Computer Vision project

Pedestrian Intention Estimation on JAAD Dataset

Paradiso Emiliano, 1940454

10 September 2024

DIPARTIMENTO DI INGEGNERIA INFORMATICA
AUTOMATICA E GESTIONALE ANTONIO RUBERTI

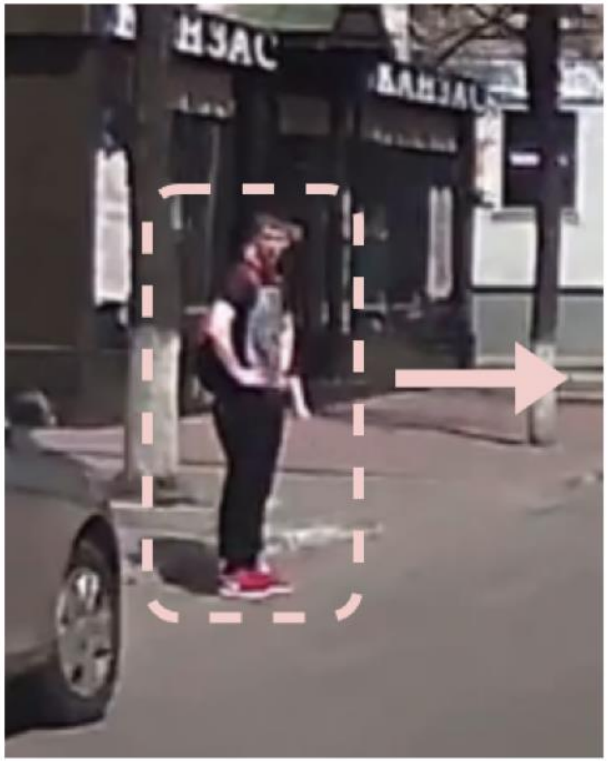
SAPIENZA
UNIVERSITÀ DI ROMA

Outline

- Introduction
- Related work
- Proposed method
- Dataset and metrics
- Implementation details
- Experimental results
- Conclusion and future works

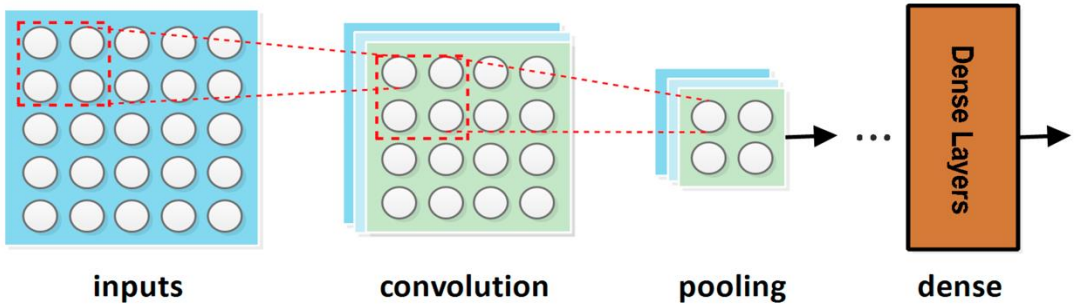
Introduction

- Why? ->Safe autonomous driving system (ADS)
- How?-> Vision-based pedestrian crossing intention prediction algorithm
- So?-> ADS can generate more safe and effcient manouver



cross or not?

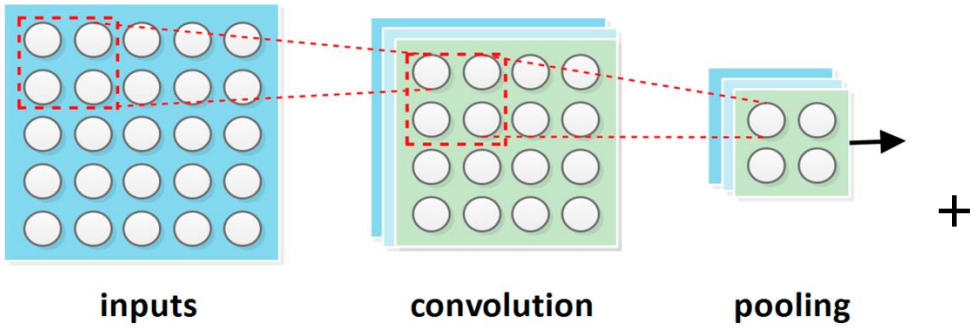
Related works



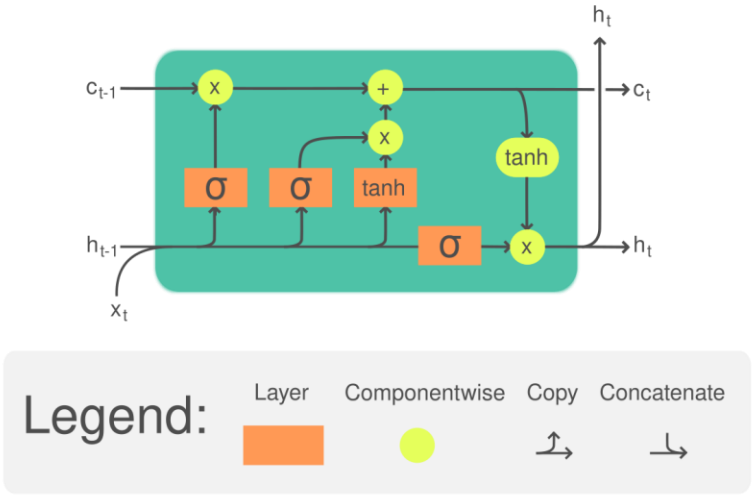
cross or not?

- Feature extraction from one frame using 2D CNN

Related works

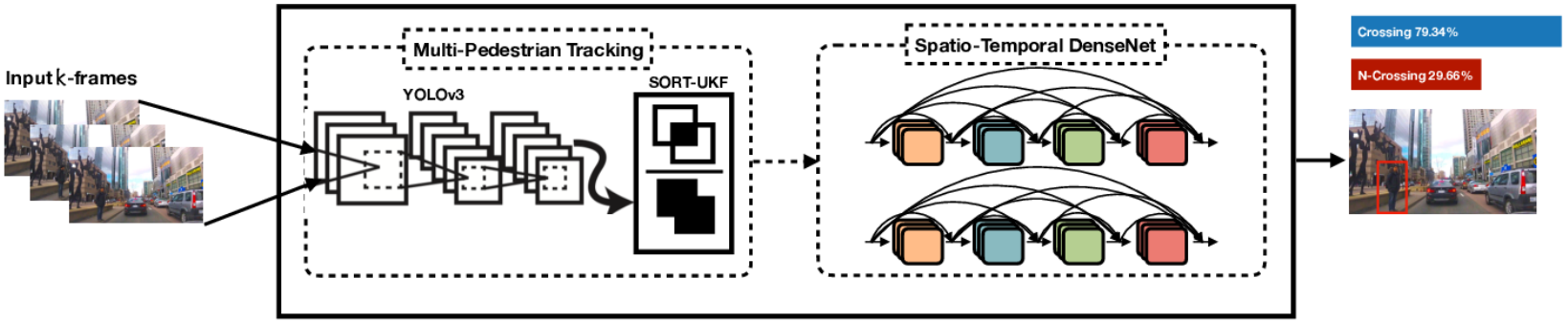


+



- Spatio temporal modeling using image sequences
  2D CNN+ RNN

Related works



- 3D CNN

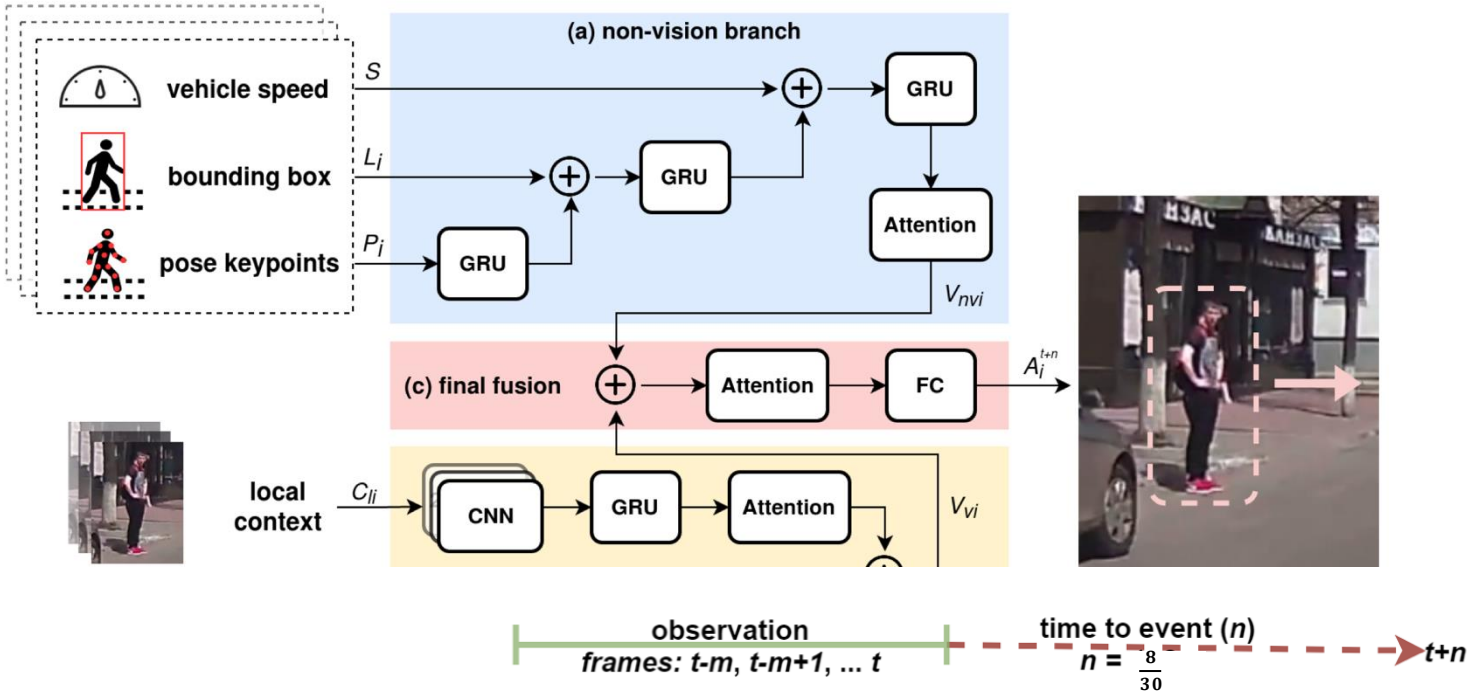
Previous works limitations

- But…

i. Pedestrian's bounding box, body- pose keypoints, vehicle motion, and the ocal context can be modeled as separate channels as inputs to the prediction model.

ii. This requires a proper way of fusing the above information.

Proposed method



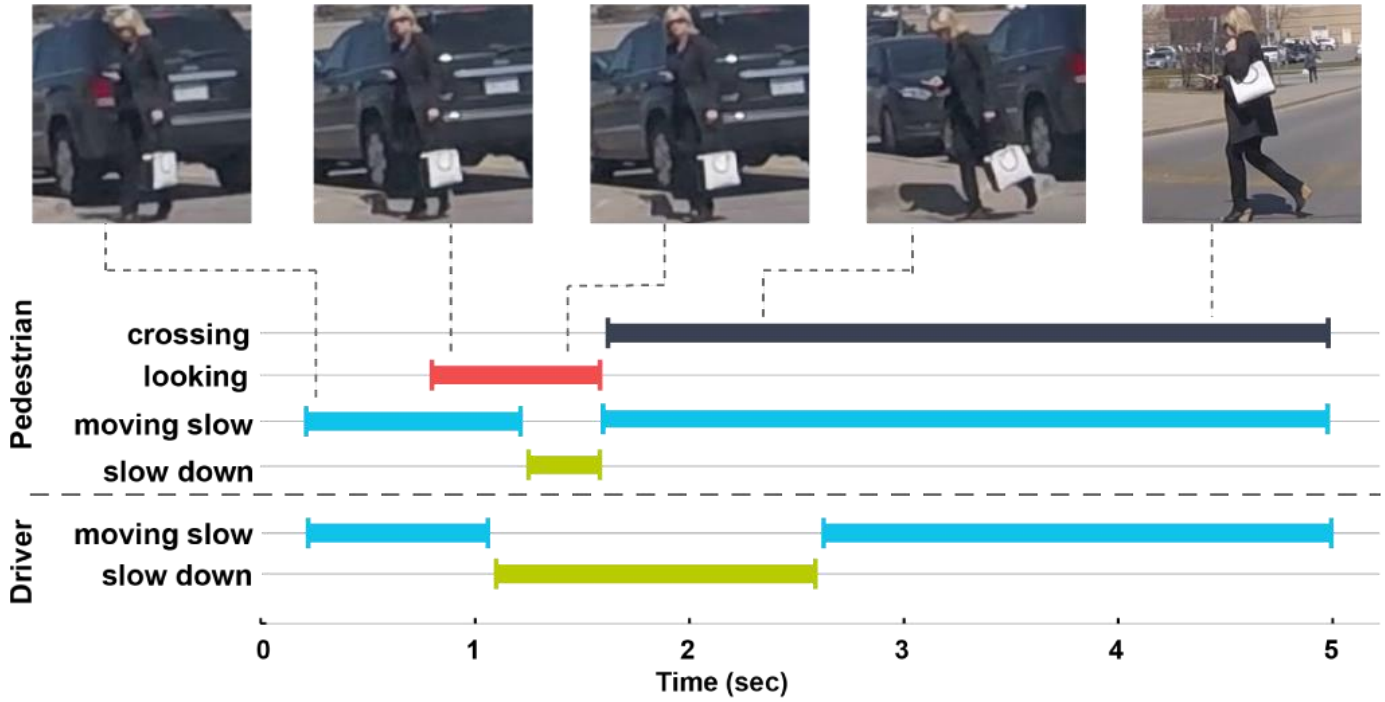
This represent a lighter version of the inspiring work:

[1]*"Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention"* ,Dongfang Yang1,2,*, Haolin Zhang1,*, Ekim Yurtsever1, Keith Redmill1, *Senior Member, IEEE*, and U¨mit O¨zgu¨ner1, *Life Fellow, IEEE*

Lighter version motivations

- I proposed a more lightweight architecture compared to the cited work, designed to efficiently adapt to the available resources.
- This approach addresses challenges in managing large datasets, but requires making certain restrictive assumptions regarding inputs, model architecture, and test reliability.
- Specifically, due to limited access to high-end resources, I opted for a simplified version as a trade-off.

JAAD Dataset



- 346 short video clips (5-10 sec long) extracted from over 240 hours
- Annotations including: bounding boxes, vehicle speed, pedestrian id, occlusion, intent for behavioral data.

Evaluation metrics

- Accuracy: the correct predictions over the whole predicted sample

$$\frac{\text{correct predictions}}{\text{total predictions}}$$

- Precision: the ratio of the correct predictions over the whole correct samples

$$\frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall: the ratio of correct predictions for a class to the total number of cases in which it occurs

$$\frac{\text{TP}}{\text{TP} + \text{FN}}$$

Evaluation metrics

- F1-score: the Harmonic mean between Precision and Recall

$$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

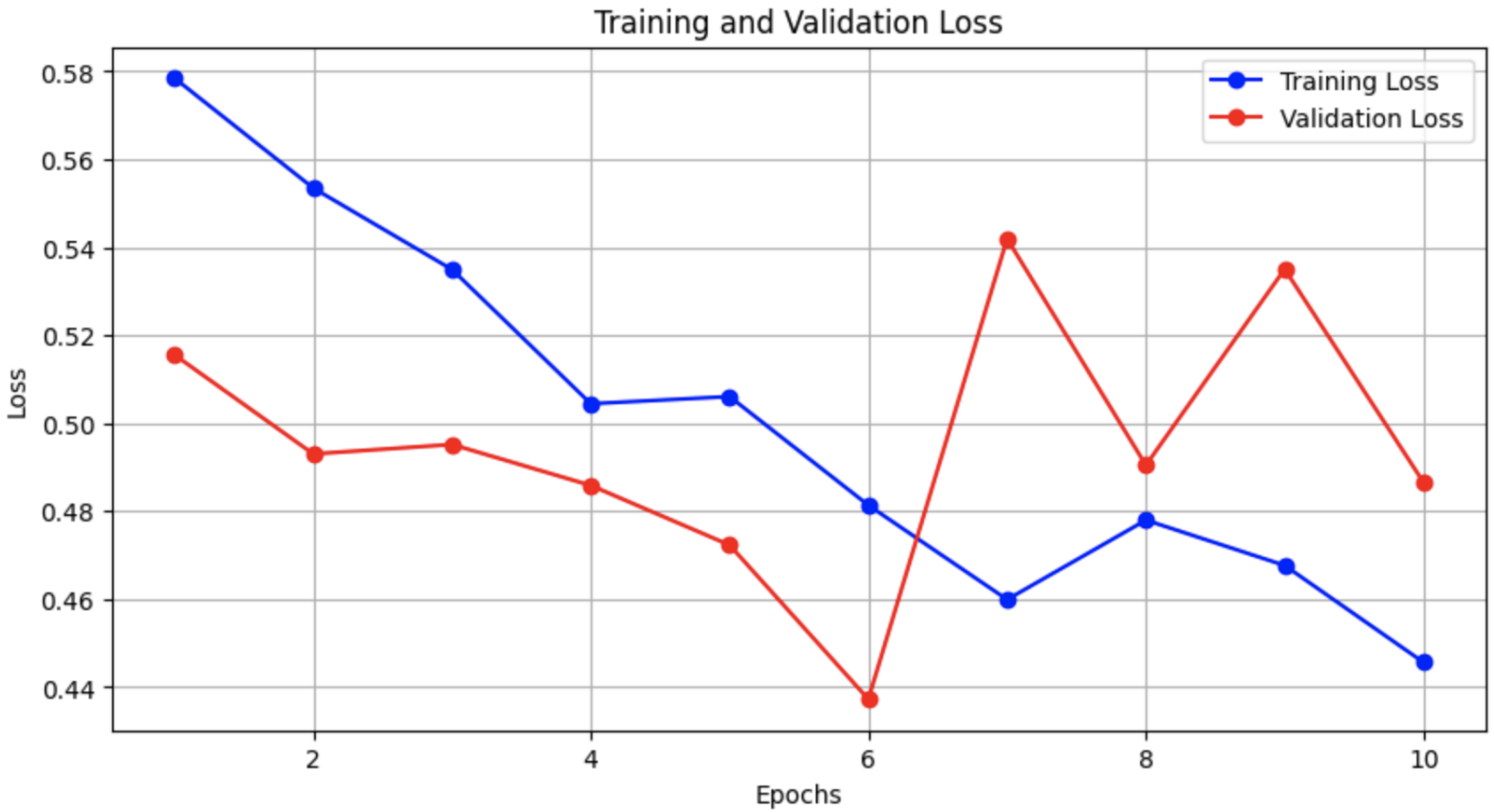
- Loss: Binary cross entropy

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \cdot log\hat{Y}_i + (1 - Y_i) \cdot log(1 - \hat{Y}_i))$$

Implementation details

- Pre-trained VGG-19 is used in the Visual branch to extract local context features
- Attention mechanisms are used for better memorizing sequential sources
- Hierarchical fusion
- Regularization techniques: Dropout, L2 penalization,Data augmentation
- Prediction step fixed to n=8 frame, corresponding to 0,2666 s

Results


Training and Validation Loss
Training Loss
Validation Loss
Loss
Epochs
0.58
0.56
0.54
0.52
0.50
0.48
0.46
0.44
2
4
6
8
10

Results

Label	Precision	Recall	F1-score	Test Samples
Not-crossing	0.81	0.94	0.87	54
Crossing	0.62	0.29	0.40	17

	Predicted NC	Predicted C
Actual NC	51	3
Actual C	12	5

Results

Model	Global context	Precision	Recall	F1-score	Dataset
Ours	✗	0.62	0.29	0.40	Reduced
Baseline	✓	0.51	0.81	0.63	All

Conclusion

- Based on the results, we can conclude that the model performs poorly on positive labels due to significant overfitting, as demonstrated by the Losses. Despite applying regularization techniques such as L2 norm penalization, Dropout, and Data Augmentation, the issue persists, likely due to the limited number of samples extracted in my work, which is insufficient given the large request of data by the implemented model.

- Overfitting appears to be the primary reason for the poor performance.

- Although i wanted to use cross-validation to perform a more reliable evaluation, it is impractical to apply it to this dataset given my available resources.

Future works

- Lighten the model in order to make it able to produce acceptable performance despite of data paucity
- Explore the effect of different kind of fusion strategies