Deep Multiple-Instance Learning for Stable Attribute Classification in Classroom Video

by

Jiani Wang

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

April 27th 2023

APPROVED:

Professor Jacob Richard Whitehill, Thesis Advisor

Professor Neil Heffernan, Reader

Professor Craig Shue, Head of Department

Abstract

In this thesis, we explored how to classify multiple attributes of each person in a classroom video that are stable over short periods of time, such as their gender, role (student vs. teacher), and skin tone. This can benefit the field of automatic classroom analysis by giving teachers better feedback about their teaching and about possible biases they may have towards certain students. We tackled this problem using a deep Multiple-Instance Learning (MIL) method. Our experimental results on a video dataset of real classroom videos suggest that the MIL strategy is useful for classifying the stable attributes of the people in classroom videos and can improve the accuracy especially in the binary classification tasks. In addition, the model MIL_MAX always performances best for all of the tasks among all the models. Finally, data augmentation and data oversampling were helpful in our experiment for solving poor model performance problem due to data imbalance.

Contents

1 Introduction **1**

2 Related works **5**

2.1 Classroom Analysis by Computer Vision 5

2.2 Multiple Instance Learning 6

2.3 Skin Tone Classification 6

2.4 Gender Classification 7

2.5 Role Classification 8

3 Proposed Research **9**

3.1 Research Question 9

3.2 Baseline 10

4 Methodology **11**

4.1 Multiple-Instance Learning 11

4.2 Model 12

4.2.1 NO_MIL 12

4.2.2 MIL_MAX 16

4.2.3 MIL_MAX_yhat 18

4.2.4 MIL_ATT 18

4.3 Model Comparison 19
4.3.1 NO_MIL vs. MIL_MAX_yhat 21
4.3.2 MIL_MAX vs MIL_MAX_yhat 21
4.3.3 MIL_MAX vs MIL_ATT 22

5 Data **23**
5.1 Data 23
5.1.1 Raw Data 23
5.1.2 Frame 24
5.1.3 Bounding Box 25
5.1.4 Track 25
5.1.5 Split Strategy 25
5.2 Dataset 26
5.2.1 Gender Classification 26
5.2.2 Role Classification 27
5.2.3 Skin Tone Classification 29

6 Experiments and Results **32**
6.1 Research Question 1 33
6.1.1 Gender Classification 33
6.1.2 Role Classification 33
6.1.3 Skin Tone Classification 36
6.2 Research Question 2 38
6.2.1 Role Classification 38
6.2.2 Skin Tone Classification 40

7 Conclusion **42**

List of Figures

1.1 Example of a classroom image.
URL: https://www.youtube.com/watch?v=pgk-719mTxM 3
1.2 Demo Frames of a Person at Different Timesteps 4

4.1 NO_MIL Model Structure for Training. 13
4.2 NO_MIL Model Structure for Evaluation. 15
4.3 MIL_MAX Model Structure. 17
4.4 MIL_MAX_yhat Model Structure. 18
4.5 MIL_ATT Model Structure. 20

5.1 A demo of a frame with several bounding boxes. 24

List of Tables

3.1 Accuracy Baseline. 10

4.1 Model Comparison. 20

5.1 Gender Data Distribution 27

5.2 Original Role Data Distribution 28

5.3 Balanced Role Training Data Distribution 29

5.4 Original Skin Tone Data Distribution 30

5.5 Frame Balanced Skin Tone Training Data Distribution. 30

5.6 Track Balanced Skin Tone Training Data Distribution 31

6.1 The Accuracy Results for Gender Classification 33

6.2 Standard Error of Testing 33

6.3 The Results of the Models Trained on Original Data 34

6.4 Standard Error of Testing 34

6.5 The Results of the Models Trained on Balanced Data 35

6.6 Standard Error of Testing 35

6.7 The Results of the Models Trained on Original Data 36

6.8 Standard Error of Testing 36

6.9 The Results of the Models Trained on Track Balanced Data 37

6.10 Standard Error of Testing 37

6.11 The Results of NO_MIL Trained on different datasets 38
6.12 The Results of MIL_MAX Trained on different datasets 38
6.13 The Results of MIL_MAX_yhat Trained on different datasets 39
6.14 The Results of MIL_ATT Trained on different datasets 39
6.15 The Results of NO_MIL Trained on different datasets 40
6.16 The Results of MIL_MAX Trained on different datasets 40
6.17 The Results of MIL_MAX_yhat Trained on different datasets 41
6.18 The Results of MIL_ATT Trained on different datasets 41

Chapter 1

Introduction

Classroom instruction is an important medium to pass on knowledge. It is a necessary approach for students to learn new knowledge and it's also significant for the teachers to get feedback from the class so that they can find out if their teaching approaches are effective or not. Teachers have different requirements for feedback based on different grade levels and subjects. However, there is some general information that is worth providing feedback on. One such aspect is whether teachers are giving almost equal attention to students of different gender and races during the class. Many studies have shown that the amount of attention which teachers give to students has a significant influence on the students academic abilities. According to Lavy's research, primary school teacher's gender biases will effect the students academic achievement during middle and high school [14]. In Copur-Gencturk and Cimpian's paper, they found that some teachers have biases against Black, Hispanic and female students when assessing students' mathematical ability [7]. In Terrier's paper, it's also shown that teachers' gender biases have a high and significant effect on girls' progress relative to boys' in both math and French [18].

In order to help teachers to avoid possible teaching bias, there are many methods

we can try. For instance, we can do it by human like ask pedagogy experts to come to the class and take notes. However, this approach takes a lot of time and is not very accurate. We want a more accurate and convenient way to achieve this goal, therefore, we decide to use computer vision methods which can save time and manpower and have unified standard which makes the results more accurate and believable.

As the first step, our work uses skin tone and perceived gender instead of race and gender since we cannot get people's race and true gender. We can get these information from the classroom videos and then use them to calculate the teacher's attention allocated to students in the classroom, and to measure whether there is too much attention or too little attention to students of a certain race or gender. This thesis discusses the application of skin tone and some other 'stable' (i.e., will not change over the course of a video) attributes classification in the classroom videos.

Our long-term vision is to build a system which take in the classroom videos and output a feedback report which include the information teacher may concern about, for example, how many time the teacher asked the girls/boys to answer questions during a class. In this thesis, we'd like to finish one little step of this whole system, which is improving the accuracy of classifying the skin tone, gender and (teacher vs. student) role of the people in the classroom video.

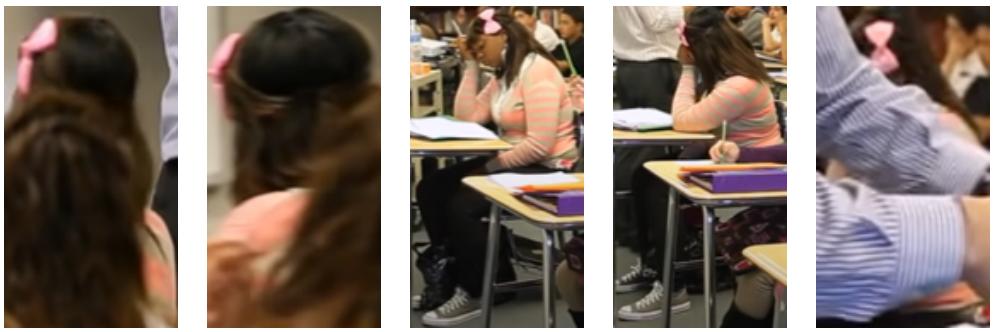
The data in our thesis comes from videos of school classrooms, and its particularity is mainly in the following aspects: First, there are changes in the angle of the camera; Second, limited skin area in the images due to the occlusion (i.e., desks and chairs); Third, People are moving and changing positions and do not always face to the camera. Thus, the environment in the classroom is quite complicated which makes it difficult to perceive skin tone, gender, and role with high accuracy in a single frame and it's also the main challenge of our experiment. The Figure 1.1 shown a

Figure 1.1: Example of a classroom image.
URL: https://www.youtube.com/watch?v=pgk-719mTxM

demo of a complicated classroom with people turned around and occlusion. Another challenge is the Unbalanced dataset for teacher vs. student role classification and skin tone classification.

Consequently, the existing methods employed in previous research does not performance well when applied to our dataset. Therefore, in order to address the aforementioned challenges, we propose the utilization of a novel method known as Multiple Instance Learning (MIL). Our objective is to extract accurate information regarding the skin tone, gender, and (teacher vs. student) role of each individual within a given video. Importantly, these attributes remain constant throughout the entire video duration, thereby enabling us to make only one decision for the whole for each person. Therefore, our focus lies in identifying a specific timeframe, as short as one second, wherein the individual is sufficiently clear for us to extract their pertinent information. MIL represents a suitable approach in accomplishing this objective. By leveraging the capabilities of MIL, we can effectively identify such

critical timeframes. The Figure 1.2 shows a demo of a person at different timesteps. Obviously, the Figure 1.2c shows more of the girl's face than the other images.



(a) Frame1 (b) Frame2 (c) Frame3 (d) Frame4 (e) Frame5

Figure 1.2: Demo Frames of a Person at Different Timesteps

The rest of the thesis is organized in the following order: Chapter 2 provides an overview of previous research and related work in this field. Chapter 3 outlines the proposed research, including the research question and the established baseline. Chapter 4 describes the methodology employed in this study. Subsequently, Chapter 5 presents the experimental setup and the corresponding results. Finally, Chapter 6 concludes this paper by summarizing the findings and discussing their implications.

Chapter 2

Related works

In this chapter, we will introduce the previous research regarding applying computer vision methods to analyze classroom, MIL, skin tone classification, gender classification and role classification respectively.

2.1 Classroom Analysis by Computer Vision

Computer vision methods have been applied to analyse classroom by many researchers. Baker and D'Mello applied three different computer-based learning environment to study students' cognitive-affective states [1]. D'Mello also used AutoTutor, an intelligent rutoring system to explore the reliability of detecting a learning's affect [8]. Monkaresi and Bosch used computer vision techniques to detect engagement while students completed a structured writing activity [16]. Bosch and D'Mello used computer vision and machine learning techniques to detect students' affect during interactions with an educational physics game [4].

2.2 Multiple Instance Learning

MIL is a well-established method applied in the field of machine learning, especially for classification problems in computer vision. Carbonneau, Cheplygina. [5] did a survey regarding the four key characteristics that affect the MIL algorithm, and analyzed and classified which problems are appropriate for using the MIL methods. At the same time, they explained in detail the 'bag' in the MIL method for different types of problems, such as how it should be created and how it should be labeled. Ilse, Tomczak and Welling [11] applied attention based MIL and gated attention based MIL methods for binary classification on some image datasets, such as real-life histopathology datasets and MNIST dataset. Sikka and Dhall [17] provided the framework multiple-segment MIL which incorporates with MIL and a dynamic extension of concept frames. This framework solves the problem that the ground-truth is only on the sequence-level but not on the frame-level. It gives us the inspiration that we can treat the bounding boxes belong to the same track as a 'sequence' in this paper and use these bounding boxes to create a 'bag'. Meanwhile, we have a 'sequence-level' ground-truth that in this bag there must have at least one bounding box which can make the annotator to classify the skin tone, the gender and the role of this person but we do not know which one it is.

2.3 Skin Tone Classification

There are several previous research which also regarding skin tone classification but not based on the classroom videos. For example, Hazirbas and Bitton conducted an experiment in the article Casual Conversations: A dataset for measuring fairness in AI which the data is basically a clear photo of the person's face [10]. In this paper, they define the skin tone into six levels which from light to dark are Type 1 to Type

6.

In this experiment, we will classify the skin tone of the people appeared in the classroom video. And the result of it can be used in some other research later, for example, to verify whether there is a bias against race in the class. In the previous research, both neural networks and some other methods have been applied to classify the skin tone a lot. Jmal et al. [12] used the RGB model to classify skin color into white and black, with an accuracy rate of 87Borza et al. [3] achieved accuracy rates of 86.67% and 91.29% respectively when using the Support Vector Machine (SVM) and CNN models to classify the skin tone of humans' front face images into light, medium and dark. Bevan et al. [2] tried to remove the bias caused by the skin tone in detecting melanoma and proposed an efficient algorithm for automatically labelling the skin tone.

2.4 Gender Classification

In this experiment, we will classify the gender of the people appeared in the classroom video. Since we do not have access to people's actual gender, we instead label each person's perceived sex, e.g., whether a person appears to be female based on the observable cues in the video – while imperfect, this approach can still give useful information for assessing bias. There are many different methods which based on various principles to estimate people's gender [15], such as using neural network trained on a small set of near-frontal face images [9], and use the Webers Local texture Descriptor for gender recognition [19].

2.5 Role Classification

In this experiment, we define the 'role classification' as classifying the person appears in the videos is a teacher or a student and it's based on the age classification. There are many different methods which based on various principles to estimate people's age [15], such as calculating ratios between different measurements of facial features [13], using local features for representing face images [21] and an improved version of relevant component analysis and locally preserving projections [6].

Chapter 3

Proposed Research

This chapter will introduce the research questions of this thesis and show the baseline from the previous research.

As we mentioned above, we have two main challenges in this experiment, the first one is the limited skin area in some of the images and the other is the unbalanced dataset for role and skin tone classification.

For the first challenge, since the attributes in our classification tasks are stable which means they will now change during the whole video, thus, we only need to do one decision for the whole video for each person. So, for each person, as long as we find one frame which this person is clear, it's enough for us to get his/her attributes. MIL is such a method that takes in a track of images and uses the combination of these images to update the parameters of the model.

For the second challenge, we plan to apply data augmentation and data balance to increase the amount of the fewer categories.

3.1 Research Question

In this research, we focus on the following questions:

	Gender	Role	Skin Tone
ResNet18	0.5685	0.8195	0.3765
ResNet50	0.5739	0.8377	0.3916
ResNet101	0.6138	0.8113	0.3291

Table 3.1: Accuracy Baseline.

1. Is MIL helpful for improving the accuracy of the gender, (teacher vs. student) role and skin tone classification?

2. Does data augmentation and data balance help for improving the performance of the model?

3.2 Baseline

Before our experiment in this paper, we applied the ResNet which is frequently used in classification tasks to classify the skin tone, the gender and the role in order to find out the result of one of the state-of-the-art models and the baseline of our result. The data used to get the baseline and also used in the subsequent experiments are the bounding boxes got from videos and frames. We introduce this process in detail in the Section 5.1.3.

The Table 3.1 shows the accuracy of the skin tone classification, gender classification and role classification respectively.

Chapter 4

Methodology

4.1 Multiple-Instance Learning

In this thesis, the most significant methodology is Multiple-Instance Learning (MIL). Here is an introduction of MIL. In MIL, examples are partitioned into 'bags' and each bag will have a new label $\hat{Y}$. Each example x has an associated label y, just like in standard supervised learning, but y is typically unobserved at training time. Instead, the model is trained using bag labels, whereby the label of a bag is determined by applying an aggregation function to the labels of the examples contained within it. For instance, the aggregation function might be the *max* function: if any of the examples in the bag are labeled positive, then the bag is labeled positive; else, the bag is labeled negative. MIL is thus useful when labeling a bag is easier than labeling each example individually.

In our experiment, each person has several images. Although not all of them are clear, at least one of them can let us classify the gender/role/skin tone of him/her. Since the attribute we want to classify is stable and will not change through the whole video, thus, as long as we find at least one clear image, it enough for us

to do the classification. Thus, we use MIL with different pooling layers to assign different weights to the images of a track, so that the image which is more clear and provides more information can contribute more to the model. Each 'bag' in our experiment is formed by the images of each track and the label of the bag is the person's gender/role/skin tone.

4.2 Model

In the following section, we will introduce the structures and principles of the three models we used in this experiment in detail, including NO_MIL, MIL_MAX, MIL_MAX_yhat and MIL_ATT. Here we indicate that the variables which use capital letters represent track-wise information, while the variables which use the lowercase represent frame-wise information. For example, $\hat{Y}$ represents the track-wise prediction, while $\hat{y}$ represent the frame-wise prediction.

4.2.1 NO_MIL

The NO_MIL model is a model that estimates the label of each frame independently, and then takes the max over all the frame-level predictions. We use it to get the baseline of the three classification tasks, including skin tone classification, gender classification and role classification. We also use its performance to compare with other models which contain MIL strategy.

Training Process

The training process of the NO_MIL model is on frame-level as opposed to track-level which we will introduce in detail in the MIL_MAX and MIL_ATT section. We use ResNet18 as our training model and we train it from scratch. The structure of

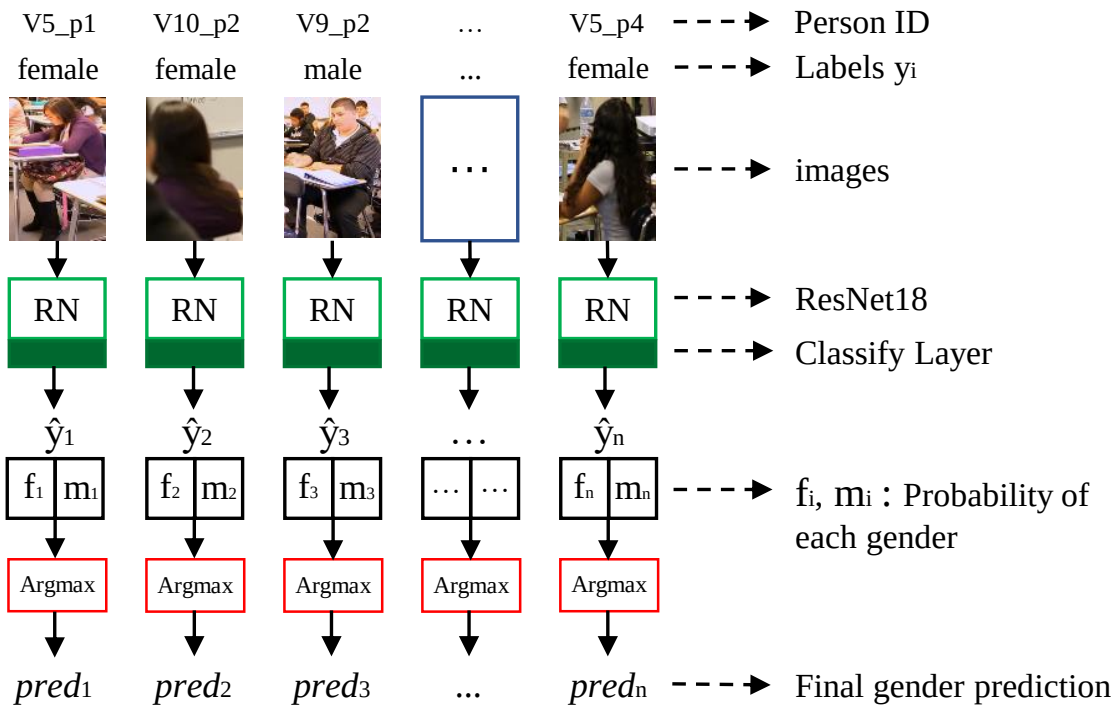


Figure 4.1: NO_MIL Model Structure for Training.

the whole NO_MIL models is shown in the Figure 4.1.

Suppose the batch size is equal to n, thus, we give n images as the one iteration's input to the model. As it's the frame-level training, there is no relationship among these images which means they are loaded randomly. Then, each image is passed into ResNet18 (including the classifier) and get a corresponding output which we call it $\hat{y}$. The shape of $\hat{y}$ is $1 * m$, where m represents the total number of categories in the classification task and each value represents the probability of the corresponding category. Next, we use Argmax to get the category with the highest probability as the *pred*. Eventually, we use the *pred* to calculate the loss value and update the weights and bias for each layer. Here we choose cross entropy as our loss function.

Evaluation Process

The evaluation process is on track-level which is different from the training process and the model structure is shown in the Figure 4.2. Since our goal is to classify a person's skin tone/gender/role but not a image's, we only need to obtain one prediction result for each person, although he/she may have many images. Thus, it's not necessary for each image to get a result. As a result, in the evaluation process, instead of loading images randomly, we load all images from one person each time. Suppose there are n images for this person, correspondingly, instead of getting n prediction results for the n images, we only get one prediction result for this person. We call this processing strategy track-level which is opposed frame-level.

Thus, compare with the model in the training process, we need to add a function to calculate the final prediction result from the n prediction results of n images. We choose Max to achieve this goal and we add it between the output of the model and the $Argmax$ function.

Therefore, in the evaluation process we pass n images which belong to the same person to the ResNet18 each time and will get n outputs each of which has the shape $1 * m$, where m represents the number of the classes. We call these outputs $\hat{y_i}, i \in [0, n]$. Then we use Max to get the maximum value for each category in all n images which means how likely the picture that most resembles this category in these n pictures is to present this category. The max values will form a new vector which we call $\hat{Y}_{final}$. This is the key step for us to get the final result from n results. Next, we pass this $\hat{Y}_{final}$ to the $Argmax$ function and get the category with the highest probability value as the training process did. And we call this result *pred*. Eventually, we use this *pred* together with the label to calculate accuracy, area under the curve (AUC), f1 score, Pearson Correlation Coefficient, confusion matrix and so on.

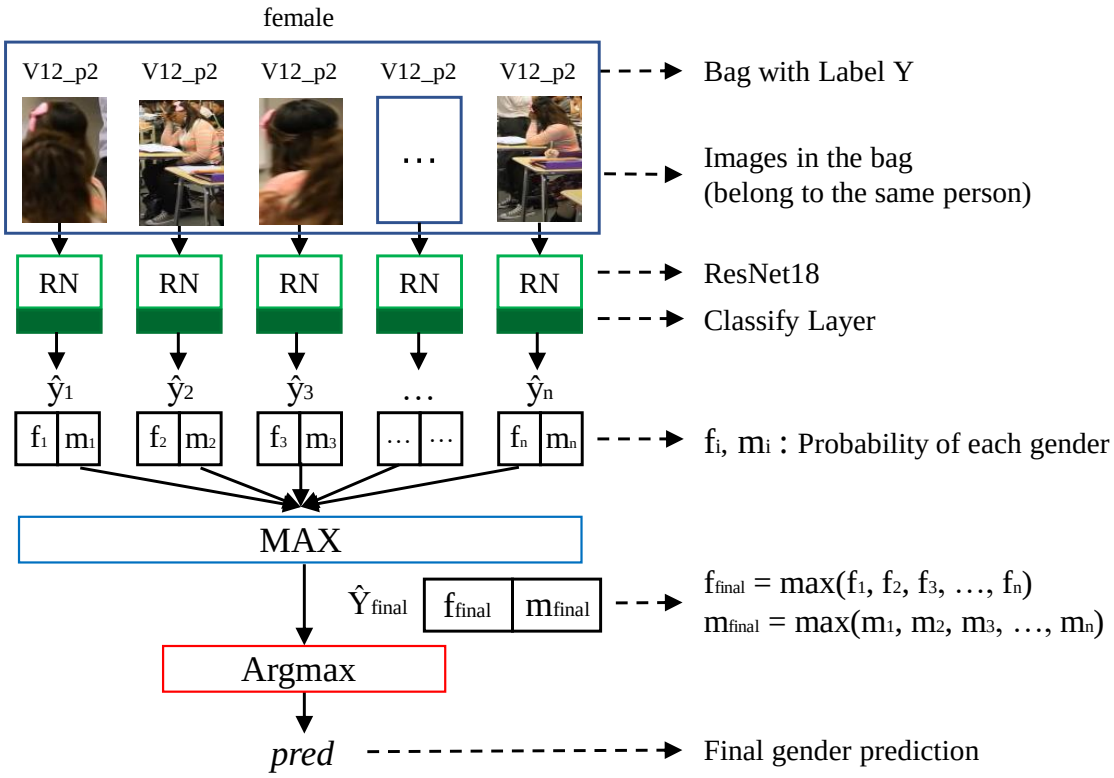


Figure 4.2: NO_MIL Model Structure for Evaluation.

4.2.2 MIL_MAX

The MIL_MAX model is a model that contains MIL strategy and its structure is shown in the Figure 4.3. Compared with the NO_MIL model, it trained on track-level but not on frame-level. Besides, it also includes a max pooling layer which between the ResNet18 and classifier. We use this model to test the results of MIL strategy compared to NO_MIL model. The MIL strategy is implemented by training and evaluating the model in track-level and the max pooling layer will assign different weights to the images. In this model, both the training process and the evaluating process are on track-level, thus, we load all images belong to one person at each time and get only one prediction result for this person.

We first load all images of one person, suppose the number is n, and pass these images into ResNet18. Then we will get n outputs, which we call $h_i, i \in [0, n]$. The shape of h_i is $1 * feature\_num$, which $feature\_num$ represents the total number of the features after the image is processed by ResNet18. Next, the max pooling layer will calculate the maximum value of each feature from the n images, which represents the value of the feature as presented in the best image among the n images. These maximum values will form a new vector which we call Z. The shape of Z is $1 * feature\_num$. Z is then passed into the classifier and we will get a vector $\hat{Y}$ which contain the probability of each category. Finally, we use $Argmax$ to get the category with the highest probability and we call it *pred*. For the training process, we then use the *pred* to calculate the loss value and update the weights. We still use cross-entropy as the loss function in this model. For the evaluation process, we use the *pred* to calculate accuracy, AUC, f1 score, Pearson Correlation Coefficient, confusion matrix and so on.

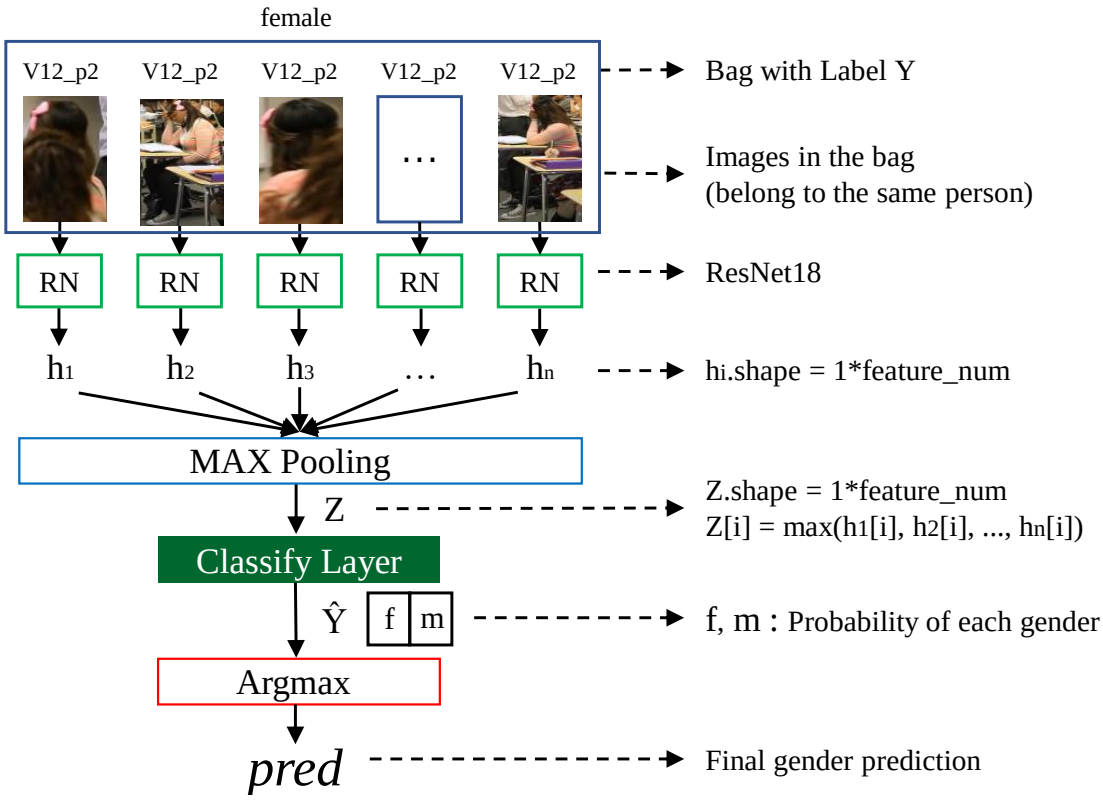


Figure 4.3: MIL_MAX Model Structure.

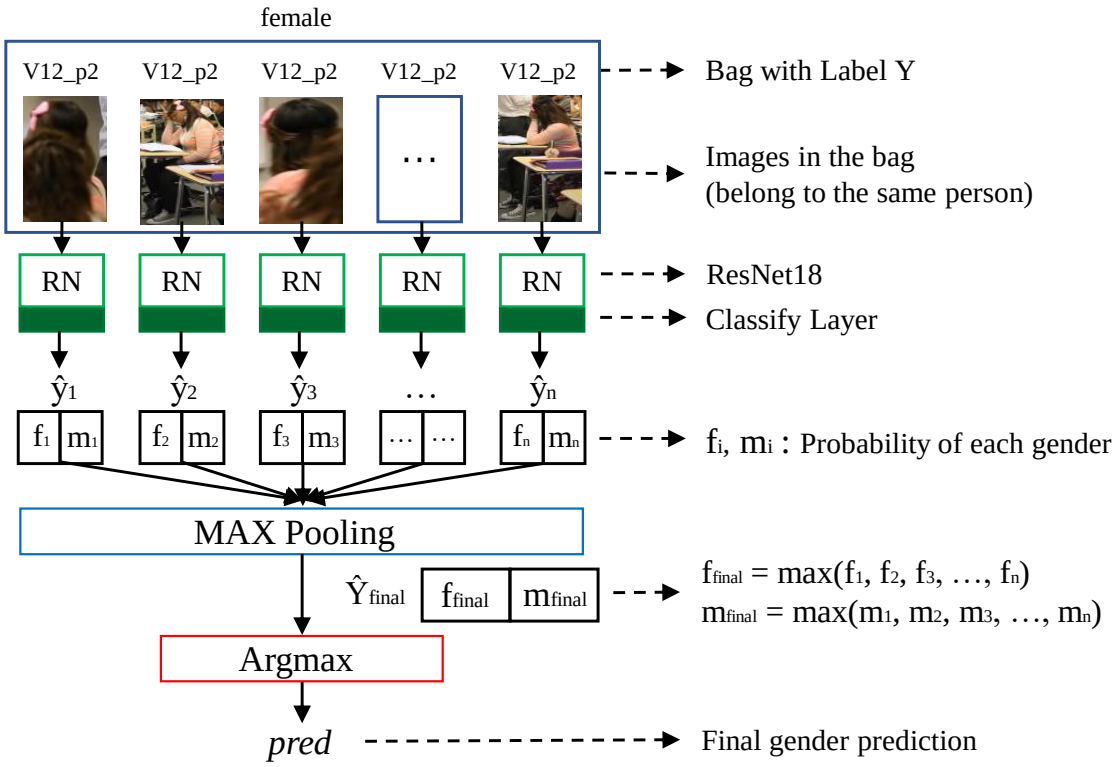


Figure 4.4: MIL_MAX_yhat Model Structure.

4.2.3 MIL_MAX_yhat

The MIL_MAX_yhat model is a variant of the MIL_MAX model. Based on the MIL_MAX model, we change the order of the Classifier and the Max Pooling layer and first go through the Classifier and then do the max pooling over the n classify results instead of n feature vectors in the MIL_MAX model. The Figure 4.4 shows the MIL_MAX_yhat's structure.

4.2.4 MIL_ATT

The MIL_ATT model is an improvement version of the MIL_MAX model by applying attention mechanism. It changes the strategy of assigning the weights to the images by replacing the max pooling layer with an attention pooling layer. Its

structure is shown in the Figure 4.5. In the MIL_MAX model, for each feature, we only select one image from the n images and assign a weight of 1 to it, while the weights of all other images are set to 0. However, in the MIL_ATT model, the weight assigned to each image for each feature is learned during training. We use this model to test the result of the attention mechanism.

The training and evaluation processes of the model are both on the track level, which is the same as the MIL_MAX model. As we mentioned in the MIL_MAX section, we first load all n images of a person and feed them to the ResNet18, and it generates $H_i, i \in [0, n]$. After that, we use attention pooling to calculate the weight vector A using H as input. The attention pooling layer consists three layers, including a linear layer, a tanh layer, and another linear layer. After calculated by the attention pooling layer, we get the weight vector A which shape is $n * 1$. We then transpose A and do the $Softmax$ through n images, and then multiply it with H and we get a vector Z which shape is $1 * feature_num$. We input Z into the classifier and get a $1 * m$ shaped vector, where m represents the number of the classes. We name this vector as $\hat{Y}$ and it contains the possibility of each class. Eventually, we use $Argmax$ to obtain the class with the highest probability as the *pred*. Then we use the cross-entropy to calculate the loss and update the weights for the training process while calculating accuracy, auc, f1 score, Pearson Correlation Coefficient, confusion matrix and so on for the evaluation process.

4.3 Model Comparison

In this section, we will show the different structures and training processes among different models. The Table 4.1 shows the overview comparison of the four models.

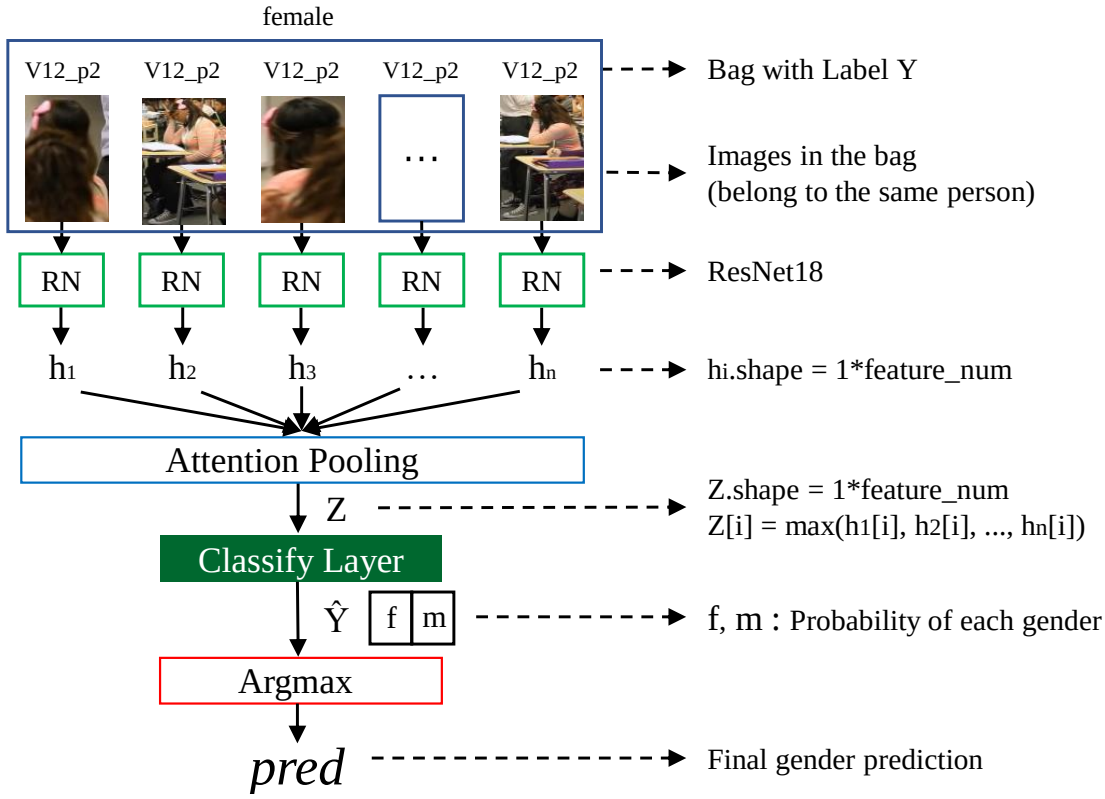


Figure 4.5: MIL_ATT Model Structure.

Model	With MIL	Pooling	Train Pooling Layer	Train Level
NO_MIL	False	On probabilities	False	Frame
MIL_MAX	True	On Features	False	Track
MIL_MAX_yhat	True	On probabilities	False	Track
MIL_ATT	True	On Features	True	Track

Table 4.1: Model Comparison.

4.3.1 NO_MIL vs. MIL_MAX_yhat

In this section, we will compare the differences between the model without MIL strategy which is the NO_MIL model and the MIL_MAX_yhat model which contains the MIL strategy.

The first difference between the two models is the input. The NO_MIL model is a frame-wise model and it's trained on the frame level, thus, it loaded the images randomly which means there's no relation between all images. While for the MIL_MAX_yhat model, it's a track-wise model and it's trained on the track level, thus, it loaded a track of images each time. The second difference is the MIL_MAX_yhat has an additional max pooling layer between the classifier and the $Argmax$. Thus, the MIL_MAX_yhat model will get only one prediction results for the n input images. While the NO_MIL model will get the n prediction results for n input images.

4.3.2 MIL_MAX vs MIL_MAX_yhat

In this section, we will compare the differences between the MIL_MAX model and the MIL_MAX_yhat model.

Both of these two models are track-wise model, thus, they both train on track level and load a track of images each time. The difference between the two models is the order of the classifier and the max pooling layer. In the MIL_MAX_yhat model, the ResNet18 first passes the result feature vectors to the classifier and then do the max pooling. While in the MIL_MAX model, the ResNet18 first passes the result feature vectors to the max pooling layer and then pass the result after the max pooling to the classifier.

4.3.3 MIL_MAX vs MIL_ATT

In this section, we will compare the differences between the MIL_MAX model and the MIL_ATT model.

Both of these two models are track-wise model, thus, they both train on track level and load a track of images each time. The difference between the two model is how to do pooling. In the MIL_MAX model, we use the max pooling while in the MIL_ATT model, we use the attention pooling layer. The attention pooling layer uses attention mechanism to do pooling. It's realised by three layers, including a linear layer, a tanh layer and another linear layer. And the parameters of these two linear layers should be trained. The following formula shows the principle of the attention pooling layer.

$$z = \sum_{k=1}^{K} a_k \mathbf{h}_k$$

$$a_k = \frac{\exp(\mathbf{w^T} \tanh(\mathbf{V}\mathbf{h}_k^T))}{\sum_{j=1}^{K} \exp(\mathbf{w^T} \tanh(\mathbf{V}\mathbf{h}_j^T))}$$

[11]

Chapter 5

Data

In this chapter, we will first introduce the data and its processing; then, we will introduce the data augmentation and data balance for the three classification tasks.

5.1 Data

In this thesis, the raw data we use are the videos which were recorded in the classroom. We then obtain the frames from the videos and processed them so that they can be used as the input of the model. And finally we split them into training, validation and test datasets. The following sections will introduce the data, its processing, and the split strategy in detail.

5.1.1 Raw Data

The dataset we used in our experiments was shared with our research group by a California-based startup company for teacher training. It consists of 957 classroom observation videos (20min long each) , which format are .mp4, ranging from kindergarten through middle school in a Midwestern state in the United States.

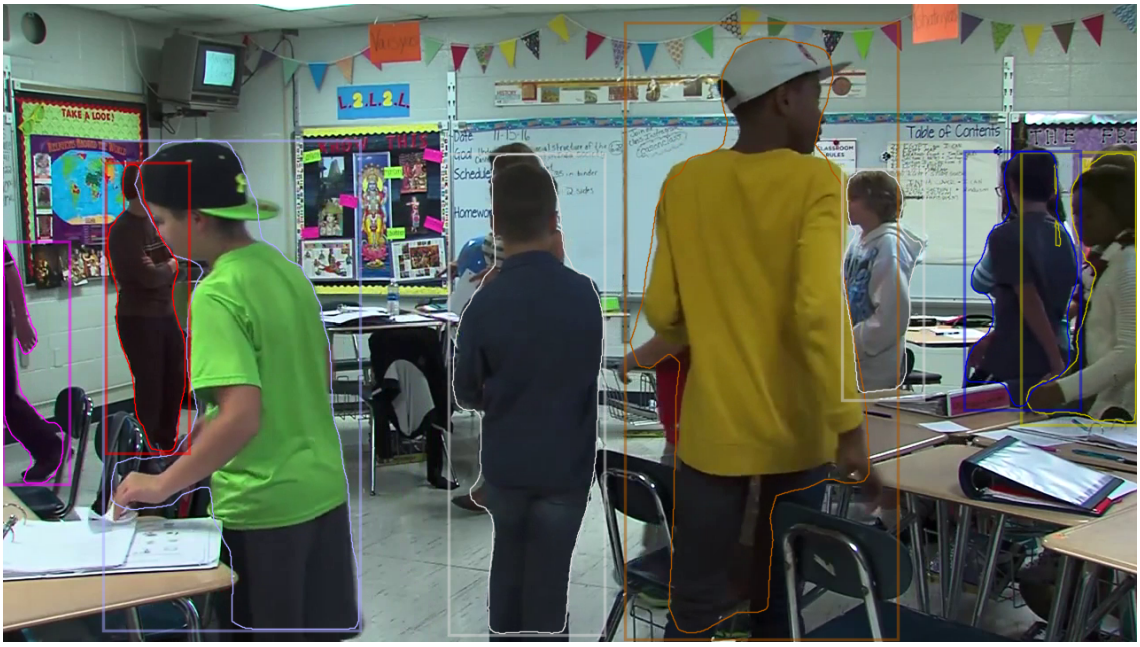


Figure 5.1: A demo of a frame with several bounding boxes.

Each video contains a different teacher and set of students. The videos were recorded by the teachers themselves for the purposes of obtaining feedback on their teaching; hence, the video recording conditions due to camera model, placement, lighting, etc., can vary strongly between videos. In most videos, the camera was placed to capture the teacher's face and speech; hence, the students are often shown from the back, and not all students may be captured in the camera's field of view.

5.1.2 Frame

In our experiment, we used 16 of the above videos for skin tone classification and 44 videos for both gender classification and role classification, and each video will have a unique name. For each video, we extract a frame about every 10 seconds, and each video will be extracted 100 to 200 frames.

5.1.3 Bounding Box

'Bounding box' is the rectangular which enclose each person in each frame. We use the Detectron2 [20] to draw these bounding boxes. The ID labels are from human annotation which means the human annotators labels who is where in different frames. Each bounding box on the frame will become an independent image containing only one person after the crop operation. Therefore, a frame will generate one or more bounding boxes, meanwhile, a track will also correspond to multiple bounding boxes.

5.1.4 Track

In this experiment, we use 'track' to represent a person. There will be multiple tracks in a video, and each track may appear in several frames, thus corresponding to several bounding boxes.

5.1.5 Split Strategy

In our experiment, we hope that the number of bounding boxes in the training set, validation set, and test set account for 70%, 10%, and 20% of the total number of bounding boxes, respectively. In order to avoid cheating machine learning models, we need to ensure that the bounding boxes from the same person are put into the same dataset when doing the splitting process, so as to avoid the model from 'known' a person's skin tone during the training process. And because the number of people appearing in each video is not equal, and the number of frames each person appears in a video is also not equal, resulting in the number of bounding box pictures included in each video is not equal. Therefore, when we split the dataset, we cannot split the video in proportion directly, since doing so is high likely to cause the proportion of the bounding boxes to be inconsistent with the expected proportion. However, it

would be quite complex to split the data based on the bounding boxes level, since we have to guarantee that the bounding boxes of the same person are put into the same dataset, although split on bounding boxes level would be more accurate. Based on the above requirements, we adopt a proportional split strategy for the track level. Although the number of bounding boxes corresponding to each track is also different, we can achieve an appropriate result after multiple shuffle operations, which is better than split on the video-level or bounding box-level.

5.2 Dataset

This section will introduce the data augmentation and data balance for the three classification tasks.

5.2.1 Gender Classification

In our experiment, different models are trained and load data in different ways. The model without MIL strategy are trained based on the frame level, which means that we load the images randomly and the images in one batch has no relationship. While for the model with MIL strategy, we train the model on track level, which means that we load the images by track and each time we load all the images belong to one track. Thus, there are two ways for us to calculate the data, one is on the frame level and the other is on the track level.

We define the data as the frame level data or the track level data according to the following strategy. We define the data loaded randomly as the frame level data, and we use this data only in the training process of the NO_MIL model since it does not care whether the loaded images are belong to the same track or not. We define the data loaded by track as the track level data, and we use this data in the models

	Train	Validation	Test	Total
Female	8996	3358	4036	16390
Male	7504	2649	3454	13607
Total	16500	6007	7490	29997

(a) Data Distribution on Frame Level

	Train	Validation	Test	Total
Female	159	23	46	228
Male	138	20	40	198
Total	297	43	86	426

(b) Data Distribution on Track Level

Table 5.1: Gender Data Distribution

with MIL strategy, including MIL_MAX, MIL_MAX_yhat and MIL_ATT and also the evaluation process of all the model.

The count of the frame level data represent the number of the images while the count of the track level data represent the number of the people. Since not everyone has the same number of images, the category with the most tracks are not always the category with the most images and vice versa. Here is a simple example to help the readers to understand the difference between data at the frame and track levels. For example, there are two people in the female category, which we call $f1$ and $f2$. Assuming that $f1$ has five images and $f2$ has three images, thus, the amount of data for the female category is 8 at the frame level and 2 at the track level.

The data distributions on both frame level and track level are shown in the Figure 5.1.

5.2.2 Role Classification

Due to the unbalanced number of the students and teachers in a real class, the images of the students are much more than that of the teachers as shown in the Figure 5.2.

	Train	Validation	Test	Total
Student	15488	5299	6838	27625
Teacher	2650	923	1464	5037
Total	18138	6222	8302	32662

(a) Original Data Distribution on Frame Level

	Train	Validation	Test	Total
Student	289	42	78	409
Teacher	34	5	10	49
Total	323	47	88	458

(b) Original Data Distribution on Track Level

Table 5.2: Original Role Data Distribution

Due to the unbalanced training data, the model does not perform well. We found that, the model always predict any samples as a 'student'. Therefore, we can conclude that the model cannot distinguish between 'student' and 'teacher' very well although the accuracy can achieve 87.5%. The accuracy cannot reflect the model's performance very well, thus, we will use AUC(Area under curve) to evaluate model's performance as well.

To solve the problem caused by unbalanced data, we use the following data augmentation strategy to increase the number of teacher's images. First, we randomly select images from the original dataset which label is 'teacher'. Second, we rotate the images by randomly set a degree $\alpha, \alpha \in [-5, 5]$. Then we will flip the image randomly and finally we get a new image. The newly created images will be treated as a new track, thus, the number of the tracks will also increase along with the increasing of the images. The newly created images from the same person will be treated as the same new track.

Here is a simple example to help the readers to understand this process. Suppose there are two teachers named Ta and Tb. Ta has 2 images named $Ta1$ and $Ta2$, while Tb has 1 image named $Tb1$, thus, there are 3 images which labels are 'teacher'.

	Frame Level	Track Level
Student	15488	289
Teacher	15488	285
Total	30976	574

Table 5.3: Balanced Role Training Data Distribution

Suppose the target number of the teaches' images is 5. One of the increasing process is as following: choose $Ta1$ and $Tb1$ and randomly rotate and flip them, and we get 2 new images named $Ta1'$ and $Tb1'$ which belong to 2 new tracks named Ta' and Tb' respectively. Now we have 5 images which labels are 'teacher', including $Ta1$, $Ta2$, $Tb1$, $Ta1'$, $Tb1'$. And we have 4 tracks who are teachers, including Ta, Tb, Ta' and Tb'.

We call the new dataset after the data augmentation the Balanced Data and its distributions on both frame level and track level are shown in the Table 5.3.

5.2.3 Skin Tone Classification

The skin tone data is very unbalanced as Figure 5.4 shows, among which the Type2 is the majority while the Type3 is the most rare. As a result, the performance of the models trained on this dataset is not good, and the accuracy is shown in the Figure 6.7a.

We also tried the data augmentation strategy mentioned in the Role Classification Section and get a new dataset which named Frame Balanced Data. The data distribution of the Frame Balanced Data is shown in the Figure 5.5.

However, not like the Role Classification, the tracks' amounts are not balanced even after the images amounts are balanced. Thus, the Frame Balanced data is not an effective augmentation for the models which contain the MIL strategy, and it's only useful for the NO_MIL model. Thus, we carry out another data augmentation

	Train	Validation	Test	Total
Type1	1692	310	300	2302
Type2	2504	145	391	3040
Type3	270	96	10	376
Type4	430	136	58	624
Type5	1102	331	333	1766
Type6	995	183	236	1414
Total	6993	1201	1328	9522

(a) Original Data Distribution on Frame Level

	Train	Validation	Test	Total
Type1	30	5	9	44
Type2	34	6	10	50
Type3	4	1	2	7
Type4	9	2	3	14
Type5	20	4	6	30
Type6	14	3	4	21
Total	111	21	34	166

(b) Original Data Distribution on Track Level

Table 5.4: Original Skin Tone Data Distribution

	Frame Level	Track Level
Type1	2504	32
Type2	2504	34
Type3	2504	28
Type4	2504	31
Type5	2504	28
Type6	2504	27
Total	15024	180

Table 5.5: Frame Balanced Skin Tone Training Data Distribution.

	Frame Level	Track Level
Type1	1923	34
Type2	2504	34
Type3	2319	34
Type4	1689	34
Type5	1912	34
Type6	2369	34
Total	12716	204

Table 5.6: Track Balanced Skin Tone Training Data Distribution

as following.

The goal of this data augmentation is to make the data balanced on the track level. Our strategy is as following: First, take the track amount of the category which has the most tracks among the six categories as the increase target. Second, for each category that needs to be increased, randomly select a track in the same category of the original dataset. Then rotate each image by a randomly select degree $\alpha, \alpha \in [-5, 5]$. Next, randomly flip each image and finally we get a new track with new images.

In other words, suppose there are three people who's skin tone is a certain type and named Ta, Tb and Tc respectively. Suppose Ta has 2 images which named $Ta1$ and $Ta2$, Ta has 1 image which named $Tb1$ and Tc has 2 images which named $Tc1$ and $Tc2$. Assume our target track number is 5. One of the increasing process is as following: Firstly, choose Ta and Tb. Secondly, rotate $Ta1$, $Ta2$ and $Tb1$ randomly and flip them randomly. Now we get two new tracks which named $Ta\prime$ and $Tb\prime$ and contain 2 images (named $Ta1\prime$ and $Ta2\prime$) and 1 image (named $Tb\prime$) respectively. The new dataset contain 5 tracks and 8 images in total.

We call the new dataset after this data augmentation the Track Balanced Data and its distributions on both frame level and track level are shown in the Figure 5.6.

Chapter 6

Experiments and Results

This chapter will introduce the experiment for the two research questions. For each of them, we will show the result by the order of gender classification, (teacher vs. students) role classification and skin tone classification. For each classification task, we apply NO_MIL, MIL_MAX, MIL_MAX_yhat and MIL_ATT models. For each model, the following hyperparameters have the same value: $batch\_size$=4; $epoch$=20; $initial\_learning\_rate$=0.001.

The Results of each research question and each classification task contain tables with both mean accuracies as well as their standard errors. The standard errors are calculated in the following formula:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

where SE represents standard error, p is the percent correct and n in the number of test tracks.

Model	Train	Validation	Test
NO_MIL	0.8781	0.8293	0.5952
MIL_MAX	0.9796	1.0000	1.0000
MIL_MAX_yhat	0.7755	1.0000	0.9881
MIL_ATT	0.5306	0.5366	0.5233

Table 6.1: The Accuracy Results for Gender Classification

Model	Standard Error
NO_MIL	0.0529
MIL_MAX	0.0000
MIL_MAX_yhat	0.0117
MIL_ATT	0.0539

Table 6.2: Standard Error of Testing

6.1 Research Question 1

This section will show the results of the three classification tasks for the first research question: Is MIL helpful for improving the accuracy of the gender, (teacher vs. student) role and skin tone classification?

6.1.1 Gender Classification

In this task, our aim is to classify if a person's perceived gender is male or female. The Table 6.1 shows the accuracy results, from where we can find out that, the MIL_MAX model gets the highest accuracy in the test process which is 100.00%. The Table 6.2 shows the standard errors of the test results.

6.1.2 Role Classification

In this task, our aim is to classify if a person is a student or a teacher. We trained the models with two different datasets which we introduced in Section 5.2.2. The Table 6.3 shows the results of the models trained on the original data, including

Model	Train	Validation	Test
NO_MIL	0.9508	0.9318	0.8621
MIL_MAX	0.8959	0.8864	0.8750
MIL_MAX_yhat	0.8959	0.8864	0.8750
MIL_ATT	0.8959	0.8864	0.8750

(a) The Accuracy Results for Role Classification

Model	Test AUC
NO_MIL	0.9065
MIL_MAX	0.2680
MIL_MAX_yhat	0.6564
MIL_ATT	0.4817

(b) Testing AUC Results

Table 6.3: The Results of the Models Trained on Original Data

Model	Standard Error
NO_MIL	0.0368
MIL_MAX	0.0353
MIL_MAX_yhat	0.0353
MIL_ATT	0.0353

Table 6.4: Standard Error of Testing

the accuracy results and the testing AUC results. In these tables, we can find out that although the models with MIL get higher accuracy than NO_MIL model, their AUC results are low. We print out their prediction results and find the models always predict all the samples as student. Thus, we trained the models again on the Balanced Dataset. The Table 6.4 shows the standard errors of the test results.

The Table 6.5 shows the results of the models trained on the balanced data, from where we can find out that the MIL_MAX model gets the highest result on both accuracy and AUC. The Table 6.6 shows the standard errors of the test results.

Model	Train	Validation	Test
NO_MIL	0.9525	0.9615	0.8391
MIL_MAX	0.9701	1.0000	0.9773
MIL_MAX_yhat	0.9542	1.0000	0.9659
MIL_ATT	0.5000	0.7955	0.6591

(a) The Accuracy Results for Role Classification

Model	Test AUC
NO_MIL	0.8506
MIL_MAX	0.9091
MIL_MAX_yhat	0.7893
MIL_ATT	0.4734

(b) Testing AUC Results

Table 6.5: The Results of the Models Trained on Balanced Data

Model	Standard Error
NO_MIL	0.0392
MIL_MAX	0.0159
MIL_MAX_yhat	0.0193
MIL_ATT	0.0505

Table 6.6: Standard Error of Testing

Model	Train	Validation	Test
NO_MIL	0.8042	0.2632	0.2903
MIL_MAX	0.3063	0.2105	0.2903
MIL_MAX_yhat	0.2883	0.3158	0.2581
MIL_ATT	0.3153	0.2105	0.2903

(a) The Accuracy Results for Skin Tone Classification

Model	Test PCC
NO_MIL	0.1917
MIL_MAX	NAN
MIL_MAX_yhat	0.3645
MIL_ATT	NAN

(b) Testing PCC Results

Table 6.7: The Results of the Models Trained on Original Data

Model	Standard Error
NO_MIL	0.0778
MIL_MAX	0.0778
MIL_MAX_yhat	0.0750
MIL_ATT	0.0778

Table 6.8: Standard Error of Testing

6.1.3 Skin Tone Classification

In this task, our aim is to classify a person's skin tone type. The skin tone types are divided into 6 levels from light to dark. And we introduced the datasets in the Section 5.2.3 The Table 6.7 shows the accuracy results of the models trained on the original dataset. The Table 6.8 shows the standard errors of the test results.

The Table 6.9 shows the accuracy results of the models trained on the Track Balanced dataset. The Table 6.10 shows the standard errors of the test results.

Model	Train	Validation	Test
NO_MIL	0.8507	0.2632	0.3548
MIL_MAX	0.6716	0.4211	0.4194
MIL_MAX_yhat	0.7059	0.3158	0.3226
MIL_ATT	0.3824	0.2105	0.2903

(a) The Accuracy Results for Skin Tone Classification

Model	Test PCC
NO_MIL	0.7057
MIL_MAX	0.6508
MIL_MAX_yhat	0.3248
MIL_ATT	NAN

(b) Testing PCC Results

Table 6.9: The Results of the Models Trained on Track Balanced Data

Model	Standard Error
NO_MIL	0.0821
MIL_MAX	0.0846
MIL_MAX_yhat	0.0802
MIL_ATT	0.0778

Table 6.10: Standard Error of Testing

Dataset	Train	Validation	Test
Original Data	0.9508	0.9318	0.8621
Balanced Data	0.9525	0.9615	0.8391

(a) Accuracy

Dataset	Test AUC
Original Data	0.9065
Balanced Data	0.8506

(b) AUC

Table 6.11: The Results of NO_MIL Trained on different datasets

Dataset	Train	Validation	Test
Original Data	0.8959	0.8864	0.8750
Balanced Data	0.9701	1.0000	0.9773

(a) Accuracy

Dataset	Test AUC
Original Data	0.2680
Balanced Data	0.9091

(b) AUC

Table 6.12: The Results of MIL_MAX Trained on different datasets

6.2 Research Question 2

This section will show the results of the three classification tasks for the second research question: Does data augmentation and data balance help for improving the performance of the model? As the data for gender classification is already balanced, we only do experiments on role classification and skin tone classification.

6.2.1 Role Classification

The Table 6.11, Table 6.12, Table 6.13 and Table 6.14 show the accuracy results and AUC results comparison of each model trained on the original dataset and on the balanced dataset respectively.

Dataset	Train	Validation	Test
Original Data	0.8959	0.8864	0.8750
Balanced Data	0.9542	1.0000	0.9659

(a) Accuracy

Dataset	Test AUC
Original Data	0.6564
Balanced Data	0.7893

(b) AUC

Table 6.13: The Results of MIL_MAX_yhat Trained on different datasets

Dataset	Train	Validation	Test
Original Data	0.8959	0.8864	0.8750
Balanced Data	0.5000	0.7955	0.6591

(a) Accuracy

Dataset	Test AUC
Original Data	0.4817
Balanced Data	0.4734

(b) AUC

Table 6.14: The Results of MIL_ATT Trained on different datasets

Dataset	Train	Validation	Test
Original Data	0.8042	0.2632	0.2903
Track Balanced Data	0.8507	0.2632	0.3548

(a) Accuracy

Dataset	Test PCC
Original Data	0.1917
Balanced Data	0.7057

(b) PCC

Table 6.15: The Results of NO_MIL Trained on different datasets

Dataset	Train	Validation	Test
Original Data	0.3063	0.2105	0.2903
Track Balanced Data	0.6716	0.4211	0.4194

(a) Accuracy

Dataset	Test PCC
Original Data	NAN
Balanced Data	0.6508

(b) PCC

Table 6.16: The Results of MIL_MAX Trained on different datasets

6.2.2 Skin Tone Classification

The Table 6.15, Table 6.16, Table 6.17 and Table 6.18 show the accuracy results and pearson correlation coefficient (PCC) comparison of each model trained on the original dataset and on the track balanced dataset.

Dataset	Train	Validation	Test
Original Data	0.2883	0.3158	0.2581
Track Balanced Data	0.7059	0.3158	0.3226

(a) Accuracy

Dataset	Test PCC
Original Data	0.3645
Balanced Data	0.3248

(b) PCC

Table 6.17: The Results of MIL_MAX_yhat Trained on different datasets

Dataset	Train	Validation	Test
Original Data	0.3153	0.2105	0.2903
Track Balanced Data	0.3824	0.2105	0.2903

(a) Accuracy

Dataset	Test PCC
Original Data	NAN
Balanced Data	NAN

(b) PCC

Table 6.18: The Results of MIL_ATT Trained on different datasets

Chapter 7

Conclusion

In this thesis, we explore a new method to classify people's stable attributes in the classroom video. There are three classification tasks in our experiment and we tried four models for each of them, including NO_MIL, MIL_MAX , MIL_MAX_yhat and MIL_ATT.

For the first research question, we have the following conclusion: Based on the result in Table 6.1 and Table 6.5, we conclude that with appropriate pooling layer, MIL is helpful for improving the accuracy of the stable attribute classification in classroom videos. For all the tasks the MIL_MAX model always did the best than the other three models.

For the second research question, we have the following conclusion: Based on the result in Table 6.12, Table 6.13, Table 6.15, Table 6.16 and Table 6.17 we conclude that data augmentation and data balance are helpful for improving the performance of the models. But since we apply data augmentation and data balance at the same time, we do not know which one plays a more significant role.

For the future exploration, we will first figure out why attention mechanism does not work well for these tasks. Also, we would like to find out if there's any other

models which can do better in the skin tone classification. What's more, we also want to explore how the number of images in a track will affect the performance of the model. Last but not least, for the (teacher vs. student) role classification and skin tone classification, we also plan to figure out which method plays a more important role in these tasks, data augmentation or data balance.

Bibliography

[1] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010.

[2] P. J. Bevan and A. Atapour-Abarghouei. Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. In *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop*.

[3] D. Borza, A. S. Darabant, and R. Danescu. Automatic skin tone extraction for visagism applications. In *VISIGRAPP (4: VISAPP)*, pages 466–473, 2018.

[4] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(2):1–26, 2016.

[5] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77.

[6] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013.

[7] Y. Copur-Gencturk, J. R. Cimpian, S. T. Lubienski, and I. Thacker. Teachers' bias against the mathematical ability of female, black, and hispanic students. *Educational Researcher*, 49(1):30–43, 2020.

[8] S. K. D'mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser. Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, 18:45–80, 2008.

[9] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, volume 1, page 2, 1990.

[10] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer. Casual conversations: A dataset for measuring fairness in ai. In *CVPR*, pages 2289–2293, 2021.

[11] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136, 2018.

[12] M. Jmal, W. S. Mseddi, R. Attia, and A. Youssef. Classification of human skin color and its application to face recognition. In *MMEDIA*, 2014.

[13] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer vision and image understanding*, 74(1):1–21, 1999.

[14] V. Lavy and E. Sand. On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263–279, 2018.

[15] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *CVPR workshops*.

[16] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2016.

[17] K. Sikka, A. Dhall, and M. Bartlett. Weakly supervised pain localization using multiple instance learning. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*.

[18] C. Terrier. Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, 77:101981, 2020.

[19] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, G. Bebis, and A. M. Mirza. Gender recognition from face images with local wld descriptor. In *2012 19th international conference on systems, signals and image processing (IWSSIP)*, pages 417–420. IEEE, 2012.

[20] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[21] S. Yan, M. Liu, and T. S. Huang. Extracting age information from local spatially flexible patches. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 737–740. IEEE, 2008.