Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 2, 2024

Enhancing Kitchen Independence: Deep Learning-Based Object Detection for Visually Impaired Assistance

Bo Dang[1, *], Danqing Ma[2], Shaojie Li[3], Xinqi Dong[4], Hengyi Zang[5], Rui Ding[6]

[1] Computer Science, San Francisco Bay University, Fremont CA, U.S.
[2] Computer Science, Stevens Institute of Technology, Hoboken NJ, U.S.
[3] Computer Technology, Huacong Qingjiao Information Technology Co., Ltd, Beijing, China
[4] Computer Science, University of Maine at Presque Isle, Presque Isle ME, U.S.
[5] Physics and Mathematics, Universitario Tecnológico Universitam, Tijuana, Mexico
[6] Independent Research, US

* Bo Dang is the corresponding author, E-mail: dangdaxia@gmail.com

Abstract: Visually impaired individuals face substantial challenges in kitchens, where identifying objects accurately is crucial yet difficult due to the complexity and variability of the environment. Traditional object detection[1] methods fall short in these settings, struggling with the assortment of items. This research highlights the need for advanced, kitchen-specific solutions that leverage deep learning to improve detection accuracy and offer real-time, interactive guidance through speech technologies. By focusing on the unique demands of kitchen environments, the proposed system aims to significantly enhance the autonomy and safety of visually impaired users, presenting a notable advancement in assistive technology. The effectiveness of this approach is assessed by its ability to accurately identify kitchen items for visually impaired individuals.

Keywords: Machine Learning, Object Detection, MobileNet SSD, Tensorflow Lite, Deep Learning, Transfer Learning, Text to Speech.

1. Introduction

The everyday kitchen becomes a challenging place for visually impaired individuals due to the complex task of identifying objects. For those without full visual capabilities, distinguishing between various kitchen items is not just difficult; it's a barrier to independence and safety. Current object detection technologies, while advanced, often fail to cater to the specific needs of visually impaired users in kitchen settings. These methods typically struggle with identifying kitchens items, where objects can be obscured, posing a significant challenge to achieving the level of accuracy and responsiveness needed for real-time assistance.

In response to these challenges, there has been a shift towards leveraging mobile-based object detection technologies that are both efficient and capable of operating within the constrained resources of mobile devices. MobileNet SSD, optimized within the TensorFlow Lite framework, exemplifies this shift. It provides a solution that balances the need for speed and accuracy, crucial for real-time assistive technologies. TensorFlow Lite enables the deployment of these complex models directly on mobile devices, ensuring fast, on-device processing that respects user privacy and operates independently of network constraints.

This paper introduces an approach that utilizes transfer learning to fine-tune a pre-trained MobileNet SSD model, making it more adept at recognizing kitchen-specific items [2]. By curating a dataset tailored for the kitchen environment, we enhance the model's ability to accurately identify objects. Additionally, the integration of Automatic Speech Recognition (ASR) [3] and Text-to-Speech (TTS) technologies provides an interactive layer, enabling users to receive auditory feedback and commands, facilitating object localization and navigation within the kitchen. Our research evaluates the system's performance, focusing on detection accuracy. By addressing these specific needs, our study contributes significant advancements to the field of assistive technology, emphasizing the importance of developing accessible, inclusive tools that empower visually impaired users to navigate their kitchens—and by extension, their lives—with greater independence and safety.

2. Background Technologies

2.1. Machine Learning in Mobile Systems

Machine Learning (ML) has evolved to become a cornerstone of mobile applications, enabling devices to learn from data and make intelligent decisions without explicit programming [4, 5]. In the context of mobile device implementation, ML techniques are optimized for efficiency and performance, catering to the limited computational resources available [6]. This is particularly evident in applications such as real-time object detection, where ML models [7], trained either in a supervised, unsupervised, or reinforcement learning setting, are deployed to perform tasks directly on the device. The integration of ML in mobile systems empowers devices with capabilities like predictive text input, voice recognition, and context-aware recommendations, making them smarter and more interactive. The adaptability of ML models, combined with their ability to learn from data [8], makes them ideal for applications requiring real-time processing and decision-making, further enhancing user experience and device functionality.

2.2. Deep Learning for On-Device Intelligence

Deep Learning (DL), a specialized subset of ML, has been instrumental in pushing the boundaries of what's possible with mobile computing [9]. Through the use of multi-layered neural networks [10], DL models excel at processing and interpreting complex, unstructured data directly on devices [11]. This ability is critical for implementing advanced object detection systems like MobileNet SSD within mobile platforms [12]. DL's capability for automatic feature extraction means that models can learn to identify relevant patterns and information from raw data without manual intervention [13], optimizing both the accuracy and efficiency of on-device inference. The advent of TensorFlow Lite has further facilitated the deployment of DL models on mobile devices, ensuring they run smoothly without compromising system resources. This makes DL an indispensable tool for developers seeking to incorporate sophisticated AI functionalities into mobile devices, delivering high-performance applications that respond and adapt to their environment in real time.

2.3. Transfer Learning

Transfer learning offers a transformative approach to object detection on customized image datasets by leveraging the capabilities of pre-trained models to enhance accuracy for new, task-specific challenges [14]. By fine-tuning models [15] that have been pre-trained on large, comprehensive datasets like ImageNet, researchers can jumpstart the object detection process on their specialized datasets. This fine-tuning, which involves modifying the number of classes and their labels in the pre-trained model to match the customized dataset, ensures that the learned features are more relevant to the specific detection tasks at hand.

3. Proposed System

3.1. Overall Architecture

The process begins when a user clicks the record button, which activates the voice input. As ASR requires decode speech with high speed, we sent captured user's spoken words to Google's Speech To Text service. This service translates the spoken words into text by recognizing and processing the user's speech.

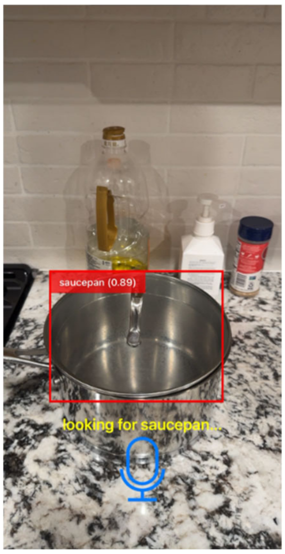

Figure 1. Voice-Controlled Object Detection System Screenshot

Once the spoken words are converted into text [16], this text is sent to a server where it undergoes natural language processing (NLP). The server evaluates the user's intention behind the spoken words [17]. This step is crucial for understanding what the user wants to do [18] - for instance, whether they are asking for information, giving a command, or requesting assistance in finding an object [19, 20].

If the user's intention is identified as something not related to object detection (for example, asking a general question or making a non-relevant request), the system responds accordingly. It might remind the user to change their request content to something that aligns with its capabilities, particularly focusing on object detection tasks.

In cases where the user's intention is to find a specific object [21], like “find the sauce can,” the system proceeds to the object detection phase [22]. Here, real-time video feed from the user's camera is processed using TensorFlow Lite. TensorFlow Lite analyzes the video feed to detect and locate the requested object within the camera's view.

After successfully identifying the position of the object, the system uses Google's Text to Speech AI to generate a spoken response. This AI converts the system's findings into speech, informing the user about the location or status of the requested object. For example, it might say, "The sauce can is on the second shelf on your right."

Throughout this process, the system leverages advanced technologies like speech recognition, NLP, machine learning, and AI to interact seamlessly with the user and provide assistance in real-time.

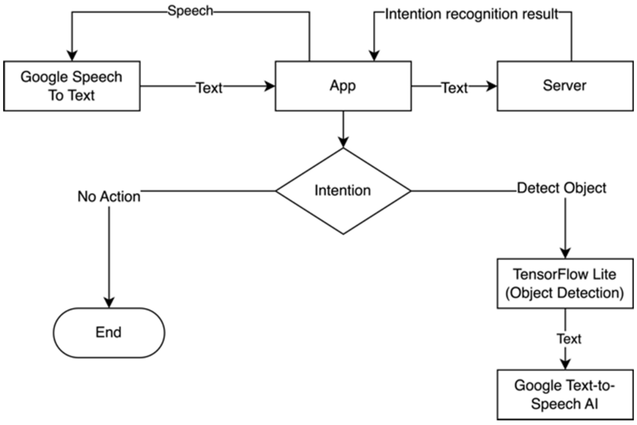

Figure 2. Voice-Controlled Object Detection System Workflow

3.2. Object Detection Module

The proposed object detection framework involves carefully tailored adjustments and refinements to the standard MobileNet SSD architecture. The training process utilized a transfer learning approach [23]. The experiment is based on the CMU Kitchen Occlusion Dataset, which consisted of 1,600 images featuring eight distinct items: cups, pitchers, shakers, thermoses, saucepans, scissors, and baking pans.

4. Experiment

4.1. Experimental Settings

During the fine-tuning process [24] of the MobileNet SSD model, we made several key adjustments. This included revising full connect and softmax neural network layers, reducing the learning rate by 40%. We employed weighted loss functions to manage the imbalanced data [25], considering the varying frequency of objects in the dataset. We also utilized a variety of data augmentation techniques to enhance the model's robustness. This involved rotating images [26, 27] by up to 30 degrees, introducing Gaussian noise with a variance of 0.01, thus providing a broader range of training examples to the model.

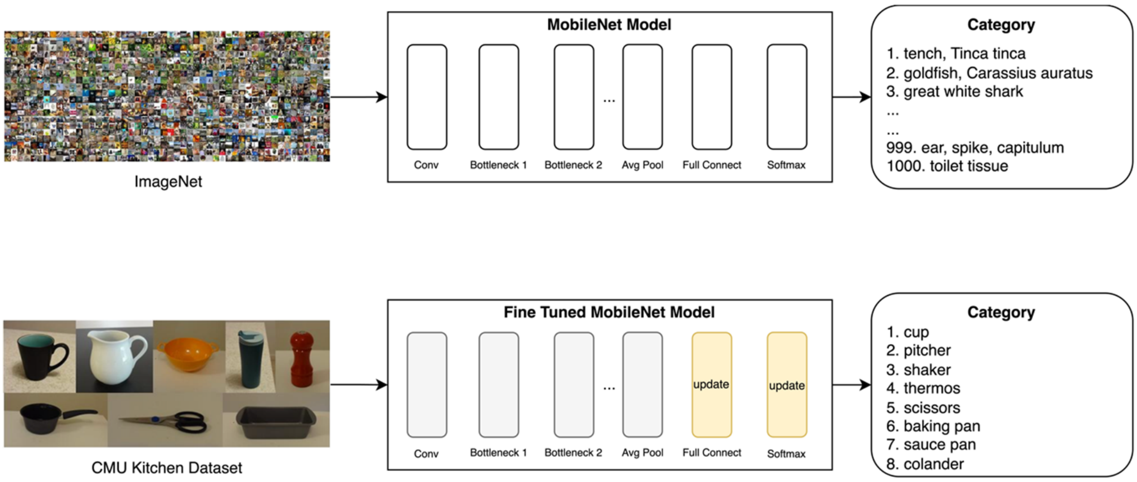

Figure 3. Transfer Learning for Object Detection: From General to Specific Datasets

Hyperparameters were meticulously optimized [28], featuring a starting learning rate of 0.001, decaying by 0.1 every 3,000 steps, a batch size of 32, and a total of 50 training epochs. The loss function was modified to give greater emphasis to localization accuracy, changing the weight ratio from 1:1 to 2:1 in favor of localization loss over confidence loss. A scale-invariant component was also introduced to enhance the detection of smaller objects. Evaluation metrics reflected the effectiveness of these extensive efforts; the model's mean Average Precision (mAP) improved significantly from a baseline of 68% to an impressive 82% post-fine-tuning. This improvement was particularly evident in the detection of kitchen items like scissors, where the mAP increased from 78% to 87%. Remarkably, these enhancements in detection accuracy were achieved with only a 10% increase in inference time, highlighting the efficiency and effectiveness of the fine-tuning process for this specialized application.

4.2. Experiment Result

This table provides a comparison of different object detection models in terms of their accuracy in detecting kitchen objects, inference speed and model size. The table is designed to highlight the strengths and weaknesses of each model within the context of a kitchen environment, where factors like accuracy for various kitchen items are considered. MobileNet SSD is selected for its efficient operation on mobile devices with a good balance of speed and accuracy. Other models like EfficientDet offer higher accuracy but at different trade-offs in speed and size, which could impact their practicality for real-time applications on mobile devices. Faster R-CNN provides high precision but may be too slow for real-time interaction. SqueezeNet is the quickest and has the smallest model size, which might be advantageous for some real-time applications with less demand for high accuracy. RetinaNet is known for handling class imbalance with its focal loss function, which could be particularly useful in kitchens where some objects are much rarer than others.

Table 1. Comparative Analysis of Object Detection Model Performance

Model	Accuracy	Speed(ms)	Model Size (MB)
MobileNet SSD	88.5%	45	32
Faster R-CNN	89.8%	120	150
SqueezeNet	84.6%	30	48
EfficientDet	89.4%	62	17
RetinaNet	90.1%	65	145

The results we obtained demonstrated varying levels of success across different objects. For instance, the model exhibited a high precision of 95% and recall of 88% for shakers, indicating its effectiveness in detecting these items even when occluded. Conversely, items like baking pans showed lower metrics, with a precision of 89% and recall of 73%, highlighting certain challenges in their detection under similar conditions.

Table 2. Performance Metrics for Object Recognition in Various Kitchen Items

Category	Precision	Recall	F1-Score	Accuracy
cup	0.72	0.86	0.78	0.81
pitcher	0.95	0.98	0.96	0.75
shaker	0.94	0.88	0.91	0.99
thermos	0.99	0.87	0.93	0.78
sauce pan	0.84	0.75	0.79	0.83
scissors	0.86	0.75	0.8	0.87
baking pan	0.89	0.73	0.8	0.75

5. Conclusion

This research presents a comprehensive and innovative approach to assist visually impaired individuals in kitchen environments [29]. The core of this system lies in the utilization of MobileNet SSD within the TensorFlow Lite framework, optimized for efficient operation on mobile devices. By employing transfer learning techniques, the pre-trained MobileNet SSD model has been adeptly fine-tuned on a kitchen-specific dataset, significantly enhancing its accuracy and responsiveness in identifying common kitchen items.

The integration of ASR and TTS technology forms a crucial part of the proposed solution, enabling visually impaired users to interact with the system through voice commands and receive auditory guidance. This feature is particularly vital for visually impaired individuals. The system's ability to provide real-time object detection and vocal instructions empowers users to navigate kitchen spaces more safely and independently, addressing a significant challenge in their daily lives.

The performance evaluation of our system highlights its potential in enhancing the quality of life for visually impaired individuals. The proposed framework demonstrates high accuracy in kitchen environments.

This research contributes to the field of deep learning by offering a viable and user-friendly solution for visually impaired individuals in kitchen environments [30]. The methodology and findings of this study not only pave the way for further advancements in object detection technologies for assistive purposes but also underscore the importance of incorporating accessibility features into everyday technology. As technology continues to advance, it holds great promise for further empowering individuals with disabilities, enabling them to navigate and interact with the world in ways that were previously challenging or impossible.

References

[1] Liu, Y., Yang, H. & Wu, C. Unveiling Patterns: A Study on Semi-Supervised Classification of Strip Surface Defects. IEEE Access 11, 119933–119946 (2023).

[2] Qiao, Y., Ni, F., Xia, T., Chen, W. & Xiong, J. AUTOMATIC RECOGNITION OF STATIC PHENOMENA IN RETOUCHED IMAGES: A NOVEL APPROACH. in The 1st International scientific and practical conference "Advanced technologies for the implementation of new ideas"(January 09-12, 2024) Brussels, Belgium. International Science Group. 2024. 349 p. 287 (2024).

[3] Liang, Z. et al. Improving Code-Switching and Named Entity Recognition in ASR with Speech Editing based Data Augmentation. Preprint at (2023).

[4] Ni, F., Zang, H. & Qiao, Y. SMARTFIX: LEVERAGING MACHINE LEARNING FOR PROACTIVE EQUIPMENT MAINTENANCE IN INDUSTRY 4.0. in The 2nd International scientific and practical conference "Innovations in education: prospects and challenges of today"(January 16-19, 2024) Sofia, Bulgaria. International Science Group. 2024. 389 p. 313 (2024).

[5] Pan, Z. et al. Ising-Traffic: Using Ising Machine Learning to Predict Traffic Congestion under Uncertainty. Proceedings of the AAAI Conference on Artificial Intelligence 37, 9354–9363 (2023).

[6] Liu, S., Wu, K., Jiang, C., Huang, B. & Ma, D. Financial Time-Series Forecasting: Towards Synergizing Performance And Interpretability Within a Hybrid Machine Learning Approach. Preprint at (2023).

[7] Li, S. S. et al. A Quantitative Approach to Understand Self-Supervised Models as Cross-lingual Feature Extractors. Preprint at (2023).

[8] Qiao, Y., Jin, J., Ni, F., Yu, J. & Chen, W. APPLICATION OF MACHINE LEARNING IN FINANCIAL RISK EARLY WARNING AND REGIONAL PREVENTION AND CONTROL: A SYSTEMATIC ANALYSIS BASED ON SHAP. WORLD TRENDS, REALITIES AND ACCOMPANYING PROBLEMS OF DEVELOPMENT 331, (2023).

[9] Wei, J., Zhang, Y., Zhou, Z., Li, Z. & Al Faruque, M. A. Leaky DNN: Stealing Deep-Learning Model Secret with GPU Context-Switching Side-Channel. in 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN) 125–137 (2020). doi:10.1109/DSN48063.2020.00031.

[10] Xiao, T., Zeng, L., Shi, X., Zhu, X. & Wu, G. Dual-Graph Learning Convolutional Networks for Interpretable Alzheimer's Disease Diagnosis. in International Conference on Medical Image Computing and Computer-Assisted Intervention 406–415 (2022).

[11] Zhang, X. et al. A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research. Journal of Theoretical and Applied Electronic Commerce Research 18, 2188–2216 (2023).

[12] Mittal, G. et al. HyperSTAR: Task-Aware Hyperparameters for Deep Networks. Preprint at (2020).

[13] Su, J., Nair, S. & Popokh, L. Optimal Resource Allocation in SDN/NFV-Enabled Networks via Deep Reinforcement Learning. in 2022 IEEE Ninth International Conference on Communications and Networking (ComNet) 1–7 (2022). doi:10.1109/ComNet55492.2022.9998475.

[14] Li, Y., Liu, T., Jiang, D. & Meng, T. Transfer-learning-based Network Traffic Automatic Generation Framework. in 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP) 851–854 (2021). doi:10.1109/ICSP51882.2021.9408767.

[15] Liu, K., Han, Y., Gong, Z. & Xu, H. Low-Data Drug Design with Few-Shot Generative Domain Adaptation. Bioengineering 10, (2023).

[16] Si, S. et al. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. in Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023).

[17] Jin, X., Larson, J., Yang, W. & Lin, Z. Binary Code Summarization: Benchmarking ChatGPT/GPT-4 and Other Large Language Models. Preprint at (2023).

[18] Gu, Y. et al. Mutual Correlation Attentive Factors in Dyadic Fusion Networks for Speech Emotion Recognition. in Proceedings of the 27th ACM International Conference on Multimedia 157–166 (Association for Computing Machinery, New York, NY, USA, 2019). doi:10.1145/3343031.3351039.

[19] Jin, X. & Wang, Y. Understand Legal Documents with Contextualized Large Language Models. arXiv preprint arXiv:2303.12135 (2023).

[20] Chen, K. et al. Chemist-X: Large Language Model-empowered Agent for Reaction Condition Recommendation in Chemical Synthesis. Preprint at (2024).

[21] Chen, Y., Arkin, J., Zhang, Y., Roy, N. & Fan, C. Scalable Multi-Robot Collaboration with Large Language Models: Centralized or Decentralized Systems? Preprint at (2023).

[22] Guo, Z. & Cao, Y. SA-CNN: Application to text categorization issues using simulated annealing-based convolutional neural network optimization. in Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering (ACM, 2022). doi:10.1145/3573428.3573788.

[23] Li, Q., Hu, Y., Dong, Y., Zhang, D. & Chen, Y. Focus on Hiders: Exploring Hidden Threats for Enhancing Adversarial Training. arXiv preprint arXiv:2312.07067 (2023).

[24] Xiao, Y. & Alam, F. Nexus at ArAIEval Shared Task: Fine-Tuning Arabic Language Models for Propaganda and Disinformation Detection. Preprint at (2023).

[25] Ukey, N. et al. Survey on Exact kNN Queries over High-Dimensional Data Space. Sensors 23, (2023).

[26] Wantlin, K. et al. Benchmd: A benchmark for modality-agnostic learning on medical images and sensors. arXiv preprint arXiv:2304.08486 (2023).

[27] Li, L. CPSeg: Finer-grained Image Semantic Segmentation via Chain-of-Thought Language Prompting. Preprint at (2023).

[28] Popokh, L., Su, J., Nair, S. & Olinick, E. IllumiCore: Optimization Modeling and Implementation for Efficient VNF Placement. in 2021 International Conference on Software, Telecommunications and Computer Networks (SoftCOM) 1–7 (2021).

[29] Wang, H. et al. Quantpipe: Applying Adaptive Post-Training Quantization For Distributed Transformer Pipelines In Dynamic Edge Environments. in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1–5 (2023).

[30] Xu, C., Yu, J., Chen, W. & Xiong, J. DEEP LEARNING IN PHOTOVOLTAIC POWER GENERATION FORECASTING: CNN-LSTM HYBRID NEURAL NETWORK EXPLORATION AND RESEARCH. in The 3rd International scientific and practical conference "Technologies in education in schools and universities"(January 23-26, 2024) Athens, Greece. International Science Group. 2024. 363 p. 295 (2024).