Cross-Modal Attention for Accurate Pedestrian Trajectory Prediction

Mayssa Zaier*, Hazem Wannous, Hassen Drira, Jacques Boonaert
IMT Nord Europe, Univ. Lille, CNRS, UMR 9189 - CRIStAL,
F-59000 Lille, France
**mayssa.zaier@imt-nord-europe.fr*

Introduction

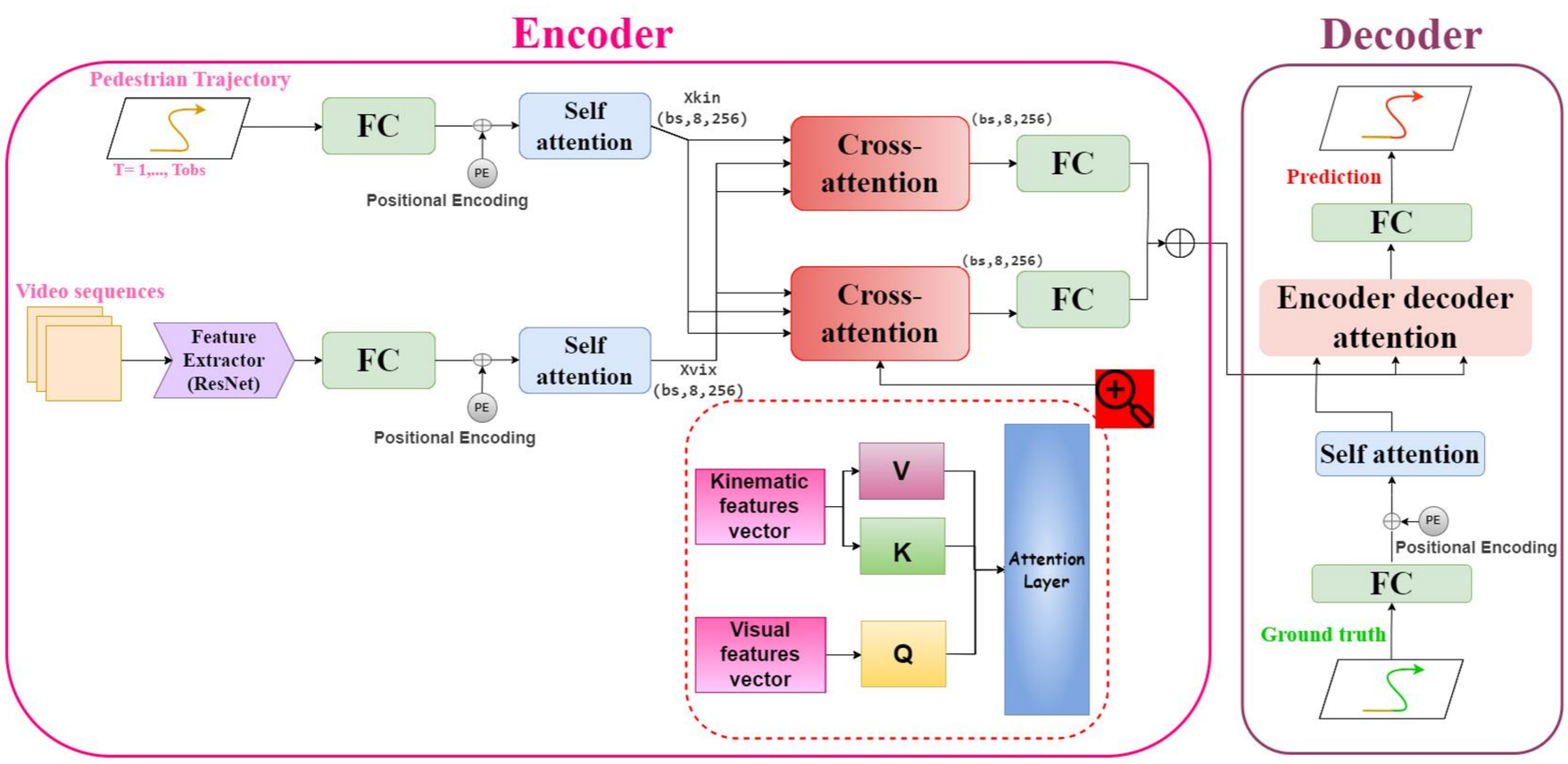
Context and Issues:

- Accurate prediction of human movement is crucial for applications, like autonomous driving and surveillance.
- Challenges include dynamic interactions between agents, complex environments, long-term dependencies, and the multimodality of human motion.
- Recent research has leveraged DL to enhance prediction accuracy but still faces difficulties, especially in capturing unexpected scenarios.
- Having a simple architecture using a 2D CNN combined with a transformer allows us to capture the dynamic scene context, understand both static and dynamic elements, and achieve better performance in scenarios with rapid motion changes and moving obstacles.

Contributions:

- We proposed CMATP, a novel approach leveraging Transformer Networks with attentive focus for trajectory prediction.
- CMATP fully leverages coordinates and video context through multimodal transformers, to capture dynamic visual information.
- Our architecture includes a cross-attention module for temporal consistency.
- It employs convolutional feature extraction and a bimodal transformer, enhancing prediction accuracy while maintaining computational efficiency by leveraging two input modalities.
- It efficiently captures intricate spatio-temporal interactions in dynamic scene contexts without additional computational power.

Approach



Results

Datasets:

- **ETH and UCY**: acquired from surveillance videos capturing pedestrians on sidewalks and annotated with location coordinates. They contain real-world pedestrian trajectories with rich human-human and human-object interaction scenarios.
  - **ETH [1]:** two scenes (ETH and Hotel) taken from a bird's eye view, with hundreds of pedestrian trajectories engaged in walking activities.
  - **UCY [2]:** three scenes (Zara1, Zara2, and Univ) taken from a bird's eye view with standing/walking activities.

Metrics:

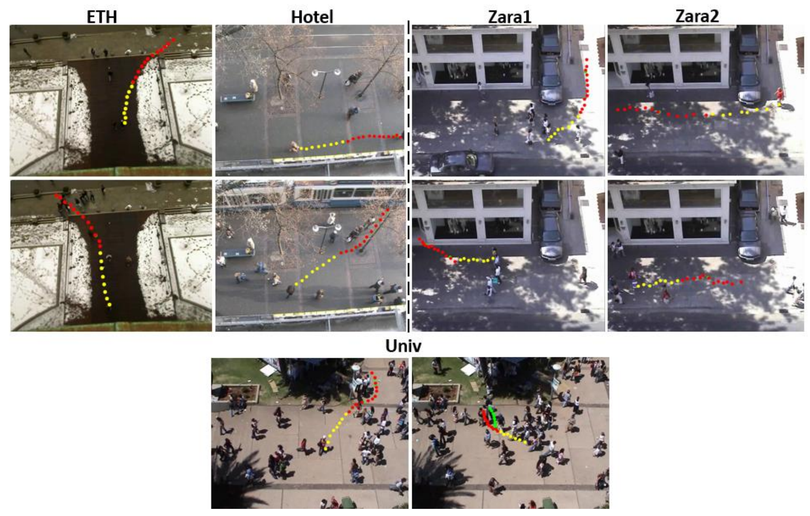
- Average Displacement Error (ADE)
- Final Displacement Error (FDE)

Method	Performance ADE/FDE ↓ (m)					
	Univ	Zara1	Zara2	Hotel	ETH	Avg
Linear*	0.82/1.59	0.62/1.21	0.77/1.48	0.39/0.72	1.33/2.94	0.79/1.59
LSTM*	0.61/1.31	0.41/0.88	0.52/1.11	0.86/1.91	1.09/2.41	0.70/1.52
Social-LSTM* [2]	0.67/1.40	0.47/1.00	0.56/1.17	0.79/1.76	1.09/2.35	0.72/1.54
Social-ATTN* [29]	0.33/3.92	0.20/0.52	0.30/2.13	0.29/2.64	0.39/3.74	0.30/2.59
TrafficPredict* [17]	3.31/6.37	4.32/8.00	3.76/7.20	2.55/3.57	5.46/9.73	3.88/6.97
SR-LSTM* [37]	0.51/1.10	0.41/0.90	0.32/0.70	0.37/0.74	0.63/1.25	0.45/0.94
DESIRE [13]	0.59/1.27	0.41/0.86	0.33/0.72	0.52/1.03	0.93/1.94	0.53/1.11
Social-GAN [8]	0.60/1.26	0.34/0.69	0.42/0.84	0.72/1.61	0.81/1.52	0.58/1.18
FSGAN [12]	0.54/1.14	0.35/0.71	0.32/0.67	0.43/0.89	0.68/1.16	0.46/0.91
SoPhie [23]	0.54/1.24	0.30/0.63	0.38/0.78	0.76/1.67	0.70/1.43	0.54/1.15
Trajectron [10]	0.54/1.13	0.43/0.83	0.43/0.85	0.35/0.66	0.59/1.14	0.47/0.92
MATF [38]	0.44/0.91	0.26/0.45	0.26/0.57	0.43/0.80	1.01/1.75	0.48/0.90
Next [16]	0.60/1.27	0.38/0.81	0.31/0.60	0.30/0.59	0.73/1.65	0.46/1.00
Social-BiGAT [11]	0.55/1.32	0.30/0.62	0.36/0.75	0.49/1.01	0.69/1.29	0.48/1.00
Social-STGCNN [19]	0.44/0.79	0.34/0.53	0.30/0.48	0.49/0.85	0.64/1.11	0.44 / 0.75
Social Ways [3]	0.55/1.31	0.44/0.64	0.51/0.92	0.39/0.66	0.39/0.64	0.46/0.83
PECNet [18]	0.35/0.60	0.22/0.39	0.17/0.30	0.18/0.24	0.54/0.87	0.29/0.48
M2P3 [21]	0.64/1.34	0.45/0.95	0.37/0.79	0.54/1.13	1.04/2.16	0.60/1.27
Transformer-TF [7]	0.35/0.65	0.22/0.38	0.17/0.32	0.18/0.30	0.61/1.12	0.31/0.55
STAR [34]	0.31/0.62	0.26/0.55	0.22/0.46	0.17/0.36	0.36/0.65	0.26/0.53
AgentFormer [35]	0.25/0.45	0.18/0.30	**0.14**/0.24	0.14/0.22	0.45/0.75	0.23/0.39
Trajectron++ [24]	0.30/0.54	0.25/0.41	0.18/0.32	0.18/0.28	0.67/1.18	0.32/0.55
SGN LSTM [36]	0.48/1.08	0.30/0.65	0.26/0.57	0.63/1.01	0.75/1.63	0.48/0.99
Introvert [26]	**0.20/0.32**	**0.16/0.27**	0.16/0.25	**0.11**/0.17	0.42/0.70	**0.21**/0.34
GroupNet [31]	0.26/0.49	0.21/0.39	0.17/0.33	0.15/0.25	0.46/0.73	0.25/0.44
Our model (CMATP)	0.37/0.52	0.19/**0.27**	**0.14/0.21**	**0.11/0.16**	**0.32/0.51**	0.22/**0.33**

ADE/FDE metrics obtained. Lower is better

Method	Performance ADE/FDE ↓ (m)					
	Univ	Zara1	Zara2	Hotel	ETH	Avg
Ours w/o CA (*BTT*)	0.36/0.52	0.19/0.29	0.15/0.23	0.12/0.17	0.48/0.81	0.26/0.40
Ours (*CMATP*)	0.37/0.52	0.19/**0.27**	**0.14/0.21**	**0.11/0.16**	**0.33/0.53**	**0.22/0.33**

Ablation study : Effect of Cross attention



Predicted trajectories: most predictions align closely with ground-truth data

Yellow dots: observation
Red dots: prediction
Green dots: ground truth

Conclusion

- CMATP: an attention-based Transformer Network for pedestrian trajectory prediction.
- Using attention mechanisms for dynamic scene context and a cross-attention mechanism to capture complex relationships, improving performance.
- Generating future-conditional predictions while respecting dynamic constraints and providing full probability distributions, making it suitable for robotic tasks.
- Demonstration of the effectiveness of Cross Attention in enhancing the model's performance.
- CMATP has significant potential for advancing pedestrian trajectory prediction and transportation safety.

References

[1] S Pellegrini, A Ess, K Schindler, and L van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In IEEE 12th International Conference on Computer Vision, pages 261–268, Kyoto, Sept. 2009. IEEE

[2] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by Example. Computer Graphics Forum, 26(3):655–664, 2007.