede signal strength which causes inaccuracy in conventional trilateration approaches, so locations can be fingerprinted based on the signal profile. They introduce a TinyML model to move computation of the location from edge servers to the device.

[54] A. Guna, P. Ganeriwala and S. Bhattacharyya, "Exploring Machine Learning Engineering for Object Detection and Tracking by Unmanned Aerial Vehicle (UAV)," 2024 International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2024, pp. 1001-1004, doi: 10.1109/ICMLA61862.2024.00149. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/stamp/stamp.jsp?tp=&arnumber=10903281

Comment: A study comparing the accuracy of two ML models deployed on a consumer grade drone for the application of object detection and tracking. They employed YOLOv4 and Mask R-CNN on a Parrot Mambo drone. Transfer learning was done on a YOLOv4 model pretrained on the COCO

dataset with a small subset of hand labeled images specific to the task. Automated labeling was done for the rest of the study's custom data set using this model.

[55] J. Xu, Q. Guo, L. Xiao, Z. Li and G. Zhang, "Autonomous Decision-Making Method for Combat Mission of UAV based on Deep Reinforcement Learning," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 538-544, doi: 10.1109/IAEAC47372.2019.8998066. Available: https://ieeexplore-ieee-org.libweb.lib.utsa.edu/stamp/stamp.jsp?tp=&arnumber=8998066

Comment: Integration of Deep Belief Networks (DBN) and Q-Learning to construct a mission decision-making model. A common battlefield scenario is a situation with a lot of unknowns. They seek to implement a ML model to allow for decision making adaptations based on real-time ground truth rather than mission pre-planning objectives. The specific tasks studied are reconnaissance and air-to-air confrontation.

[56] Hongxing Zhang, Hui Gao, and Xin Su. 2020. Channel prediction based on adaptive structure extreme learning machine for UAV mmWave communications. In Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '19). Association for Computing Machinery, New York, NY, USA, 492–497. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3360774.3368199

Comment: This paper proposes the Adaptive-Structure Extreme Learning Machine (ASELM) algorithm for optimizing inter-UAV communication in a dynamic environment. The algorithm accurately predicts future characteristics of the communication channel to minimize signal loss and retransmission, thus improving energy efficiency.

[57] Hengpeng Guo, Bo Zhang, and Yuanzhong Fu. 2025. Multi-Agent Deep Reinforcement Learning-Based UAV Trajectory Optimization for Data Collection. In Proceedings of the 3rd International Conference on Signal Processing, Computer Networks and Communications (SPCNC '24). Association for Computing Machinery, New York, NY, USA, 394–399. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3712335.3712404

Comment: A study on employing multiple UAVs to perform data collection on field sensors. A novel metric is introduced that measures Age of Information vs unit of energy consumption to optimize the flight path for each UAV. And a deep reinforcement learning algorithm is used to optimize collaboration amongst all UAVs.

[58] Chao Zhang, Haipeng Yao, and Tianle Mai. 2024. Graph Transformer Aided Resource Virtualization Embedding in UAV Swarm Networks. In Proceedings of the International Conference on Computing, Machine Learning and Data Science (CMLDS '24). Association for Computing Machinery, New York, NY, USA, Article 36, 1–6. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3661725.3661763

Comment: This paper showcases a graph transformer network approach to enhance virtual network (VNE) on a UAV swarm network. They are able to show significant improvements in embedding efficiency.

[59] Le Qi and Wanyang Wang. 2024. Integrating Deep Learning Techniques for Enhanced Multi-Target Tracking in UAV Fire Control Systems. In Proceedings of the 2024 9th International Conference on Cyber Security and Information Engineering (ICCSIE '24). Association for Computing Machinery, New York, NY, USA, 865–870. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3689236.3696038

Comment: This paper introduces a multi-module deep learning network for multi-target tracking in UAV fire control systems. The proposed architecture improves handling of challenges such as rapid movement and occlusion. They also achieve a better latency value versus YOLOv4-DeepSORT and Faster R-CNN.

[60] Islam Güven and Evsen Yanmaz. 2023. Maintaining Connectivity for Multi-UAV Multi-Target Search Using Reinforcement Learning. In Proceedings of the Int'l ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications (DIVANet '23). Association for Computing Machinery, New York, NY, USA, 109–114. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3616392.3623414

Comment: The paper proposes a reinforcement learning framework for maintaining connectivity among multiple UAVs during multi-target search and rescue missions. It highlights the importance of balancing search efficiency with connectivity. And demonstrates that this approach is scalable and adaptable to various mission scenarios.

[61] Ning Wu, Li Li, Xingyu Liu, Ziwen Wang, and Tongyao Jia. 2024. Enhancing UAV Swarm Routing with Multi-Agent Attention Reinforcement Learning. In Proceedings of the 2023 13th International Conference on Communication and Network Security (ICCNS '23). Association for Computing Machinery, New York, NY, USA, 258–264. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3638782.3638822

Comment: This paper presents attention mechanisms to enhance a multi-agent reinforcement learning (MARL) algorithm to optimize routing strategies for UAV swarms. Attention mechanisms allow UAVs to focus on pertinent information from neighboring agents leading to more efficient and effective communication. This also enhances scalability due to the reduction in communication overhead.

[62] Yanfan Zhang, Hongyuan Zheng, and Xiangping Zhai. 2023. Deep Reinforcement Learning Based UAV Mission Planning with Charging Module. In Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things (CNIOT '23). Association for Computing Machinery, New York, NY, USA, 658–662. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3603781.3603897

Comment: This paper integrates charging considerations into the DRL to incorporate charging stops in route planning for UAVs. The DRL model allows the UAV to make dynamic decisions in real-time based on unforeseen events that require more energy usage, leading to path or charging schedule adjustments.

[63] Tuan Do Trong, Tran Bao Duy, Vu Dinh Khai, and Hoang-Anh Pham. 2023. Applying Deep Learning for UAV Obstacle Avoidance: A Case Study in High-Rise Fire Victim Search. In Proceedings of the 12th International Symposium on Information and Communication Technology (SOICT '23).

Association for Computing Machinery, New York, NY, USA, 831–837. Available: https://doi-org.libweb.lib.utsa.edu/10.1145/3628797.3628813

Comment: This study introduces a two-phase methodology for SAR missions. In phase one the UAV navigates to a particular location in a building leveraging a DL model for obstacle detection and avoidance. Stage two the UAV employs a brute force searching algorithm to locate potential victims in the structure.

[64] Zhao, Y., Nie, Z., Dong, K., Huang, Q. and Li, X., 2024. Autonomous Decision Making for UAV Cooperative Pursuit-Evasion Game with Reinforcement Learning. Available: *https://arxiv.org/html/2411.02983v1*

Comment: Introduction of the Multi-Environment Asynchronous Double Deep Q-Network (MEADDQN) algorithm, which integrates prioritized experience replay (PER) to improve data efficiency and training speed. Emphasizes role specialization in multi-UAV systems to enhance collaboration in a multi-objective mission.

[65] Jun Zhang and Khaled B. Letaief. 2019. Mobile edge intelligence and computing for the internet of vehicles. Proceed ings of the IEEE 108, 2 (2019), 246–261.

Comment: Survey on mobile edge AI in vehicular networks

[66] Jihong Park, Sumudu Samarakoon, Mehdi Bennis, and Mérouane Debbah. 2019. Wireless network intelligence at the edge. Proceedings of the IEEE 107, 11 (2019), 2204–2239.

Comment: Survey on edge intelligence in wireless networks

[67] X. Wang, Z. Tang, J. Guo, T. Meng, C. Wang, T. Wang and W. Jia, "Empowering Edge Intelligence: A comprehensive Survey on On-Device AI Models," *ACM Comput. Surv.,* vol. 57, no. 9, pp. 228:1 - 228:39, April 2025.

Comment: Survey on On-Device AI models

[68] Yuanming Shi, Kai Yang, Tao Jiang, Jun Zhang, and Khaled B. Letaief. 2020. Communication-efficient edge AI: Algo rithms and systems. IEEE Communications Surveys & Tutorials 22, 4 (2020), 2167–2191.

[69] Yueyue Dai, Ke Zhang, Sabita Maharjan, and Yan Zhang. 2020. Edge intelligence for energy-efficient computa tion offloading and resource allocation in 5G beyond. IEEE Transactions on Vehicular Technology 69, 10 (2020), 12175–12186.

[70] G. Kendall, "Real Clear Science," RealClear Media Group, 02 07 2019. [Online]. Available: https://www.realclearscience.com/articles/2019/07/02/your_mobile_phone_vs_apollo_11s_guidance_computer_111026.html. [Accessed 19 04 2025].

[71] S. S, U. A, P. N. Srivatsa, S. B and S. S, "Artificial Intelligence-based Voice Assistant," in *Smart Trends in Systems, Security, and Sustainability*, 2020.

[72] Linjuan Ma and Fuquan Zhang. 2021. End-to-end predictive intelligence diagnosis in brain tumor using lightweight neural network. Applied Soft Computing 111 (2021), 107666.

[73] He, K., Zhang, X., Ren, S., & Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[74] Y. Roh, G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," *IEEE Transactions on Knowledge and Data Engineering,* vol. 33, no. 4, pp. 1328-1348, 2021.

[75] Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. 2019. Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In *Proceedings of the 3rd international workshop on data management for end-to-end machine learning* pp. 1-4.

[76] B. Zoph and Q. V. Le, "Neural Architecture Search With Reinforcement Learning," in *ICLR*, 2017.

[77] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, "Language Models are Few-Shot Learners," Open AI, 2020.

[82] S. Gao, F. Huang, J. Pei and H. Huang, "Discrete Model Compression with Resource Constraint for Deep Neural Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, 2020.

[83] "Coral," Google, 2020. [Online]. Available: https://coral.ai/products/. [Accessed April 2025].

[84] Seyedehfaezeh Hosseininoorbin, Siamak Layeghy, Brano Kusy, Raja Jurdak, Marius Portmann, “Exploring Edge TPU for deep feed-forward neural networks”, in Internet of Things, vol. 22, 2023. Available: https://doi.org/10.1016/j.iot.2023.100749

[85] "Robotics and Edge AI," Nvidia, 2025. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/. [Accessed April 2025].

[86] D. Narayanan, A. Harlap, A. Phanishayee, V. Seshadri, N. R. Devanur, G. R. Ganger, P. B. Gibbons and M. Zaharia, "PipeDream: Generalized Pipeline Parallelism for DNN Training," in *SOSP*, Huntsville, ON, Canada, 2019.