RESEARCH FORUM 19TH UGRF

AIS411-5

Comparative Analysis of Video Classification Models Using the TikHarm Dataset

Abstract

Video classification is essential for content moderation to tag harmful content and sensitive content found mainly in social media. This work contains a comparison among four contemporary deep learning models, namely, Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Model(SSM), on the TikHarm dataset, a pre-curated collection of videos labeled as Harmful Content, Safe, and Suicide. The pre-processing pipeline consisted of taking key frames from videos and generating captions from them as inputs for these models with the help of the BLIP model. Bi-LSTM models were less efficient in capturing sequential patterns, with attention-elevated Bi-LSTM demonstrating greater contextual focus but a much higher number of false positives between overlapping categories. The transformer was the most well-performing architecture, scoring a final training accuracy of 92.09

Introduction

The rapid growth of social media platforms has significantly increased the volume of user-generated video content, raising concerns about harmful or sensitive material that could impact user safety and violate platform policies. Automated video classification systems have become crucial for identifying and moderating such content to ensure a safe online environment. These systems must not only handle the complexities of video data but also accurately classify content into categories like harmful, safe, or suicidal, given the potential consequences of misclassification. Traditional approaches to video classification have relied on handcrafted features and shallow learning models, which often do not work well and generalize across diverse datasets. Advances in deep learning, particularly in the field of sequence modeling and attention mechanisms, have paved the way for developing more robust and accurate models capable of handling the temporal and contextual aspects of video data. However, if these were to be more effective for specific classification tasks, it is still a challenge since that happens usually in every scenario regarding overlapping features among classes. The paper focuses on comparing four latest deep learning models—Bidirectional Long Short-Term Memory (Bi-LSTM),Bi-LSTM with Attention, Transformer, and State Space Model(SSM)—in the experiment on the TikHarm dataset that isa designed dataset for addressing harmful content in video. The preprocessing pipeline consists of extracting key videoframes and captions generated using the BLIP model, so the models would focus on textual context for classification. Each model was evaluated to classify captions into classes: Harmful Content, Safe, and Suicid

Methodology

The process began with the selection of the TikHarm dataset, a curated collection of videos categorized into three classes: Harmful Content, Safe, and Suicide. Each video was annotated with one of these labels, providing a structured foundation for supervised learning. The dataset was chosen for its relevance to content moderation tasks, ensuring real-world applicability.

To prepare the dataset for training and evaluation, several preprocessing steps were undertaken. First, we extracted five representative key frames from each video using OpenCV. These frames were selected at regular intervals to ensure diverse visual coverage of the video content. Next, we used the BLIP (Bootstrapped Language-Image Pretraining) model to generate captions for each frame.

The captions for all frames were then concatenated to create a single textual representation for each video, capturing the essence of the visual content in text form. This textual representation served as the primary input to our models.

We evaluated the performance of four state-of-the-art deep learning models: Bidirectional LSTM (Bi-LSTM), Bi-LSTM with Attention, Transformer, and State Space Model (SSM). The Bi-LSTM model was used to capture sequential dependencies from both past and future contexts, leveraging pre-trained Word2Vec embeddings for input representation. The Bi-LSTM with Attention extended this approach by incorporating an attention mechanism that assigned weights to input features, enhancing the model's focus on contextually relevant parts of the input sequence. The Transformer model, known for its ability to model long-range dependencies, employed self-attention mechanisms and positional encodings to process the video captions effectively. Lastly, the State Space Model (SSM) used state-space representations to model temporal relationships efficiently, offering a computationally lightweight alternative.

The dataset was split into training (70%), validation (20%), and test (10%) sets for a fair evaluation of model performance. All models were trained using the Adam optimizer with a learning rate of 0.001, and cross-entropy loss was used as the objective function for the multi-class classification task. To ensure optimal performance, hyperparameters such as batch size (32) and the number of epochs (10) were carefully selected, with early stopping implemented to prevent overfitting. Finally, model performance was evaluated using several metrics. Overall accuracy was used to assess classification performance on the validation set, while confusion matrices provided deeper insights into misclassifications and trends across the three classes. Precision, recall, and F1-scores were also calculated, with particular emphasis on the recall for the "Suicide" category, where accurate identification was critical. This methodological framework ensured a thorough evaluation of the models, providing insights into their strengths and limitations for video classification.

TABLE I
SUMMARY OF MODEL ACCURACIES

Model	Training Accuracy	Validation Accuracy
Bi-LSTM	90.43%	89.03%
Bi-LSTM with Attention	87.63%	84.44%
Transformer	92.09%	91.84%
State Space Model	89.47%	84.95%

Results

This study highlights the comparative strengths and weak-nesses of four deep learning architectures—Bi-LSTM, Bi-LSTM with Attention, Transformer, and State Space Model(SSM)—applied to video classification tasks on the TikHarmdataset. The results reveal key insights into the performance of each model and their suitability for automated content moderation. Considering overall accuracies, Transformer gave high training accuracy at 92.09Bi-LSTM models show how important time-wise modeling is when it comes to video classification. The Bi-LSTM be with Attention mechanism focused on important discriminatory features based on their assigned times and provides better input variable importance. The catch is that it tends to give slightly more misclassifications compared to the rest between "Harmful Content" and "Safe." The transformed Bi-LSTM,simpler in a way, performs considerably on suicide cases but has a greater struggle in the other instances that have overlapping characteristics. The results suggest that attention mechanisms are very powerfully effective, but careful tuning would be necessary as far as the complexity-performance trade-off is concerned. As a competitive alternative, the State Space Model (SSM)has proven exceptional in detecting suicidal content, manifest-ing a very high recall rate for the "Suicide" class. This highlights the model's application in specialized scenarios where sensitivity to such high-risk categories is most important. In contrast, the SSM displayed a lower overall accuracy level compared to the Transformer, indicating that it may not bathe best option in general for classification tasks. However, its robustness and reliability in detecting sensitive content indicate its value as a component in hybrid systems, whereat could work alongside other models to bring about better performance. Misclassifications across all these models were mainly pronounced in the "Harmful Content" and "Safe" label categories, which would seem to mirror feature overlap and the subjectivity of these labels. This calls for more refined feature extraction and possible incorporation of multimodal data-audio with visual cues-to draw sharper delineation in classification. The results also show how important preprocessing is, especially in using the BLIP model for creating accurate, context-enriched captions. High-quality input made possible strong baseline performances across models, helping underscore the power of productive preprocessing pipelines within video classifications.

Conclusion

This study compared four deep learning architectures-Bi-LSTM, Bi-LSTM with Attention, Transformer, and StateSpace Model (SSM)-in the classification of videos on the TikHarm dataset. All the architectures were outperformed by the transformer model, achieving the highest accuracy scores of 92.09 and 91.84This tells us which model is best in a particular context: the Transformer generalizes for most classification tasks, while the SSM hones into real niche safety-related jobs. Future work can introduce the multimodality in combination with enhanced attention mechanisms for fine-tuning by which the performance may further scale up and minor problems indistinguishing overlapping categories are addressed.

Malak Elsamman 211001045
Salma hesham 211000069
Khadiga Nasser 211000738
Rovan ehab 211000559
Mohamed hafez 211000565