Energy-Efficient YOLO with Knowledge Distillation and Dynamic Energy Control for Edge Devices

Anggi Andriyadi[1,3], Chandra Wijaya[1,4], Shih-Yen Chen[2], Ding-Hsiang Huang[1], and Chao-Tung Yang[2,5](✉)

[1] Department of Industrial Engineering and Enterprise Information, Tunghai University, No. 1727, Sec. 4 Boulevard, Xitun District, Taichung 407224, Taiwan

[2] Department of Computer Science, Tunghai University, No. 1727, Sec. 4 Boulevard, Xitun District, Taichung 407224, Taiwan
ctyang@thu.edu.tw

[3] Information System Department, Institut Informatika dan Bisnis Darmajaya, Bandar Lampung 35141, Lampung, Indonesia

[4] Center for Data Science and Artificial Intelligence System, Informatics Department, Parahyangan Catholic University, Bandung 40141, West Java, Indonesia

[5] Research Center for Smart Sustainable Circular Economy, Tunghai University, No. 1727, Sec. 4 Boulevard, Xitun District, Taichung 407224, Taiwan

Abstract. This paper discusses the ongoing development of an energy-efficient YOLO-based fire detection system optimized for edge devices. Using Knowledge Distillation, we compress the YOLOv8m model into YOLOv8n, making it more suitable for deployment on energy-constrained edge devices while maintaining its accuracy. Additionally, we are designing a real-time dynamic energy control mechanism to manage energy usage during the inference process based on real-time power monitoring. Initial results demonstrate that the proposed method reduces model size and power consumption without compromising performance.

Keywords: YOLO · Knowledge Distillation · Dynamic Energy Control · Edge AI

1 Introduction

Intelligent, energy-efficient infrastructure monitoring for fire detection [1] requires integrating AIoT with smart grids. However, deploying models such as YOLO (You Only Look Once) on edge devices is hindered by high computational complexity and energy demands [2]. While YOLO's real-time capabilities make it ideal for fire detection, its power consumption renders it unsuitable for energy-constrained edge devices in smart grids [3].

D.-J. Deng and J.-C. Chen (Eds.): SGIoT 2024, LNICST 633, pp. 11–18, 2026.
https://doi.org/10.1007/978-3-031-93825-2_2

In this study, we compress YOLOv8m into a more efficient model (YOLOv8n Distill) using Knowledge Distillation (KD) while maintaining accuracy [4,5]. With this reduction in size and complexity, it is perfect for deployment on edge devices such as Jetson Nano for fire detection. We also present ongoing a dynamic energy control mechanism using real-time power monitoring via the Tegrastats API. It optimizes energy consumption during fire detection inference by dynamically adjusting model activity according to real-time power data.

2 Related Work

Previous research has explored Knowledge Distillation (KD) to optimize object detection models for edge devices. Liu et al. (2023) improved YOLO's robustness in foggy environments using KD with transformers [6], while Zhou et al. (2022) adapted YOLOv5 for domain-specific detection [7]. Li et al. (2023) compressed YOLOv5 for agricultural use, showcasing its real-time optimization [8]. Chan et al. (2024) applied Kafka for flame and smoke detection on edge devices [9]. Talaat and ZainEldin (2023) enhanced fire detection with YOLOv8 for smart cities [10].

3 Methodology

3.1 Knowledge Distillation

Stochastic gradient descent (SGD) with a 0.01 learning rate was used to train the YOLOv8m model on a fire and non-fire dataset [11]. Then, As illustrated on Fig. 1, knowledge distillation was used to transfer knowledge from the YOLOv8m (teacher) model to the smaller YOLOv8s (student) model [12]. The student model used soft labels from the teacher and complex labels from the dataset to mimic the teacher's predictions with reduced computational and memory demands [13]. This approach achieves high accuracy and efficiency, making it suitable for energy-constrained edge devices [13].

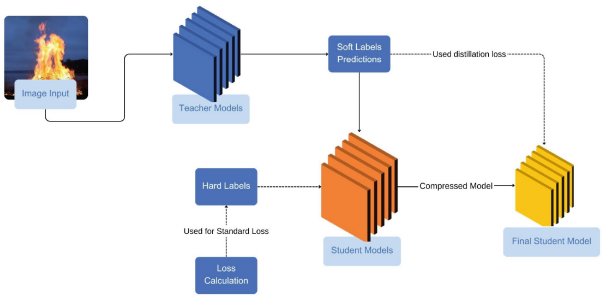

Fig. 1. Knowledge Distillation Process for YOLO Optimization.

3.2 Dynamic Energy Control Concept

We use K-Means clustering to group energy data into low, average, and peak clusters. The system pauses object detection during peak usage and resumes when consumption normalizes. YOLO adapts to CPU and GPU power levels, stopping inference at peak usage to save energy and prevent overheating, as shown in Fig. 2.

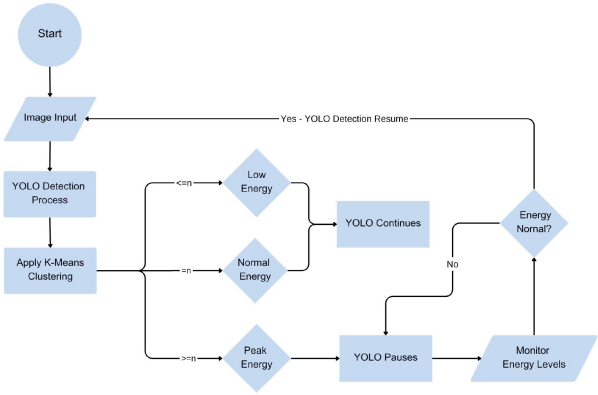

Fig. 2. Dynamic Energy Control Flowchart Propose.

4 Results and Discussions

4.1 Energy, Resource, and Performance Analysis

During YOLOv8m and YOLOv8n inference, we monitored energy consumption and system utilization. As presented in Table 1, YOLOv8n Distill is 60% more energy-efficient, using 6.58 mW of CPU+GPU power compared to 16.59 mW for YOLOv8m.

Table 1. Summary of Resource Utilization for YOLOv8m and YOLOv8n Distill States

Metric	YOLOv8m (Avg)	YOLOv8n Distill (Avg)
RAM Usage (MB)	13 MB	13 MB
CPU Usage (%)	23.49%	10.32%
GPU Usage (%)	2.60%	3.00%
Power VDD_IN (mW)	0 mW	0 mW
Power CPU+GPU (mW)	16.59 mW	6.58 mW
CPU Temperature (C)	55.42 °C	54.35 °C
GPU Temperature (C)	51.12 °C	50.06 °C

Furthermore, YOLOv8n Distill uses only 10.32% of CPU compared to YOLOv8m's 23.49%, and the same RAM, GPU usage, and temperatures, as can be seen in Fig. 3, indicating energy savings without performance loss.

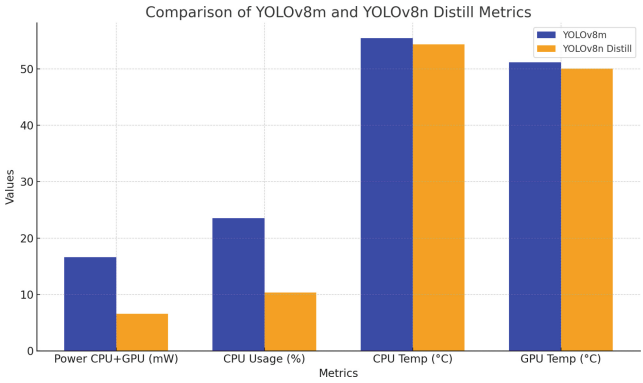

Fig. 3. Comparison of YOLOv8m and YOLOv8n Distill Energy Consumption

Additionally, as visualized in Fig. 4, YOLOv8n Distill performs better than YOLOv8m across the board: a higher FPS (31.43 compared to 13.33) and faster inference time (0.032 s compared to 0.090 s). These improvements make YOLOv8n Distill more suitable for edge devices in real-time applications.

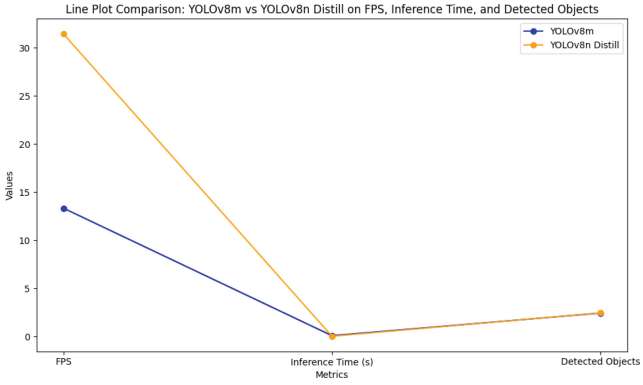

Fig. 4. Comparison of YOLOv8m and YOLOv8n Distill Detection Performance

As listed in Table 2, both models achieve similar detection accuracy, with YOLOv8n Distill detecting an average of 2.46 objects, compared to YOLOv8m's 2.44. Although YOLOv8n Distill has a slightly lower F1 score (0.92 vs. 0.94 for YOLOv8m), the substantial gains in speed and power consumption make it

an optimal choice for energy-constrained environments, such as edge AI deployments. Meanwhile, Fig. 5 shows that YOLOv8n Distill achieves significantly better FPS, with 30.43 compared to 13.06 for YOLOv8m, and has a faster inference time of 0.0329 s compared to 0.0765 s.

Table 2. Average Performance Comparison

Metric	YOLOv8m (Avg)	YOLOv8n Distill (Avg)
FPS	13.33	31.43
Inference Time (s)	0.090	0.032
Detected Objects	2.44	2.46
F1-Score	0.94	0.92

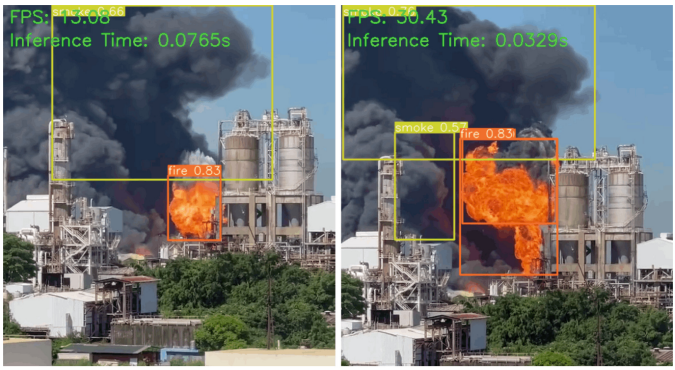

Fig. 5. Comparison of YOLOv8m and YOLOv8n Comparison Performance

4.2 Dynamic Energy Control Progress

To optimize energy consumption during YOLO inference, we have made significant progress in developing a dynamic energy control system that leverages K-Means clustering for real-time power usage data. The system currently categorizes power consumption into three clusters: low, normal, and peak. Although the system does not yet include an alerting mechanism, it detects changes in energy consumption and displays the current energy state (low, normal, or peak) on the screen in real time. This allows for immediate awareness of energy use during inference tasks.

At this stage, K-Means clustering has been successfully integrated into the system, enabling real-time classification of energy states during video inference, as As demonstrated in Fig. 6. However, the YOLO object detection model has not

been fully integrated into this energy control framework. In upcoming studies, YOLO will be integrated with the dynamic energy control system to enable real-time object detection and efficient energy management on edge devices such as Jetson.

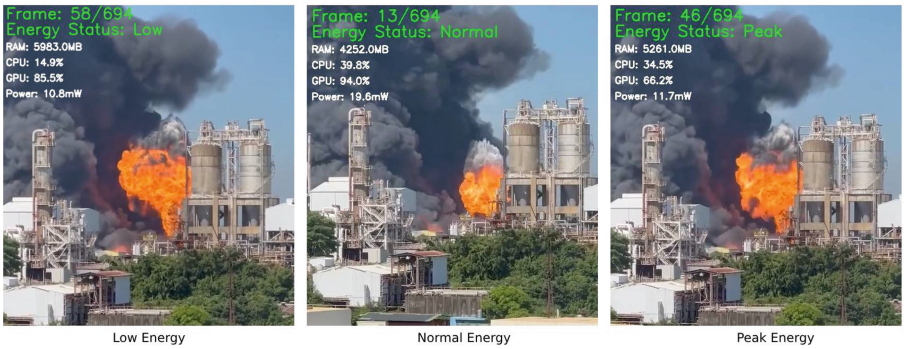

Fig. 6. Energy Status Classification Using K-Means Clustering

The pseudo-code for the dynamic energy control algorithm is provided below, illustrating the core logic of the energy status classification and control mechanism.

Algorithm 1. Dynamic Energy Control with K-Means

```
Initialize K-Means clustering with three clusters: Low, Normal, Peak
Collect power consumption data from the Tegrastats API
while object detection is running do
    Classify current energy status using K-Means
    if energy status is Peak then
        Pause YOLO object detection
    else
        Continue YOLO object detection
    end if
end while
```

5 Conclusion and Future Work

This paper presents the progress of integrating energy-efficient mechanisms into fire detection models for edge devices. Based on real-time data from edge devices, we implemented a dynamic energy control system using K-means clustering to classify energy usage into low, normal, and peak categories. This enables adaptive

control of system performance, which is important for resource-limited environments. Although the YOLO model has yet to be fully implemented with this system, the energy classification framework provides a basis for further optimizations.

Future work will focus on integrating our YOLO Distillation Model with the energy control mechanism. The existing K-Means-based energy classification system will be further refined, enabling real-time fire and smoke detection while maintaining energy efficiency. This enhancement will improve the model's suitability for applications in energy-constrained environments such as smart grids, industrial safety systems, and IoT-based fire monitoring.

Acknowledgements. This work was sponsored by the National Science and Technology Council (NSTC), Taiwan, under Grant No. 113-2622-E-029-003, and 113-2221-E-029-028-MY3.

References

1. Kumar, N.M., et al.: Distributed energy resources and the application of AI, IoT, and blockchain in smart grids. Energies **13**(21), art. no. 5739 (2020). https://doi.org/10.3390/en13215739
2. Li, C., Xu, R., Lv, Y., Zhao, Y., Jing, W.: Edge real-time object detection and DPU-Based hardware implementation for optical remote sensing images. Remote Sens. **15**(16), art. no. 3975 (2023). https://doi.org/10.3390/rs15163975
3. Shin, D.-J., Kim, J.-J.: A deep learning framework performance evaluation to use YOLO in Nvidia jetson platform. Appl. Sci. **12**(8), art. no. 3734 (2022). https://doi.org/10.3390/app12083734
4. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: a survey. Int. J. Comput. Vision **129**(6), 1789–1819 (2021). https://doi.org/10.1007/s11263-021-01453-z
5. Zhai, Z., Liang, J., Cheng, B., Zhao, L., Qian, J.: Strengthening attention: knowledge distillation via cross-layer feature fusion for image classification. Int. J. Multimed. Info. Retr. **13**, art. no. 23 (2024). https://doi.org/10.1007/s13735-024-00332-w
6. Liu, X., Zhang, B., Liu, N.: CAST-YOLO: an improved YOLO based on a cross-attention strategy transformer for foggy weather adaptive detection. Appl. Sci. **13**(2), art. no. 1176 (2023). https://doi.org/10.3390/app13021176
7. Zhou, H., Jiang, F., Lu, H.: SSDA-YOLO: semi-supervised domain adaptive YOLO for cross-domain object detection. Comput. Vis. Image Underst. **229**, art. no. 103649 (2023). https://doi.org/10.1016/j.cviu.2023.103649
8. Li, Y., Gong, Z., Zhou, Y., He, Y., Huang, R.: Production evaluation of citrus fruits based on the yolov5 compressed by knowledge distillation. In: 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Rio de Janeiro, Brazil, pp. 1938–1943 (2023). https://doi.org/10.1109/CSCWD57460.2023.10152740
9. Chan, Y.-W., Liu, J.-C., Kristiani, E., Lien, K.-Y., Yang, C.-T.: Flame and smoke detection using Kafka on edge devices. Internet Things (Netherlands) **27**, art. no. 101309 (2024). https://doi.org/10.1016/j.iot.2024.101309

10. Talaat, F.M., ZainEldin, H.: An improved fire detection approach based on YOLO-v8 for smart cities. Neural Comput. Appl. **35**, 20939–20954 (2023). https://doi.org/10.1007/s00521-023-08809-1
11. Tian, Y., Zhang, Y., Zhang, H.: Recent advances in stochastic gradient descent in deep learning. Mathematics **11**(3), art. no. 682 (2023). https://doi.org/10.3390/math11030682
12. Yang, J., Zhu, X., Bulat, A., Martinez, B., Tzimiropoulos, G.: Knowledge distillation meets open-set semi-supervised learning. Int. J. Comput. Vis. **2024**(8), art. no. 02192–7 (2024). https://doi.org/10.1007/s11263-024-02192-7
13. Khaider, Y., Rahhali, D., Alami, H., En Nahnahi, N.: Exploring the knowledge distillation. In: Tabaa, M., Badir, H., Bellatreche, L., Boulmakoul, A., Lbath, A., Monteiro, F. (eds) New Technologies, Artificial Intelligence and Smart Data. INTIS INTIS 2022 2023. Communications in Computer and Information Science, vol 1728. Springer, Cham. **21**(11), art. no. 47366–1 (2023). https://doi.org/10.1007/978-3-031-47366-1_3