Event Recognition in Hockey with a Multi-Task Learning Approach

Pascale Walters
20566177
CS 886-002: Deep Learning and Natural Language Processing

April 15, 2020

Abstract

A critical step towards generating automatic analytics for hockey games from broadcast video is classifying the events that occur. These events over the course of the game allow for an understanding of how each of the teams are playing and the flow of game play. Typically, event classification requires large amounts of annotated data from game video, which can be time-consuming and expensive. Hockey broadcasts come with commentary provided by broadcasters with a deep understanding of the game. This can be thought as a form of weak supervision, which would reduce the requirement of manual annotation. Inspired by this, ChirpNet is a semi-supervised model for event classification in hockey games. It is evaluated on a new ice hockey dataset. Experimental results show that the model works well for classifying eleven types of hockey events.

1 Introduction

Ice hockey is a popular sport, with several professional leagues operating, such as the NHL (National Hockey League) and CHL (Canadian Hockey League). Video analytics of hockey games can be used to provide teams with an advantage over their competitors, whereby they can gather more data about game events. These data can be used to influence coaching strategies and management decisions. In addition, the data can increase fan engagement as sports consumption becomes more digital [1].

With many recent developments in the field of computer vision, automatic generation of hockey analytics data from video has become possible. To develop effective methods for understanding hockey games with deep learning, large quantities of annotated training data are required for fully-supervised methods. Collecting this data is very time consuming and can be very expensive. Many broadcast networks provide commentary from hockey experts while the game is playing. Since the commentary provides a description of what is happening during the game, it can be thought of as weak supervision.

Previous research into including video captions for action classification has found that additional information can improve the performance of the method [2, 3]. Furthermore, action classification has been researched for hockey and other sports, but it frequently does not use the commentary from broadcast video [4, 5]. The sports for which there exist methods for action classification including commentary differ at a fundamental level from the game of hockey [6, 7]. Hockey games are very fast-paced and are dense in terms of the events that occur. There are also line changes, where the number of players on the ice is variable. This paper aims to address the lack of methods for semi-supervised hockey action classification from broadcast footage.

This paper introduces ChirpNet, a method for performing action classification for hockey games. It achieves an average accuracy of 66.39% classifying eleven hockey event types. This network is tested and trained on a novel dataset of one NHL game including commentary and event annotations. This opens the door into further research into semi-supervised methods for hockey analytics.

2 Related Work

Representing captioned videos and images is an established field of research. Since sports broadcast video is frequently accompanied by commentary, there has been much exploration of event detection.

2.1 Event detection in sports

There are several approaches for identifying and understanding what is going on in sporting events using visual and other approaches. Sports broadcast video is challenging for action recognition. This footage can include varied camera viewpoints, camera motion (panning

(a) Hockey (b) Soccer (c) Cricket (d) Diving

Figure 1: Example frames from various sports broadcasts. The different natures of each of the sports mean that there are significant differences in the broadcasts.

and zooming), and rapid transitions between events [4, 6]. It has been noted that it is difficult to create a sport-agnostic method for action recognition due to the many differences between sports [4, 5].

For example, broadcast footage of hockey games is very different from that of soccer, diving, or cricket. Example frames of each of these sports can be seen in Fig. 1. In hockey games, the game is very fast-paced, which means that events are dense over the course of the broadcast. Hockey also has line changes, where the players on the ice can change rapidly. An additional challenge is that the boards on the nearside of the broadcast can occlude players and the puck if they are in this area. Soccer features a playing field that is very large compared to the scale of the players. Events are also sparse, with no breaks in the play [5]. In diving, there is one athlete completing a dive at a time, and each dive is shown several times from different camera angles [8]. Cricket games can be very long, on the scale of several hours [7]. The events also tend to be repetitive (i.e., players bowling to a batsman), due to the nature of the game. These differences in sports mean that domain knowledge is required to develop an effective method for event detection in sports.

Team sports face unique challenges compared to individual sports [9]. Players have complex relationships within a team (e.g., goalie vs. defense) and between teams (offense and defense). In addition, these relations can change rapidly and the play can switch frequently between offensive and defensive zones. There may be players who do not take part in the play at a particular instance, but have a particular role in another situation. The challenge is to attend to the correct components of the video and ignore those that are not important to the action.

Tora *et al.* attempt to classify puck posession events from broadcast footage of hockey games [4]. The use a purely visual solution with uses several inputs to classify group activities: frames from the input video and bounding boxes of the players, with features from both extracted with a pretrained AlexNet. To model the temporal aspects of the task, they use a basic LSTM module. It was found that adding player-level features significantly improves the performance of the network, rather than solely relying on frame-level features.

Due to the sparse nature of events and the ambiguity in defining temporal boundaries for events in soccer games, Giancola *et al.* propose the task of spotting [5]. They attempt to identify an anchor frame that represents an event within a margin of error of a target anchor

frame, rather than defining a start and end frame for a sequence. They introduce SoccerNet, a new dataset for this task.

Many broadcasts of sports include a commentary feed, where experts in the sport describe what is going on in the game [8]. This additional feed can be thought of as a form of weak supervision for the visual information feed [6, 7].

A method for video retrieval from soccer games is proposed by Gupta *et al.* [6]. They note that captions are useful clues as to what is happening, but they do not necessarily give an explicit description. For example, commentators can talk about statistics for a particular player, which references their entire career or a season, or how a team performed in a previous game. This data is therefore very noisy, however there is enough signal to gain insight from the commentary stream. Their method includes a word sub-sequence kernel classifier to predict whether the contents of the captions actually refer to the current events. They rank retrieved video clips from soccer games with a weighted sum of the visual and caption features.

Sharma *et al.* use the text commentary to generate fine-grained annotations of cricket videos for retrieval of specific actions from an entire game [7]. Their method comprises two steps: scene segmentation, in which anchor frames that best identify the scene are identified and minute-long clips are aligned with a paragraph of commentary; and shot/phrase alignment, where the actions in the shot are identified and aligned with sentences of commentary.

In a similar task to action recognition, Parmar *et al.* propose a novel multi-task learning method for performing action quality assessment for diving videos [8]. Diving is evaluated by judges providing an overall score. Their method extracts visual features from the frames of the video. The loss function of this method is a weighted sum of a regression component for the score of the dive, in addition to a caption generation and action recognition. It is shown that adding additional tasks to the method increases the performance, and the network can be thought as "reasoning" like a diving judge.

2.2 Commentary generation in sports

To extend action recognition from sports videos, Yu *et al.* propose a method for generating fine-grained video captions from basketball videos. They first describe a method for understanding what is happening in the game. This involves several components, including segmenting the pixels of each frame into one of four classes: background, ball, team 1 or team 2. Motion is modeled by using optical flow and pose estimation techniques, then group relationships are modeled. A narrative is generated by first fusing the action and relationship features across the frames of the video with a two-layer bidirectional LSTM encoder. Sentences and paragraphs are decoded with two additional LSTMs.

Sukhwani proposes a method for finding the sentences that best match the visual feed for tennis matches [10]. A sliding window approach over the frame is used to generate the sentences.

2.3 Multi-modal learning

Quattoni *et al.* have found that including captions in image representations speeds up training [2]. Gupta *et al.* provide motivation for including captions as a form of weak supervision in learning representations for images and videos [3]. To achieve high performance in visual recognition tasks, thousands of labeled images and video are required. This can require a significant amount of time and resources, which may not be possible. In addition, visual cues may be ambiguous. This can include changes in illumination and members of the same class that have different appearances, such as a class for dogs.

Gupta *et al.* use the captions of images and videos as an additional view of the data, in additional to the visual view [3]. The two views are conditionally independent, but complimentary, and each view is sufficient. They use both views in co-training, which is an approach that requires two distinct views of the data. The method trains two classifiers, one for each view, and perform inference by selecting the result from the classifier with the higher confidence.

Convolutional neural networks (CNNs) have emerged as the state-of-the-art method for representing visual information [11]. Tran *et al.* introduce the C3D architecture, which can represent the visual and temporal information from videos. This 3D CNN performs convolution across the frames and time to efficiently produce an output volume.

In the literature, there is a lack of semi-supervised methods that deal with hockey broadcast footage and incorporate commentary. This paper attempts to address this.

3 Methodology

ChirpNet is a method for classifying events in hockey game with a multi-task learning approach. The code for this project is available online[1].

3.1 Dataset

Annotations of event data were obtained from one NHL game. The annotations are at a one second resolution and were provided by Stathletes Inc., a company that provides annotated games for clients in the NHL. Annotations were generated for ten event types: switch, advance, faceoff, play make, play receive, whistle, shot, shot block, hit, penalty, and ricochet. The distribution of the events for this game are shown in Fig. 2.

The selected game occurred on January 3, 2019 between the Toronto Maple Leafs and Minnesota Wild. The video of the whole game includes commentary from the broadcast network. It also has replays of certain plays to add to the understanding of the game.

The game is divided into short video clips that each contain ten events. The average clip length is 22.6 seconds, with a minimum clip length of 4 seconds and a maximum clip length

[1] `https://github.com/pascalewalters/multitask-action-recognition`

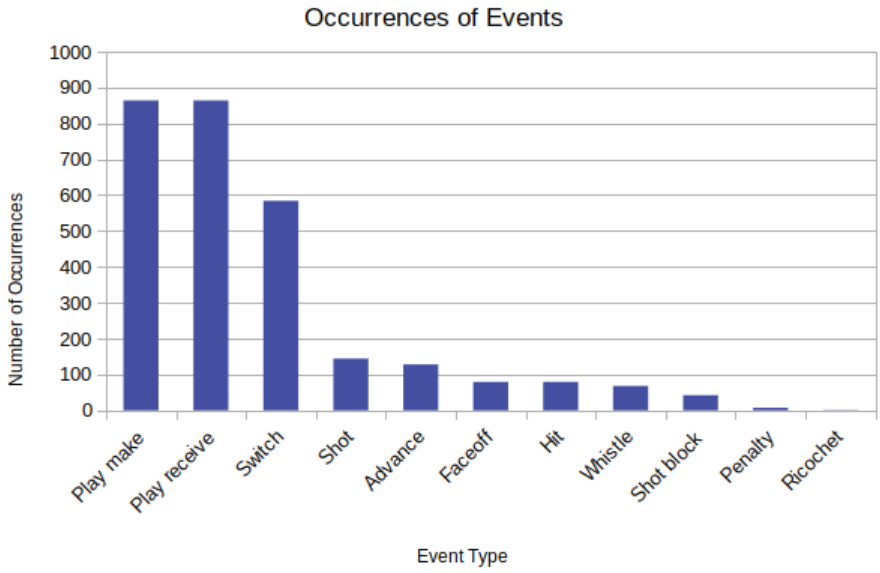


Figure 2: Distribution of event types in the NHL game dataset.

of 181 seconds. Each clip then had its audio transcribed using AWS Transcription service[2]. A custom vocabulary that contained the names of all players on the rosters of both teams. This was because the players' names were frequently mistranscribed.

The dataset was then divided into train and test splits, with the train split representing approximately 70% of the video clips. This gave a train split of 198 clips and a test set of 86 clips. The clips were shuffled before splitting to ensure that there was as much variety in possible between the two splits.

The original broadcast videos were provided at 30 fps. They were downsampled to 6 fps due to storage limitations.

For each clip, the number of each type of event occurring is predicted with a classification task. Since the events in the dataset happen with a high frequency, it is difficult to assign a ground truth start and end time to each event [5]. In addition, events in hockey occur at a very short time scale. In order to capture a reasonable length of video clip for commentary transcription, clips that each contain ten events were generated.

Example frames from two clips, as well as the ground truth event labels and transcribed captions are shown in Table 1.

3.2 Network architecture

The architecture for ChirpNet is shown in Fig. 3. The network takes as input a series of frames from a clip and the transcribed commentary. The input frames are first grouped into series of 16 frames and features are extracted with a C3D network [11] pretrained on the Sports-1M

[2] `https://aws.amazon.com/transcribe/`

	**Events:** play make: 2, play receive: 3, switch: 1, shot: 1, faceoff: 2, whistle: 1 **Commentary:** Hands it over top. Quick, Quick goal here for Toronto throws a tour, and Zucker nearly got there before Hutchinson could cover. Let's watch it again. Just seven seconds after the opening face off. Well, it was a quick little bank off. It looked like it went off of suitor scape. Came back to martyr who cut back around. Spurgeon gets this backhand shot off, and Devon Dupnik did not look like he had stepped out to the top of his crease. There went for the butterfly, and it was a well placed shot. And he's a big goaltender, but probably needed to be out another foot. There. Toe. Get a piece of it. Great individual effort by martyr. My goodness.
	**Events:** play make: 3, play receive: 2, switch: 4, advance: 1 **Commentary:** trying to throw it towards the air trying for

Table 1: Example clips from the NHL dataset. Each clip has counts for all of the events, as well as a transcription of the commentary.

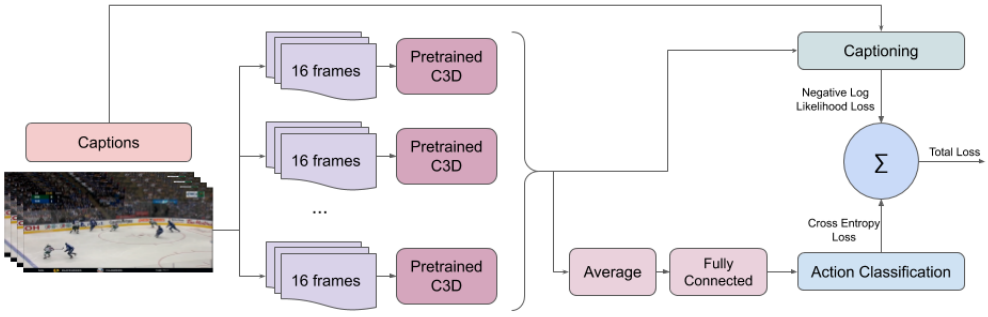


Figure 3: Architecture for ChirpNet.

video classification dataset [12]. The captions and the C3D features are concatenated and encoded and decoded with gated recurrent units [13].

For action classification, the average of the C3D features is taken and they are then passed into a fully connected layer then a ReLU layer. Finally, a fully connected layer for each of the event types, each with ten output nodes to represent the possible number of class occurrences. Inference is performed by taking a softmax of these values.

The overall loss function is a weighted sum of the loss functions for the caption and action classification loss functions. For the captions, a negative log likelihood loss function is used (Eq. 1). N is the number of samples and sl is the sentence length [8].

$$\mathscr{L}_{\text{caption}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{sl}\ln(x^{cap}y^{cap}) \tag{1}$$

A cross entropy loss is used for the action classification (Eq. 2), where $n_{cl} = 11$ is the number of classes and $n_{val} = 10$ is the number of values that each class can take.

$$\mathscr{L}_{\text{action classification}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_{cl}}\sum_{k=1}^{n_{val}} = y_{i,k}^{j}\log x_{i,k}^{j} \tag{2}$$

The total loss function for training shown in Eq. 3. The factor of 0.01 is added to make the two loss values on the same order of magnitude.

$$\mathscr{L} = \mathscr{L}_{\text{action classification}} + 0.01\mathscr{L}_{\text{caption}} \tag{3}$$

3.3 Experiments

The performance of ChirpNet was evaluated on the train/test split of the dataset as described in this report. In addition, the performance of ChirpNet when only visual input is provided was also investigated.

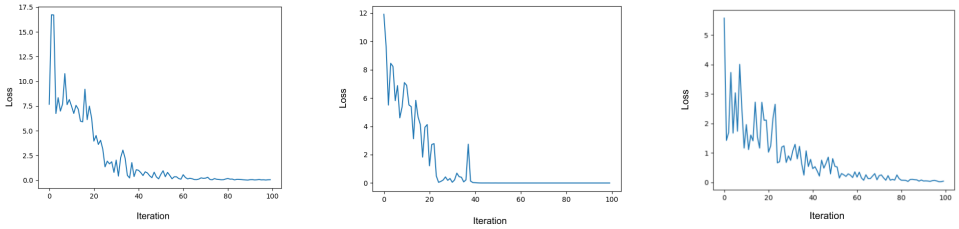
As seen in Fig. 2, there is a significant amount of class imbalance. Only one ricochet occurs in the whole game, while almost 900 play make and play receive events occur. In an additional experiment, the five least frequent event types were ignored (hit, whistle, shot block, penalty, ricochet) during testing and training. Furthermore, an inverse class frequency weighted loss was used for the remaining classes.

Training for all experiments was performed for 100 epochs with a GeForce RTX 2070 graphics card.

4 Results

Fig. 4 shows the loss functions of the three methods during training over the course of 100 epochs.

The average training classification accuracies for all the methods are reported in Table 2. ChirpNet uses visual and linguistic input to perform event classification. The second



(a) ChirpNet (b) Only visual input (c) Weighted loss

Figure 4: Loss during training of each of the three methods. Training was performed for 100 iterations.

Method	ChirpNet	Only Visual Input	Weighted Loss
Overall accuracy	66.39	68.39	-
Top class accuracy	49.61	52.32	46.71

Table 2: Accuracy of each method averaged across all classes. The overall accuracy refers to the classification task with all classes present. The top class accuracy is the average accuracy of the classification method on the six most prevalent classes. Refer to Fig. 5 for the classes.

method has the same architecture as ChirpNet, however it only uses the visual input. The weighted loss method weighs the cross entropy loss function with inverse class frequency and only considers the six most prevalent classes during testing and training. Fig. 5 shows the performance of the three methods for each class.

5 Discussion

Unfortunately, I was unable to achieve higher performance with the multitask learning of ChirpNet than using the visual features alone. This may have occurred for several reasons. First, the dataset was quite small, which means that the training data may not be a good representation of the distributions in the whole dataset. In Parmar *et al.*'s multitask learning action quality assessment dataset of diving videos, there are 1412 clips [8]. The NHL dataset used with ChirpNet only has a total of 284 clips.

Additionally, poor quality annotations and transcriptions may be interfering with the performance of ChirpNet. The event annotations are provided at a second level, rather that at a frame level which would be more precise. Giancola *et al.* also note that selecting beginning time stamps may be ambiguous. The hockey events have some subtleties that may be interpreted differently between annotators. For example, the difference between the switch, advance, play make, and ricochet events is not necessarily clear, as they all refers to events in which the puck is moved down the ice.

Transcription of the audio broadcast commentary was obtained with the AWS Transcription service. The service recommends using a custom vocabulary for words that are domain

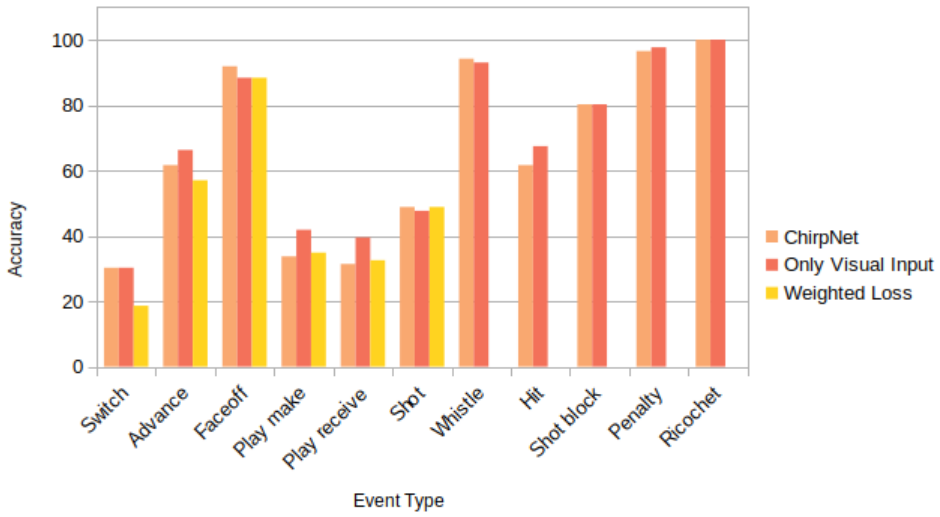


Figure 5: Classification accuracy of ChirpNet, ChirpNet with only visual input, and ChirpNet with weighted loss. The weighted loss method only includes the top six most prevalent classes.

specific, such as the names of players and teams. Despite using a custom vocabulary based on the roster information of the two teams, the transcription service frequently misspelled player names. For example, the first clip in Table 1 refers to Mitch Marner of the Toronto Maple Leafs as "martyr" and Ryan Suter of the Minnesota Wild as "suitor" in the commentary. These transcriptions may also cause the network to lose meaning of the sentences.

ChirpNet assumes that the events are independent, which is not necessarily true. For example, the play make and play receive events are mutually dependent. Information is therefore lost in not including the relationships between event types.

The distribution of the events in this dataset, which includes only one game, is shown in Fig. 2. It is also unknown if this distribution can be generalized to other games. It is expected that the distribution cannot, since there are some games that high-scoring and others that have many penalties. Further research in this area with more included in the testing and training datasets is required.

In their methods for classifying events in hockey and basketball games, respectively, Tora *et al.* and Yu *et al.* use bounding boxes of the players to achieve higher classification accuracy [4, 9]. I tried to obtain player bounding boxes with YOLOv3 pretrained on the COCO dataset [14, 15]. Unfortunately, the network was not successful in detecting hockey players. I suspect that this is due to the fact that they are wearing equipment and therefore do not resemble the humans in the training set.

I also attempted to predict commentary with ChirpNet, but I obtained BLEU scores that were very close to zero [16]. This is likely due to the use of a small training dataset that may contain inaccurate transcriptions. Pretraining of the language model on another

text corpus, such as the commentary from SoccerNet [5]. Due to the prevalence of domain specific terminology in broadcasts of hockey games, fine-tuning would be required on the hockey dataset.

5.1 Future research

Performance of the network could be improved by using a dataset that uses more than one game. Commentary provided by varying networks and varying commentators within a network could have differences that ChirpNet cannot handle well. There are also visual differences between arenas in which NHL games are played, in addition to the teams' home and away jerseys. To implement this improvement, an infrastructure with a larger capacity for storage would be required, as the computer used for ChirpNet's training and testing did not have enough memory to use more than one game.

Another interesting avenue of research would be to obtain high quality player detections and include them in the input [4, 9]. This would require an annotated dataset and a trained object detection network, such as Faster R-CNN [17].

Finally, an implementation of the word sub-sequence kernel classifier as described by Gupta *et al.* could improve the performance of ChirpNet [6]. By ignoring commentary that does not directly refer to the actions that are occurring visually, there could be a better representation in the latent space of the combined visual and linguistic features. Again, this would require an annotated dataset and a trained classifier, such as an SVM string classifier [6].

6 Conclusion

ChirpNet is a network for classifying events from broadcast video of hockey games. It uses a multi-task learning approach to incorporate both visual and linguistic information provided by commentary. Despite a small dataset, the method achieves good performance on event classification. There are opportunities for further research by expanding the dataset with more games and player bounding boxes. ChirpNet could be used as a semi-supervised learning method for classifying hockey events from broadcast video without requiring manual annotation.

References

[1] V. Viswanathan, "Why ai is the next frontier in sports fan engagement and revenue," Aug 2019.

[2] A. Quattoni, M. Collins, and T. Darrell, "Learning visual representations using images with captions," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[3] S. Gupta, J. Kim, K. Grauman, and R. Mooney, "Watch, listen & learn: Co-training on captioned images and videos," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 457–472, Springer, 2008.

[4] M. R. Tora, J. Chen, and J. J. Little, "Classification of puck possession events in ice hockey," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 147–154, IEEE, 2017.

[5] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1711–1721, 2018.

[6] S. Gupta and R. J. Mooney, "Using closed captions to train activity recognizers that improve video retrieval," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 30–37, IEEE, 2009.

[7] R. A. Sharma, K. P. Sankar, and C. Jawahar, "Fine-grain annotation of cricket videos," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 421–425, IEEE, 2015.

[8] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 304–313, 2019.

[9] H. Yu, S. Cheng, B. Ni, M. Wang, J. Zhang, and X. Yang, "Fine-grained video captioning for sports narrative," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6006–6015, 2018.

[10] M. K. Sukhwani, "Understanding and describing tennis videos," *PhD thesis*, 2016.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[13] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[14] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015.