Opleiding Informatica

Using image captioning model for

event and object recognition in football images

Max Visser

Supervisors:
Dr. Gijs Wijnholds
Dr. Arno Knobbe

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

30/06/2024

Abstract

Football is the most-watched sport in the world with hundreds of games being played every week. Games are 90 minutes and watching them can be time-consuming, therefore watching summaries or reading commentaries are great alternatives for most people. Creating summaries and commentaries takes a long time, so creating them automatically would save a lot of time, but to do so, event and object recognition is crucial. There are a lot of models that can be used for event and object recognition. Finding the best one is important to make the best summaries possible. In this thesis, we compare the VvgNet-13 and ResNet-18 models to the Inception V3 model based on a few tasks and test the LLaVa model for generating commentary using different prompts. We found that the Inception v3 model performed better than the VvgNet-13 and ResNet-18 models in certain tasks when the class imbalance in the dataset was limited. Also, the Inception v3 model correctly labeled all events and objects of the images with an accuracy of 80.32%. The LLaVa model was not fine-tuned for football images, so it only correctly labeled 10.1% of the images for the best prompt but showed more promise for correctly describing an image. We argue this shows that the Inception v3 model is a great model to use for event and object detection for football images, while the LLaVa model still needs fine-tuning.

Contents

1 Introduction **1**
1.1 Thesis overview 2

2 Related Work **3**
2.1 Event and object detection 3
2.1.1 Inception v3 3
2.1.2 VvgNet-13 and ResNet-18 3
2.1.3 IAUFD database 4
2.2 Automatically generating commentary 5
2.2.1 LLaVa model 5

3 Methodology **6**
3.1 Classification 6
3.1.1 Two-class classification 7
3.1.2 Seven-class classification 7
3.1.3 Scenario classification 7
3.1.4 Multi-label classification 7
3.2 Commentary 8

4 Results **9**
4.1 Classification 9
4.1.1 Two-class classification 9
4.1.2 Seven-class classification 10
4.1.3 Scenario classification 12

4.1.4 Multi-label classification . . . 12
4.2 Commentary . . . 12
4.2.1 Prompt 1 . . . 12
4.2.2 Prompt 2 . . . 13
4.2.3 Prompt 3 . . . 13

5 Discussion **16**

6 Conclusion **17**

References **18**

A Observation measures two-class classifiction **19**

1 Introduction

Football is the most popular sport in the world. There are hundreds of matches each week, making it impossible to watch them all. Watching summaries back is a great alternative to watching the full 90 minutes. Making these summaries is a process that takes a long time since you need to watch the full 90 minutes and determine what the most important moments are that need to be in the summary. Automating this process can save a lot of time. A second alternative to watching the full 90 minutes is reading the commentary of a match, but creating commentary is a process that takes a long time too, and could be automated. Event and object detection is a crucial part of the automatization process for both alternatives, to decide when a moment is important enough for the summary or commentary, like when a player receives a penalty or a red card.

Most of the time, event and object recognition is done by convolutional neural networks like in Zabganeh et al. [ZJL22] where they used the VvgNet-13 and ResNet-18 models to identify events and objects in football images. The problem is there are a lot of other models and finding the best model is crucial for preventing mistakes in the summary. A convolutional neural network that hasn't been tested on the identification of events and objects in football images is the Inception v3 model. In this thesis, we compare the Inception v3 model with the VvgNet-13 and ResNet-18 models.
Another important aspect is the generation of commentary. At the moment there is only a study regarding the generation of commentary based on football videos [MCG+23]. In this study, they used a spotting model in combination with a captioning model. Since we use images instead of videos we don't need the spotting model and only use an image captioning model.

The research question this thesis tries to answer is:
To what extent can image captioning models be used for creating summaries and commentary automatically based on football images?
To answer this research question, we divide it up into two subquestions. The first subquestion focuses on the first part of an image captioning model where the event and object recognition happens. The first subquestion is as follows:
When comparing the Inception v3 model to the VvgNet-13 and ResNet-18 models, which is better at recognizing important events and objects in football images?
The second subquestion focuses on the whole image captioning model, to test if it can create correct commentary. The second subquestion is as follows:
To what extent can an image captioning model be used to make automated commentary based on football images?

In the first part of the thesis, we focus on the first subquestion. We compare the Inception v3 model to the VvgNet-13 and ResNet-18 models based on four different tests. These tests are the two-class-, seven-class-, scenario-, and multi-label classification. Since the Inception v3 model has a lot of parameters, we decided to freeze most of the model, so we only need to train the classification part of the model. After training the model, we calculate the results and compare them based on different measures to see which model performed better on a certain test.
For the second part of the thesis, we try to answer the second subquestion. We use the LLaVa model to generate commentary based on 3 different prompts, which test the model in different ways. The first prompt tests if the model can label the events and objects. The second prompt is to

test if it can generate commentary in combination with naming the events and objects. The third prompt is to see if the model can generate correct commentary without restrictions. We use these three prompts on a dataset of 99 images where every event and object occurs at least 11 times. We then analyze the first two prompts and compare the results with the results of the Inception v3 model to see which performed better. For the final prompt, we analyze if the model describes the image correct, only partly correct, or wrong and where the model has difficulties.

We found that the Inception v3 model performed better than the ResNet-13 and VvgNet-18 models on some tests when the class imbalance of a dataset was limited. The Inception v3 model correctly labeled the events and objects of the images with an accuracy of 80.32%, which is pretty high but isn't on the level yet to automatically create summaries unsupervised. The LLaVa model isn't fine-tuned on football images, so it has difficulties correctly labeling events and objects only achieving an accuracy of 10.1% on the best-performing prompt. The creation of commentary went a bit better, describing a total of 44.4% of the images correct and a total of 27.3% of the images partly correct. The model had a few difficulties causing the commentary to not be on the same level as a human commentator.

In this paper, we first make a comparison between the Inception v3 model and the VvgNet-13 and ResNet-18 models. We then make a comparison between the Inception v3 model and the LLaVa model based on event and object detection. at last, we make an analysis of the LLaVa model based on the ability to create commentary.

1.1 Thesis overview

This chapter contains the introduction; Section 2 contains an overview of the related work and a description of the experiments; Section 3 contains the methodology; Section 4 contains the results of the experiments; Section 5 discusses the results and answers the research question and subquestions; Section 6 concludes the thesis.
This thesis is made for the bachelor thesis at LIACS under the supervision of Dr G.J. Wijnholds and Dr. A.J Knobbe.

2 Related Work

2.1 Event and object detection

As mentioned in the introduction, the first and main problem we deal with in this thesis is event and object detection in sports. Most studies use a convolutional neural network for this. Joshi et al. [JTBB20] got great results using the Inception v3 model to classify different images based on the sport depicted, but this paper did not include football images. When focussing on event and object detection in football images Karimi et al. [KTA21] used the EfficientNetB0 architecture to label images of the SEV database based on the event that was depicted. Zanganeh et al. [ZJL22] created their own database of labeled football images and tested the VvgNet-13 and ResNet-18 models on this database based on 4 different tests. These studies fine-tune the pre-trained convolutional neural networks, which were trained on the ImageNet database, to fit their database. Compared to these methods we also use a convolutional neural network for event and object detection and fine-tune it on our database.

2.1.1 Inception v3

The Inception v3 model was first introduced in Szegedy et al. [SVI+16] and is a convolutional neural network pre-trained on the ImageNet database. The network consists of multiple convolutional layers through which an image is passed. In a convolution layer, a raster goes over every pixel of the image. It then takes the weighted average, which is determined by a filter, of the pixels in the raster and forms a new image that accentuates some features of the image. Which features are accentuated depends on the filter. At the end of the neural network, the classification happens where the model translates the features into actual events and objects. What's different about Inception v3 than other convolutional neural networks is that it consists of multiple inception blocks. In contrast to other models which focus more on getting deeper using more layers, without increasing the training time too much, the inception model focuses on getting wider in one layer. By this, we mean that in one layer multiple convolutions are done after which they are concatenated. This allows the model to choose the most useful convolution. This is done multiple times after which the classification happens.

2.1.2 VvgNet-13 and ResNet-18

The VvgNet-13 is another convolutional neural network, first introduced in Simonyan et al. [SZ14]. The structure consists of a linear path of convolutional layers translating the image into features, after which they are classified into events and objects. The 13 corresponds to the number of convolutional and classifying layers in the model. In this case 10 convolutional and 3 classifying layers. The ResNet-18 model introduced in He et al. [HZRS16] builds further upon the VvgNet model. The problem with the VvgNet model is that when the total amount of layers increases, the vanishing gradient problem happens. This means that when back-propagating a bigger model, the gradient that updates the weights gets smaller and smaller when it gets to the deeper layers, causing the weights deeper in the model to update with a small margin. To fix this, the ResNet model implemented skip connections which essentially resets the gradient back to the initial weight after a few layers. This made sure the model could increase in size without increasing the training time exponentially.

2.1.3 IAUFD database

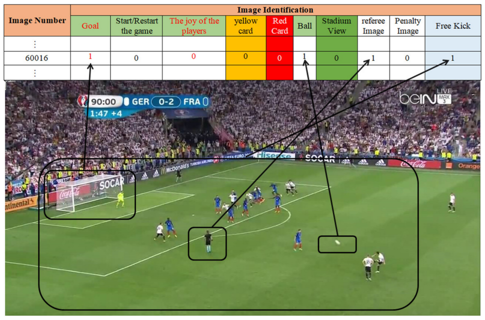


Figure 1: An example of an image in the IAUFD database.

We use the mentioned models on the database created in Zanganch et al [ZJL22]. In this thesis, we refer to this database as the IAUFD database and to the paper as the IAUFD paper. The IAUFD database consists of 100000 images containing 10 different events and objects. These events and objects are the ball, the goal, the referee, a kick off, a free kick, a penalty, a celebration, a view of the stadium, a yellow card, and a red card. Since the red and yellow cards are similar, we combine them into one. Every image in the IAUFD database is labeled with all the events and objects depicted in the image. An event or object is labeled with a 1 if it is in the image and a 0 if it's not. An example is shown in figure 1. Alongside this database is the IAUFD paper. In this paper, they explain what the database consists of and test the VvgNet-13 and ResNet-18 models on it. They do this based on 4 different tests.

Two-class classification

The first test is the two-class classification. In this test, we only look at one event or object at a time. The purpose of this test is to decide if the image given as input has the event or object in it or not. How this test works is visualized in figure 2.

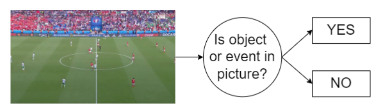


Figure 2: Working of the two-class classification.

Seven-class classification

The second test is the seven-class classification. In this test, we only use images that depict one event or object. With this image as input, the model has to decide which event or object occurs in the image. The model can choose between the goal, the referee, a free kick, a celebration, a stadium view, a kick-off, and the cards. The ball and penalty kick are not included, because the ball is a common factor in the images which means that only a few images have one event or object and there isn't an image in which the penalty kick is the only event or object in that image since a penalty kick always co-occurres with the goal and the ball. How this test works is visualized in figure 3.

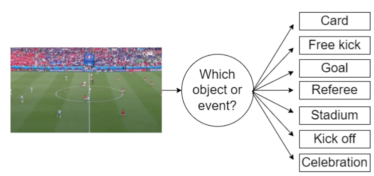


Figure 3: Working of the seven-class classification.

Scenario classification

The third test is the scenario classification. The purpose of this test is for the model to predict if a specific scenario, consisting of three events or objects, occurs in the image given as input. For this test, we use a scenario that consists of the ball, the goal, and a free kick. How this test works is visualized in figure 4.

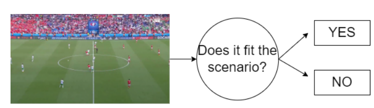


Figure 4: Working of the scenario classification.

Multi-label classification

The final test is the multi-label classification. In this test, the purpose is to identify every event and object in the image. This test consists of two parts. The first part is generating the probabilities of every event and object occurring in the image. The second part is determining if the probability is high enough, and if that's the case labeling the image with the event or object. How this test works is visualized in figure 5.

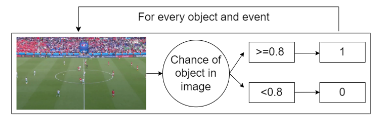


Figure 5: Working of the multi-label classification.

2.2 Automatically generating commentary

The second problem is automatically generating commentary. Most studies do this using video games like Zheng et al.[ZK10] where they use data from the game championship manager to generate commentary. The positive of doing this is that all the data is available, like the precise position of the players and the ball, while this isn't the case in normal broadcasts. The paper Mkhallati et al[MCG+23] tried to generate commentary based on broadcasts from the SoccerNet-Caption database consisting of videos of football matches and commentaries with correlating time-stamps. They did this by creating a single-anchored Dense Video Captioning model which consisted of two steps. First, the model determined the features of a few frames to spot if those frames needed commentary. If they required commentary, the extracted features were fed into a captioning model which generated commentary. Compared to this thesis we also use a model that determines the features present in an image and translates them into commentary. In contrast, we use images instead of videos so we don't need the first part of the model.

2.2.1 LLaVa model

The LLaVa model is an image captioning model first proposed by Liu et al[LLWL24]. It consists of a CLIP encoder and vicuna as a large language model. The LLaVa model takes an image and a prompt as input. It then passes the image through the CLIP encoder to get the features of the image. CLIP is a zero-shot classifier, which tries to classify objects that weren't present when training the model. The model then passes these features with the given prompt through the vicuna model and generates a caption on the image based on the prompt.

3 Methodology

3.1 Classification

In the first part of the thesis, we look at the first part of an image captioning model, the convolutional neural network. For this, we use the 4 tests mentioned in the background and train the Inception v3 model on them. We choose to train a model for each of the four tests. For training these models we use a 50/25/25 split for the training, test, and validation set respectively, since the IAUFD paper also used this split. This means each epoch, we train the model with the train set which is 50% of the dataset. After each epoch, we calculate the results of the current model state on the validation set which is 25% of the dataset. When the model finishes the desired epochs, the final results of the model are calculated using the test set, which is 25% of the dataset, using the model state that got the best result on the validation set. We also used a learning rate of 0.01 and the Adam optimizer.

Since the inception v3 model has almost 22 million parameters, it takes a long time for the model to run 1 epoch. Because of this, we decided to freeze part of the model and only train a small portion of the parameters. Since the Inception v3 model is trained on the ImageNet database we assumed the first part of the model (the feature extraction part) was already trained well to extract features of an image. For that reason, we decided to freeze that part of the model, so we only need to train the classification part.

To compare the different models to each other we use 4 different measures. Recall, Precision, F1 score, and Accuracy. Per test, it is decided which measure(s) is used to compare them based on the context of the test and the length of the database that is used.
We calculate these measures the following way:

$$Recall = \frac{TP}{TP + FP} \tag{1}$$

$$Precision = \frac{TP}{TP + FN} \tag{2}$$

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

3.1.1 Two-class classification

The first test is the two-class classification. We decided to train a model for every event and object for this test and train them for 4 epochs. Since some of the events and objects were depicted substantially more in the images than others, we decided to use a reduced dataset. In this dataset, we only use the first 10000 images of every event or object. It may be the case that after a specific event or object has reached the cap of 10000, additional images with that event or object got added because they occurred in a picture with another event or object that hadn't reached the cap yet. Despite this, the database was more in balance than before the reduction. This reduced dataset consisted of 39821 images.
To compare the results of this test of the Inception v3 model with the other two models we use the 4 different measures 1, 2, 3, 4, but ultimately compare the models using the F1 score 3.

3.1.2 Seven-class classification

The second test is the seven-class classification. In this test, we also used a reduced database consisting of images that only contain one event or object. This database consisted of 38090 images. Because we only need to train one model for this test we decided to train the model for 10 epochs. To compare the results of the models we calculate the accuracy4. Since the test isn't a binary classification, we need to alter the formula. For TP+TN we use the number of times the model correctly identified the event or object and for TP+FP+TN+FN we use the number of times the event or object occurred. We do this for every event and object.

3.1.3 Scenario classification

The third test is the scenario classification. To avoid class imbalance we decided to use a database consisting of all the images that fit the scenario description (Ball, goal and free kick) and the same amount of images that don't fit the scenario. This way the class distribution was 50/50. This left us with a database consisting of 2958 images. Since the database is quite small, we trained the model for 100 epochs. We compare the results of the models based on the accuracy 4.

3.1.4 Multi-label classification

The last test is the multi-class classification. In this test, we use the same reduced database used in the first test. Because of the complex nature of this test, we decided to train the model for 100 epochs. Unfortunately, the IAUFD database didn't report any findings, so we can't compare the different models. Instead, we compare the percentage of the fully correct labeled images of this test with the percentage of the fully correct labeled images of the first two prompts of the image captioning model: LLaVa.

3.2 Commentary

For the second part of the thesis, we look at the image captioning model as a whole. For this, we use the LLaVa model. For this model, we used a total of 3 different prompts.

"If you can only choose between the following events or objects, which are depicted in the image? You can choose between a ball, the goal, the referee, a free kick, a penalty kick, a kick off, a card, a celebration or a view of the stadium."

Figure 6: First prompt to test if the image captioning model is capable of recognizing different events and objects.

"Mimic a soccer commentator and give a commentary based on the football image. In this commentary, mention the following events and objects, if they are in the picture: a Ball, the Goal, the Referee, a free kick, a penalty kick, a kick off, a card, a celebration or a view of the stadium."

Figure 7: The second prompt to test if the model is capable of generating commentary with those events and objects.

"Mimic a soccer commentator and give a commentary based on the football image."

Figure 8: The third prompt to test if the model is capable of creating commentary that represents what is happening in the image.

We used these three prompts on a total of 99 images. To make sure every event and object was in the images, per event or object we chose 11 random images which contained that event or object. This way we could test the model's capability to recognize every event or object.

For the first prompt 6, to determine the model's capability to recognize the events and objects we analyze the output to see if it mentions a specific event or object. We then compare those findings with the data from the IAUFD database and calculate for every event and object the accuracy 4, F1 score 3, and in what percentage of the images the event or object was mentioned. We then compare these results with the results of the Inception v3 model.

For the second prompt 7, we do the same as for the first prompt. Since we ask the model to do the same as for the first prompt but in the form of commentary we predict that these results are worse than those of the first prompt.

For the final prompt 8, we go over all the prompts and determine if the commentary reflects what's going on in the image. We choose between 3 different classifications per commentary. They can either be fully correct, partly correct, or wrong. A commentary is fully correct when what is described in the commentary happens. It is partly correct if one aspect, like the position on the field or which direction it is going, of the commentary is mixed up, and it is wrong if more than one aspect or one big aspect, like when it says there is a kick off instead of a free kick, is wrong. Finally, we go over some difficulties the model has and give some corresponding examples.

4 Results

4.1 Classification

4.1.1 Two-class classification

The Inception v3 model outperformed the other models on most observation measures. The tables with the different observation measures of all the models are shown in appendix A. We noticed that the accuracy of some events and objects was really high because of the class imbalance. As shown in table 1 if the occurrence rate of an event or object is lower, most of the time the accuracy is higher. This is because when an event or object isn't depicted in a lot of images, the model is well-trained to detect when the event or object isn't in the image so it gets that correct most of the time. This causes the accuracy to skyrocket, in some cases to nearly 100%.

Event or object	*Occurrence rate*	*Accuracy*
Ball	67.92%	81.58%
Goal	54.60%	94.41%
Referee	54.12%	85.05%
Stadium view	14.72%	96.05%
Free kick	12.07%	96.39%
Celebration	3.71%	98.69%
Kick off	2.32%	99.65%
Cards	2.21%	99.52%
Penalty kick	1.75%	99.64%

Table 1: The occurrence rate vs accuracy per event or object.

Since the accuracy in this case isn't a fair measurement of comparison, we decided to compare the different models based on the F1 score instead of accuracy. Table 2 shows how the different models compare to each other based on the F1 score.

Event or object	*Inception v3*	*VvgNet-13*	*ResNet-18*
Ball	**86.76%**	72.53%	71.26%
Cards	**86.61%**	85.11%	75.06%
Free kick	83.22%	**85.12%**	66.60%
Goal	**94.94%**	90.19%	83.31%
Penalty kick	**88.60%**	80.43%	84.35%
Referee	**86.97%**	72.20%	55.81%
Stadium view	85.77%	86.66%	**89.07%**
Kick off	**92.47%**	87.04%	81.00%
Celebration	79.81%	**81.16%**	73.81%

Table 2: Results of the two-class classification test per model based on F1 score.

The Inception v3 model outperformed the other 2 models on 6 of the 9 events or objects.

Inception v3 was outperformed by VvgNet-13 twice, on the free kick and the celebration, and was outperformed by ResNet only on stadium view.

4.1.2 Seven-class classification

Event or object	*Inception v3*	*VvgNet-13*	*ResNet-18*
Cards	**91.67%**	63.75%	30.00%
Free kick	**85.00%**	42.00%	28.00%
Goal	40.48%	**90.90%**	86.50%
Referee	12.83%	84.00%	**89.50%**
Stadium view	**87.00%**	69.33%	31.33%
Kick off	53.13%	**56.66%**	10.00%
Celebration	80.14%	**85.33%**	37.66%

Table 3: Results of the seven-class classification test per model based on accuracy.

Table 3 shows the results of the model on the seven-class classification. It performed better than the other two models on 3 of the 7 events and objects. The model got unexpected results, performing great on the events and objects where the VvgNet and ResNet performed worse and performed terribly on the events and objects where the other models performed great. In contrast to the two-class classification test, the classes which have a higher occurrence rate performed worse. The referee, which has the highest occurrence rate, performed the worst with an accuracy of only 12.83%.

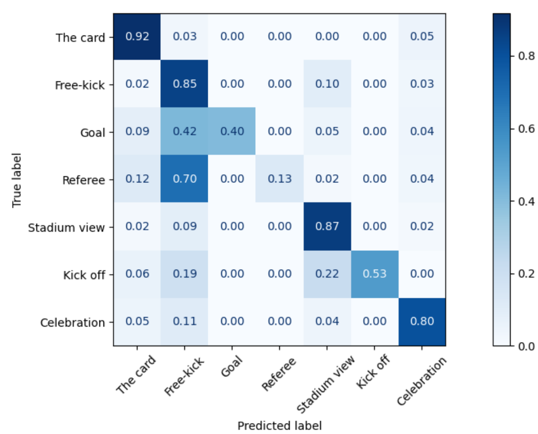


Figure 9: The confusion matrix of the seven-class classification of Inception v3.

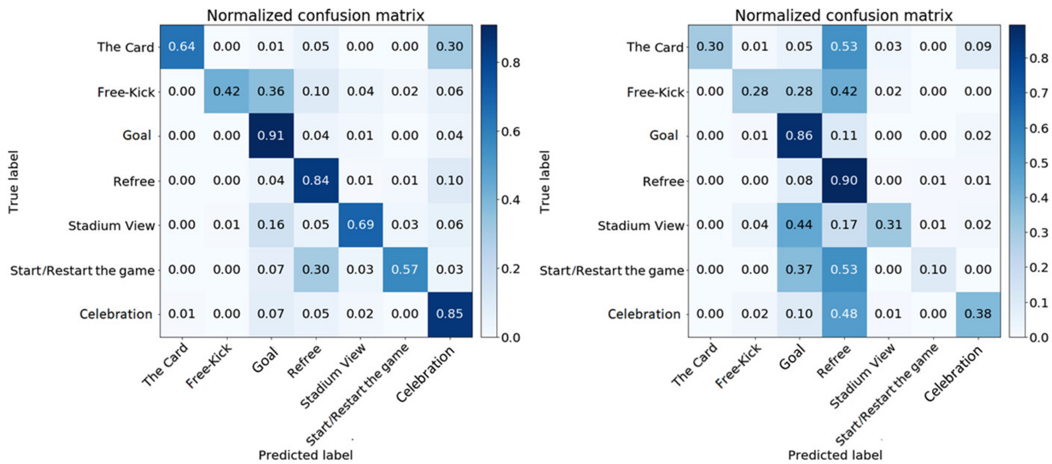


Figure 10: The confusion matrices of the seven-class classification test of VvgNet-13(left) and ResNet-18(right).

Figure 9 displays the confusion matrix that belongs to the results of the seven-class classification of the Inception v3, so we can see why the goal and referee have low accuracies. The model frequently predicted that the image was of a free kick when in reality it was an image of the referee or a goal. Figure 10 shows the confusion matrices of the other models. The ResNet-18 model, in contrast to the Inception v3 model, frequently predicted it was a goal or a referee when in reality it was a different event or object.

We assumed Inception v3 made this mistake a lot because of the class imbalance in the dataset. To test this we also trained the model using a more balanced dataset where the cap of each event and object was at a total of 2000 images. The results of this test are shown in table 4. The results are much more balanced with the lowest accuracy being 71.34% compared to the 12.83% in the normal dataset. Also, the model trained on the reduced dataset performed better than the model trained on the normal dataset regarding 4 of the 7 events or objects, despite being trained with less data.

Event or object	*normal dataset*	*reduced dataset*
Card	**91.67%**	90.80%
Free kick	**85.00%**	83.33%
Goal	40.48%	**71.45%**
Referee	12.83%	**71.34%**
Stadium view	**87.00%**	82.53%
Kick off	53.13%	**88.89%**
Celebration	80.14%	**89.74%**

Table 4: Results of the seven-class classification of the reduced vs normal dataset based on accuracy.

4.1.3 Scenario classification

Model	*Accuracy*
Inception v3	**89.18%**
VvgNet-13	78.86%
ResNet-18	62.24%

Table 5: The results of the scenario classification based on accuracy.

In table 5 the results of the scenario classification test are shown. The Inception v3 model outperformed the other two models by quite a lot. We assume the model performed well on this test because there is no class imbalance. Since there is an equal amount of images that fit the scenario, as images that don't fit the scenario the model isn't trained with any bias.

4.1.4 Multi-label classification

The inception v3 model performed quite well, labeling a total of 80.32% of the pictures fully correct. The IAUFD paper didn't report any findings regarding this test. For that reason, we can't compare the different models with each other.

4.2 Commentary

4.2.1 Prompt 1

Observation measurement	*Ball*	*Cards*	*Free-kick*	*Goal*	*Penalty-kick*	*Referee*	*Stadium view*	*Kick-off*	*Celebration*
Mentioned in output	73.74	12.12	16.16	73.74	15.15	72.73	45.45	10.10	20.20
F1 score	81.25	43.48	20.00	67.83	37.04	72.00	39.39	36.36	60.06
Accuracy	75.75	86.87	75.76	62.63	82.83	64.65	59.60	85.86	86.87

Table 6: The results of the first prompt based on F1 score, Accuracy and mentioned rate in percentages.

For the first prompt, we asked the model to give us the events and objects corresponding to the images. The model correctly mentioned all events and objects in 10,1% of the images. The observation measures of all the events and objects are depicted in Table 6.

We noticed that the model mentioned the ball, the goal, and the referee the most in their prompts, mentioning them in over 70% of the output. These events and objects also have the highest F1 score with the ball having the highest with an F1 score of 81.25%, followed by the referee with an F1 score of 72% and the goal with an F1 score of 67.83%. We also noticed that the model was quite good at recognizing if a celebration was going on, achieving the highest accuracy and the highest F1 score after the ball, goal, and referee. Compared to the Inception v3 model, the LLaVa model in combination with the first prompt did substantially worse at correctly identifying events and objects in images.

4.2.2 Prompt 2

Observation measurement	*Ball*	*Cards*	*Free-kick*	*Goal*	*Penalty-kick*	*Referee*	*Stadium view*	*Kick-off*	*Celebration*
Mentioned in output	89.90	6.06	85.86	94.94	53.54	97.98	96.97	17.17	12.12
F1 score	75.0	47.06	24.24	58.82	12.31	70.67	35.90	13.79	8.00
Accuracy	63.64	90.90	24.24	43.43	42.42	55.55	24.24	74.75	76.76

Table 7: The results of the second prompt based on F1 score, Accuracy and mentioned rate in percentages.

For the second prompt, we asked the model to give us a commentary including the events and objects that are depicted in the image. The model mentioned all events and objects correctly for 1.0% of the images. The observation measures of all the events and objects are depicted in Table 7. We noticed the model had a tendency to mention certain events and objects in almost every commentary. The model mentioned the goal, the referee, and the stadium view in more than 90% of the images, and the ball and a free kick were also mentioned more than 85% of the time.

Similar to the first prompt the model was best at recognizing whether there was a ball with an F1 score of 75%, followed by the referee with an F1 score of 70.67 and the goal with an F1 score of 58.82%. The LLaVa model in combination with the second prompt did not only perform substantially worse than the Inception v3 model but also than the LLaVa model in combination with the first prompt.

4.2.3 Prompt 3

Category	*frequency*
Incorrect	28
Partly correct	27
Fully correct	44

Table 8: The results of the third prompt based on frequency per category.

For the final prompt, we went over the commentaries generated on all of the images and determined if they were correct in regards to what happens in the image. We did this manually so the results of this are subjective. The results of this are shown in table 8. The model got a total of 44 of the 99 commentaries fully correct. We noticed a few things when we went over the commentaries.

Firstly the model had problems with determining the correct position on the field. It would often say that a specific image depicted the middle of the field when it was actually in the 16-meter box or when a corner was taken. We suspect this is because the model sees the half circle on the end of the box or the corner circle and thinks it is the center circle. Figure 11 shows such an example where it thinks the ball is in the middle of the field, while a corner is being taken.

”And now, we have a player from the team in white, bending down to pick up the ball. The crowd is watching intently as the player prepares to make a play. The atmosphere is electric, as the players on the field are fully engaged in the game. The soccer ball is located near the center of the field, and the player is in a ready stance, anticipating the next move. The excitement is palpable, as the players and the audience alike are fully immersed in the action on the field.”

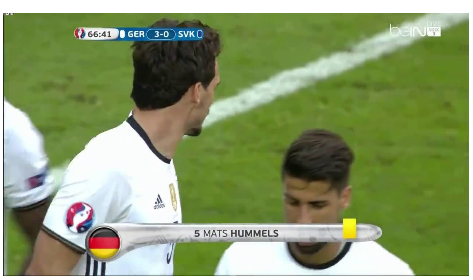
Figure 11: Example of a wrong position.

Another difficulty the model had is determining which direction a play is going. By this we mean that for example if the defending team is passing the ball inside their own half the model thinks there is an attack happening from the opponent or when a keeper is kicking a ball away the model thinks that the keeper is preparing to catch the ball. Figure 12 is an example of this.

”And now, we have a great opportunity for a goal. The soccer players are spread out across the field, with one player in the center, another on the right side, and the third one on the left side. The ball is in the middle of the field, and the players are ready to make a move. The crowd is eagerly watching the game, and the atmosphere is electric. The players are focused and determined, and we can expect an exciting moment in this match.”

Figure 12: Example of the wrong direction.

The model also wasn’t able to spot a card when it was not actually the referee showing a player a card, but a pop-up showing that a player just received a card. It would ignore the pop-up and just describe what happened in the background. Figure 13 shows this.



”In the image, we can see two soccer players on the field, with one player being taller than the other. The taller player appears to be engaged in a conversation with the shorter player. The field is marked with a white line, indicating the boundaries of the playing area. The players seem to be focused on the game, with one player possibly giving instructions or discussing strategies with his teammate. The scene captures the essence of teamwork and communication in soccer, as players often need to work together to achieve success on the field.”

Figure 13: Example of a missing card detection.

The model would often give really basic commentaries, which would fit every random soccer image and technically still be correct. This is alright if nothing special went on in the image, but it also happened when there was a bit of action that could be described. Figure 14 shows this where the commentary is really basic, despite the image showing a free kick.

"In the image, we see a group of soccer players gathered on the field, possibly discussing their next move or strategizing. The players are spread out across the field, with some closer to the foreground and others further back. The scene captures the essence of teamwork and communication in the sport of soccer. The players' positions and interactions suggest that they are actively engaged in the game, showcasing their skills and coordination. Overall, the image portrays a dynamic and exciting moment in the soccer match."

Figure 14: Example of a basic commentary.

Finally, the model had difficulty determining whether the soccer match was still in progress and the ball was in play or the match still needed to start, already ended or the ball was out of play. This is shown in figure 15 where the model thinks the game still needs to start when it has already begun.

"And now, the teams are lined up, ready to kick off the match. The players are spread out across the field, with some closer to the foreground and others further back. The atmosphere is intense, as the crowd eagerly awaits the start of the game. The referee is in position, preparing to make the call. The teams are poised, anticipating the next move. The tension is palpable as the players are about to engage in a high-stakes competition. The excitement is building, as the crowd eagerly awaits the start of the match."

Figure 15: Example of a situation not in play.

The model was good at telling when a celebration was going on, correctly mentioning it at every image a celebration was happening except one, where it thought the player celebrating was injured instead.

5 Discussion

So first of all, the Inception v3 model got good results on the tests. In the first test, it performed better in 6 out of the 9 events and objects than the VvgNet-13 and ResNet-18 models based on the F1 score. In the second test, the model only outperformed the other models in 3 out of the 7 events and objects. In the third test, the model outperformed both models, getting a substantially higher accuracy. In the fourth test, the model got an accuracy of 80.32%, but there were no results from the other models, so we have nothing to compare to.
The model performed best in the scenario classification where there was no class imbalance. Also, the results of the seven-class classification got better when the class imbalance was reduced. Because of this, we believe the Inception v3 model is vulnerable to a class imbalance in the dataset. In future work, it might be interesting to see if the parameters of the Inception v3 model can be altered so a class imbalance has a lesser effect on the results. To answer the first subquestion, the Inception v3 model performs better than the VvgNet-13 and ResNet-18 models on certain tasks depending on the class imbalance of the dataset on which the model was trained.

In the second part of the thesis, the LLaVa model didn't perform well on the event and object detection, only achieving an accuracy of 10.1% in the first prompt and 1.0% in the second prompt. This is much worse compared to the Inception V3 model with an accuracy of 80.32%. We noticed that the LLaVa model had difficulties detecting the events (Free kick, Kick off, and penalty kick), but was better at recognizing the objects. This is probably because the events are much more difficult to correctly identify since the model needs to look at how players are standing and deduce from that which event, if any, is depicted. This task is too difficult for a model like LLaVa which isn't fine-tuned on football data. In future work, it might be interesting to fine-tune an image captioning model like LLaVa on football images to see if it gets better results. The results of the final prompt were a bit better, getting 44.4% of the photos fully correct and 27.3% of the photos partly correct. Despite this, the model still had difficulties. It would often give the wrong position or direction of the play, it couldn't detect card notifications or when the ball was in play or not. Finally, the model would just give really basic commentary which could be placed under any football image and still be right. Again it would be interesting to fine-tune an image captioning model like LLaVa on football images to see if it gets better results. To answer the second subquestion, an image captioning model, like LLaVa can be used to create commentary, but is very limiting and susceptible to making mistakes. The commentary is far away from the level of actual commentators and still needs work.

When answering the research question: "To what extent can image captioning models be used for creating summaries and commentary automatically based on football images?" we concluded that for event and object detection a fine-tuned model like the Inception v3 model can get pretty high results, but is not perfect yet. It can't be used to automatically make summaries unsupervised, but it can help make the process of creating summaries a bit easier. We also concluded that an image captioning model like LLaVa which isn't fine-tuned on football images is not great for object and event detection, but can generate correct commentary in some cases. Unfortunately, the level of the commentary isn't high and is susceptible to mistakes, so again isn't able to automatically make commentary unsupervised, but can be used for inspiration.

6 Conclusion

In this thesis, we tried to see if image captioning models can be used for the automatic generation of summaries and commentary. We first compared to Inception v3 model to the VvgNet-13 and ResNet-18 models based on 4 different tests. We conclude that the model performs better in some tasks and the results are dependable on the class imbalance. We also found that the Inception v3 model labeled 80.32% of the photos fully correct. For the second part of the thesis, we used the LLaVa model to generate commentary based on 3 different prompts. The first two prompts were focused on event and object detection and labeled respectively 10.1% and 1.0% of the photos fully correct. The third prompt focused on the generation of commentary. The model performed better than the first two prompts but still had a few difficulties with a few aspects. Overall we concluded that image captioning models can't be used for automatically creating summaries and commentaries, but can be used to make the process a bit easier. We mentioned that in future work it might be interesting to fine-tune the LLaVa model on football images, so it performs better at event and object detection and generating correct commentary. We would do this by using or even creating a dataset that consists of commentaries that are focused on the events and objects that are in them. That way when the model is fine-tuned it tends to automatically mention certain events and objects.

References

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JTBB20] Ketan Joshi, Vikas Tripathi, Chitransh Bose, and Chaitanya Bhardwaj. Robust sports image classification using inceptionv3 and neural networks. *Procedia Computer Science*, 167:2374–2381, 2020.

[KTA21] Ali Karimi, Ramin Toosi, and Mohammad Ali Akhaee. Soccer event detection using deep learning. *arXiv e-prints*, pages arXiv–2102, 2021.

[LLWL24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[MCG+23] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5073–5084, 2023.

[SVI+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[ZJL22] Amirhosein Zanganeh, Mahdi Jampour, and Kamran Layeghi. Iaufd: A 100k images dataset for automatic football image/video analysis. *IET Image processing*, 16(12):3133–3142, 2022.

[ZK10] Maliang Zheng and Daniel Kudenko. Automated event recognition for football commentary generation. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, 2(4):67–84, 2010.

A Observation measures two-class classifiction

Observation measurement	*Ball*	*Cards*	*Free-kick*	*Goal*	*Penalty-kick*	*Referee*	*Stadium view*	*Kick-off*	*Celebration*
TP	88.77	80.85	77.87	96.11	80.47	91.52	79.30	93.48	72.39
TN	66.31	99.89	98.80	92.38	99.98	77.31	99.01	99.79	99.67
FP	11.23	19.15	22.13	3.89	19.5	8.48	20.70	6.52	27.61
FN	33.69	0.11	1.20	7.62	0.02	22.69	0.99	0.21	0.33
Recall	**84.83**	**93.25**	**89.36**	**93.80**	**98.55**	**82.84**	**93.38**	**91.49**	**88.93**
Precision	**88.77**	80.85	77.87	**96.11**	80.47	**91.52**	79.30	**93.48**	72.40
F1 score	**86.76**	**86.61**	83.22	**94.94**	**88.60**	**86.97**	85.77	**92.47**	79.81
Accuracy	**81.58**	**99.52**	**96.39**	**94.41**	**99.64**	**85.05**	**96.05**	**99.65**	**98.69**

Table 9: The observation measures of the inception v3 model on the IAUFD database in percentages.

Observation measurement	*Ball*	*Cards*	*Free-kick*	*Goal*	*Penalty-kick*	*Referee*	*Stadium view*	*Kick-off*	*Celebration*
TP	73.96	86.66	86.00	92.00	82.22	72.93	87.97	87.68	82.00
TN	70.02	83.03	83.95	88.00	77.77	70.93	84.96	86.23	79.94
FP	26.04	13.34	14.00	8.00	17.78	27.07	12.03	12.32	18.00
FN	29.98	16.97	16.05	12.00	22.23	29.07	15.04	13.77	20.06
Recall	71.15	83.62	84.27	88.46	78.71	71.50	85.39	86.42	80.34
Precision	73.96	**86.66**	**86.00**	92.00	82.22	72.93	87.97	87.68	**82.00**
F1 score	72.53	85.11	**85.12**	90.19	80.43	72.20	86.66	87.04	**81.16**
Accuracy	71.99	84.84	84.97	90.00	79.99	71.93	86.46	86.95	80.97

Table 10: The observation measures of the VvgNet-13 model on the IAUFD database in percentages.

Observation measurement	*Ball*	*Cards*	*Free-kick*	*Goal*	*Penalty-kick*	*Referee*	*Stadium view*	*Kick-off*	*Celebration*
TP	71.96	75.75	67.91	84.98	85.92	56.93	89.97	81.88	76.10
TN	70.02	73.93	63.99	80.98	82.22	52.93	87.97	79.71	69.91
FP	28.04	24.25	32.09	15.02	14.08	43.07	10.03	18.12	23.90
FN	29.98	26.07	36.01	19.02	17.78	47.07	12.03	20.29	30.09
Recall	70.59	74.39	65.34	81.71	82.85	54.74	88.20	80.14	71.66
Precision	71.96	75.75	67.91	84.98	**85.92**	56.93	**89.97**	81.88	76.10
F1 score	71.26	75.06	66.60	83.31	84.35	55.81	**89.07**	81.00	73.81
Accuracy	70.95	74.82	65.95	82.98	84.07	54.93	88.97	80.79	73.00

Table 11: The observation measures of the ResNet-18 model on the IAUFD database in percentages.