**Innovation on the Field: Merging Deep Learning and Sequential Analysis for**

**Advanced Play Classification in College Football**

Paper Track: Football

Paper ID: 193944

**Abstract**

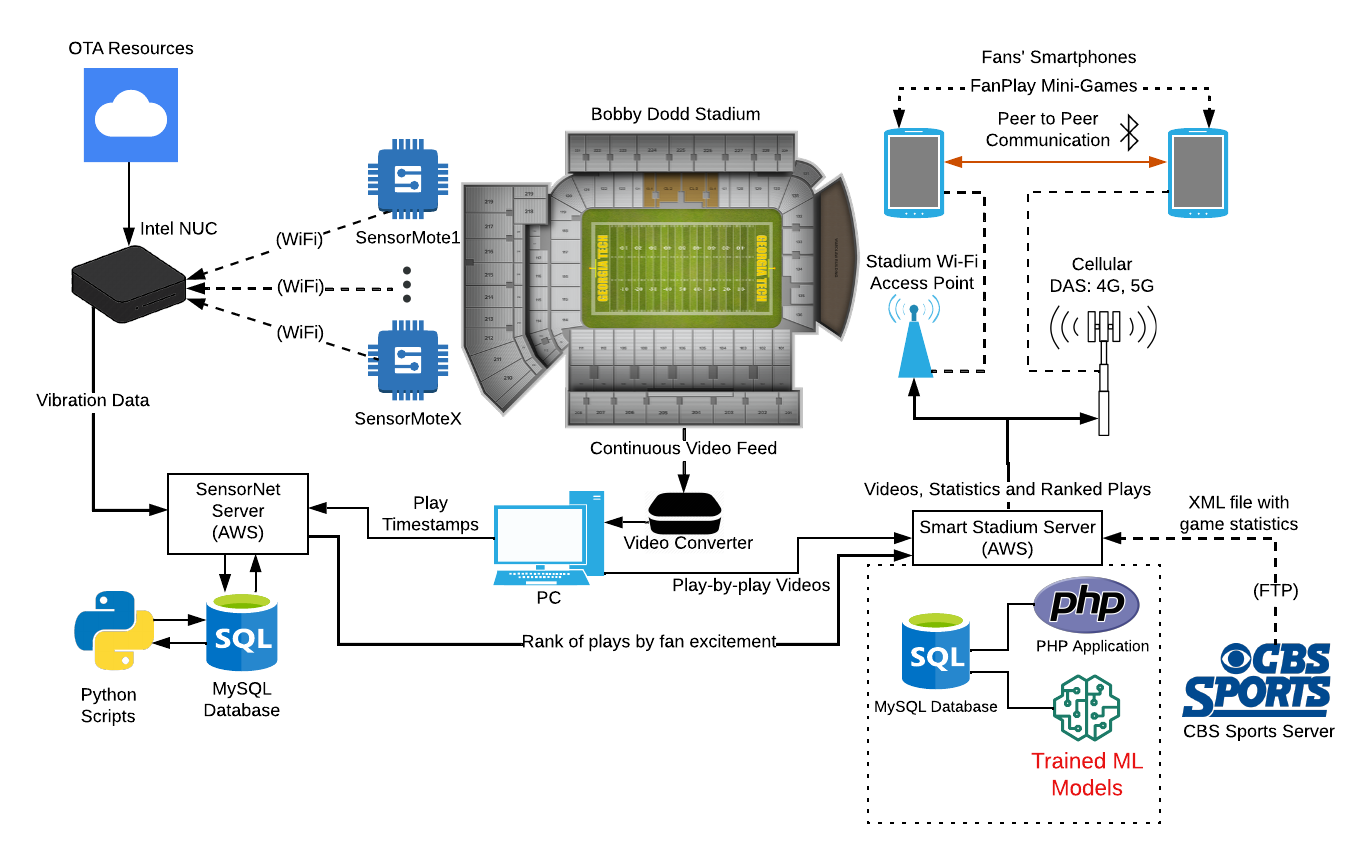
Fast, accurate, and automated classification of a football play that has just occurred is a fundamental task in the *Smart Stadium* system (Figure 1) that we are developing. This will enable remote, hands-off operation of the entire system, thus making commercialization of the system viable. To achieve this, we have developed an approach that, via conflation, fuses: (1) the play classification probabilities of a Computer Vision model trained with ground truth from more than 6,000 annotated video clips of plays, and (2) the play classification probabilities from a Markovian model that was derived from more than 270 football games to predict the probabilities that each type of play will occur given the previous play. This approach has improved play classification accuracy from the prior best of 79.65% to 87.26%.

1. **Introduction**

Every Saturday, millions of college football fans around the country attend over 400 football games to watch more than 80,000 plays. Millions of additional fans follow these live college football games on alternative mediums, such as the radio, television, and sports apps such as Yahoo Sports. Just as many people want to catch up with or review the results of these games at later times. As a result, there is a growing demand to produce synchronous, instantaneous, and searchable records of each play, including a video clip of the play and fans’ reaction to the play. This aligns with our permanent goal to maximize fan engagement with the game.

This demand led our team to create and deploy the Smart Stadium System [1-9] that is depicted in Figure 1 below. It consists of

* The subsystem that captures video clips, matches them with the play-by-play annotations, and makes them available on-demand to fans in the stands or elsewhere [1-3]. This system was deployed and operated within two football stadiums for many years, most recently from 2009 through 2017 at Georgia Tech’s Bobby Dodd stadium.
* A wireless sensor network that monitors the vibrations of the stands to measure the excitement of fans as they react to each play [4-6, 8, 9]. Fans cheering, jumping up and down, stomping, etc. cause the structure to vibrate and these vibrations can be captured and turned into a measure of the energy that fans expended. This energy measure is time stamped for each play and automatically matched with the video of the play captured at that time. The result is a ranking, for each quarter, of plays from the most exciting to least exciting [6].
* FanPlay, a suite of entertaining, sports-related games that are currently being developed for fans to play before (tailgating, pregame), after, or during longer intermissions such as halftime. These games use game statistics, videos, and sensor data from the venue in unique ways.



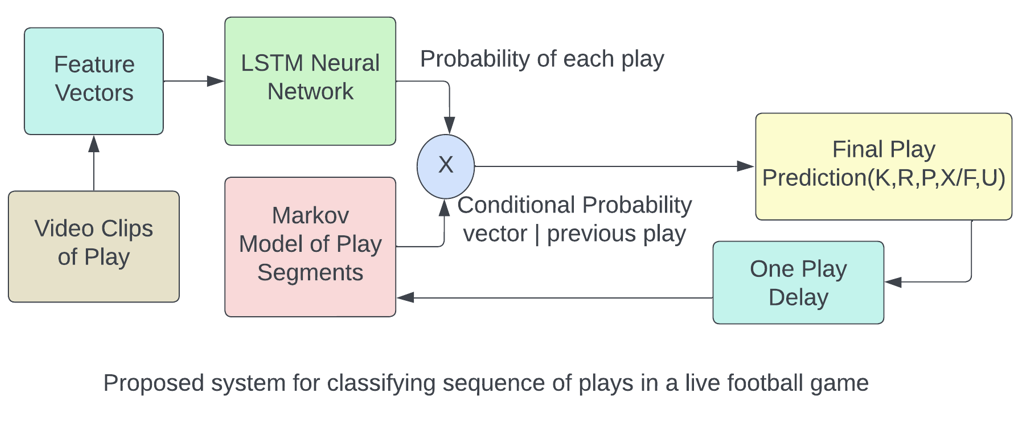
**Figure 1**: This is the architecture of the Smart Stadium System developed over the past 20 years. It shows the video capture and annotation subsystem, the vibration sensor network subsystem, the FanPlay subsystem, and the delivery of content to fans via Wi-Fi/4G+5G Cellular in the stadium. The trained Machine Learning pipeline, shown in red in the lower right, is the focus of this paper and resides in the Smart Stadium server. It autonomously matches the video clips of plays gathered during games in a MySQL database with the official annotations of the plays from an xml file that is available from CBS Sports.

For each of the above subsystems, there are opportunities to generate revenue by: (1) embedding ads in user interfaces; (2) making the system available for a subscription fee to fans outside the stadium; (3) including per game or per season ticket purchase for a fee; etc.

Currently, the entire system can be operated remotely during a game ***except*** for the capture and matching of video clips of plays with their appropriate annotations. This requires a person at the stadium who is familiar with football and focuses solely on this task whenever the game clock is running. Beyond human errors, a major challenge is that the annotations for plays are often delayed, created out of order, or need to be revised after their initial posting. This capture and matching task **must** be automated if the system above is to be scalable to 400 games and 80,000 plays per week in a cost-effective way.

The purpose of the work reported in this paper is thus to automate this capture and annotation task and eliminate the need for a skilled person in the stadium. That will allow the system to operate fully remotely and economically for many simultaneous games in many stadiums from a single command center. The result will be a commercially viable system that does not require a human to attend each football game for our system to provide live updates to fans during every game.

To achieve this, we have developed a system whose inputs are the video clips from the game, the play annotations (which may be out of order), and a Markovian model of the sequence of football plays throughout a game. A machine learning approach is used to train a Computer Vision (CV) model to recognize plays (e.g., run, pass, punt, P.A.T., kickoff, and field goal) from the video clip of the play. To catch classification errors made by this system, we use the grammar of a football game embodied in a Markov Chain whose states are labeled with relevant information including the down, field position, type of play, etc. Given the previous play, the Computer Vision model and the Markov Chain each produce probabilities of each type of play occurring in the current video. These predictions are then conflated [10] to improve the accuracy of the overall system. This approach is shown in Figure 2:



**Figure 2**: Architecture of the Smart Stadium video classification system. The video of the current play is the input to the system (lower left). The initial vector of classification probabilities for each type of play is made by a neural network trained on thousands of annotated plays from our dataset. This probability vector and the vector of conditional probabilities of the different play types given the previous play details (down, field position, etc.) are conflated in a final neural network [10]. This outputs a final 5-entry probability vector that gives us the likelihood of the specific play instance being each of the 5 categories, and the play type with the highest probability is then chosen as the classification.

We have achieved classification accuracies of 87.26% even with a limited training set and with a simplified Markov Chain. This heavily eases the task of matching plays to their annotations. This system performs the task of understanding the content of the video play clip, a task previously performed by humans. As discussed later in this paper, additional training videos and a more comprehensive Markov Chain – developed from play videos and sequences from many more teams and seasons – this system will surpass human performance in this task.

This paper is organized as follows: Section 2 details the current state of research regarding classifying sports plays from video clips. Section 3 describes how we obtained, cleaned, and explored our videos and associated Datasets. Section 4 delves into how we conflate the outputs of the Computer Vision Model and the Markov Chain Model to obtain the final, most accurate classification. Conclusions and a description of Future Work are provided in Section 5. We end with some key acknowledgements in Section 6.

1. **Related Work**

American football teams at all levels are constantly adopting innovative technologies to enhance fans’ experience of the game. From the Yellow Line developed by Sportvision to Machine Learning powered statistics like AWS’ Next GenStats, technological innovation has become an integral part of the football experience. More specifically, there has been an increase in the use of machine learning (ML) algorithms to analyze and classify video footage of sports. This can be partially attributed to the increase in computing resources that has enabled the convenient storage and processing of exceptionally large sets of video footage of all types.

Several studies have explored the application of Machine Learning in Sports Analytics, including action recognition and classification tasks of multiple sports activities. Other studies have used the predictive nature of machine learning algorithms to predict sequences of events within sports, most relevant to the task we tackle. However, compared to individual sports, classifying plays of team sports is often more difficult due to the involvement of actions from many individuals. This section provides an overview of the prior literature, with a focus on play classification and sports video analysis using deep learning techniques.

One successful use of machine learning for American football was the use of a deep learning model to identify play formations of the offense immediately before the snap of the ball. An accuracy of 95.4% was achieved, as described in the paper by Zhou, et. al. in 2023 [11]. A dataset of 1400 video clips of plays were used to train the model. This automatic offensive line-up identification is of use in play analytics and scouting for coaching staff. This study thus showed the ability of a Machine Learning model to aid in analytics. Identifying formations, however, is not directly comparable to play classification – the former is the classification of a snapshot taken at a certain time before the snap; the latter is classification of a video sequence of the entire play, which can last up to 20 seconds.

Another seminal work relating to play classification is the 2017 study by Chen et al. titled "Play Type Recognition in Real-World Football Video" [12]. The authors proposed a two-stage Convolutional Neural Network (CNN) architecture to recognize offensive and defensive play types in American Football videos. In the first stage, they used a CNN to extract spatial features from video frames, while the second stage utilized temporal CNNs to capture the temporal relationships between frames. The study achieved an overall recognition accuracy of **75.3%** in distinguishing between offensive and defensive play types. Our work expands upon this work, performs a more complex task with more classification categories, and utilizes additional data about football to achieve an accuracy of **87.26%.**

A key reference in the domain play recognition is the paper by Siddiquie, et al., titled "Recognizing Plays in American Football" [13]. The authors presented a novel method for recognizing football plays by utilizing a combination of player tracking and modeling techniques. They developed a probabilistic framework that incorporated spatial and temporal relationships among players, enabling the identification of distinctive play patterns that achieved an average accuracy of **71.9%**. This approach laid the groundwork for understanding the complex nature of American Football plays and their recognition using machine learning techniques. Furthermore, this study was performed on a dataset which included clips of football plays from NCAA teams, including our video clips at Georgia Tech that were available on the web. We initially faced the difficulties discussed in [6]: video clips with different time intervals, different camera angles, and sometimes faulty capture of the video.

Unfortunately, the accuracy of these works is not high enough for reliable analytical use, especially for commercial purposes. In addition to these studies, several other works have explored the use of ML algorithms for sports clip analysis, including action recognition and event detection to improve accuracy. For example, Li et al. proposed a novel framework based on CNNs and Long Short-Term Memory (LSTM) networks for recognizing and extracting key features relating to human actions, which has been a fundamental challenge in the world of computer vision [14]. The training set used included all types of sports, which is great input for modeling complex and active human interactions. With the two-layer LSTM structure, their model was able to represent the temporal pattern of visual input much better than the prior LSTM model. The study was able to achieve a feature extraction accuracy rate of up to **89%** based on action recognition benchmarks. They were able to extract and highlight important aspects of actions from specific frames, which would need to be scaled up in a significant way as in our classification task, there are 22 players on the football field, and they are not operating independently. This approach is thus very interesting but not immediately relevant to our task of classifying plays.

Another approach uses a Markov Chain to predict the next play. An example of this approach within sports is Ötting’s 2021 paper [15]. It studies the use of Hidden Markov Models to predict play calls within the NFL. Using a dataset of almost 300k observations, this study was able to get a prediction accuracy of **71.5%** for play calls in the 2018 NFL season. We use conflation to fuse this Markov Chain approach with our Machine Learning approach to maximize accuracy.

In another work using fusion, Baccouche, et al. proposed a sequential deep learning method for human action recognition in videos using a combination of 3D Convolutional Neural Networks (3D CNNs) and Long Short-Term Memory (LSTM) networks [16]. This fusion of 3D CNNs and LSTM networks allowed for efficient handling of both spatial and temporal information in the video sequence, leading to improved performance in human action classification tasks by over 10% compared to related studies. The study demonstrated the benefits of conflating different machine learning models to effectively address the challenges of capturing complex patterns in real-world data. This is the approach utilized in this paper for our neural net for play classification, prior to conflation with the Markov Chain prediction of the next play.

The paper "Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features" [17] is a significant contribution to the field of computer vision and action recognition. This study utilized a combination of Convolutional Neural Networks (CNNs) and Deep Bidirectional Long Short-Term Memory (DB-LSTM) networks to create a novel method for action recognition in video sequences. By extracting deep features from every sixth frame of the video and then employing a DB-LSTM network with multiple layers in both the forward and backward passes, the researchers were able to process lengthy videos and learn the sequential information among frame features, resulting in significant improvements in action recognition, and a maximum accuracy of **92%**. This study describes the success that can be achieved with model fusion, and how we can fuse different models to achieve higher accuracy than they would achieve independently.

Finally, in the study "Event Recognition and Classification in Sports Video,” [18] the authors use a Hidden Markov Model (HMM) to analyze ball trajectories and classify different events in sports videos. This Markov model can deal with unobserved (hidden) states and represents a dynamic Bayesian network. In contrast, our study employs a Markov Chain model, a simpler and more direct approach that considers the states of the game to be fully observable. While both are stochastic processes used to model sequences of events, our choice of a Markov Chain was guided by its simplicity and effectiveness in predicting plays in American football. While the previous study's use of a Hidden Markov Model was effective for their specific application of classifying diverse sports events based on ball trajectories, our approach with the Markov Chain model demonstrates its efficacy in successfully predicting plays.

Overall, these studies demonstrate the potential of machine learning techniques, specifically Convolutional Neural Networks and Markov Chains, in classifying football plays and analyzing sports video footage. They also give us an insight into the many ways we can fuse different models to achieve better accuracy and get the best out of each standalone model. In our study, we use these principles to develop a *fast* classification system for plays in American football. And, comparing the above work that is most closely related to the task of classifying plays, the approach in this paper increases the classification accuracy from **75.6%** to **87.2%.**

**3. Datasets**

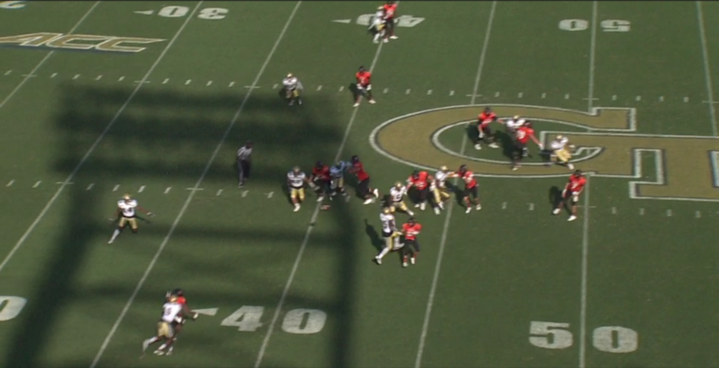
Our dataset for training our computer vision model was built completely in-house in with the cooperation of Georgia Tech Athletics. During Georgia Tech home football games for the NCAA Football seasons from 2010 through 2017, past and current members of our team manually recorded video clips containing each play. This gave our team a dataset of **6,056 football play videos,** which we manually labeled with the official NCAA annotation of the play. Exploring this dataset, we immediately noticed errors with some recordings, including some that began significantly before or after the play, others with overlaid graphics/ads, some with changing camera angles during the play, and even a few recordings that contained no play.



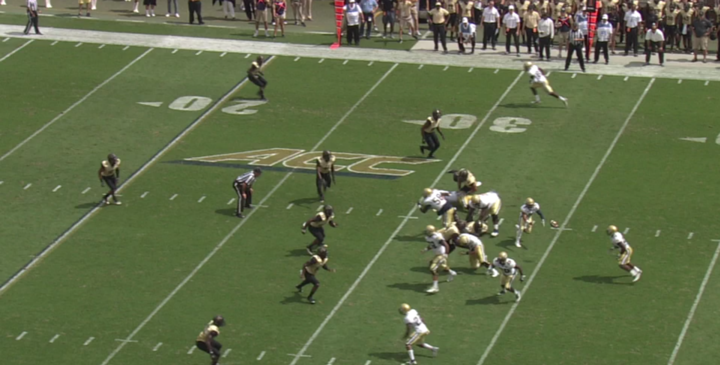
**Figure 3**: A frame from a video clip with an overlaid graphic during a play. This is an example of an unusable play clip that was removed from our original dataset.

Figure 3 shows one of the 1,000+ play clips we cleaned from our original dataset. Like this recording error, many of the errors found in our independent manual recording of videos reflect common problems associated with the current state of college football play delivery and analysis. Specifically, our data cleaning process included visual inspection of all 6,056 Georgia Tech football plays in our dataset to identify recording errors. Each video clip is typically 7-15 seconds long and consists of one single football play, such as a completed pass or a field goal kick (or potentially no play at all). The result of this intensive process was a cleaned dataset of usable plays for training.

Plays with fundamental recording errors, like “No play”, or “Duplicate”, that were deemed unusable were excluded from the dataset. Minor errors like “Late” or “Minor graphic” were reviewed on a case-by-case basis: in some cases, the error was deemed not significant enough to affect how the play was classified. For the plays labeled “Early”, the video was simply trimmed to the actual start of the play to mitigate the issue. After scrubbing the data set of all recording errors, the size of the cleaned dataset was **4,722 plays**.

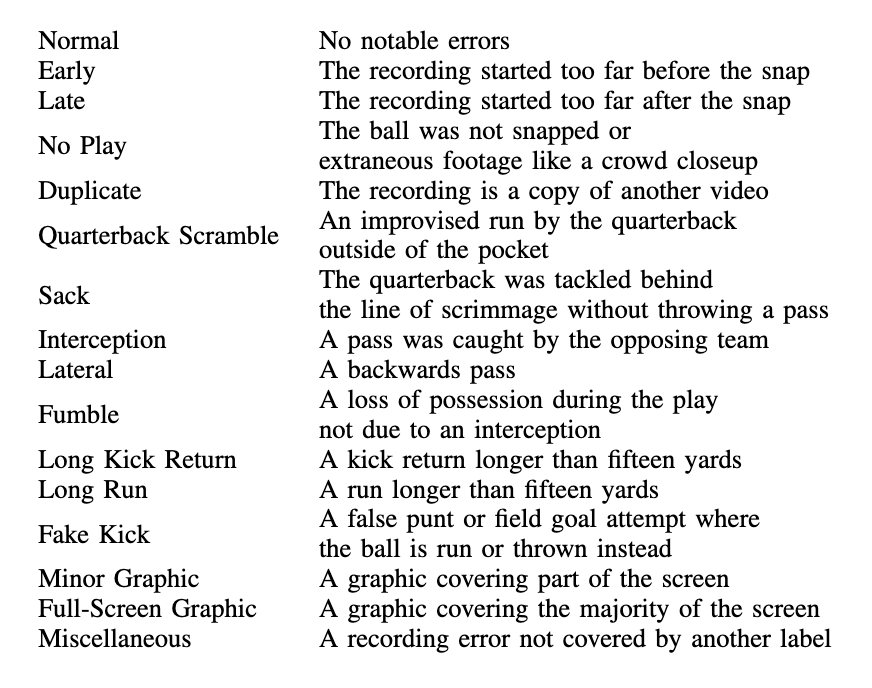


**Figure 4:** This is a frame from video of a play that our model correctly classified as a pass.



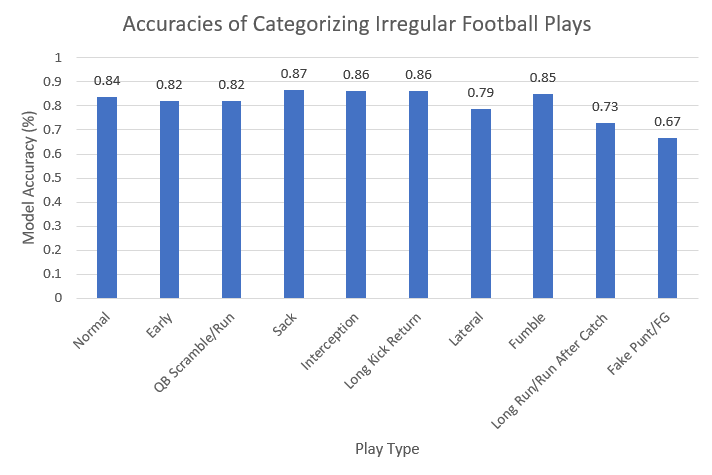
**Figure 5:** This is a frame from a video of a pitch run that our model incorrectly classified as a pass.

To gain a better general understanding of our dataset, while watching each play clip to identify recording errors, we also identified complex/rare play clips that would be harder to classify. For example, Figures 4 and 5 both show frames from video clips that were classified as a pass; while the play in Figure 4 was actually a forward pass, the play in Figure 5 was a backwards pitch, where the ball is “passed” backwards. While this is technically a run play, it was understandably misclassified as a pass. Figure 6 below lists the possible labels for each play clip, with some labels corresponding to recording errors and others corresponding to complex plays that we believed to be inherently difficult to classify.



**Figure 6**: The labels used for video clips during the Data Cleaning process.

The motivation to label and review these rare play types, such as Lateral and Interception, was to identify potential weaknesses of the existing Computer Vision pipeline we had in place. To test our hypotheses, we measured the model accuracy for each irregular play type, for which results are shown in Figure 7.



**Figure 7**: Overall Model Accuracy for Problematic/Complex Play Types.

As shown in Figure 7, we observed that the accuracy of the model on play clips containing most irregular play types performs similarly, if not better, than the accuracy of the model on play clips containing normal (Simple Run, Simple Pass, etc.) plays. We noticed that our standalone Computer Vision model performs worse on the Long Run, Lateral, and Fake Punt/FG play types, which are all inherently confounding play types, even to a trained human eye. As the play develops, Laterals and Long Runs show many similarities to a typical pass play and Fake Punts/FGs may appear similar to realistic Punts/FGs. From this experiment, we hypothesized that the sequential context provided by the Markov Model can assist in our classification task. For example, it is much more likely for a Fake Punt to occur while on the opponent’s 40-yard line than on a team’s own 10-yard line.

To build our Markov model of play sequences, we used a set of XML files containing play annotations and game statistics created by the NCAA and deployed on a CBS Sports server. This dataset included XML data for each play that occurred during each of the **277** football games played by Georgia Tech from 2000 to 2021. We have automated parsing of these files to extract contextual and statistical data for the **56,596** plays that occurred in this time frame.

Given that our dataset consists solely of Georgia Tech plays, to ensure generalizability of our model, we performed a comparison between the play distribution of Georgia Tech’s football team from 2010-2017, the timeframe for which we have play recordings for, and the play distribution of all other NCAA College Football teams during the same timeframe.

|  |  |  |
| --- | --- | --- |
| **Play Type** | **Georgia Tech** | **NCAA Average** |
| Run | 58.18 | 44.3 |
| Pass | 23.28 | 38.65 |
| Punt | 4.29 | 5.74 |
| Extra Point | 5.44 | 3.58 |
| Field Goal | 1.5 | 1.98 |

**Figure 8**: Average distribution of play types per game: Georgia Tech vs. NCAA average

As listed in Figure 8, we also observed that Georgia Tech football, specifically during the years 2010-2017, had a significant number of Run plays, 10 plays per game above the national average. This is due to the Triple Option being a popular play design under Coach Paul Johnson, Georgia Tech’s head coach during that time span. These results supported our hypothesis that our model showed overfitting on run plays. We noticed that recordings that started late or contained no play were classified as run plays, meaning the model classified a play clip to be run even when it could not gather any information from the video of that play. To prevent this over-prediction of run plays, we are currently building a more representative training dataset, one that contains play clips from all NCAA football teams. We discovered a csv dataset of NCAA College Football plays from all teams during 2021-2023 [19]. The dataset was formatted differently than the NCAA XML dataset, so further data manipulation was required. The original all-teams dataset had information for the same play split across rows, where there was one row per key player involved on the play (i.e. for a completed pass play, there was one row each for the quarterback, the skill position player who caught the ball, and the defender who made the tackle). We adjusted the dataset so that each play corresponds to a singular row, and we created new columns calculating the current down and distance, metrics that are useful in creating the Markov Chain.

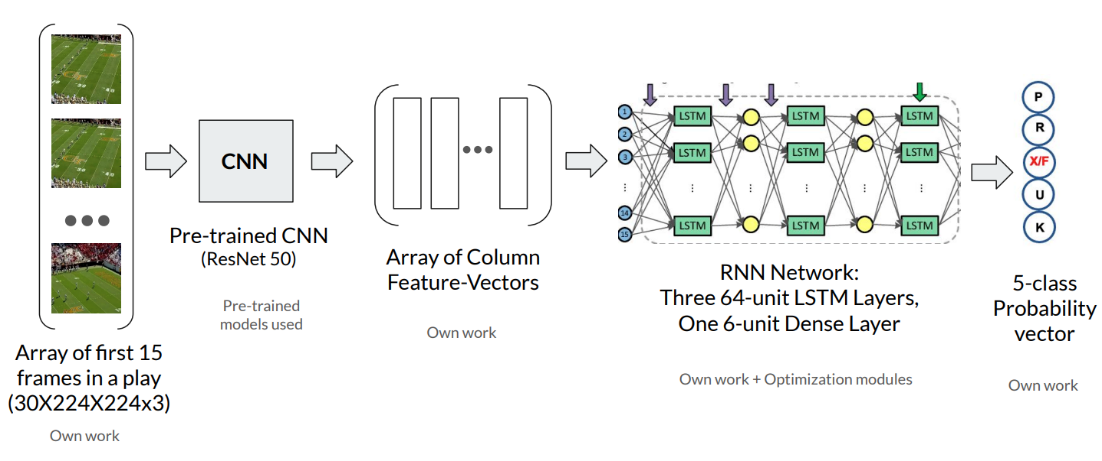
**4. Methods and Results**

**4.1 Computer Vision Model**

The main objective of our effort was to develop a model to automatically classify videos of football plays into one of 5 categories:

* ‘R’: Rushing Play
* ‘P’: Passing Play
* ‘X/F’: Extra Point or Field Goal
* ‘U’: Punt
* ‘K’: Kickoff

We combined the Extra Point and Field Goal categories due to the similarity in play type and to mitigate small sample size issues we would face if we separated them.



**Figure 9**: Base Model Architecture description.

We built a Computer Vision pipeline to classify the play type given a football play clip. Our model (Figure 9) utilizes a pre-trained Convolutional Neural Network, ResNet 50, to extract feature vectors from static frames of the play clip. ResNet50 is trained on over 1 billion images from the ImageNet database. The early layers of the ResNet50 architecture can detect lines, edges and hence, general objects from an image, making each of the hidden units in the network semantically significant [20]. We leveraged this ability to generate feature vectors for each static frame of our play clip. Our system then inputs the collection of feature vectors to a custom Recurrent Neural Network (RNN) trained on our dataset of labeled plays. Essentially, we follow the popular Computer Vision process of fine-tuning a pretrained network to perform a certain task: Football Play Clip Classification.

Initially, we created and tuned our Computer Vision pipeline on our local systems with an 800-play sample. However, to utilize the entire dataset, we migrated all our models to the Partnership for an Advanced Computing Environment (PACE) [9] Instructional Cluster Environment (ICE) at Georgia Tech. We used The Open OnDemand interface for our Jupyter notebooks, because it allows access to PACE clusters through a web browser. On this cluster, our model is trained using a 24-core node hosting Dual Intel Xeon Gold 6226 CPUs @ 2.7 GHz, 192GB RAM and the Nvidia Tesla V-100 GPU. Our software stack mirrors commercial setups as it consists of locally built Anaconda3 Python 3.9 environments with OpenCV, FFmpeg, Pandas, TensorFlow, Keras, ScikitLearn, and CUDA Toolkit.

Wanting to maximize the performance of our Computer Vision architecture, we implemented a Grid Search algorithm for hyperparameter tuning. The specific parameters we included were the duration of the play clip, batch size, frames per second, epochs, activation function, loss function, and optimization algorithm. We tested all combinations of these hyperparameters on both the pre-trained ResNet50 and VGG16 Convolutional Neural Network architectures, along with the images in Grayscale or RGB, respectively. Hence, there were a total of 9 parameters that we tuned in combination. We performed this extensive hyperparameter tuning process locally on the 800-play sample dataset, and eventually on the entire dataset.

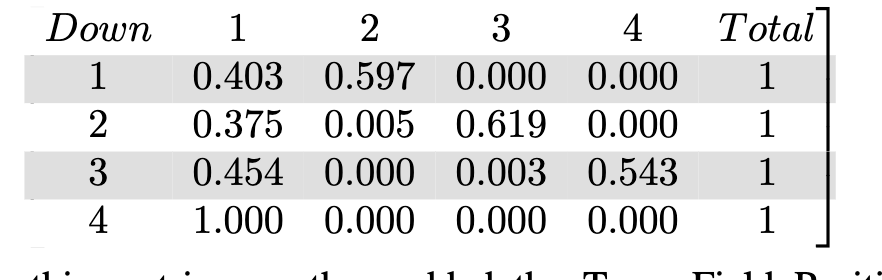
**4.2 Markov Model**

To ensure that the classification probabilities from the Computer Vision model were consistent with the “grammar” of play sequences in football games, we developed a Markov Chain that contains the probabilities associated with all possible state transitions that could occur from one play to the next. The main characteristics that comprise each state are:

1. **Down** (4): 1, 2, 3, 4
2. **Play type** (7): run (R), pass (P), PAT (X), field goal (F), punt (U), kickoff (K), other (O)
3. **Field Position** (10): 100-yard field divided into 10 10-yd buckets on the home side: [0,9], [10,19] . . . [40,50] and opposition side: [50,40], [39,30] . . . [9,0]
4. **Distance (yards) to 1st-down** (6): [1,3], [4,6], [7,9], [10,14], [15,19], [20,99]

We have a different number of options for each state: Down (4), Type (7), Field Position (10), Distance to 1st (6). This yields 4 \* 7 \* 10 \* 6 = 1680 distinct possible states that are grouped by the main characteristics. Thus, we have a 1680 x 1680 square matrix with state transition probabilities recorded at each entry; the matrix is built in row-major order, meaning that an entry M[i, j] represents P(i → j), the probability of going from state i to state j, and the sum of each row must be 1. Each row maps a current state to a list of possible next states with the corresponding probability of moving to each state. Rows will have entries that are 0 because certain play transitions, such as a kickoff on 2nd down, are impossible. Other play transitions that are unlikely, such as a field goal or punt before 4th down, will generally have lower transition probabilities.

Starting with the Down matrix, which is just a 4x4 matrix of transition probabilities between the 4 Down states, we gradually added new states to our matrix in order of decreasing relevance. From the Down matrix, we then built the Down-Type, Down-Type-FieldPosition, and finally the complete 1680x1680 Down-Type-FieldPosition-DistanceToGo matrix. For example, below we display the down matrix, which is a simple 4x4 matrix with labels/indices. From our row-major order, the probability of going from a 1st-down play to a 2nd-down play is 0.597 as denoted in the 1st row and 2nd column; naturally, it is impossible to go from 1st down to 3rd down, so that probability is 0.

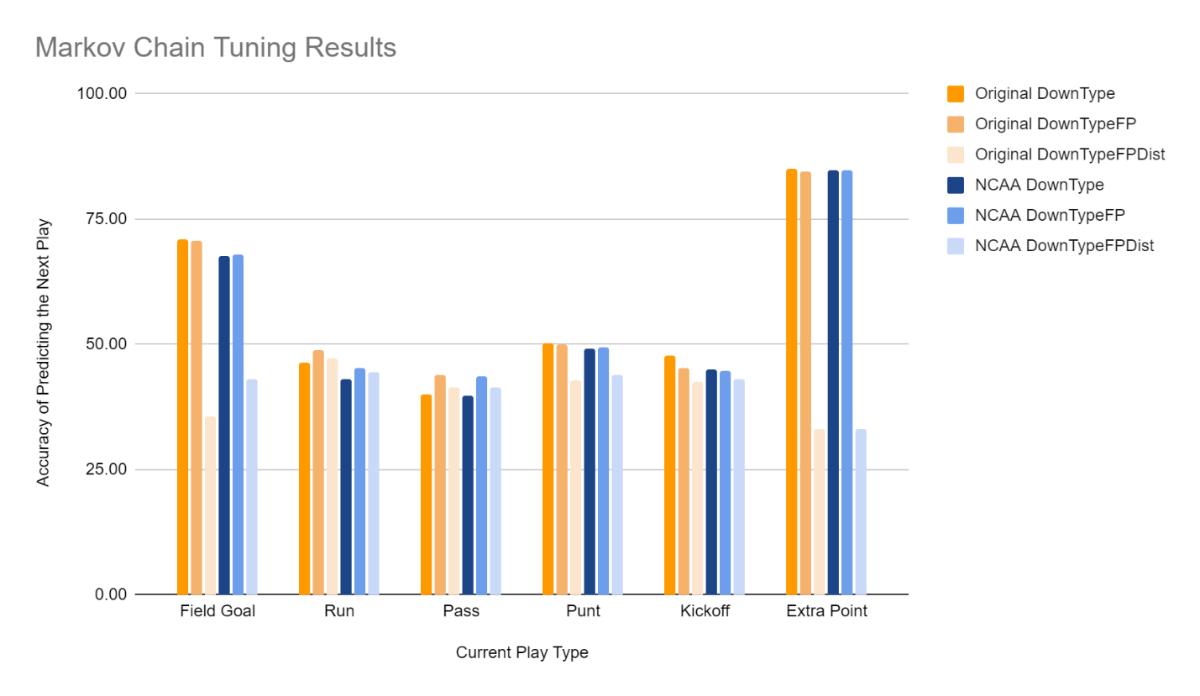


**Figure 10**: Markov Chain for Downs. Ex. P (1st → 2nd) = 0.597; P (1st → 3rd) = 0.000

Note that Kickoffs and PATs are recorded as 1st-down plays, which inflates P (1st → 1st) to 0.403.

From this base matrix, we then added the Type, Field Position, and Distance fields consecutively to the respective matrices. These matrices are only differentiated from each other based on the specific input parameters that are used to define a state. Alongside the Computer Vision model, the matrix is then used to influence the final decision by providing calculated probabilities of the most likely next state/play type.

To address the run bias due to GT’s run-heavy offense, we tested the use of the aforementioned general Markov Model, which was trained on all NCAA plays from 2021-2023.



**Figure 11**: Graph comparing the performance of 6 different Markov Chains independently predicting the next play, based on the six possible categories of the current play.

As shown in Figure 11, we compared the performance of three varieties of the Original (Georgia Tech data only) Markov Chain with the same three corresponding Markov chains trained on all NCAA plays from the past 3 seasons. We tested the performance of these Markov chains on solely Georgia Tech statistical data to ensure consistency with the Computer Vision model, in which we only have Georgia Tech video data to work with. It is important to note that as we only have video data from Georgia Tech games, we cannot accurately gauge the performance of NCAA Markov chains as part of our overall Conflated model until our model incorporates video recordings representative of the entire NCAA’s plays. Ideally, we would develop unique Markov chains for each NCAA football team to accurately represent their unique tendencies. For teams unable to develop their own Markov chains, we suggest the general NCAA chain be used as a baseline approximation. For our Conflation process, we solely use the original (Georgia Tech) Markov chains.

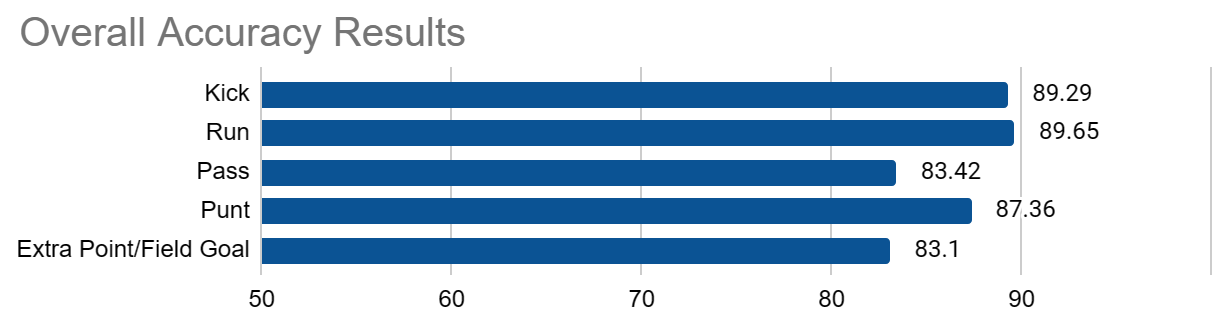
The Markov Model is not designed to be used as a standalone model in our process, rather, it is designed to be used in tandem with our CV model to harness sequential football knowledge when possible. This is why the results demonstrated in Figure 11 are primarily for informational purposes, rather than information we use to derive key conclusions. Nonetheless, the Markov Model particularly excels in predicting certain play types such as Extra Point. This is a quality we used to increase the accuracy of predicting the Extra Point type in our final conflated model.

**4.3 Conflation**

We explored several strategies for fusing the classification probabilities produced from the video clips by our Neural Network model and the vector of conditional probabilities from our Markov model.

For the Markov model, regardless of the particular State Transition Matrix (STM) used, we condensed the columns of the matrix into predicting a Run, Pass, Punt, Kickoff, or Extra Point/Field Goal. For example, the 1680x1680 full matrix is converted into a 1680x5 matrix. Hence, given the full details about the previous play, we can obtain a normalized 5-element prediction vector from the relevant row of the matrix, representing the probabilities of the current play being each of the 5 play types given the *prior* play (which corresponds to a row in the STM). The Neural Network model also provides a normalized 5-element prediction vector, given the video clip of the *current* play.

We explored multiple fusion strategies to fuse these two play prediction vectors. First, we performed a manual tuning process to find the optimal weights for each play type when taking a weighted average of the prediction vectors. We then experimented with specific confidence thresholds. Next, we explored adding previous play information as an input to our Recurrent Neural Network itself. As visualized in the Computer Vision pipeline diagram (Figure 9), an array of column feature vectors is passed to the RNN. Specifically, the input is a 30x2048 array, representing 2048 pixels in each of the 30 frames. Representing the previous play information as a 30-element vector, we were able to append a column to the input. While the accuracy of this approach improved the prediction accuracies of the modified neural network, we observed that that the input to the architecture still heavily contains information regarding the video clip, only slightly representing prior play information (2048 vs 1 column). In other words, we hypothesized that the model simply treats this additional information as noise.



**Figure 12**: Accuracy results for the final Neural Net + Markov Model system.

We discovered that the best approach to fusion was to regard the Neural Net’s prediction vector and the Markov Model’s prediction vector as independent predictors of the same phenomenon: namely, the current play type. Under this interpretation, conflation is a fusion approach that is optimal under several criteria, as proven previously [10]. The Conflation strategy solved the imbalance of information problem between the video clip and Markov Model discussed in the previous paragraph. Specifically, we create a final Neural Network that optimally calculates the specific weights to assign to each element of the 5-class probability vectors. This approach allowed us to evenly use the entirety of the dataset and incorporate Markov information in a unified model that performs optimal conflation. As for the tuning process, we performed hyperparameter tuning once again to obtain the final conflated output. This led to a final overall accuracy of **87%,** with play specific accuracies of [89.29, 89.65, 83.42, 87.36, 83.10] shown in Figure 12.

**5. Conclusion and Future Work**

While previous classification models for American football have proposed either computer vision or Markovian models for play classification, we combined these approaches using conflation to significantly improve classification accuracy. We used a set of 6,056 annotated videos of Georgia Tech football plays that were manually recorded by our team to train our LSTM model, along with an NCAA dataset of 56,596 play annotations from 277 Georgia Tech games ranging from 2000 to 2021 to calculate the state transition probabilities for our Markov Chain. By advancing from relying solely on the Computer Vision model to using conflation to fuse our Computer Vision and Markov model, we have increased overall classification accuracy from **79.65%** to **87.26%**, respectively.

While these Georgia Tech football videos have been adequate for training and testing our model, our current work incorporates annotated video data from other NCAA (and later NFL) teams across the country to diversify our model and improve classification results during live games. Secondly, for plays that consistently confuse our visual model – such as laterals or short passes – we will explore enhancing our system with object detection and object tracking algorithms. Thirdly, we will develop specific models for each team by training our Markov and Computer Vision models on their historical data to account for team-specific strategy and tendencies. Finally, we will optimize our neural network to ensure quick and accurate classification of plays during live games, thus enabling live updates for fans. This allows our Smart Stadium system to operate fully remotely and economically for all College Football games across the country. The vision for Smart Stadium is to expand to fit the entire NCAA, NFL, and eventually other collegiate and professional sports.

**6. Acknowledgements**

We thank prior members of our [Smart Stadium VIP team](https://vip.gatech.edu/teams/vp3) for gathering the ground truth data we used and for their initial efforts on the Computer Vision model. We also thank the leadership and staff of Georgia Tech Athletics for allowing our team access to the stadium to deploy our system to gather video and other data during games. This work was supported in part by Texas Instruments, Verizon, and Cisco Systems.

**7. References**

[1] X. Zhong, H.-H. Chan, T.J. Rogers, C. Rosenberg, and E.J. Coyle, “The Development and eStadium Testbeds for the Research and Development of Wireless Services for Large-scale Sports Venues,” *2nd IEEE/Create-Net Intl. Conf. on Testbeds and Research Infrastructures for the Development of Networks and Communities: TridentCom 2006,* Barcelona, Spain, March 1-3, 2006.

[2] X. Zhong, H.-H. Chan, T.J. Rogers, C. Rosenberg, and E.J. Coyle, “eStadium – The Living Lab,” *Proceedings of Infocom 2006*, Barcelona, Spain, April 23-28, 2006.

[3] A.A. Ault, J.V. Krogmeier, S.R. Dunlop, and E.J. Coyle, "eStadium: The Mobile Wireless Football Experience," *Third International Conference on Internet and Web Applications and Services (ICIW 2008),* Athens, Greece, June 8-13, 2008.

[4] D. Phanish, P. Garver, G. Matalkah, T. Landes, F. Shen, J. Dumond, R. Abler, D. Zhu, X. Dong, Y. Wang, and E.J. Coyle, “Design, Development and Deployment of a Sensor Network for Monitoring the Structural Behavior and Health of a Football Stadium,” *Proceedings of the 2015 IEEE World Forum on the Internet of Things,* Milan, Italy, Dec. 14-16, 2015.

[5] S. Laitrakun, D. Phanish, and E.J. Coyle, “Collision-Aware Distributed Detection and Estimation in Random-Access-Based Wireless Sensor Networks,” invited chapter in *Data Fusion in Wireless Sensor Networks: A Statistical Signal Processing Perspective*, The Institution of Engineering and Technology (IET), March 2019.

[6] H. Li, P.H.R. Pinto, D. Phanish, R.T. Abler, E.J. Coyle, “Good Vibrations: An IoPT Testbed for Infotainment in a Football Stadium,” *2020 IEEE World Forum on the Internet of Things*, New Orleans, LA, April 5-9, 2020.

[7] Team Leaders: S. Challa, G. Garre, R. Nibhanupudi, G. Reidy; Team Members: A. Agarwal, N. Candello, A. Chauhan, E. Galluzzi, V. Ganesh, C. Hu, S. Kapoor, N. Karthik, U. Kumar, M. Meng, J. De Ocampo, S. Pamidi, K. Parameswaran, N. Patel, J. Purkar, M. Schlicht, B. Tan, L. Ulku, N. Wang, D. Yang, S. Zhang; Graduate Students: K. Watters (PhD), N. Meyer (MS); Adviser: E.J. Coyle, “Cognitive Wireless Networks: Research, Applications, and Testbed,” *2023 IEEE World Forum on the Internet of Things,* Aveiro, Portugal, Oct. 12-27, 2023.

[8] S. Bandyopadhyay and E.J. Coyle, “An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks,” Proceedings of the *22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003),* San Francisco, CA, March 30 – April 3, 2003.

[9] D. Phanish and E.J. Coyle, “Application-based optimization of multi-level hierarchical clusters in ad hoc and sensor networks,” *IEEE Transactions on Wireless Communications*, Vol. 16, #7, pp. 4460-4475, July 2017.

[10] Hill, T. P., & Miller, J. (2011, July 20). How to combine independent data sets for the same quantity. AIP Publishing.

[11] Zhou, K., & Galbraith, J. (2023). Automatically labeling offensive formations in American football film using Deep Learning. Journal of Student Research, 12(1).

[12] S. Chen, Z. Feng, Q. Lu, B. Mahasseni, T. Fiez, A. Fern, and S. Todorovic, “Play type recognition in real-world football video,” in IEEE Winter Conference on Applications of Computer Vision, pp. 652–659, 2014.

[13] B. Siddiquie, Y. Yacoob, and L. Davis, “Recognizing plays in American football videos,” 01 2009.

[14] W. Li, W. Nie, and Y. Su, “Human action recognition based on selected spatio-temporal features via bidirectional lstm,” IEEE Access, vol. 6, pp. 44211–44220, 2018.

[15] M. Ötting, “Predicting play calls in the national football league using hidden markov models,” IMA Journal of Management Mathematics, vol. 32, no. 4, pp. 535–545, 2021.

[16] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, “Sequential deep learning for human action recognition,” in Human Behavior Understanding (A. A. Salah and B. Lepri, eds.), (Berlin, Heidelberg), pp. 29–39, Springer Berlin Heidelberg, 2011.

[17] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," in IEEE Access, vol. 6, pp. 1155-1166, 2018.

[18] V. Ellappan and R. Rajasekaran, "Event Recognition and Classification in Sports Video," 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, India, 2017, pp. 182-187.

[19] SportsDataStuff. "College Football Play-by-Play Data." Accessed November 15, 2023. <https://sportsdatastuff.com/cfb_pbpdata>.

[20] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. arXiv preprint arXiv:1412.6856.