Benchmarking Trajectory Prediction Models

Semester project, VITA lab

École Polytechnique Fédérale de Lausanne

Student: Gojko Čutura
Mentor: Parth Kothari
Professor: Alexandre Alahi

June, 2022

1 Introduction

Understanding the social etiquette of human movement is a natural ability of humans, such as respecting personal space, yielding right-of-way, and avoiding passing across persons in the same group. Our social interactions cause a variety of complicated pattern-formation phenomena in crowds, such as the establishment of pedestrian lanes with uniform walking directions and pedestrian flow oscillations around bottlenecks. The ability to model social interactions and thus forecast crowd dynamics in real-world environments is extremely useful for a variety of applications, including infrastructure design, traffic operations, crowd abnormality detection systems, evacuation situation analysis, deployment of intelligent transport systems, and assisting in the deployment of intelligent transport systems. Modeling social interactions, on the other hand, is a difficult task because there is no set of rules that govern human movements. Forecasting the movement of the surrounding people who adhere to common social norms is a task directly related to studying human social interactions. This task is referred to as *human motion forecasting* (HTF).

Let us first define the necessary terminology of HTF, i.e. the *Trajectory* and *Scene*. A *Trajectory* of a single pedestrian is defined as a set of its attributes (coordinates, velocity, or more complicated ones) collected over a span of time. Furthermore, a *Scene* is defined as a collection of trajectories of multiple pedestrians interacting in a social setting. A scene may also comprise physical objects and non-navigable areas that affect the human trajectories, e.g., walls, doors, and elevators. Now, having defined these important elements, we define human trajectory forecasting as:

Given past trajectories of all humans in a scene, forecast the future trajectories which conform to the social norms.

Before moving on to specific contributions of this project, we first define the precise problem statement of HTF as well as the most important metrics.

1.1 Problem Statement

As previously explained, the main goal is to forecast the future trajectories of all the pedestrians in a scene. A given trajectory forecasting model takes as input the trajectories of all the people in a scene denoted by $\mathbf{X} = \{X_1, X_2, ..., X_n\}$, and its task is to forecast future trajectories $\mathbf{Y} = \{Y_1, Y_2, ..., Y_n\}$. The position and velocity of pedestrian i at time-step t is denoted by $\mathbf{x}_i^t = (x_i^t, y_i^t)$ and $\mathbf{v}_i^t$ respectively. The model has access to the attributes of all pedestrians at time-steps $t = 1, ..., T_{obs}$ and wants to predict the future positions from time-steps $t = T_{obs+1}, ..., T_{pred}$. We denote the predictions of the model using $\hat{\mathbf{Y}}$. To view the problem a bit more generally, we denote the state of pedestrian i at time-step t as $\mathbf{s}_i^t$. the state can refer to different attributes of a person other than coordinates, such as the body pose. Additionally, trajectory forecasting models can also be design to predict the k most likely future trajectories.

2 TrajNet++

To demonstrate the efficacy of a trajectory forecasting model, the standard practice is to evaluate these models against baselines on a standard benchmark. However, current methods have been evaluated on different subsets of available data without proper sampling of scenes in which social interactions occur. In other words, a data-driven method cannot learn to model agent-agent interactions if the benchmark comprises primarily of scenes where the agents are static or move linearly. TrajNet++ [1] is an interaction-centric human trajectory forecasting benchmark that tries to alleviate these issues by focusing on a dataset that comprises largely of scenes where social interactions occur.

2.1 Evaluation Metrics

In addition to comprising well-sampled trajectories, TrajNet++ provides an extensive evaluation system to understand model performance better.

Unimodal Evaluation. Unimodal Evaluation: Unimodal evaluation refers to the evaluation of models that propose a single future mode for a given past observation. The most commonly used metrics of human trajectory forecasting in the unimodal setting are Average Displacement Error (ADE) and Final Displacement Error (FDE) defined as follows:

- **Average Displacement Error (ADE)**: Average L_2 distance between ground-truth and model prediction over all predicted time steps.
- **Final Displacement Error (FDE)**: The distance between the predicted final destination and the ground truth final destination at the end of the prediction period T_{pred}.

Multimodal Evaluation. For models performing multimodal forecasting, i.e., outputting a future trajectory distribution, the following metrics are used to measure their performance:

- **Top-k ADE**: Given k output predictions for an observed scene, this metric calculates the ADE of the prediction closest to the ground truth trajectory.
- **Top-k FDE**: Given k output predictions for an observed scene, this metric calculates the FDE of the prediction closest to the ground truth trajectory.

Collision metrics: Having in mind the nature of the task - forecasting trajectories in a social setting, one of the most important aspects of human behavior in crowded spaces is collision avoidance. To ensure that models forecast feasible collision-free trajectories, TrajNet++ proposes two new collision-based metrics:

- **Collision I - Prediction collision (Col-I)**: This metric calculates the percentage of collision between the primary pedestrian and the neighbors in the forecasted future scene. This metric indicates whether the predicted model trajectories collide, i.e., whether the model learns the notion of collision avoidance.
- **Collision II - Ground truth collision (Col-II)**: This metric calculates the percentage of collision between the primary pedestrian's prediction and the neighbors in the groundtruth future scene.

Note: please note that in this project we only report the Col-I collision metrics.

3 TrajNet++ Model Zoo

With the goal of utilizing TrajNet++ in order to objectively and thoroughly compare different relevant models from the field of human trajectory forecasting, we propose the TrajNet++ Model Zoo, which is in fact the primary focus of this project. Namely, we provide a public github repository in which HTF methods are compared to each other, along with links to our modified repositories (in order to adapt the original implementation for the TrajNet++ benchmark loaders and evaluators). As of the moment of writing this report, we have evaluated 6 external models, each of which will be briefly covered in the following section.

4 Related Work

In this section, we summarize the main ideas and contributions from the methods that were evaluated on TrajNet++ and added to the Model Zoo during this semester project. The paper Social Ways [2] will be excluded both from the related work and the evaluation sections due to its poor performance on our benchmark, potentially due to the unreliable official implementation of the method.

Trajectory Tranformer. [3] The authors of the Trajectory Transformer bring into question the wide use of LSTM models for the task of human trajectory forecasting. Instead of that approach, they propose a novel use of Transformer Networks and argue that this is a fundamental switch from the sequential step-by-step processing of LSTMs to the only-attention-based memory mechanisms of Transformers. Finally, they claim that their TF model "without bells and whistles" yields the best score on the TrajNet benchmark.

STGAT. [4] In STGAT, it is argued that most of the (at the time) existing methods ignore the temporal correlations of interactions with other pedestrians involved in a scene. To address this issue, they propose a Spatial-Temporal Graph Attention network (STGAT), based on a sequence-to-sequence architecture to predict future trajectories of pedestrians. Besides the spatial interactions captured by the graph attention mechanism at each time-step,

they adopt an extra LSTM to encode the temporal correlations of interactions. Finally, they claim that their method outperforms the state of the art methods on the ETH and UCY datasets, and produces more "socially" plausible trajectories for pedestrians.

Social-STGCNN. [5] Similarly to STGAT, the main contribution of Social-STGCNN is incorporating the interactions between pedestrians in the forecasting model. Thay propose the Social Spatio-Temporal Graph Convolutional Neural Network (Social-STGCNN), which substitutes the need of aggregation methods by modeling the interactions as a graph. They claim that their results show an improvement over the state of the art by 20% on FDE and an improvement on the ADE with 8.5 times less parameters and up to 48 times faster inference speed than previously reported methods.

Causal-HTP. [6] In this paper, it is stated that most existing methods learn to predict future trajectories by behavior clues from history trajectories and interaction clues from environments. However, the authors argue that the inherent bias between training and deployment environments is ignored. Hence, they propose a counterfactual analysis method for human trajectory prediction to investigate the causality between the predicted trajectories and input clues and alleviate the negative effects brought by the environment bias. Their main idea is to cut off the inference from environment to trajectory by constructing the counterfactual intervention on the trajectory itself. Finally, it is claimed that their method achieves consistent improvement for different baselines and obtains the state-of-the-art results on public pedestrian trajectory forecasting benchmarks.

SR-LSTM. [7] The authors of SR-LSTM state that (at the time) recent studies based on LSTM networks have shown great ability to learn social behaviors. However, they argue that many of these methods rely on previous neighboring hidden states but ignore the important current intention of the neighbors. In order to address this issue, they propose a data-driven state refinement module for LSTM network (SR-LSTM), which activates the utilization of the current intention of neighbors, and jointly and iteratively refines the current states of all participants in the crowd through a message passing mechanism. They claim to have achieved state-of-the-art results at the time of publishing, on ETH and UCY datasets.

5 Evaluation and Results

In table 1 we summarize the performances that were reported for each of the methods in their original papers. It is important to note that, due to the fact that we considered datasets ETH, UNIV and ZARA2 while evaluating on TrajNet++, we took the average reported performances (ADE/FDE) on these particular datasets from the papers as well, and those are the values reported in 1.
In table 2 we report the performances of the evaluated models on TrajNet++. For the sake of comparison with the original papers, we report the Top-20 ADE/FDE metric along with the main metrics that Trajnet++ focuses on: Top-3 ADE/FDE and collision (Col-I). Finally, please note that, for easier analysis, all the rows are sorted based on their performances, from the best performing to the worst performing.

	Method Name	Performance
Top-20 ADE/FDE	Trajectory Tranformer	0.38 / 0.70
	Social-STGCNN	0.46 / 0.79
	Causal-STGAT	0.47 / 0.89
	STGAT	0.49 / 0.94
	SR-LSTM	0.49 / 1.02

Table 1. Results of evaluated methods as reported in the original papers.

	Method Name	Performance
Top-20 ADE/FDE	STGAT	0.34 / 0.68
	Trajectory Transformer	0.36 / 0.72
	Social-STGCNN	0.42 / 0.84
	Causal-STGAT	0.46 / 0.97
	SR-LSTM *	0.80 / 1.58
Top-3 ADE/FDE	Social-STGCNN	0.61 / 1.18
	STGAT	0.65 / 1.35
	Causal-STGAT	0.65 / 1.35
	Trajectory Transformer	0.70 / 1.45
	SR-LSTM *	0.80 / 1.58
Collision (Col-I)	SR-LSTM	10.84
	Causal-STGAT	10.87
	STGAT	11.28
	Trajectory Transformer	13.86
	Social-STGCNN	16.78

Table 2. Results of evaluated models as performed on the TrajNet++ benchmark

References

[1] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," 2020.

[2] J. Amirian, J.-B. Hayet, and J. Pettre, "Social ways: Learning multi-modal distributions of pedestrian trajectories with gans," 2019.

[3] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," 2020.

[4] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," 2019.

[5] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction," 2020.

[6] G. Chen, J. Li, J. Lu, and J. Zhou, "Human trajectory prediction via counterfactual analysis," 2021.

[7] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," 2019.

6 Appendix

- Official implementation - Github link
- Trained models - Google Drive link