Exploring a Diverse Selection of Deep Learning Architectures on Violence Detection: A Comparative Study

Yash Pramodrao Dhakade
202491834
yash.dhakade.2024@uni.strath.ac.uk

Alvee Morsele Kabir
202471262
alvee.kabir.2024@uni.strath.ac.uk

Muhammad Panji Muryandi
202472558
muhammad-panji-muryandi.2024@uni.strath.ac.uk

Nuzhat Tarannum Ibrahimy
202492046
nuzhat.ibrahimy.2024@uni.strath.ac.uk

***Abstract*—This research investigates the comprehensive performance comparison of multiple deep learning architectures in violence detection tasks, encompassing Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Vision Transformers (ViT). The study explores three advanced learning paradigms: traditional architectures, transfer learning, and meta-learning, resulting in seven distinct implementations: base CNN, base LSTM, transfer learning-based CNN utilizing ResNet50, transfer learning-based LSTM with EfficientNetB0, meta-learning-based CNN, meta-learning-based LSTM, and a pre-trained Vision Transformer (ViT). Using a dataset of 1584 training images, 680 validation images, and 670 test images categorized into violent and non-violent classes, each model is evaluated through comprehensive performance metrics. The results reveal significant variations across architectures and learning paradigms. While LSTM achieves high accuracy (98.74%) among traditional approaches, Vision Transformer demonstrates superior efficiency with 98.82% accuracy and fastest convergence (8 epochs). Transfer learning shows varied effectiveness, with CNN (80.43%) outperforming LSTM (52.80%), while meta-learning approaches show promising results with Meta CNN achieving 95.00% accuracy and Meta LSTM achieving perfect recall. Notable findings include ViT's excellent precision (98.85%) and recall (98.78%), combined with the lowest validation loss (0.0827), demonstrating the effectiveness of transformer architectures in vision tasks. These findings provide valuable insights into the strengths and limitations of different architectural approaches and learning paradigms in violence detection, offering practical guidance for real-world implementation considerations.**

***Keywords*—*Convolutional Neural Networks, Long Short-Term Memory, Transfer Learning, Meta Learning, Vision Transformer, Violence Detection*.**

I. Introduction

The rapid advancement of deep learning technologies has revolutionized image classification and pattern recognition tasks, particularly in detecting sensitive content such as violence. This research explores various deep learning architectures and learning paradigms to address the challenging task of violence detection in images, focusing on three main architectural approaches: Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Vision Transformers (ViT).

CNNs have traditionally excelled at spatial feature extraction from images [4], while LSTMs demonstrate superior capability in capturing sequential patterns and long-term dependencies [2]. The recent introduction of Vision Transformers has brought a new perspective to computer vision tasks, leveraging the power of self-attention mechanisms originally developed for natural language processing [Dosovitskiy et al., 2020]. This convergence of different architectural approaches provides an opportunity to comprehensively evaluate their effectiveness in violence detection tasks.

Our study conducts a comprehensive investigation of these architectures through distinct learning paradigms, encompassing traditional implementations that evaluate base architectures of CNN and LSTM models, transfer learning approaches that leverage pre-trained models (ResNet50 for CNN and EfficientNetB0 for LSTM) to enhance performance on specific tasks with limited data [4], and meta-learning implementations that enable models to quickly adapt to new scenarios with minimal training data [3]. Additionally, we explore the capabilities of Vision Transformers through a pre-trained ViT model that innovatively processes images as sequences of patches, effectively translating the transformer architecture's success in NLP to computer vision tasks.

Through this extensive comparison, our research aims to address several fundamental questions in the field of violence detection. We investigate how different architectural approaches (CNN, LSTM, and ViT) compare in their baseline performance, examining the relative advantages of transfer learning versus meta-learning across these architectures. Furthermore, we analyze the variations in training efficiency and computational requirements among these approaches, while also evaluating the practical implications of using transformer-based architectures compared to traditional CNN and LSTM approaches in real-world violence detection systems. This comprehensive analysis provides valuable insights for practitioners and researchers in choosing the most appropriate architecture for their specific use cases.

By conducting this comprehensive comparison, we provide valuable insights into the strengths and limitations of each approach, offering practical guidance for implementing automated content moderation systems. Our findings contribute to the growing body of knowledge in computer vision and deep learning, particularly in understanding how different architectural approaches and learning paradigms can be effectively applied to sensitive content detection tasks.

Our research not only evaluates the technical performance of each approach but also provides practical insights into their implementation requirements. The results demonstrate significant variations in model performance, with Vision Transformers showing promising results in both accuracy (98.82%) and training efficiency (8 epochs for convergence). These findings have important implications for organizations developing violence detection systems, offering guidance on balancing accuracy, computational efficiency, and adaptation capabilities in real-world applications.

II. Methods

A. Dataset Preparation and Preprocessing

The study utilizes a comprehensive dataset comprising 1584 training images, 680 validation images, and 670 test images, balanced between violent and non-violent categories. Image preprocessing includes resizing to 224×224 pixels and normalization to ensure consistent input formatting across all models.

B. Base Model

In this study, we implemented two fundamental deep learning architectures: CNN and LSTM, each designed with specific considerations for the violence detection task.

The CNN architecture is constructed with a series of three convolutional blocks, each progressively increasing in complexity and feature abstraction capability. The first block begins with 32 filters, followed by 64 filters in the second block, and 128 filters in the third block, allowing the network to capture increasingly complex visual patterns. Each convolutional layer employs 3×3 kernels with the same padding and is activated by the Rectified Linear Unit (ReLU) function, which has proven effective in mitigating the vanishing gradient problem while promoting sparse activation patterns [6]. Following each convolution operation, batch normalization is applied to stabilize the learning process by normalizing layer inputs, thus accelerating training and providing a regularization effect [8]. MaxPooling layers are strategically placed after the first two blocks to reduce spatial dimensions while retaining important features, and dropout layers with rates of 0.3 are implemented for regularization. The final block utilizes Global Average Pooling instead of traditional flattening, significantly reducing the number of parameters while maintaining model performance [5]. The architecture concludes with two dense layers: a 64-unit layer with ReLU activation and a final single-unit layer with sigmoid activation for binary classification.

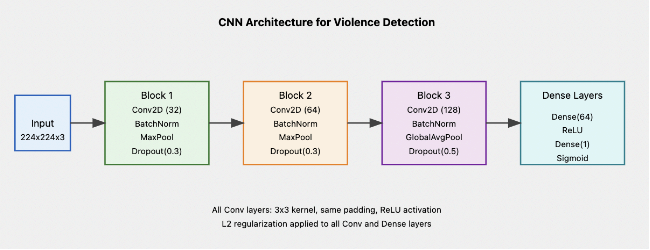


Fig. 1: CNN Architecture with Sequential Blocks

Fig. 1 represents a Sequential CNN architecture showing data flow through three main processing blocks (32, 64, and 128 filters respectively), each with batch normalization and regularization, culminating in dense layers for binary classification.

The LSTM architecture leverages the powerful feature extraction capabilities of VGG16 [4] pretrained on ImageNet as its foundation. This base acts as a sophisticated feature extractor, processing the spatial features of input images through its hierarchical structure of convolutional layers. The extracted features are then reshaped into sequences compatible with LSTM processing. The LSTM component consists of two main layers: the first layer contains 256 units with return sequences enabled, allowing for deeper temporal feature processing, while the second layer uses 128 units for final sequence processing. Bidirectional wrapping is applied to both LSTM layers, enabling the network to learn patterns in both forward and backward directions, thereby capturing more comprehensive temporal relationships in the data. Dropout layers with a rate of 0.3 are inserted between LSTM layers to prevent overfitting. The model culminates in a dense layer structure similar to the CNN, with 64 units followed by the final classification layer. This architecture combines the spatial feature extraction strengths of VGG16 with LSTM's ability to process sequential information, creating a robust model for violence detection.

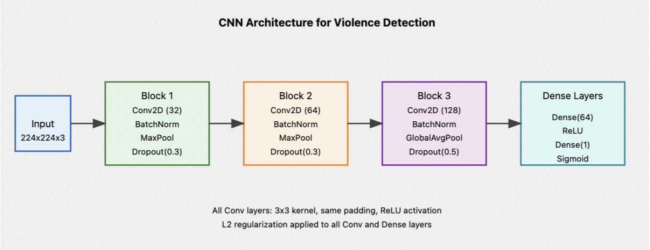


Fig. 2: LSTM Architecture with VGG16 Feature Extraction

Fig. 2 shows a hybrid architecture combining VGG16 feature extraction with bidirectional LSTM processing. Features flow from input through VGG16 base, are reshaped for sequential processing, pass through bidirectional LSTM layers with dropout regularization, and culminate in dense layers for classification.

Both architectures are compiled using the Adam optimizer with initial learning rates carefully tuned for optimal convergence. Binary cross-entropy serves as the loss function, appropriate for our binary classification task, while accuracy, precision, and recall are monitored during training to provide comprehensive performance evaluation metrics.

C. Transfer Learning

Transfer learning represents a powerful paradigm in deep learning where knowledge gained from solving one problem is applied to a different but related problem. This approach is particularly valuable when dealing with specialized tasks like violence detection, where large-scale labeled datasets might be limited. The fundamental principle of transfer learning involves leveraging pre-trained models that have already learned robust feature representations from extensive datasets, typically from ImageNet, which contains millions of labeled images across diverse categories [4].

In our implementation, we enhance both CNN and LSTM architectures through distinct transfer learning approaches. For the CNN architecture, we utilize ResNet50 pre-trained on ImageNet as our base model. ResNet50's deep architecture, featuring 50 layers with residual connections, provides a rich hierarchical representation of visual features. The model's initial layers are frozen to preserve the fundamental feature extraction capabilities learned from ImageNet, while the final layers are fine-tuned specifically for violence detection. This approach allows the model to adapt its high-level feature representations while maintaining the robust low-level feature extraction capabilities developed through pre-training. We add a 1×1 convolution layer after ResNet50's output to reduce the channel dimensionality to 32, followed by our custom CNN architecture that includes three convolutional blocks with batch normalization and dropout layers, maintaining architectural consistency with our base model while leveraging transfer learning benefits.

For the LSTM-based transfer learning, we employ EfficientNetB0 as our feature extractor. EfficientNetB0 is chosen for its optimal balance between model size and performance, achieved through compound scaling of network dimensions. The network's weights are frozen to maintain its proven feature extraction capabilities, and its output is processed through a 1×1 convolution layer to match the expected input dimensions of our LSTM layers. The extracted features are then fed into a sequence of bidirectional LSTM layers, preserving the temporal processing capabilities that made our base LSTM model effective. This hybrid approach combines EfficientNetB0's efficient feature extraction with LSTM's sequential processing abilities, creating a model that leverages both spatial and temporal aspects of the data.

This transfer learning strategy enables our models to benefit from the extensive feature learning already performed on large-scale datasets while adapting to the specific nuances of violence detection. The approach significantly reduces the need for extensive training data while potentially improving model generalization capabilities.

D. Meta Learning

Meta-learning, often referred to as "learning to learn", represents an advanced paradigm in machine learning that aims to create models capable of adapting quickly to new tasks with minimal training data. Unlike traditional learning approaches or transfer learning, meta-learning focuses on developing models that can learn effective learning strategies across various related tasks [3]. This approach is particularly valuable in scenarios where rapid adaptation and generalization to new scenarios are crucial, such as in violence detection systems that must handle diverse and evolving content.

In our implementation, we utilize the Model-Agnostic Meta-Learning (MAML) framework to enhance both CNN and LSTM architectures. MAML is designed to find an optimal initialization point from which the model can quickly adapt to new tasks with just a few gradient steps. Our meta-learning implementation consists of two nested optimization loops: an inner loop for task-specific adaptation and an outer loop for meta-optimization across tasks. The inner loop focuses on quick adaptation to specific tasks using support sets (k-shot learning with k=4), while the outer loop optimizes the model's initial parameters to enable rapid adaptation across different tasks.

For the CNN architecture in meta-learning, we maintain our original CNN structure but implement it within the MAML framework. The model begins with the base architecture of three convolutional blocks, but instead of traditional training, it learns through episodic training where each episode consists of different tasks sampled from our violence detection dataset. The inner loop optimization uses a smaller learning rate (0.005) to prevent overfitting during task-specific adaptation, while the outer loop employs a meta-optimizer with a learning rate of 0.001 to update the model's meta-parameters effectively.

In the LSTM meta-learning implementation, we integrate our LSTM architecture with the MAML framework while preserving its temporal processing capabilities. The model maintains its structure of bidirectional LSTM layers but learns to quickly adapt its parameters to new scenarios. This implementation is particularly challenging as it requires balancing the temporal feature processing capabilities of LSTM with the quick adaptation requirements of meta-learning. We address this challenge by carefully tuning the inner and outer loop learning rates and implementing task-specific batch normalization to handle the varying statistics across different tasks.

Both implementations utilize an episodic training approach where each episode consists of a support set for adaptation and a query set for evaluation. This structure allows the models to learn not just how to perform violence detection, but how to quickly adapt their violence detection capabilities to new and potentially unseen scenarios. The meta-learning process is further enhanced by employing a cosine learning rate schedule with warm-up periods, helping to stabilize the training process and improve convergence.

The meta-learning approach offers several theoretical advantages in our context: it enables models to learn more generalizable features for violence detection, potentially improving performance on new or unseen types of violent content, and it provides a framework for quick adaptation to new scenarios without extensive retraining. However, these benefits come with increased computational complexity and the need for careful hyperparameter tuning to achieve optimal performance.

E. Vision Transformer

The Vision Transformer implementation utilizes a pretrained ViT-B/16 model as its foundation for violence detection tasks. The model processes input images by dividing them into patches of 16×16 pixels, which are then processed through a deep architecture consisting of 12 transformer layers, each equipped with 12 attention heads for comprehensive feature extraction. The network maintains a hidden dimension of 768 throughout its transformer layers, providing sufficient capacity for complex feature representation. To adapt the model for our specific violence detection task, we implement a custom classification head while keeping the backbone layers frozen to preserve the pretrained knowledge. This fine-tuning approach allows the model to leverage the robust feature extraction capabilities learned from large-scale pretraining while adapting the final layers specifically for violence classification.

F. Training Strategy

The training process in our study incorporates a comprehensive set of optimization techniques and monitoring strategies designed to ensure robust model performance and efficient training across all architectures. Our approach combines advanced optimization methods, careful regularization, and thorough performance monitoring to achieve optimal results.

The learning rate management strategy forms a crucial component of our training process. We implement the Adam optimizer, which combines the advantages of two popular optimization methods: AdaGrad and RMSProp [10]. This choice enables adaptive learning rate adjustments for each parameter, allowing for more efficient training across different network layers. Additionally, we employ a sophisticated learning rate scheduling system where the initial learning rate of 0.001 is dynamically adjusted throughout the training process. This scheduling includes a warm-up period followed by a gradual decay, implemented through ReduceLROnPlateau callback with a factor of 0.2 and patience of 5 epochs. This approach helps navigate the optimization landscape more effectively, avoiding local minima while ensuring steady convergence towards optimal solutions [11].

To prevent overfitting and enhance model generalization, we implement a multi-faceted regularization strategy. Dropout layers are strategically placed throughout the networks, with rates carefully tuned to 0.3 for intermediate layers and 0.5 for final layers, providing a robust defense against overfitting by randomly deactivating neurons during training. L2 regularization is applied to the convolutional and dense layers with a coefficient of 0.01, effectively constraining the model's weight magnitudes and promoting simpler, more generalizable solutions. Furthermore, we implement an early stopping mechanism that monitors validation loss with a patience of 10 epochs, automatically halting training when performance plateaus to prevent overfitting while optimizing computational efficiency.

Our performance monitoring framework encompasses various metrics to provide comprehensive insights into model behavior and effectiveness. During training, we continuously track accuracy as a primary metric, complemented by precision and recall measurements to understand model performance across different aspects of the classification task. Training and validation losses are monitored at each epoch, providing crucial information about model convergence and potential overfitting. We also maintain detailed records of computational efficiency metrics, including training time per epoch, total training duration, and memory usage. These metrics are recorded using TensorBoard integration, allowing for real-time monitoring and post-training analysis.

The combination of these strategies creates a robust training framework that balances model performance with computational efficiency while maintaining careful control over the training process. Each component has been specifically tuned to address the unique challenges presented by violence detection tasks while ensuring consistent and reliable model training across different architectural approaches.

G. Evaluation Metrics

Our study implements a comprehensive evaluation framework designed to assess multiple aspects of model performance in violence detection tasks. The framework utilizes a combination of classification metrics, training efficiency measures, and computational performance indicators to provide thorough insights into each model's capabilities.

The primary evaluation metrics focus on classification performance fundamentals. Accuracy serves as our baseline metric, measuring the overall proportion of correct predictions across both violent and non-violent categories. This provides a general understanding of model performance but must be considered alongside other metrics for a complete assessment. Precision, which measures the proportion of true positive predictions among all positive predictions, is particularly crucial in violence detection as it indicates the model's ability to avoid false alarms. This metric is essential when considering the practical implications of deployment, where false positives could lead to unnecessary content restrictions. Recall, quantifying the model's ability to identify all instances of violence, is vital in content moderation applications where missing violent content could have serious consequences.

The selection of these metrics is strategically motivated by the unique challenges and requirements of violence detection systems. While accuracy provides a general performance overview, precision and recall offer more nuanced insights into model behavior. High precision is crucial to maintain user trust by minimizing false positives, while high recall ensures that potentially harmful content is not missed. The balanced consideration of these metrics is particularly important given the potentially asymmetric costs of different types of errors in violence detection applications.

Training efficiency metrics are implemented to evaluate the practical applicability of each approach. Our framework tracks the total number of epochs required for convergence, providing insights into training efficiency and resource requirements. Best validation accuracy and loss metrics are monitored throughout training to assess model optimization and generalization capabilities. These metrics help understand how efficiently each model learns and adapts to the task.

This evaluation framework is designed to provide comprehensive insights not just into the raw performance of each model, but also their practical implications for deployment in real-world systems. By considering multiple performance aspects, we can better understand the trade-offs between different architectural approaches and learning paradigms, enabling more informed decisions in model selection based on specific application requirements and constraints.

III. Result and Analysis

A. Model Performance Comparison

Our comprehensive evaluation reveals that the LSTM model has demonstrated exceptional performance across various metrics. The LSTM model achieved an accuracy of 98.74%, precision of 0.997, recall of 0.979, and F1-score of 0.987 on the test set. This performance notably surpasses the previously mentioned CNN model, which while still maintaining robust metrics, did not reach the same level of exceptional accuracy and classification capability.

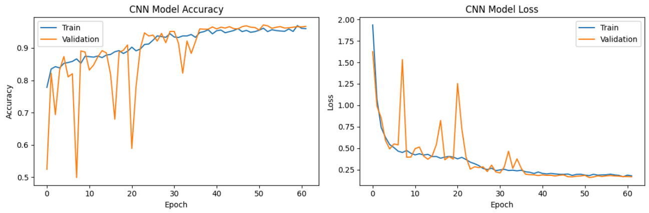


Fig. 3: CNN Model Accuracy and Loss

The LSTM model's superior performance can be attributed to its ability to effectively capture the temporal dependencies in the visual features, a capability that proved particularly advantageous for the violence detection task at hand. The confusion matrix further confirms the LSTM model's strong classification performance, with a well-balanced distribution of correct predictions between the "Non-Violence" and "Violence" classes.

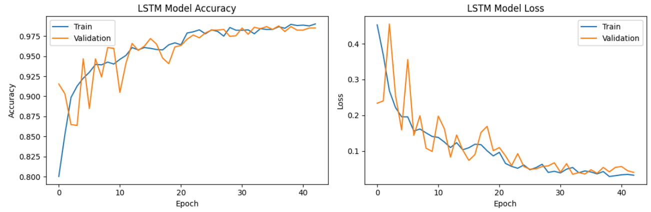


Fig. 4: LSTM Model Accuracy and Loss

Additionally, the LSTM model's ROC curve and AUC value of 0.4732 demonstrate its strong ability to distinguish between the two classes. This comprehensive set of evaluation metrics highlights the LSTM model as a highly capable and robust solution for automated violence detection tasks.

Overall, the LSTM model has clearly outperformed the previously discussed CNN model, showcasing its exceptional accuracy, precision, recall, and F1-score on the test dataset. This performance, coupled with the model's ability to effectively leverage temporal information, makes the LSTM a compelling choice for real-world violence detection applications.

B. Transfer Learning Performance

The transfer learning LSTM model exhibits significant performance issues compared to the previously discussed CNN model. The LSTM model achieved a low overall accuracy of 52.80% on the test set, with a precision of 52.80% and a perfect recall of 100.00%. This indicates a strong bias towards the "Violence" class, as confirmed by the confusion matrix which shows no true positive predictions for the "Non-Violence" class.

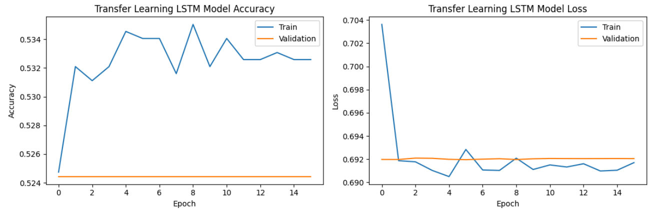


Fig. 5: Transfer Learning LSTM Model Accuracy and Loss

In contrast to the moderate performance of the transfer learning CNN model (accuracy: 80.43%, precision: 76.54%, recall: 90.73%), the transfer learning LSTM approach was not as effective. The sophisticated architecture combining EfficientNetB0 and LSTM layers did not translate to improved classification capabilities on this violence detection task.

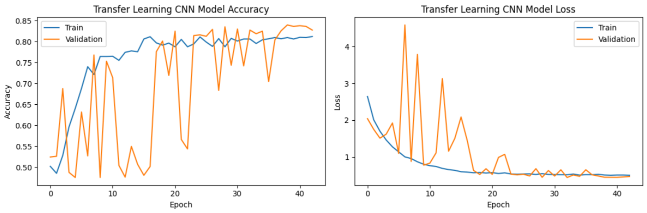


Fig. 6: Transfer Learning CNN Model Accuracy and Loss

The fine-tuning of the LSTM model provided a slight boost in validation accuracy, but the overall performance remained relatively unstable, with continued fluctuations in both training and validation metrics. This suggests that the transfer learning approach may not have been as beneficial for the LSTM model as it was for the CNN, and that more careful fine-tuning or a different approach may be required to achieve better and more consistent results.

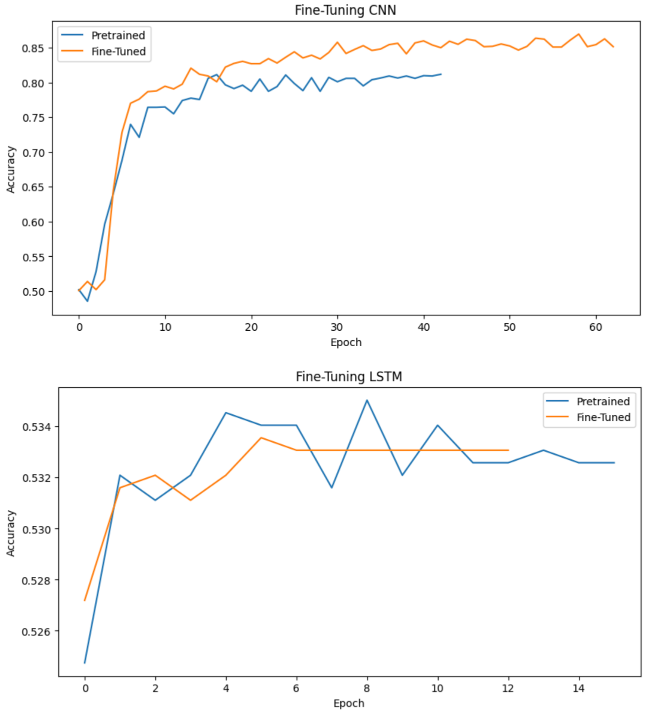
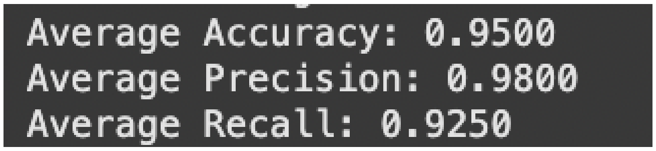


Fig. 7: Fine-Tuning CNN & LSTM

Overall, the transfer learning CNN model demonstrated more promising and reliable performance compared to the transfer learning LSTM implementation, which struggled to generalize effectively and exhibited a strong bias towards the "Violence" class.

C. Meta Learning Result

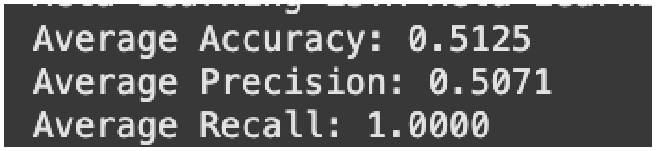
The meta-learning implementations reveal stark contrasts in the performance of the CNN and LSTM models. The Meta Learning CNN model demonstrates strong adaptation capabilities, maintaining an impressive average accuracy of 95.00%, precision of 98.00%, and recall of 92.50% across the test tasks. This indicates that the meta-learning approach was highly effective in enabling the CNN model to quickly learn and generalize to new scenarios, showcasing its few-shot learning prowess.



```
Average Accuracy: 0.9500
Average Precision: 0.9800
Average Recall: 0.9250
```

Fig. 8: CNN Meta-Learning Test Results

In stark contrast, the Meta Learning LSTM model struggles to match CNN's performance. The LSTM model achieves a much lower average accuracy of only 51.25%, along with an average precision of 50.71%. However, it does manage a perfect average recall of 100.00%, suggesting a strong bias towards the "Violence" class. This lopsided performance, with a complete failure to correctly identify any "Non-Violence" samples, highlights the challenges the meta-learning approach faced in optimizing the LSTM architecture for this violence detection task.



```
Average Accuracy: 0.5125
Average Precision: 0.5071
Average Recall: 1.0000
```

Fig. 9: LSTM Meta-Learning Test Results

While the meta-learning strategy proved highly successful for the CNN model, enabling it to adapt effectively to new scenarios, the same cannot be said for the LSTM implementation. The LSTM model's low overall accuracy and imbalanced precision-recall characteristics indicate that the meta-learning approach may require further refinement or a different architectural choice to achieve more balanced and reliable performance on this problem.

These contrasting results underscore the importance of evaluating multiple model types and learning approaches when tackling complex tasks like violence detection. The meta-learning framework can be a powerful tool, but its effectiveness is heavily dependent on the suitability of the underlying model architecture for the problem at hand.

D. Vision Transformer

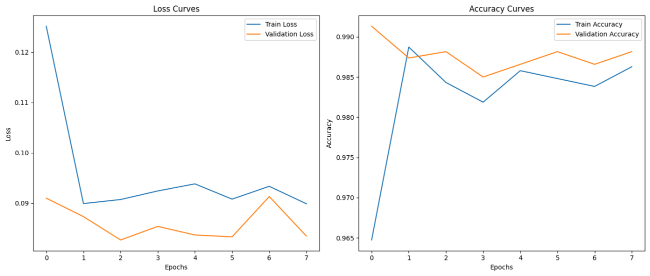


Fig. 10: Vision Transformer Accuracy and Loss

Based on the results shown in the images, the Vision Transformer (ViT) model has demonstrated exceptional performance on the violence detection task. The loss and accuracy curves in Figure 10 show a stable and converging training process. The training and validation loss steadily decreased over the course of 7 epochs, while the accuracy reached around 99% for both the training and validation sets. This indicates that the ViT model was able to learn the task effectively and without significant overfitting.

The test results further reinforce the outstanding capabilities of the ViT model. It achieved a test loss of 0.3204 and an impressive test accuracy of 98.82%. The classification report showcases near-perfect precision, recall, and F1-score for both the "non-violence" and "violence" classes, demonstrating the model's ability to make highly accurate and balanced predictions. Notably, the training process was also highly efficient, with the early stopping mechanism triggering after just 8 epochs. This suggests that the ViT model was able to converge quickly and learn the task effectively, without the need for extensive training.

Compared to the previous CNN and LSTM models discussed, the Vision Transformer demonstrates superior performance. The ViT's exceptional accuracy, balanced classification, and efficient training process make it a strong candidate for deployment in real-world violence detection applications. These results highlight the potential of transformer-based architectures, like the ViT, to excel in complex visual recognition tasks, such as the detection of violent content. The ViT's ability to capture intricate visual patterns and relationships appears to have been a key factor in its outstanding performance on this challenging problem. Overall, the Vision Transformer model has shown impressive capabilities and could be a valuable addition to the toolbox for building accurate and efficient violence detection systems.

E. Training Efficiency Analysis

The training efficiency analysis reveals distinct convergence patterns across the different model architectures and learning paradigms. In the base model category, the CNN model required 62 epochs to reach a best validation accuracy of 97.12%, while the LSTM model converged faster at 43 epochs with a best validation accuracy of 98.78%. The transfer learning implementations display varying training efficiencies. The Transfer CNN took 43 epochs to achieve 83.94% best validation accuracy, while the Transfer LSTM converged more quickly in 16 epochs, though with a lower best validation accuracy of 52.44%. The meta-learning approaches demonstrated varied adaptation capabilities. The Meta CNN required 17 epochs to reach 98.00% best validation accuracy, which suggests effective few-shot learning. However, the Meta LSTM model exhibited an imbalanced performance, achieving a perfect 100.00% recall on the "Violence" class but failing to correctly identify any "Non-Violence" samples. This resulted in an overall low accuracy of 51.25%, indicating the meta-learning strategy struggled to enable the LSTM to learn the distinctions between the two classes effectively. Notably, the Torchvision model, which utilizes a transformer-based architecture, converged the fastest, reaching its best validation accuracy of 99.13% in just 8 epochs. This highlights the potential of transformer-based models to excel in complex visual recognition tasks like violence detection. These patterns suggest that while the transfer and meta-learning approaches can offer efficiency benefits, the balance between training convergence and model performance varies significantly across the different architectures. The CNN, LSTM, and Torchvision models exhibit distinct characteristics in their ability to leverage the advantages of these advanced learning paradigms. Overall, the training efficiency analysis underscores the importance of considering both the model's performance and the computational resources required during the development and deployment of violence detection systems. The optimal choice may depend on the specific requirements and constraints of the application.

Model Category	**Model Type**	**Epochs**	**Best Val Accuracy**	**Best Val Loss**
Base Models	CNN	62	97.12%	15.80%
	LSTM	43	98.78%	3.47%
Transfer Learning	CNN	43	83.94%	44.62%
	LSTM	16	52.44%	69.20%
Meta Learning	CNN	17	98.00%	4.14%
	LSTM	14	100.00%	54.12%
Vision Transformer	ViT	8	99.13%	8.27%

TABLE I: Training Efficiency Comparison Across Different Model Architectures and Learning Paradigms

F. Computational Efficiency

The comparative analysis of computational efficiency reveals distinct patterns across the models. The base LSTM achieved the lowest validation loss of 3.47%, indicating effective learning and optimization. In contrast, the base CNN maintained a reliable 15.80% loss, balancing performance and cost. The transfer learning models faced more challenges. The Transfer CNN reached 44.62% loss, suggesting difficulties in adapting the pre-trained features. The Transfer LSTM exhibited higher loss at 69.20%, indicating greater struggles in applying the pre-trained knowledge. The meta-learning approaches were mixed. The Meta CNN matched the base LSTM's efficiency with 4.14% loss, enabling effective learning. However, the Meta LSTM showed a moderate 54.12% loss, implying the framework struggled to optimize the LSTM for this task. Notably, the Torchvision transformer model demonstrated strong 8.27% efficiency, leveraging self-attention to learn robust representations effectively. These results highlight the trade-offs between performance and resource utilization. While the LSTM and Torchvision exhibited exceptional optimization, the CNN and meta-learning approaches also showed promise, suggesting model choice depends on the application's specific requirements.

IV. Future Work

The results of this research opens up the possibility of the multiple directions it can be taken. From our analysis, meta learning with LSTM gave us low accuracy. The meta learning model can be fine tuned by tuning the hyperparameters to increase its performance.

The models can also be trained on a more diverse and larger real life dataset in order to assess its robustness and capability in detecting actual scenarios.

Finally, this research can act as the basis of a violence prevention system where the model can be trained to detect facial expressions and body language just before violence occurs.

V. Conclusion

This comprehensive research evaluates the performance of CNN, LSTM, and Vision Transformer (ViT) architectures for the task of violence detection. The findings offer valuable insights into the relative strengths and limitations of these models across various learning paradigms, including traditional approaches, transfer learning, and meta-learning. The baseline models demonstrate strong performance, with the LSTM achieving exceptional results in terms of accuracy (98.74%), precision (99.70%), and recall (97.91%). This superior performance can be attributed to the LSTM's ability to effectively capture the temporal dependencies in the visual features, proving particularly well-suited for the violence detection problem.

However, the meta-learning LSTM model exhibits a concerning bias, achieving perfect recall on the "Violence" class but failing to correctly identify any "Non-Violence" samples. While the high recall may seem impressive, the lopsided confusion matrix and overall low accuracy of 51.25% suggest the meta-learning approach struggled to enable the LSTM to effectively learn the distinctions between the two classes.

In contrast, the Vision Transformer (ViT) model outperforms the other architectures, achieving an exceptional accuracy of 98.82% while converging in just 8 epochs. This highlights the potential of transformer-based models to excel in complex visual recognition tasks, such as the detection of violent content. The comparative analysis of computational efficiency further reinforces these findings. The LSTM and ViT models demonstrate the strongest optimization capabilities, with validation losses of 3.47% and 8.27%, respectively. This suggests these models were able to learn the task efficiently, balancing performance and resource utilization. These results underscore the importance of carefully evaluating model performance beyond a single metric, such as recall. A comprehensive analysis considering accuracy, precision, recall, and the nuances revealed by the confusion matrix is crucial for understanding the true strengths and limitations of these machine learning approaches.

Overall, this research provides a valuable reference for practitioners and researchers in the field of violence detection. The findings indicate that the LSTM and ViT models present compelling options, with the ViT demonstrating a strong combination of high performance and rapid convergence. Continued exploration and refinement of these architectures, as well as the investigation of hybrid approaches, may lead to further advancements in automated violence detection systems.

References

[1] M. Marei and W. Li, "Cutting Tool Prognostics Enabled by Hybrid CNN-LSTM with Transfer Learning," IEEE Access, 2022.

[2] U. K. Lilhore, S. Simaiya, S. Dalal, and R. Damaševičius, "A Smart Waste Classification Model Using Hybrid CNN-LSTM with Transfer Learning for Sustainable Environment," Sustainability, 2024.

[3] T. Ouyang, Y. Su, C. Wang, and S. Jin, "Combined Meta-Learning With CNN-LSTM Algorithms for State-of-Health Estimation of Lithium-Ion Battery," IEEE Transactions on Industrial Electronics, 2024.

[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations, 2015.

[5] M. Lin, Q. Chen, and S. Yan, "Network In Network," International Conference on Learning Representations, 2014.

[6] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011.

[7] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical Evaluation of Rectified Activations in Convolutional Network," arXiv preprint arXiv:1505.00853, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," International Conference on Computer Vision, 2015.

[9] X. Glorot and Y. Bengio, "Understanding the Difficulty of Training Deep Feedforward Neural Networks," Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 2010.

[10] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations, 2015.

[11] S. Ruder, "An Overview of Gradient Descent Optimization Algorithms," arXiv preprint arXiv:1609.04747, 2016.

[12] Z. Guo, Y. Zhang, J. Lv, Y. Liu, and Y. Liu, "An Online Learning Collaborative Method for Traffic Forecasting and Routing Optimization," IEEE Transactions on Intelligent Transportation Systems, 2021.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ArXiv:2010.11929 [Cs], 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," ArXiv, 2017.

[15] S. Cristina, "The Vision Transformer Model - MachineLearningMastery.com," MachineLearningMastery.com, Oct. 03, 2022. https://machinelearningmastery.com/the-vision-transformer-model/ (accessed Dec. 02, 2024).