Literature Review

□ **LCRM: Layer-Wise Complexity Reduction Method for CNN Model Optimization on End Devices- Hanan Hussain, P. S. Tamizharasan, and Praveen Kumar Yadav.**

The research paper proposes an optimization algorithm called Layer-wise Complexity Reduction Method (LCRM) to address these challenges by converting accuracy-focused CNNs into lightweight models. The authors evaluate the standard convolution layers and replace them with the most efficient combination of substitutional convolutions based on the output channel size. The primary goal is to reduce the computational complexity of the parent models and the hardware requirements. The effectiveness of the framework is assessed by evaluating its performance on various standard CNN models, including AlexNet, VGG-9, U-Net, and Retinex-Net, for different applications such as image classification, optical character recognition, image segmentation, and image enhancement. The experimental results show up to a 95% reduction in inference latency and up to 93% reduction in energy consumption when deployed on GPU. Furthermore, the LCRM-optimized CNN models are compared with state-of-the-art CNN optimization methods, including pruning, quantization, clustering, and their four cascaded optimization methods, by deploying them on Raspberry Pi-4. The profiling experiments performed on each model demonstrate that the LCRM-optimized CNN models achieve comparable or better accuracy than the parent models while providing added benefits such as a 62.84% reduction in inference latency on end devices with significant memory compression and complexity reductions.

- **U-Net: Convolutional Networks for Biomedical Image Segmentation- Olaf Ronneberger, Philipp Fischer, Thomas Brox.**

In this paper, the authors present a network and training strategy that relies on the strong use of data augmentation to leverage available annotated samples more efficiently. The architecture comprises a contracting path to capture context and a symmetric expanding path that enables precise localization. The authors demonstrate that such a network can be trained end-to-end from very few images and outperforms the prior best method (a sliding-window convolutional network) on the ISBI challenge for segmentation of neuronal structures in electron microscopic stacks. Utilizing the same network trained on transmitted light microscopy images (phase contrast and DIC), the authors achieved a significant margin of victory in the ISBI cell tracking challenge 2015 in these categories. Furthermore, the network is noted for its speed, with segmentation of a 512x512 image taking less than a second on a recent GPU.

- **Very Deep Convolutional Networks for Large-Scale Image Recognition- Karen Simonyan, Andrew Zisserman**

In this study, the researchers investigate the impact of convolutional network depth on its accuracy in the context of large-scale image recognition. The primary contribution of the research is a comprehensive evaluation of networks with increasing depth, utilizing an architecture featuring small (3x3) convolution filters. The results demonstrate that a significant enhancement over prior state-of-the-art configurations can be achieved by increasing the depth to 16-19 weight layers. These findings served as the foundation for the ImageNet Challenge 2014 submission, where the research team secured the first and second places in the localization and classification tracks, respectively. The study also highlights that the representations developed in their work generalize well to other datasets, consistently achieving state-of-the-art results.

- **Robust Real-Time Violence Detection in Video Using CNN And LSTM - Al-Maamoon R. Abdali, Rana F. Al-Tuma**

Violence event detection is a critical component of surveillance systems that are essential to both public safety and law enforcement. As such, there has been a great deal of research done to improve the speed, accuracy, and generality of these detectors in a variety of video sources. Although earlier research concentrated on accuracy or speed, few examined the more general issue of compatibility with a wide range of video formats. In response, this study presents a deep learning based real-time violence detector, which uses LSTM for temporal relation learning and CNN as a spatial feature extractor. The suggested model, which emphasizes the trinity of overall generality, accuracy, and fast response time, achieves an astonishing 98% accuracy at 131 frames per second. A comparison with earlier studies on violence detection demonstrates how much better the model is in terms of accuracy and speed. The study concludes that the combined use of CNN and LSTM, leveraging transfer learning, represents the optimal approach for achieving accuracy, robustness, and speed in violence detection tasks, particularly with limited datasets and computing resources. However, the authors acknowledge the scope for further improvement and recommend future work to explore or create well-balanced, large datasets encompassing different video sources for enhanced violence detection, expanding beyond mere binary classification to identify specific violence actions.

- **A hybrid model using 2D and 3D Convolutional Neural Networks for violence detection in a video dataset – Anusha Jayasimhan, Pabitha P'**

This work provides a hybrid CNN model for violence detection in surveillance films in the modern setting of the Internet of Things (IoT), where data from varied sources such edge devices and sensors, especially movies from CCTV devices, are readily available. The difficulty is in using deep learning algorithms, like Convolutional Neural Networks (CNNs), to video categorization in a way that takes into account the temporal aspects that two-dimensional (2D) CNNs frequently ignore. The hybrid model, which combines a 2D and 3D CNN, shows impressive speed and accuracy in identifying violence, with a maximum training accuracy of 84.5 percent and a validation accuracy of 99.93 percent. The need for automated surveillance to identify and stop violent occurrences highlights how important this effort is. The absence of publicly available datasets for violence detection emphasizes the need for accurate deep learning models capable of classifying videos based on violence-related activities. The proposed hybrid CNN strategically bridges the limitations of 2D and 3D CNNs, ensuring the extraction of both temporal and spatial features for precise and swift violence recognition. This review outlines the context, challenges, motivation, methodology, and results, offering a comprehensive understanding of the proposed model's efficacy.

- **Detecting Violence in Video Based on Deep Features Fusion Technique - Heyam Bin Jahlan, Lamiaa Elrefaei**

In the realm of surveillance systems and violence detection, there is a discernible demand for automated solutions given the surge in surveillance cameras across public spaces. This necessitates systems capable of swiftly and accurately identifying violent events. Prior studies have predominantly focused on either speed or accuracy, with limited attention to generality across diverse video sources. The proposed work contributes to this domain by introducing a real-time violence detection model, combining Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) for temporal relation learning. This hybrid CNN model attains commendable accuracy of 98% and a speed of 131 frames per second, outperforming previous works in both accuracy and speed. The integration of a 3D CNN addresses the temporal aspect of video classification, while the proposed model achieves a balance between accuracy, speed, and generality across various video sources. The study underscores the importance of developing models that not only excel in accuracy and speed but also exhibit adaptability to diverse video inputs, thus contributing to the broader landscape of violence detection in surveillance scenarios.

- **Violence Recognition from Videos using Deep Learning Techniques – Mohamed Mostafa Soliman, Dina Khattab**

Within the domain of violence detection from videos, this study addresses the need for automated systems to recognize interpersonal violence, focusing on physical altercations. Traditional surveillance methods relying on human attention prove inadequate due to high costs and errors. In the pre-deep learning era, classical computer vision methods faced limitations. Leveraging the benefits of deep learning, this research proposes an end-to-end model integrating VGG-16 and Long Short-Term Memory for spatial and temporal feature extraction. A new benchmark dataset, Real-Life Violence Situations (RLVS), is introduced. The model achieves fine-tuned accuracies of 86.2%, 88.2%, and 84.0% on hockey fight, movie, and violent-flow datasets. The paper concludes with insights into related works, the proposed model, dataset descriptions, experimental results, and future directions.