FINAL PROJECT

Detection of violence against children in videos

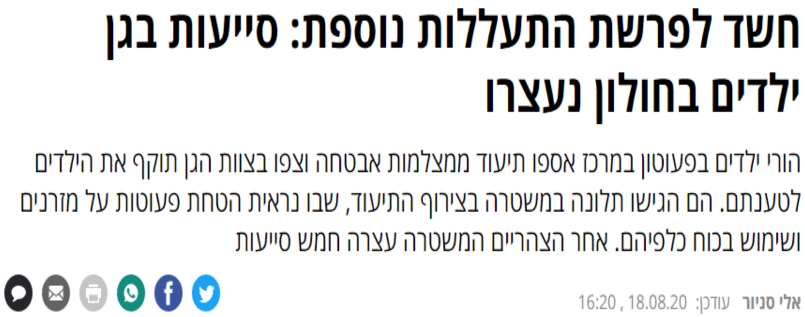
Student Name: Rakhlevski Ilia

Supervisor: Dr. Oren Dinai

Advisor: Dr. Yehudit Aperstein

Motivation

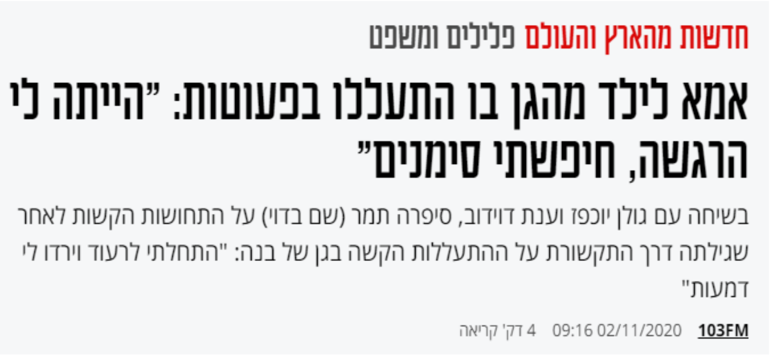
□ Violence against children is one of the biggest problems affecting families and societies. There many publications in mass media about this problem.



חשד לפרשת התעללות נוספת: סייעות בגן ילדים בחולון נעצרו

הורי ילדים בפעוטון במרכז אספו תיעוד ממצלמות אבטחה וצפו בצוות הגן תוקף את הילדים לטענתם. הם הגישו תלונה במשטרה בצירוף התיעוד, שבו נראית הטחת פעוטות על מזרנים ושימוש בכוח כלפיהם. אחר הצהריים המשטרה עצרה חמש סייעות

אלי סניור עודכן: 16:20 , 18.08.20



חדשות מהארץ והעולם פלילים ומשפט

אמא לילד מהגן בו התעללו בפעוטות: "הייתה לי הרגשה, חיפשתי סימנים"

בשיחה עם גולן יוכפז וענת דוידוב, סיפרה תמר (שם בדוי) על התחושות הקשות לאחר שגילתה דרך התקשורת על ההתעללות הקשה בגן של בנה: "התחלתי לרעוד וירדו לי דמעות"

103FM 09:16 02/11/2020 4 דק' קריאה

□ According to the World Health Organization estimation up to 1 billion children aged 2–17 years, have experienced physical, sexual, or emotional violence or neglect in the past year.

Motivation (cont.)

- When violence occurs in a Kindergarten it often remains unknown to other people.
- Our goal is development of an application that will help to detect frames of violence in video streams.
- After the violence detection, an alert or a summary could be extracted to adjacent security department (or to the parents) to yield an action.

Research problem

- The input of the application will be video stream - sequence of frames, that contains any interaction between an adult and children in Kindergarten.
- The output of the application will be localization of the detected frames with violence in the video streams.
- Training and testing data will be video streams taken from the Internet, movies or semi-synthetic created videos. Some parts of them contain scenes with violence and the other parts contain non-violent ones.
- The quality of the solution is measured by quantity of correct detected violent frames.

Prior work

- All the projects/articles that we have found are related to common violence and are not focused on a specific type of violence of adults against children.

- In these projects/articles are used both classical methods (for example: SVM, CNN, LSTM) and modern ones (for examples: VGG16, ResNet50).

- They reached various classification accuracies between 87 and 100%. Such high accuracies could be reached as a result of using very small datasets (not enough frames containing the scenes with violence).

Project novelty

- In this project we are focused on specific type of the violence of adults against children.

- There are no available data sets match this problem.

- It is very difficult to obtain a real data for privacy and legal reasons.

Data description

- Format of video streams used in the project is MP4.

- While a video stream processing the frames are grabbed from the video. It is created an array of the frames. This array serves as an input of the model.

- After converting of the video stream into array of the frames we change the frames size according to the model input.

Data description (cont.)

- Video streams contain violent/non-violent scenes.

- At this stage we defined two types of violent behavior which will be detected - hitting a child by hand and by leg .

- Non-violent actions are the actions that do not contain the actions that are defined as violent.

Data description (cont.)

Violent scenes

Data description (cont.)

Non-violent scenes

Data description (cont.)

- Because of lack of videos containing scenes of violence of adults against children we decided to create synthetical videos.

- Using reborn baby dolls – the dolls that are very similar to human infants.

Data description (cont.)

- Our goal to use such videos to train the model to detect the movements that can be considered as violence.

- The model must be trained to classify video according to violent movements only.

- Elements like background, clothes, beating adult or beaten child, lighting etc. must not have influence on classification results.

- During each video session must be made videos of both types: with violence and no violence.

- Using data augmentation.

Data description (cont.)

Labeling

- Training:

  We sort the video files according to their content: violent or non-violent and put them into the relevant directory.

- Testing:

  For each video is created a file containing ranges of the “violent” frames.

Data description (cont.)

Training data creation

For each video session we perform several actions:

- Camera must be stable during the whole video session.
- Only humans/dolls movements can change.
- Each video session must contain both types of actions: with/no violence.
- Each video stream created during the video session is divided into small videos of size 20 frames.
- All the small videos are sorted according to their content (violent/non-violent).

Methodology

Architecture

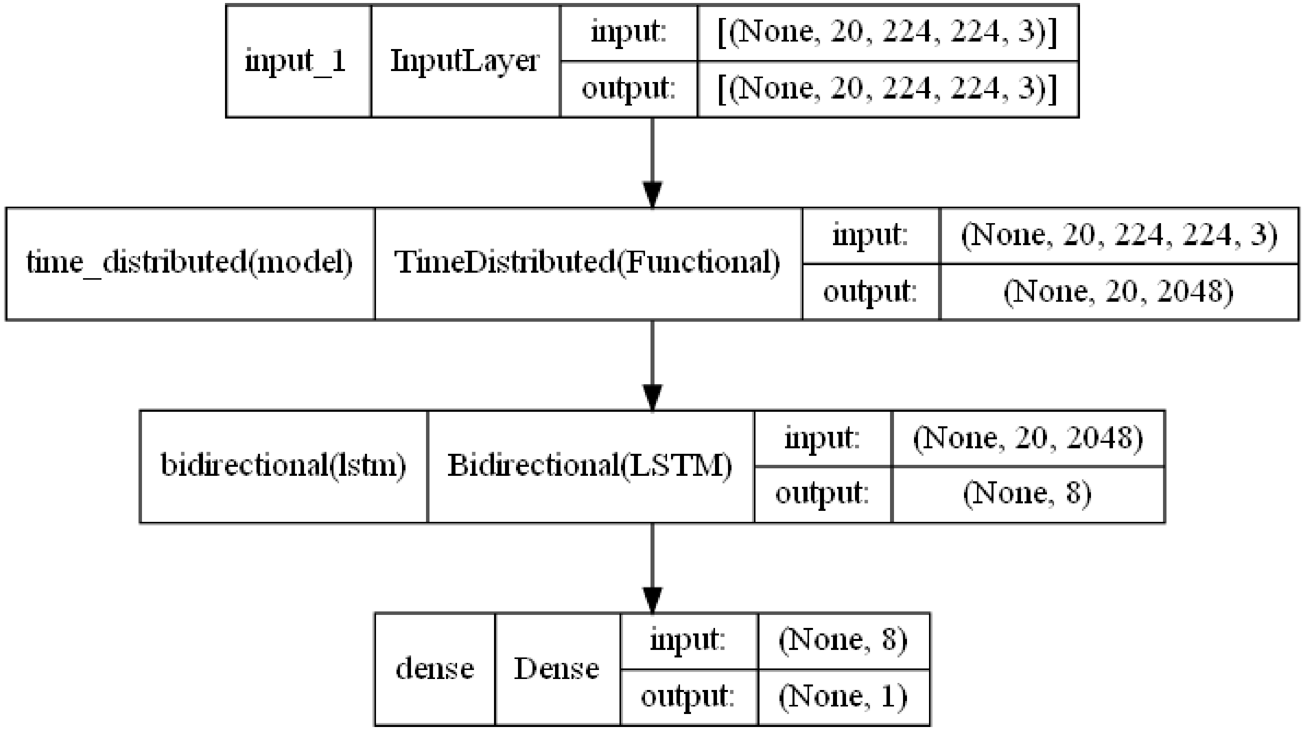
- We chose the CNN+LSTM. This is a combination of two architectures CNN and LSTM.
- The input of the network is a video, sequence of 2D images.
- The output is predicted class.
- CNN is used for feature extraction and LSTM is used to classify video based on those features.

Methodology (cont.)

Architecture Implementation

- Using Keras library.
- Using a pre-trained CNN model ResNet152V2.
- Input - short video stream 20 frames. Each frame is 224x224x3 and pixels values are scaled between -1 and 1.
- Output – predicted class. 0 – violent, 1 – non-violent.
- Using Bidirectional LSTM. It can improve accuracy when the dataset has imbalanced classes.

Methodology (cont.)



input_1	InputLayer	input:	[(None, 20, 224, 224, 3)]
		output:	[(None, 20, 224, 224, 3)]

time_distributed(model)	TimeDistributed(Functional)	input:	(None, 20, 224, 224, 3)
		output:	(None, 20, 2048)

bidirectional(lstm)	Bidirectional(LSTM)	input:	(None, 20, 2048)
		output:	(None, 8)

dense	Dense	input:	(None, 8)
		output:	(None, 1)

Methodology (cont.)

Transfer learning

Using the ResNet152V2 model.

- Create the ResNet152V2 model without an output layer.
- Load the weights for the pre-trained model (without output layer).
- Freeze all the layers of the model that they will not be trained during the whole model training.
- As the output layer is added GlobalAveragePooling2D layer.

Methodology (cont.)

Overfitting prevention

- To simplify the architecture - reducing number of layers or/and neurons.
- Increasing number of samples - creating extra videos.
- Data augmentation.
- Using Regularization: Elastic Net (combination L1 and L2).

Methodology (cont.)

Hyperparameters

Hyperparameters are set experimentally during model training.

□ Number of layers and neurons. This parameter is set for the bidirectional LSTM layer – 4 units.

□ Learning rate. It is set to 0.0001 at the start of training. At the end stages of training it can be changed to 0.00001.

□ Adam optimizer was used to train the network.

□ The batch size is 64.

□ Class weight parameter. It is calculated (for class j):

wj = n_samples / (n_classes * n_samplesj)

Methodology (cont.)

Model training process

1. We train the created/trained model.
2. During the model training we follow the parameters both “loss” and “accuracy”.
3. If the model stopped improving, then the learning rate should be updated and the training should continue.
4. If the model stopped improving finally, we stop the training.
5. We change the model: layers, number of neurons, methods of regularization and repeat the steps 1-4.

Methodology (cont.)

Testing

- □ Testing of the model is performed on real videos.
- □ Using the windowing algorithm for videos processing.
- □ Classification accuracy measurement:
  - IoU (Intersection over Union) is used for videos containing both “violent”/”non-violent” frames.
  - Also, we use statistics on correct/incorrect predicted “violent” / ”non-violent” frames.
- □ For each video is created a file containing ranges of the “violent” frames.

Methodology (cont.)

For the testing purpose we used videos of several types:

- Synthetical videos – they were using reborn dolls containing scenes with violence and without one.
- Real videos with violence – real videos containing scenes of violence .
- Real videos without violence - real videos that do not contain scenes of violence.
- Videos that do not contain humans - videos that contain rooms with furniture only.

Total: 120 video files, 45850 – frames, ~25.5 minutes duration.

Results

Summary – all types	With Violence (actual): 3230 Correct predicted: 2788 (86.32 %) Incorrect predicted: 442 (13.68 %) No Violence (actual): 42620 Correct predicted: 38852 (91.16 %) Incorrect predicted: 3768 (8.84 %)

Results (cont.)

Synthetical videos containing scenes with and without violence	With Violence (actual): 1638 Correct predicted: 1398 (85.35 %) Incorrect predicted: 240 (14.65 %) No Violence (actual): 9914 Correct predicted: 8855 (89.32 %) Incorrect predicted: 1059 (10.68 %)
Real videos with violence	With Violence (actual): 1592 Correct predicted: 1390 (87.31 %) Incorrect predicted: 202 (12.69 %) No Violence (actual): 2045 Correct predicted: 693 (33.89 %) Incorrect predicted: 1352 (66.11 %)

Results (cont.)

Real videos without violence	With Violence (actual): 0 Correct predicted: 0 (100 %) Incorrect predicted: 0 (0 %) No Violence (actual): 30477 Correct predicted: 29120 (95.55 %) Incorrect predicted: 1357 (4.45 %)
Videos that do not contain humans	With Violence (actual): 0 Correct predicted: 0 (100 %) Incorrect predicted: 0 (0 %) No Violence (actual): 184 Correct predicted: 184 (100.0 %) Incorrect predicted: 0 (0.0 %)

Results (cont.)

- 86 % “violent” and 91 % “non-violent” frames are correct predicted.
- Synthetical videos: 85 % “violent” and 89 % “non-violent” frames are correctly predicted.
- Real videos with violence: 87 % “violent” and 33 % “non-violent” frames are correctly predicted.
- Real videos without violence: 95 % “non-violent” frames are correctly predicted.
- Videos with empty rooms: 100 % “non-violent” frames are correctly predicted.

Results (cont.)

- For “non-violent” frames the results are better.
- The “non-violent” frames are found near the “violent” ones and they are predicted as “violent” too.
- The best results for “violent” frames recognition were received for those videos where were used sequences of hits: several hits, quickly following one another.
- Results of the testing are relative good, in the main for those videos that are similar to the videos from the dataset.

Discussion

Differences between our project and the existing ones

Other projects:

□ In most cases it is enough to use one frame to perform recognition.

□ In all cases it talks about certain actions. Each action is one class.

Our projects:

□ We must recognize action. We need all the frames from a video for this purpose.

□ We have two classes only, but the “violence” class can have many actions.
The “non-violence” class has an endless number of actions.

Discussion

Our goals

- Our main task is that the model will learn from the human movement only. To meet this target, during each video session we created videos containing both violent and non-violent movements.

- Another important goal is to achieve generalization. The trained model must learn to recognize humans (reborn dolls) on the video and to classify their movement under different conditions: poses, clothes, gender, hairstyle etc.

Discussion

Further work

To continue this work, we need to increase the dataset.

- □ To add new real videos containing scenes of violence against children.
- □ Using more different reborn dolls and humans, clothes, backgrounds etc. to create scenes with many humans and reborn dolls.
- □ Using extra augmentation techniques: saturation, changing brightness etc.