Journal of Research High School

Vol. 2023 (2) 115 – 121
ISSN 2688-3651 [online]

Enhancing American Football Analytics: Classifying Play Videos as Run or Pass using Deep Learning

Kyle Zhou[1] *

[1]Sunset High School, Portland, OR, USA
*Corresponding Author: kyle.yiheng.zhou@gmail.com

Advisor: Jason Galbraith, Jason_Galbraith@beaverton.k12.or.us

Received March 14, 2023; Revised July 12, 2023; Accepted, August 7, 2023

Abstract

High school football coaches often rely on reviewing past season game footage to develop winning strategies. Identifying recurring patterns in the opposing team's offensive plays through video analysis helps coaches plan effective defensive tactics. However, accurately classifying play types from video clips using automated methods remains challenging and obtaining a sizable dataset of labeled plays for training models is difficult. This paper proposes a machine learning-based approach that utilizes the MoViNets model for action recognition. To overcome the challenge of limited labeled video clips, this paper utilized transfer learning to fine-tune MoViNets models that are pre-trained using a large dataset Kinetics-600. Extensive experiments were conducted to determine the optimal sampling scheme for the videos and compared the performance of two variants of MoViNets, the smaller A0 model and the scaled-up A3 model. The results show that using 24 frames sampled at 2 frames per second yields the best classification performance, and the A3 model achieves 81% accuracy on the test dataset, outperforming the A0 model's 74%. Ongoing data collection is expected to further improve the model's accuracy, potentially enabling automated play type classification.

Keywords: Machine learning, Deep learning, Convolutional Neural Network (CNN), Video classification, Transfer learning

1. Introduction

Applying machine learning in sports analytics has gained significant momentum in recent years. However, due to the vast diversity of game plays, extracting insights from American football remains a challenging task. In American football, the game is played on a 120-yard-long field with a width of 53 ⅓ yards. It involves two teams of 11 players each, one on offense and the other on defense. The offense executes plays, either by running or passing the ball, aiming to gain yards and score. Running plays involve the quarterback handing off the ball to a running back, while passing plays involve the quarterback throwing the ball to a receiver. The objective is to move the ball towards the end zone and score points by either kicking a field goal or advancing into the end zone. The offense has four downs (chances) to move the ball at least 10 yards, and if unsuccessful, the possession of the ball changes. Game film analysis is a common practice among high school football coaches, as it allows them to gain valuable insights into their opponents' strengths and weaknesses and develop effective strategic plans. A crucial aspect of this analysis is studying the offensive play types of the opposing team, including their tendency to pass or rush the ball. Manual annotation of past play video clips is often tedious and prone to errors. Recent advancements in computer vision and video understanding have made it possible to automatically classify play types from recorded video clips.

In previous studies, such as the one conducted by (Chen et al., 2014), authors utilized several processing techniques such as noisy detector, Hidden Markov Model, and KLT tracking to identify five different play types

Vol. 2023 (2) 115 – 121
ISSN 2688-3651 [online]

(Offense, Defense, Kick off, Punting, and Field goal) with a 77% accuracy rate. Additionally, another paper (Fernandes et al., 2019) predicted whether the play was a pass, or a run based on extracted play features such as the current down, yards to go for the first down, and point differential. They compared four different models (CART, KNN, Random Forest, and neural network) and found that the neural network achieved the highest accuracy of 75.3%. More recently, (Liu et al., 2022) proposed a deep learning-based pipeline for player tracking in videos. Their end-to-end system integrates an object detection network, detection transformer, and convolutional neural networks (CNNs) for player identification and time tracking.

Deep learning models have proven to be more effective than traditional methods for recognizing intricate and nuanced patterns in video data. Among these techniques, 3D CNNs have demonstrated exceptional performance and are considered state-of-the-art for video classification tasks (Tran et al., 2018). However, the memory and computational demands of 3D CNNs can be prohibitively high, which poses challenges for their deployment. To address this issue, a family of mobile computer vision models called MoViNets has been developed (Kondratyuk et al., 2021). These models are specifically designed to overcome the limitations of 3D CNNs in terms of memory and computation, while still maintaining high accuracy.

The objective of this research is to explore the effectiveness of multiple MoViNets models in automating the classification of football play types. The study focuses on analyzing video clips obtained from 10 matches played by Sunset High School's football team during the 2022 season, specifically from September 2022 to November 2022. These video clips consist of 1039 play clips, each labeled as either a run or a pass. Given the limited amount of available data, the research employs transfer learning techniques as described by Hosna et al. (2022). Transfer learning involves leveraging pre-trained deep learning models that have been trained on extensive datasets. By utilizing these pre-existing models, the performance of the MoViNets models on smaller datasets can be enhanced.

The primary goal of the research is to achieve a reasonably high classification accuracy of at least 80%. This level of accuracy would enable the automation of the manual process involved in tagging play clips with their respective run or pass labels. This system holds enormous potential to aid coaching staff of all levels by saving them long hours spent tagging and allowing them to focus on more crucial tasks.

Initially, MoViNets base model A0 was used to conduct experiments to analyze the classification impact of video clip duration and sampling rate. The scheme that sampled a total of 24 frames at a rate of 2 frames per second outperformed other schemes, achieving a classification accuracy of 74% on the test dataset. As anticipated, the test accuracy increased to 81% by simply employing a scaled-up MoViNets A3 model. Future work includes using game clips from previous seasons, which is expected to further enhance the model's performance.

2. Data and Preprocessing

During the past football season, the Sunset High School football team played 10 games, and for each match, separate video clips were recorded and analyzed using the Hudl Assist Service (Hudl Service, n.d.). In total, there were 1226 plays, with most of them being labeled as either Run (534) or Pass (505). The footage was filmed from the press box of the stadium with a singular position camera. Each play clip was then cut from the start of the play to the end before being posted to Hudl. This study will only focus on distinguishing between Run and Pass plays by training video classification models.

Since the videos' length ranges from 12 seconds to 28 seconds and are recorded at 30 frames per second, the video pre-processing step involves selecting frames from the clip to be used as input to the model. This preparation pipeline is depicted in Figure 1. The pipeline includes frame sampling, image resizing, normalization, optional augmentation, and dataset formation and train/validation/test split. Frame sampling involves selecting a subset of frames from the video based on a specific sampling rate. Image resizing transforms the selected frames to a specific size to ensure consistent input for the model. Optional augmentation can be applied to further increase the diversity of the dataset and improve the model's performance. Finally, the dataset is formed and split into training, validation, and test sets to evaluate the model's performance.

The accuracy of the classification is significantly impacted by two key hyperparameters: the number of frames chosen, and the sampling rate used. Figure 1 also shows four sampling schemes evaluated in this study.

Vol. 2023 (2) 115 – 121
ISSN 2688-3651 [online]

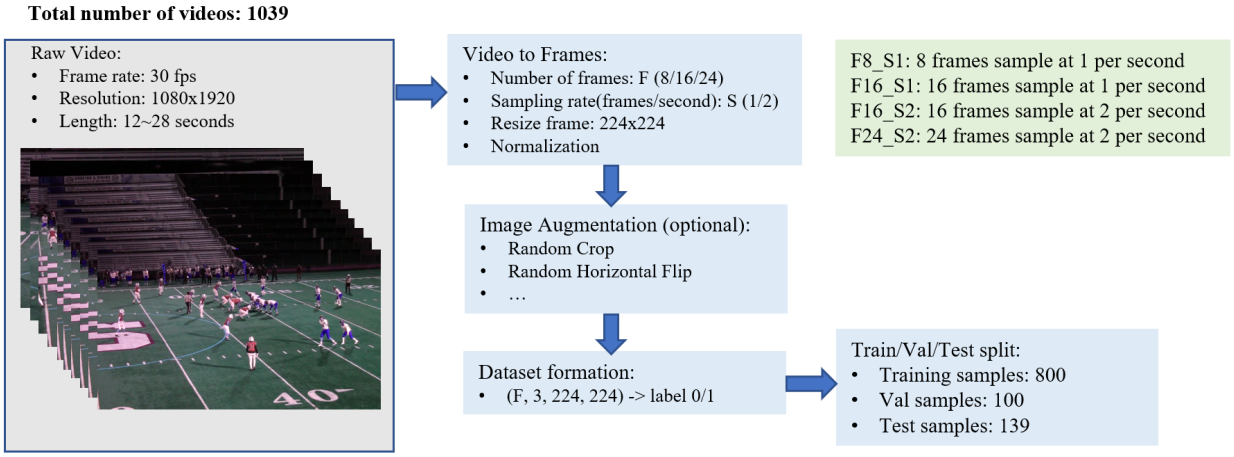

Figure 1. Video preprocessing pipeline and four frame sampling schemes.

3. Method

3.1 Video Classification Model

There are several types of video classification models, which can be broadly categorized as follows:

1. 2D CNN: These models process each frame of the video as an image using 2D CNNs and then combine the predictions from each frame to classify the video.
2. 3D CNN: These models process the video frames as a sequence and use 3D CNNs to learn spatiotemporal features which helps to capture the motion information in the video.
3. RNN: Use recurrent neural networks (RNNs) to model the temporal dependencies between frames.
4. Transformer-based models: These models use attention mechanisms to selectively focus on certain parts of the video frames, enabling them to better capture salient information.

Each type of video classification model has its own advantages and disadvantages, and the choice of model will depend on the specific task and data at hand.

This paper explored one specific video classification model called Mobile Video Networks (MoViNets), which have been designed to be efficient in terms of computation and memory usage based on 3D CNN architecture. As a result, MoViNets are suitable for deployment on mobile devices and support online inference. The MoViNets model family has been optimized using the Neural Architecture Search (NAS) technique to find the most efficient architecture. Furthermore, the introduction of Stream Buffers allows for the processing of videos in small, consecutive sub-clips, ensuring a constant memory requirement without compromising long temporal dependencies.

The MoViNets model architecture includes six different versions labeled A0 through A5, balancing model efficiency and classification accuracy. The A0, A1, and A2 variants are optimized for fast performance and are well-suited for deployment on mobile devices. In contrast, the A3 and other variants have more parameters and can achieve higher levels of accuracy. This study used the A0 model to assess the impact of video sampling rate and number of frames on model performance, as it has a shorter training time. Later, the classification accuracy was compared between the A0 and A3 models to evaluate the benefits of the higher model capacity of the A3 variant.

3.2 Transfer Learning with MoViNets

Transfer learning is a machine learning technique that involves repurposing a pre-existing model trained on a specific task as a foundation for a new, related task. This strategy enables a model to capitalize on the insights acquired while solving a problem, and apply them to a distinct but related problem, even when there is limited labeled data available for the new task. Typically, a pre-trained model is employed as a feature extractor to gather pertinent

Vol. 2023 (2) 115 – 121
ISSN 2688-3651 [online]

information from the input data. This information is then passed on to a new model to solve the new task. The new model can be fine-tuned, retrained, or modified using the extracted features and some labeled data from the new task. Employing this method can save considerable time and resources when compared to training a new model from scratch. Transfer learning has been applied successfully across various domains, including natural language processing, computer vision, and speech recognition.

TensorFlow Hub (Tensorflow Hub, n.d.) provides a range of pre-trained machine learning models that can be used as is or fine-tuned for specific tasks. One of these models is MoViNets, a 3D CNN architecture which has several versions pre-trained on Kinetics-600, a vast dataset of around 495,000 videos across 600 action categories (Carreira et al., 2018). To cater to the specific task of run and pass detection in football videos, a new model was created by taking a pre-trained MoViNets model and freezing its convolutional base. The original classifier head was then replaced with a new one that is tailored to the run and pass labels. Finally, transfer learning was performed by training the new classifier head using the pre-processed football videos.

4 Results

Although MoViNets models are designed to be more computationally and memory-efficient than other architectures, training the model can still be resource-intensive. To run all the experiments in the study, Google Colab (Google Colab, n.d.) was used with backend GPU support. The same training hyperparameters were set for all scenarios to ensure fair comparison, including using the Adam optimizer with a learning rate of 0.0005 and default values for all other parameters. A batch size of eight was used and trained the model for five epochs.

4.1 Sampling rate and frame counts

The MoViNets A0 pre-trained model was used, with its classification head replaced, as the baseline to examine the effect of frame count and sampling rate. Table 1 illustrates four distinct scenarios that were evaluated, varying in terms of total frame count and sampling rate. Covered video duration is the length of the video used in the analysis and is calculated by dividing the number of frames by the sampling rate.

Table 1. Four sampling schemes.

	Number of frames	Sampling rate (frames per second)	Covered Video duration (seconds)
F8_S1	8	1	8
F16_S1	16	1	16
F16_S2	16	2	8
F24_S2	24	2	12

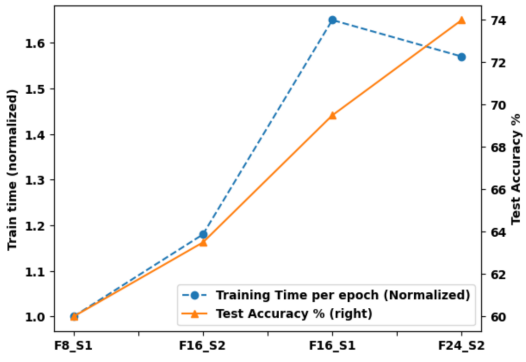

Figure 2. Training time and test accuracy for four sampling schemes.

Figure 2 shows the normalized training time and test accuracy for four different scenarios. Processing more frames leads to longer training time, as demonstrated by the F16_S1 scheme, which takes 65% more time to train each epoch than F8_S1. Higher sampling rates, on the other hand, result in shorter video length for the same number of frames. As a result, the epoch training time for F16_S2 is considerably shorter than that for F16_S1.

Capturing more temporal information by increasing the number of frames and sampling rate leads to higher test accuracy. The model F24_S2 attains the highest accuracy of 74.0%.

The confusion matrix for F16_S1 and F16_S2 is depicted in Figure 3. Both schemes extract 16 frames from the video clip, but with different sampling rates: F16_S1 samples one frame per second, while F16_S2 samples two frames per second.

Vol. 2023 (2) 115 – 121
ISSN 2688-3651 [online]

Two popular metrics used to assess the quality of a classifier's outcomes are precision and recall. Precision is calculated by dividing the true positive by the sum of true positive and false positive, while recall is determined by dividing the true positive by the sum of true positive and false negative. In simple terms, precision measures a classifier's ability to minimize the false positive rate and recall measures how well the classifier identifies all positive examples. In Figure 4, a comparison is presented between the precision and recall of the F16_S1 and F16_S2 sampling schemes. The results show that when sampling two frames per second (F16_S2), the precision for Pass is improved, but the precision for Run is negatively impacted compared to F16_S1 which samples 1 frame per second. Conversely, the recall shows the opposite pattern.

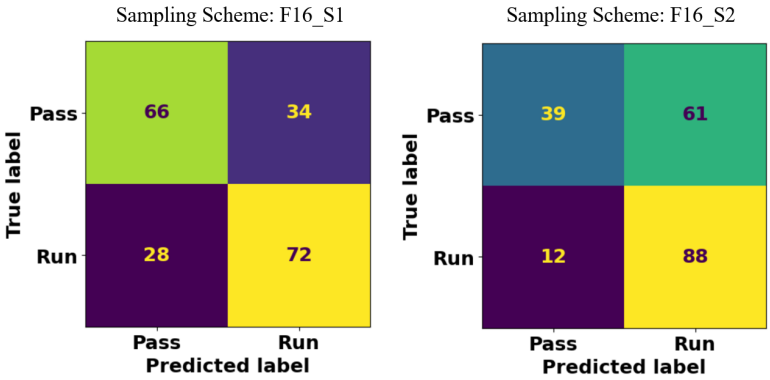

Figure 3. Confusion matrix for scheme F16_S1 and F16_S2.

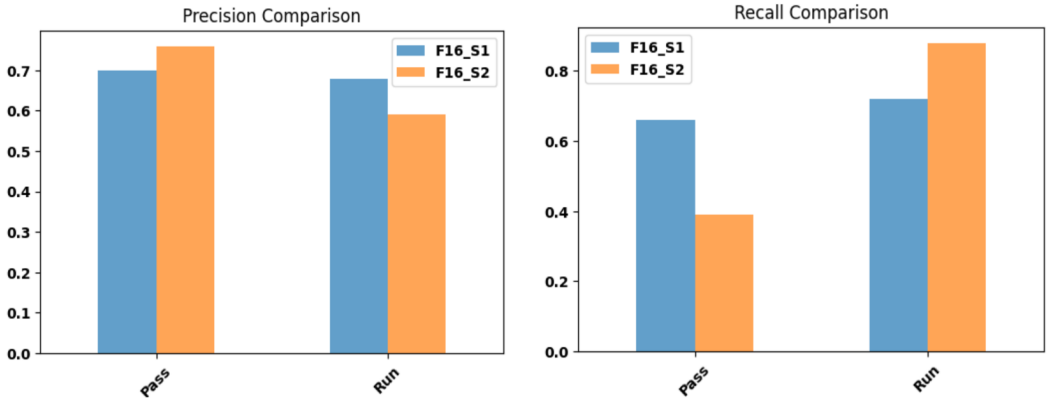

Figure 4. Precision/Recall comparison for schemes F16_S1 and F16_S2.

4.2 MoViNets A0 vs. A3

MoViNets offers a variety of models with ascending capacities, with A0 being the most efficient and A5 being the most powerful. Table 2 illustrates that MoViNets A3 has over three times more parameters than the baseline A0. Typically, the pre-trained feature extractor weights remain fixed, and only the classification head is trained. In this case, A0 has 911,583 pre-trained weights, while A3 has 4,688,129.

Table 2. MoViNets A0 vs. A3 parameter comparison.

	MoViNets A0	MoViNets A3
Total Params	1,900,769	6,217,987
Trainable Params	989,186	1,529,858
Non-trainable params	911,583	4,688,129

Generally, increasing the model capacity can enhance performance, provided there is sufficient training data. Utilizing the F24_S2 sampling scheme, the test accuracy of A3 model surged to 81.0%, compared to 74.0% achieved by the A0 model as shown in Table 3.

Table 3. Test accuracy comparison between A0 and A3 both using F24_S2 scheme.

	A0	A3
Test Accuracy (%)	74.0	81.0

Vol. 2023 (2) 115 – 121
ISSN 2688-3651 [online]

5 Discussion

This study conducted an analysis of 1039 plays taken from 10 official games played by Sunset High School's football team against other schools in the state during the 2022 season from September to November. This data set formed the basis of the analysis, and it was used to develop and refine the models. To increase the size of the data set, play clips from previous seasons between teams other than Sunset High School should be added, if possible. Incorporating additional data will improve the coverage of the data distribution and increase the accuracy of the classification algorithms. Furthermore, utilizing larger data sets will enable us to employ more advanced machine learning models, such as the scaled-up MoViNets A4 and A5, which can handle more complex and varied data sets. This should significantly improve the performance of Pass/Run classification.

As part of ongoing research, the aim is to investigate the impact of image resolution on classification accuracy. Currently, during video pre-processing, all sampled images are resized from the original 1080x1920 resolution to 224x224. However, some studies suggest that using higher-resolution images could enhance classification accuracy, albeit at the cost of higher computing resource overhead.

Additionally, the field of video classification is constantly evolving with the emergence of new algorithms. Among these, transformer-based models have gained significant attention due to their promising performance. Specifically, TimeSformer has achieved state-of-the-art results on various benchmarks (Bertasius et al., 2021). This model extends the Transformer architecture to handle spatiotemporal data and has been shown to effectively capture both spatial and temporal features in videos.

6 Conclusion

Machine learning techniques are becoming increasingly important in sports analytics to enhance team performance and gain valuable insights. By analyzing vast amounts of data, such as video footage, machine learning can identify patterns, predict outcomes, and make strategic decisions. In the context of American football, for example, analyzing the play types used by opposing teams during previous games is crucial for developing an effective defense strategy. However, manually labeling play types in video clips can be a tedious and error-prone task.

In recent years, developing effective machine learning models for action recognition in videos has become increasingly important in the field of sports analysis. However, one of the major challenges for American football has been the lack of a large dataset of labeled plays for training the models. This paper proposes a machine learning-based approach that employs the state-of-the-art MoViNets model for action recognition in videos. To overcome the issue of limited data, transfer learning is utilized, where a pre-trained MoViNets model is fine-tuned using around 1000 labeled football video clips. To determine the optimal sampling scheme for the videos, extensive experiments are conducted. The results of these experiments indicate that using 24 frames sampled at a rate of 2 frames per second produces the best classification performance.

Furthermore, the study compares the performance of two different MoViNets models, namely the efficient A0 model and the scaled-up A3 model. The results of this comparison demonstrate that the A3 model achieves higher accuracy, with a score of 81% accuracy on the test dataset. As data collection continues, the accuracy of the model is expected to improve further, potentially enabling automated play type classification.

In summary, the study proposes a hopeful method for classifying Pass/Run in football videos using machine learning. This approach could have substantial implications for game analysis and decision-making. The results of this research also demonstrate the potential for use of transfer learning and the state-of-the-art models such as MoViNets to overcome challenges related to limited data and achieve accurate classification of sports videos.

Acknowledgment

This research was mentored and supported by Sunset High School’s computer science teacher Mr. Jason Galbraith as part of the Artificial Intelligence class. I also want to thank the Sunset High School football program for providing me with labeled game film from the past season.

Vol. 2023 (2) 115 – 121
ISSN 2688-3651 [online]

References

Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? https://doi.org/10.48550/ARXIV.2102.05095

Carreira, J., et al. (2018). A Short Note about Kinetics-600.

Chen, S., et al. (2014). Play type recognition in real-world football video. 2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014. 652-659. 10.1109/WACV.2014.6836040.

Fernandes, C., et al. (2019). Predicting plays in the National Football League. Journal of Sports Analytics. 6. 1-9. 10.3233/JSA-190348.

Google Colab. (n.d.). https://colab.research.google.com/

Hosna, A., et al. (2022). Transfer learning: a friendly introduction. Journal of Big Data. 9. 10.1186/s40537-022-00652-w.

Hudl Service. (n.d.). https://www.hudl.com/

Kondratyuk, D., et al. (2021). MoViNets: Mobile Video Networks for Efficient Video Recognition. 16015-16025. 10.1109/CVPR46437.2021.01576.

Liu, H., et al. (2022). Deep Learning-based Automatic Player Identification and Logging in American Football Videos.

Tensorflow Hub. (n.d.). Collection of all MoViNet models. https://tfhub.dev/google/collections/movinet

Tran, D., et al. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.