ISSN 2616-5775 Vol. 8, Issue 2: 1-6, DOI: 10.25236/AJCIS.2025.080201

Detection of Fighting Behavior Based on an Improved SlowFast Architecture

Youwei Jia*

School of Geomatics and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China
**Corresponding author: 2998945767@qq.com*

Abstract: *The recognition of fighting behavior has been widely applied in various video-based domains. In the field of public safety, accurate and timely identification of fighting behavior in videos is crucial for making prompt decisions regarding such incidents. Among existing behavior recognition models, the SlowFast model has emerged as one of the most popular algorithms due to its dual-stream structure. However, its focus on local features limits its ability to extract global features, resulting in suboptimal classification accuracy. To address this issue, this paper proposes a fighting behavior recognition model incorporating an attention mechanism, which enhances the model's capability to identify behaviors more effectively. Compared to mainstream behavior recognition models, the proposed model demonstrates improved accuracy, offering valuable insights for addressing sudden incidents.*

Keywords: *Fighting Behavior Recognition, SlowFast, Attention Mechanism, Focal Loss Function*

1. Introduction

Behavior recognition can be applied in video surveillance systems and security domains. By automatically analyzing and identifying the behavior of individuals or objects, it enables real-time detection, anomaly identification, and event forecasting. SlowFast is a deep learning architecture designed for video understanding, aiming to enhance the efficiency and accuracy of video analysis and action recognition. Traditional video analysis methods often process each video frame independently, leading to significant computational costs when handling high-resolution videos. However, actions in videos typically occur in localized regions, while other areas remain relatively static. To process video data more effectively, researchers have explored leveraging the temporal information and varying motion speeds within videos to improve video analysis. The core idea of the SlowFast network is to introduce two pathways with different speeds to process video data. One pathway, known as the Slow Path, samples the video at a lower frame rate to capture global motion and spatial details. The other pathway, known as the Fast Path, samples the video at a higher frame rate to capture rapid actions. By simultaneously utilizing both the Slow Path and the Fast Path, the SlowFast network can fully exploit the temporal information and varying speeds of motion within the video. The Slow Path provides more contextual information and accurate pose estimation, while the Fast Path better captures instantaneous actions. This dual-path structure enables the SlowFast network to excel in video understanding tasks, including video classification, action detection, and video segmentation. By introducing the SlowFast network, researchers have achieved outstanding performance across multiple video understanding benchmark datasets. The research on the SlowFast network has provided new ideas and methods for deep learning algorithms in the video domain, making significant contributions to achieving efficient and accurate video analysis[1]. In contrast, other researchers advocate for using 3D convolutions to directly extract spatiotemporal features from video clips for action recognition.[2] A classic 2D convolution-based method for action recognition is the spatiotemporal dual-stream CNN, which utilizes optical flow to capture temporal information between video frames. Compared to 2D convolutions, 3D convolutions introduce an additional temporal dimension, facilitating end-to-end feature extraction and classification.[3]

The SlowFast model is based on convolutional neural networks, with its feature extractor modeled through multiple convolutional layers and spatial pyramid pooling layers, which excel at capturing local features. However, human actions often interact with surrounding individuals or background objects, leading to correlated features. Therefore, during the feature extraction process, it is essential to consider not only local features but also global features, interaction features, and other related

ISSN 2616-5775 Vol. 8, Issue 2: 1-6, DOI: 10.25236/AJCIS.2025.080201

information. The SlowFast model extracts local features through its temporal slow and fast channels, but it cannot capture other behavior-related features, which limits the model's recognition performance. Therefore, this paper improves the SlowFast action recognition model by enhancing its multi-feature information extraction capability. An attention mechanism is added to increase the network's receptive field and classification ability, thereby improving the model's classification accuracy and recognition performance. This improvement provides valuable reference for other behavior recognition models.[4]

2. Network Model

2.1 Slowfast

SlowFast is derived from the approach of processing visual information at different speeds in videos. In a video, some actions and detail changes occur slowly, while others happen more quickly. Based on this, the SlowFast model introduces two paths with different speeds, specifically designed to process slow and fast visual information. The network model is shown in Figure 1.By merging the slow and fast paths, the SlowFast model can better handle visual information at different temporal scales within the video, thereby improving the performance of video understanding tasks.

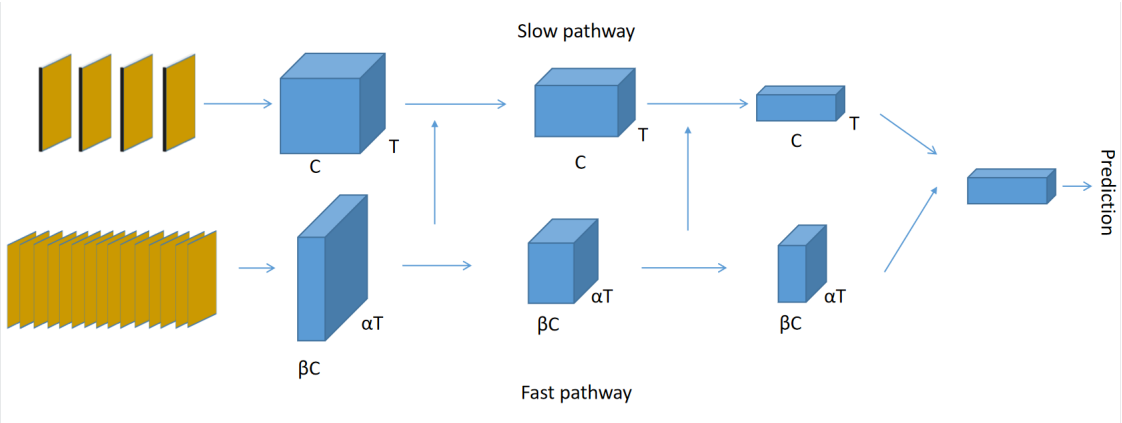

Figure 1: SlowFast Network Architecture

SlowFast is inspired by the types of retinal ganglion cells in primates, where approximately 80% of the cells (P-cells) operate at low frequencies and are capable of recognizing detailed information, while around 20% of the cells (M-cells) operate at high frequencies and are sensitive to temporal changes. SlowFast introduces a novel fast-slow network architecture, where two branches separately process and analyze the temporal and spatial dimensions. The Slow branch has fewer frames and a larger number of channels to learn spatial semantic information. The Slow branch uses a larger stride t to sample video frames, typically set to 16. For a video with a frame rate of 30, this means that approximately 2 frames can be sampled per second, i.e., T=2. The number of channels in the Slow branch is D. The Fast branch has a larger number of frames and fewer channels to learn motion information. The Fast branch uses a smaller stride of t/α, where α is typically set to 8. Therefore, for a video with a frame rate of 30, this means that 15 frames can be sampled per second (αT).The Fast branch maintains a lightweight design by using fewer channels (βD), where β is typically set to 1/8. Both the Slow and Fast branches use 3D convolutional ResNet models. At the end of each branch, SlowFast performs global average pooling, and then the features from both branches are concatenated for classification prediction.

The features extracted by the Slow and Fast branches need to be fused. SlowFast uses a lateral connection to send the features from the Fast branch into the Slow branch for mixing. However, the feature dimensions of the two branches are inconsistent (the Fast branch has the shape $\{\alpha T, S^2, \beta C\}$, while the Slow branch has $\{T, S^2, \alpha\beta C\}$). Therefore, SlowFast employs a 5×1×1 3D convolution with an output channel of 2βC and a stride of α to transform the data from the Fast branch.

2.2 Attention Mechanism

(1) Introducing Self-Attention Mechanism into the Network

Introducing the self-attention mechanism into SlowFast can capture complex temporal and spatial dependencies between video frames, helping the model better understand the relationships and motion

ISSN 2616-5775 Vol. 8, Issue 2: 1-6, DOI: 10.25236/AJCIS.2025.080201

patterns across different time frames. The principle is as follows: Let the input video sequence be X, with the slow path receiving input X_{slow} and the fast path receiving input X_{fast}. For the slow path:

$$Q_{slow} = X_{slow}W_{slow}\,, K_{slow} = X_{slow}W_{slow}^{K}\,, V_{slow} = X_{slow}W_{slow}^{V} \tag{1}$$

$$Attention_{slow}(Q_{slow}, K_{slow}, V_{slow}) = softmax\left(\frac{Q_{slow}K_{slow}^{T}}{\sqrt{d_k}}\right)V_{slow} \tag{2}$$

For the fast path:

$$Q_{fast} = X_{fast}W_{fast}\,, K_{fast} = X_{fast}W_{fast}^{K}\,, V_{fast} = X_{fast}W_{fast}^{V} \tag{3}$$

$$Attention_{slow}(Q_{slow}, K_{slow}, V_{slow}) = softmax\left(\frac{Q_{slow}K_{slow}^{T}}{\sqrt{d_k}}\right)V_{slow} \tag{4}$$

Fusion of outputs:

$$Output = Concat\big(Attention_{slow}, Attention_{fast}\big)W \tag{5}$$

(2) Introducing the ECA attention mechanism into the network.

The ECA attention mechanism is a lightweight channel attention mechanism. The SlowFast model combines information from both the slow and fast paths, and the ECA attention mechanism helps to effectively fuse features across different temporal scales, enhancing the overall model's expressive power. Let X be an input feature map with shape (C,H,W).

Through global average pooling:

$$Z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c} \tag{6}$$

A global descriptor vector Z is obtained, with a shape of (C,).

The kernel size k is determined using the following formula:

$$k = \Psi(C) = \left|\frac{log_2(C)}{2} + \frac{1}{2}\right| \tag{7}$$

Then, a 1D convolution operation is performed:

$$a = \sigma\big(Conv1D(z)\big) \tag{8}$$

The weight a is applied to the original feature map, which retains the shape (C,H,W).

The improved SlowFast model architecture is shown in Figure 2.

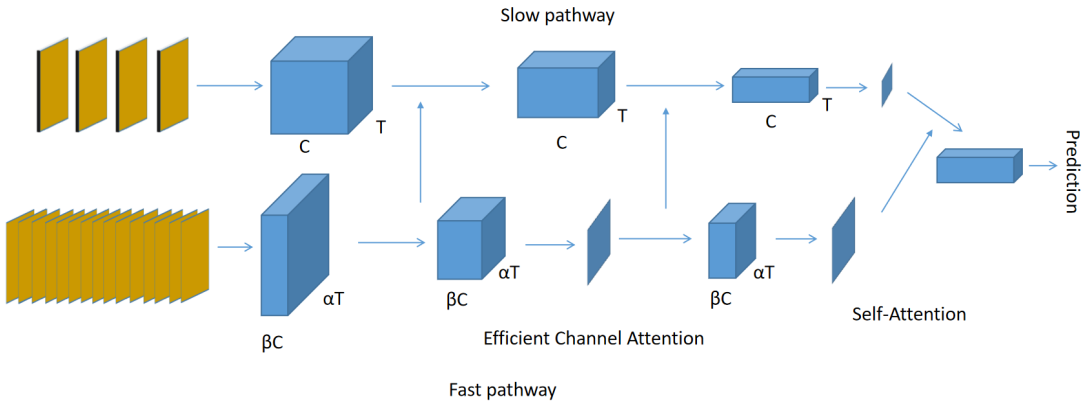

Figure 2: Improved SlowFast Network

ISSN 2616-5775 Vol. 8, Issue 2: 1-6, DOI: 10.25236/AJCIS.2025.080201

2.3 Fighting dataset

The dataset is downloaded from CSDN. It contains a total of six actions. The striking behaviors are divided into six categories as shown in Figure 3. The number of instances for each action is shown in Figure 4.

Figure 3: Dataset Overview

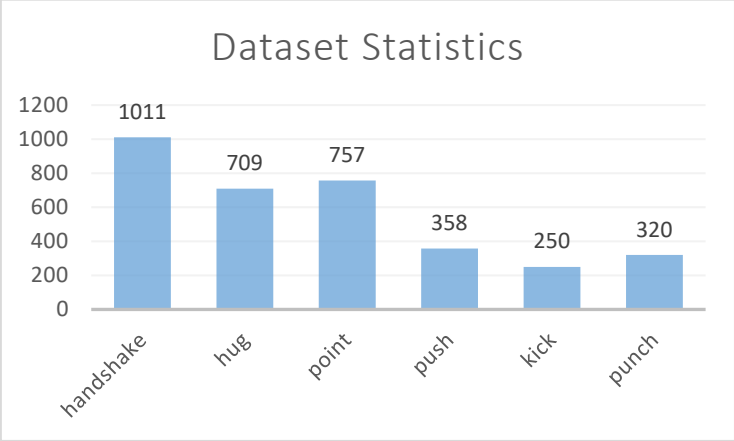

Figure 4: Statistics of the number of actions for each class in the dataset

2.4 Focal Loss Function

In the current fighting action recognition dataset, there may be significant differences in the frequency of occurrence across different action categories. Therefore, this paper applies focal loss to the aforementioned model to alleviate the class imbalance problem and improve the recognition capability for minority class actions. By adjusting the impact of easy-to-classify samples on the loss, we can make the training process more stable, leading to better convergence results. In practical applications, some actions may frequently appear in video data, while others may be relatively rare. This class imbalance can cause the model to favor predicting the common categories while neglecting the rare ones. To address this issue, we can use focal loss, a loss function specifically designed for class imbalance problems. Focal loss adjusts the impact of easy-to-classify samples on the loss, allowing the model to focus more on hard-to-classify and minority class samples. Specifically, it introduces a modulation factor and a balancing factor to adjust the weight of the loss function, reducing the contribution of easy samples and increasing the weight of hard-to-classify samples. As a result, during the training process, the model will focus more on the hard-to-classify categories, thereby improving the recognition ability for minority class actions. By applying focal loss, we can make the training process more stable and prevent the model from over-relying on samples from common categories. This helps improve the model's generalization ability and the recognition accuracy for rare categories. Ultimately, this will lead to better performance and outcomes in action recognition tasks.

ISSN 2616-5775 Vol. 8, Issue 2: 1-6, DOI: 10.25236/AJCIS.2025.080201

In the SlowFast model, the slow path and fast path extract information at different time scales. For the features output by the two paths, focal loss can be applied for supervised learning, allowing the model to focus more on hard-to-classify samples when processing information from different time scales. Suppose there are N action categories, the actual label is y_i, and the model's predicted probability is pi, then the focal loss formula can be expressed as:

$$FL(p_i) = -\alpha_i(1-p_i)^{\gamma}\log(p_i) \tag{9}$$

Here, ai and pi correspond to the category of the actual label y_i .

For each sample, the prediction probabilities p_{slow} from the slow pathway and p_{fast} from the fast pathway are first computed, followed by the calculation of the loss.

The loss for the slow pathway:

$$FL_{slow} = -\alpha(1-p_{slow})^{\gamma}\log(p_{slow}) \tag{10}$$

The loss for the fast pathway:

$$FL_{fast} = -\alpha\left(1-p_{fast}\right)^{\gamma}\log\left(p_{fast}\right) \tag{11}$$

The final total loss:

$$Total\ Loss = \lambda_{slow}.FL_{slow} + \lambda_{fast}.FL_{fast} \tag{12}$$

The above steps ultimately enhance the recognition capability for minority class actions.

3. Conclusion

The original network and the improved network were validated on the dataset processed using the above steps. Figure 5 compares the learning rate curves before and after the improvement.

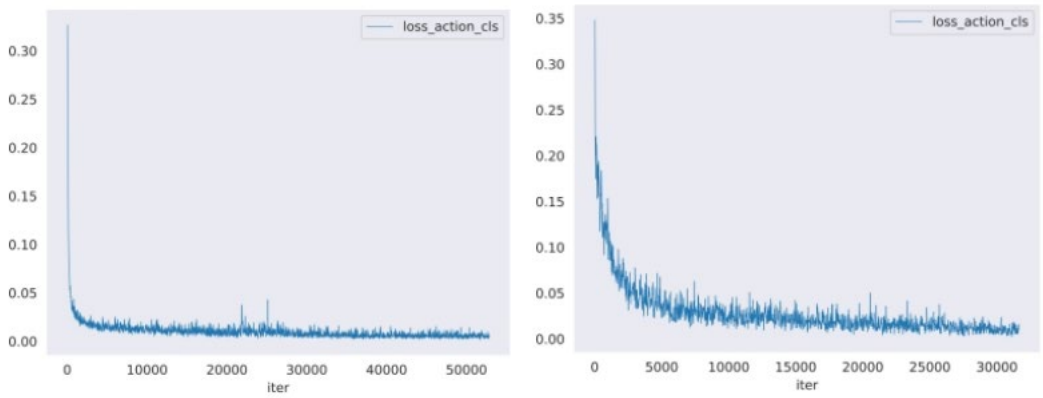

Figure 5: The learning rate curves before and after the improvement

By observing the learning rate curve after the improvement, it is evident that the model's convergence speed has significantly increased. The new learning rate strategy enables the loss function to decrease more rapidly, allowing the model to approach the optimal solution more quickly. This acceleration not only shortens the training time but also improves overall training efficiency, enabling the model to achieve higher performance levels in a shorter period. In the end, the improved learning rate adjustment method enables the model to perform outstandingly on complex tasks, quickly and robustly converging to the optimal solution.

By training the original network and the improved network on the fighting dataset, a significant improvement in accuracy is clearly observed. Specifically, the accuracy of the original network is 54.92%, while the accuracy of the improved network reaches 85.11%. This substantial increase demonstrates that incorporating the self-attention mechanism, the ECA attention mechanism, and using the Focal Loss function have greatly enhanced the network's performance for this dataset. The self-attention mechanism helps the network better capture global information, allowing for more accurate understanding of temporal and spatial features during action recognition. The ECA attention mechanism fine-tunes the inter-channel weight relationships, enabling the network to more effectively

ISSN 2616-5775 Vol. 8, Issue 2: 1-6, DOI: 10.25236/AJCIS.2025.080201

utilize information from different channels, thereby strengthening its feature representation capabilities. Moreover, the Focal Loss function mitigates the influence of easily classified samples, focusing the model's attention on difficult-to-classify and minority class samples, thus effectively addressing the class imbalance issue. With these combined improvements, the network demonstrates excellent performance in handling complex action recognition tasks, achieving an accuracy increase of nearly 30%. This not only validates the effectiveness of attention mechanisms and Focal Loss in enhancing model performance but also provides valuable insights for further research and practical applications.

References

[1] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.
[2] Zou M, Zhou Y, Jiang X, et al. Spatio-Temporal Behavior Detection in Field Manual Labor Based on Improved SlowFast Architecture[J]. Applied Sciences, 2024, 14(7): 2976.
[3] Vrskova R, Kamencay P, Hudec R, et al. A new deep-learning method for human activity recognition [J]. Sensors, 2023, 23(5): 2816.
[4] Wang Z, Zheng J, Yang M, et al. Research on human behavior recognition in factory environment based on 3-2DCNN-BIGRU fusion network [J]. Signal, Image and Video Processing, 2025, 19(1): 1-12.