Transformer Networks for Trajectory Forecasting

Francesco Giuliari, Irtiza Hasan, Marco Cristani, Fabio Galasso

Inha University, Informatics Lab
Keywoong Bae

I. Introduction

About Trajectory Forecasting

Pedestrian Forecasting, the goal of predicting future people motion given their past trajectories, has been steadily growing in attention by the research community

In Trajectory Forecasting, LSTM usually is used.

However, LSTMs have been targets of criticism

1. LSTM's **memory mechanism** has been criticized
2. LSTM's capability of **modelling social interaction** has been criticized

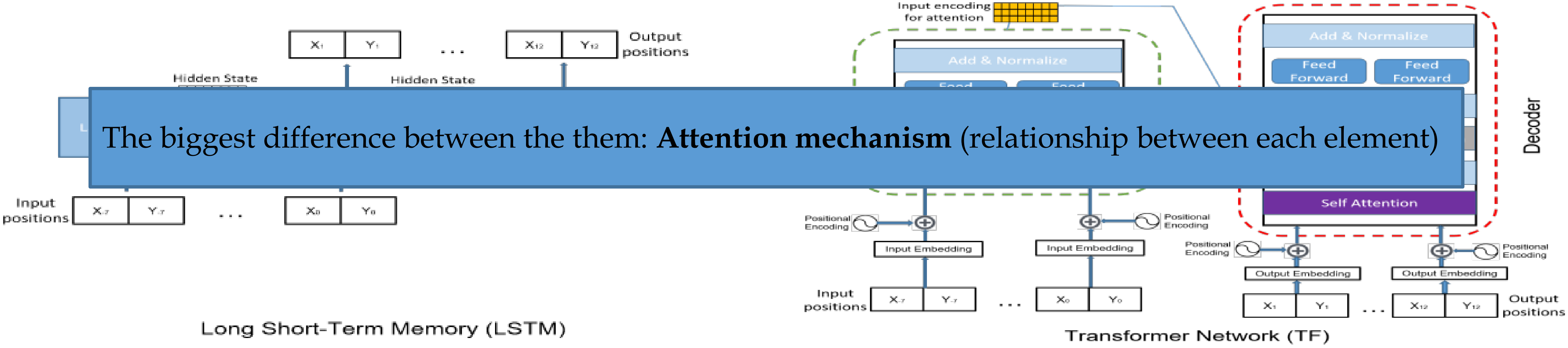
So, this paper forecast people individual's trajectories using Transformer Network

I. Introduction

What is Transformer Network?

Transformer Network

1. Used for Natural Language Processing (NLP) modeling Word Sequence.
2. Used in the process of answering questions or completing sentences using translation.



The biggest difference between the them: **Attention mechanism** (relationship between each element)

By using Original Transformer Network and Bidirectional Transformer (BERT) Model, forecast the individual's trajectory

I. Introduction

Difference between LSTM and TF in Trajectory Forecasting

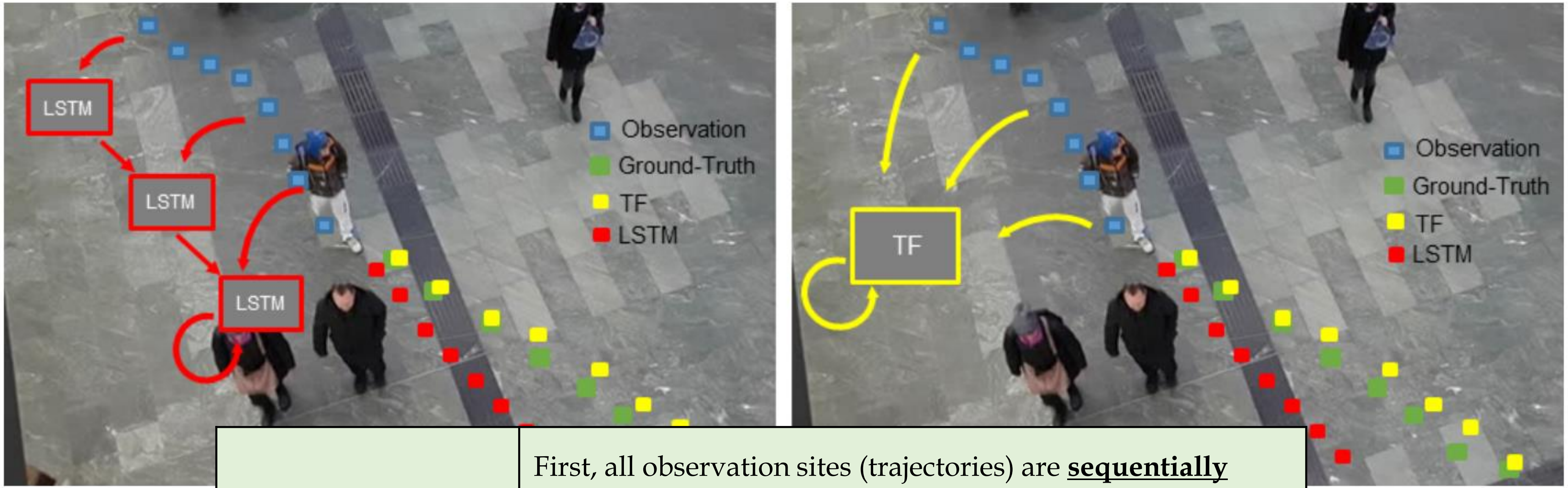


Fig. 1. People trajec ... ation interval (*blue dots*).
LSTM (left) sequenti ... ions.

LSTM	First, all observation sites (trajectories) are **<u>sequentially</u>** input and processed. After that, (automatic regression) prediction is performed.
Transformer Network	Examine all possible observations and predict the trajectories while assigning weights according to importance using an ***attention mechanism***.

II. Related Work

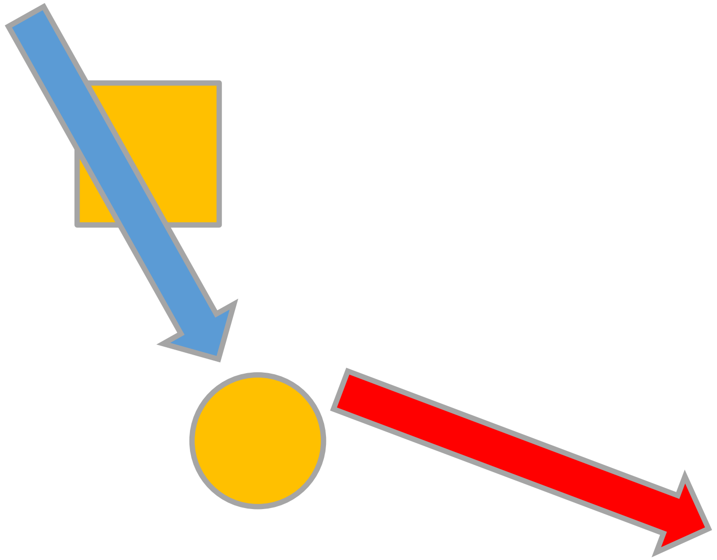
(1) Sequence modelling

Past Trajectory Forecasting	Recent Trajectory Forecasting
- Hand crafted energy-based optimization approaches	- Data driven approaches
- Linear analysis - Gaussian regression model - Time series analysis - Automatic regression analysis	- LSTM and RNN analysis techniques trained with copious amounts of data - Gaussian LSTM : regress directly the predicted value or produce mean, covariance of (x,y) to express the uncertainty of prediction

Because of TF's better capability to learn non-linear patterns, TFs are most suitable to sequence modelling

II. Related Work

(2) Social models and context



social interaction and scene context among people
(Tracking dynamics and spatio-temporal relations among people)

LSTM's ability limits the <u>generalization ability</u> of the model.

In this paper, assume that TF excludes social and environmental interactions and focuses on predicting individual movements.

IV. Experimental Evaluation

Introduction about experiment

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

Experiments		Dataset used	Test model
Exp1 TrajNet Challenge dataset		TrajNet dataset	22 models including Transformer and BERT.
Exp2 ETH+UCY dataset		ETH dataset, UCY dataset	- LSTM-based models (individual, social) - Transformer-based models (individual)
Exp3 Ablation Study	(1) Changing the Prediction Lengths	ETH and Zara01 datasets that are not part of the TrajNet training set	Transformer, LSTM
	(2) Missing and noisy data	TrajNet dataset	Transformer
	(3) Qualitative Results	TrajNet dataset	Transformer, LSTM, TF_q

IV. Experimental Evaluation

Experiment 1	Dataset used	Test Model
TrajNet Challenge dataset	TrajNet dataset	22 models including Transformer and BERT.

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

What is the TrajNet Dataset?

Trajectory Dataset	Explanation
BIWI Hotel dataset	orthogonal bird's eye flight view, moving people
Crowds UCY dataset	3 datasets, tilted bird's eye view, camera mounted on building or utility poles, moving people
MOT Pets dataset	multisensory, different human activities
Stanford Drone dataset	8 scenes, high orthogonal bird's eye flight view, different agents as people, cars etc

IV. Experimental Evaluation

Experiment 1	Dataset used	Test Model
TrajNet Challenge dataset	TrajNet dataset	22 models including Transformer and BERT.

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

How the experiment was conducted.

Observe

t-7	t-6	t-5	t-4	t-3	t-2	t-1	**t**	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12
(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)	(x,y)

Predict

IV. Experimental Evaluation

Experiment 1	Dataset used	Test Model
TrajNet Challenge dataset	TrajNet dataset	22 models including Transformer and BERT.

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

Results of the Experiments

Rank	Method	Avg	FAD	MAD	Context	Cit.	Year
2	*TF*	***0.776***	***1.197***	***0.356***	/		*2020*
3	REDv3	0.781	1.201	0.360	/	[14]	2019
4	REDv2	0.783	1.207	0.359	/	[14]	2019
6	RED	0.798	1.229	0.366	/	[14]	2018
7	SR-LSTM	0.816	1.261	0.37	s	[30]	2019
9	S.Forces (EWAP)	0.819	1.266	0.371	s	[39]	1995
12	N-Lin. RNN-Enc-MLP	0.827	1.276	0.377	/	[14]	2018
13	N-Lin. RNN	0.841	1.300	0.381	/	[14]	2018
15	Temp. ConvNet (TCN)	0.841	1.301	0.381	/	[10]	2018
16	*TF_q*	*0.858*	*1.300*	*0.416*	/		*2020*
17	N-Linear Seq2Seq	0.860	1.331	0.390	/	[14]	2018
18	MX-LSTM	0.887	1.374	0.399	s	[40]	2018
21	Lin. RNN-Enc.-MLP	0.892	1.381	0.404	/	[14]	2018
22	Lin. Interpolation	0.894	1.359	0.429	/	[14]	2018
24	Lin. MLP (Off)	0.896	1.384	0.407	/	[14]	2018
25	*BERT*	*0.897*	*1.354*	*0.440*	/	[16]	*2020*
26	*BERT_NLP_pretrained*	*0.902*	*1.357*	*0.447*	/		*2020*
27	S.Forces (ATTR)	0.904	1.395	0.412	s	[39]	1995

The transformer Network
has the smallest *Avg value*.
➔ TF network is the most accurate.

Metrics	Explanation
Rank	indicates the absolute ranking over all the approaches
Method	deep learning model name
Avg	$\frac{FAD + MAD}{2}$
FAD	(Final Average Displacement ,Final Displacement Error) check the goodness of the prediction at the last time step
MAD	(Mean Average Displacement ,Average Displacement Error) measuring the general fit of the prediction the ground truth, averaging the discrepancy at each time step
Context	Social Context, the trajectories of the other co-occurring people (‘s’ : consider, ‘/’ : don’t consider)
Year	Publishment Year

IV. Experimental Evaluation

Experiment 1	Dataset used	Test Model
TrajNet Challenge dataset	TrajNet dataset	22 models including Transformer and BERT.

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

Results of the Experiments

Rank	Method	Avg	FAD	MAD	Context	Cit.	Year
2	*TF*	*0.776*	*1.197*	*0.356*	/		*2020*
3	REDv3	0.781	1.201	0.360	/	[14]	2019
4	REDv2	0.783	1.207	0.359	/	[14]	2019
6	RED	0.798	1.229	0.366	/	[14]	2018
7	SR-LSTM	0.816	1.261	0.37	s	[30]	2019
9	S.Forces (EWAP)	0.819	1.266	0.371	s	[39]	1995
12	N-Lin. RNN-Enc-MLP	0.827	1.276	0.377	/	[14]	2018
13	N-Lin. RNN	0.841	1.300	0.381	/	[14]	2018
15	Temp. ConvNet (TCN)	0.841	1.301	0.381	/	[10]	2018
16	*TF_q*	*0.858*	*1.300*	*0.416*	/		*2020*
17	N-Linear Seq2Seq	0.860	1.331	0.390	/	[14]	2018
18	MX-LSTM	0.887	1.374	0.399	s	[40]	2018
21	Lin. RNN-Enc.-MLP	0.892	1.381	0.404	/	[14]	2018
22	Lin. Interpolation	0.894	1.359	0.429	/	[14]	2018
24	Lin. MLP (Off)	0.896	1.384	0.407	/	[14]	2018
25	*BERT*	*0.897*	*1.354*	*0.440*	/	[16]	*2020*
26	*BERT_NLP_pretrained*	*0.902*	*1.357*	*0.447*	/		*2020*
27	S.Forces (ATTR)	0.904	1.395	0.412	s	[39]	1995

For the top 4 models, they didn't consider social context(trajectories of co-occurring people)

Metrics	Explanation
Rank	indicates the absolute ranking over all the approaches
Method	deep learning model name
Avg	$\frac{FAD + MAD}{2}$
FAD	(Final Average Displacement ,Final Displacement Error) check the goodness of the prediction at the last time step
MAD	(Mean Average Displacement ,Average Displacement Error) measuring the general fit of the prediction the ground truth, averaging the discrepancy at each time step
Context	Social Context, the trajectories of the other co-occurring people ('s' : consider, '/' : don't consider)
Year	Publishment Year

IV. Experimental Evaluation

Experiment 1	Dataset used	Test Model
TrajNet Challenge dataset	TrajNet dataset	22 models including Transformer and BERT.

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

Results of the Experiments

Rank	Method	Avg	FAD	MAD	Context	Cit.	Year
2	*TF*	*0.776*	*1.197*	*0.356*	*/*		*2020*
3	REDv3	0.781	1.201	0.360	/	[14]	2019
4	REDv2	0.783	1.207	0.359	/	[14]	2019
6	RED	0.798	1.229	0.366	/	[14]	2018
7	SR-LSTM	0.816	1.261	0.37	s	[30]	2019
9	S.Forces (EWAP)	0.819	1.266	0.371	s	[39]	1995
12	N-Lin. RNN-Enc-MLP	0.827	1.276	0.377	/	[14]	2018
13	N-Lin. RNN	0.841	1.300	0.381	/	[14]	2018
15	Temp. ConvNet (TCN)	0.841	1.301	0.381	/	[10]	2018
16	*TF_q*	*0.858*	*1.300*	*0.416*	*/*		*2020*
17	N-Linear Seq2Seq	0.860	1.331	0.390	/	[14]	2018
18	MX-LSTM	0.887	1.374	0.399	s	[40]	2018
21	Lin. RNN-Enc.-MLP	0.892	1.381	0.404	/	[14]	2018
22	Lin. Interpolation	0.894	1.359	0.429	/	[14]	2018
24	Lin. MLP (Off)	0.896	1.384	0.407	/	[14]	2018
40	Gauss. Process	1.642	1.038	2.245	/	[42]	2010
42	N-Linear MLP (Off)	2.103	3.181	1.024	/	[14]	2018

The Quantized TF_q ranks 16^{th} due to quantization error.

Metrics	Explanation
Rank	indicates the absolute ranking over all the approaches
Method	deep learning model name
Avg	$\frac{FAD + MAD}{2}$
FAD	(Final Average Displacement ,Final Displacement Error) check the goodness of the prediction at the last time step
MAD	(Mean Average Displacement ,Average Displacement Error) measuring the general fit of the prediction the ground truth, averaging the discrepancy at each time step
Context	Social Context, the trajectories of the other co-occurring people ('s' : consider, '/' : don't consider)
Year	Publishment Year

IV. Experimental Evaluation

The Quantized Transformer

What is Quantization?	Converting continuous analog values to <u>discrete digital values</u>
Purpose & Advantage of Quantization	1. Improve the Compression rate 2. Can use small amounts of data 3. By reducing the number of weights, the size of the model can be significantly reduced.
Disadvantage of Quantization	1. Loss of information is inevitable 2. The signal accuracy is reduced.

24-bit(16.77 million colors)	**4-bit (16 colors)**
Although the numerical precision is incomparably reduced, the imagery in the photo is fairly well preserved.	

IV. Experimental Evaluation

Experiment 1	Dataset used	Test Model
TrajNet Challenge dataset	TrajNet dataset	22 models including Transformer and BERT.

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

Results of the Experiments

Rank	Method	Avg	FAD	MAD	Context	Cit.	Year
2	*TF*	*0.776*	*1.197*	*0.356*	/		*2020*
3	REDv3	0.781	1.201	0.360	/	[14]	2019
4	REDv2	0.783	1.207	0.359	/	[14]	2019
21	Lin. RNN-Enc.-MLP	0.892	1.381	0.404	/	[14]	2018
22	Lin. Interpolation	0.894	1.359	0.429	/	[14]	2018
24	Lin. MLP (Off)	0.896	1.384	0.407	/	[14]	2018
25	*BERT*	*0.897*	*1.354*	*0.440*	/	[16]	*2020*
26	*BERT_NLP_pretrained*	*0.902*	*1.357*	*0.447*	/		*2020*
27	S.Forces (ATTR)	0.904	1.395	0.412	s	[39]	1995
29	Lin. Seq2Seq	0.923	1.429	0.418	/	[14]	2018
30	Gated TCN	0.947	1.468	0.426	/	[10]	2018
31	Lin. RNN	0.951	1.482	0.420	/	[14]	2018
32	Lin. MLP (Pos)	1.041	1.592	0.491	/	[14]	2018
34	LSTM	1.140	1.793	0.491	/	[41]	2018
36	S-GAN	1.334	2.107	0.561	s	[5]	2018
40	Gauss. Process	1.642	1.038	2.245	/	[42]	2010
42	N-Linear MLP (Off)	2.103	3.181	1.024	/	[14]	2018

BERT is 2.2 times bigger than Transformer.
It requires a large amounts of datasets.
(The current dataset is not enough, so 25th is achieved)

Metrics	Explanation
Rank	indicates the absolute ranking over all the approaches
Method	deep learning model name
Avg	$\frac{FAD + MAD}{2}$
FAD	(Final Average Displacement ,Final Displacement Error) check the goodness of the prediction at the last time step
MAD	(Mean Average Displacement ,Average Displacement Error) measuring the general fit of the prediction the ground truth, averaging the discrepancy at each time step
Context	Social Context, the trajectories of the other co-occurring people (‘s’ : consider, ‘/’ : don’t consider)
Year	Publishment Year

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp2 ETH+UCY dataset	ETH dataset, UCY dataset	- LSTM-based models (individual, social) - Transformer-based models (individual)

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

ETH, UCY Dataset & How the experiment was conducted.

ETH dataset	UCY dataset
univ	zara01
hotel	zara02
	univ

It consists of 5 videos from 4 different scenes.
(Zara01 and Zara02 used the same camera but filmed at a different time)

1.
Every 0.4 seconds, one frame of trajectory data is generated.
Observing 8 frames (3.2 seconds) of the current reference past and predicting 12 frames (4.8 seconds) of the future.

2.
Observation and prediction by converting the original pixel position into (x, y) coordinates in meters using the isomorphic matrix announced by the author.

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp2 ETH+UCY dataset	ETH dataset, UCY dataset	- LSTM-based models (individual, social, soc+map) - Transformer-based models (individual)

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

Results of the Experiments

	LSTM-based				TF-based
	Individual	Social		Soc.+ map	Ind.
	S-GAN-ind [5]	S-GAN [5]	Trajectron++ [7]	Soc-BIGAT [6]	TF_q
ETH	0.81/1.52	*0.87/1.62*	*0.35/0.77*	*0.69/1.29*	**0.61 / 1.12**
Hotel	0.72/1.61	*0.67/1.37*	*0.18/0.38*	*0.49/1.01*	**0.18 / 0.30**
UCY	0.60/1.26	*0.76/1.52*	*0.22/0.48*	*0.55/1.32*	**0.35 / 0.65**
Zara1	0.34/0.69	*0.35/0.68*	*0.14/0.28*	*0.30/0.62*	**0.22 / 0.38**
Zara2	0.42/0.84	*0.42/0.84*	*0.14/0.30*	*0.36/0.75*	**0.17 / 0.32**
Avg	0.58/1.18	*0.61/1.21*	*0.21/0.45*	*0.48/1.00*	**0.31 / 0.55**

- TF technique yields a performance surprisingly
(more than best social techniques enclosing additional map information)

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp2 ETH+UCY dataset	ETH dataset, UCY dataset	- LSTM-based models (individual, social) - Transformer-based models (individual)

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study

Results of the Experiments

	LSTM-based				TF-based
	Individual	Social		Soc.+ map	Ind.
	S-GAN-ind [5]	S-GAN [5]	Trajectron++ [7]	Soc-BIGAT [6]	TF_q
ETH	0.81/1.52	*0.87/1.62*	*0.35/0.77*	*0.69/1.29*	**0.61 / 1.12**
Hotel	0.72/1.61	*0.67/1.37*	*0.18/0.38*	*0.49/1.01*	**0.18 / 0.30**
UCY	0.60/1.26	*0.76/1.52*	*0.22/0.43*	*0.55/1.32*	**0.35 / 0.65**
Zara1	0.34/0.69	*0.35/0.68*	*0.14/0.28*	*0.30/0.62*	**0.22 / 0.38**
Zara2	0.42/0.84	*0.42/0.84*	*0.14/0.30*	*0.36/0.75*	**0.17 / 0.32**
Avg	0.58/1.18	*0.61/1.21*	*0.21/0.45*	*0.48/1.00*	**0.31 / 0.55**

Straight Trajectory Datasets

- LSTM compares favorably with TF is on Zara1, which is the less structured of the datasets of the benchmark, mostly containing straight lines

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(1) Ablation Study Changing the Prediction Lengths	ETH and Zara01 datasets that are not part of the TrajNet training set	Transformer, LSTM

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(1/3)
Changing the Prediction Length

Results of the Experiments

Predict
from 12 steps to 32 steps

Pred.	TF (ours) MAD / FAD	LSTM [41] MAD / FAD
12	**0.71/1.56**	0.78/1.70
16	**0.95/2.15**	1.15/2.72
20	**1.27/2.90**	1.64/3.99
24	**1.66/3.76**	2.29/5.55
28	**2.27/5.09**	3.07/7.46
32	**2.98/4.52**	4.13/9.96

Performance Decreasing

from 0.71 to 2.98 (MAD)

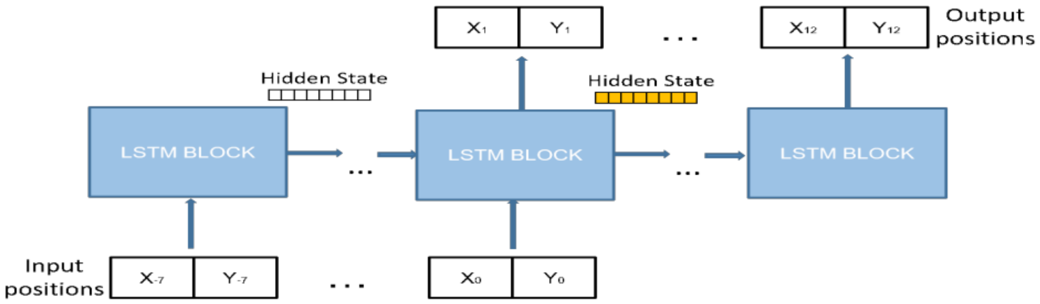
from 0.78 to 4.13 (MAD)

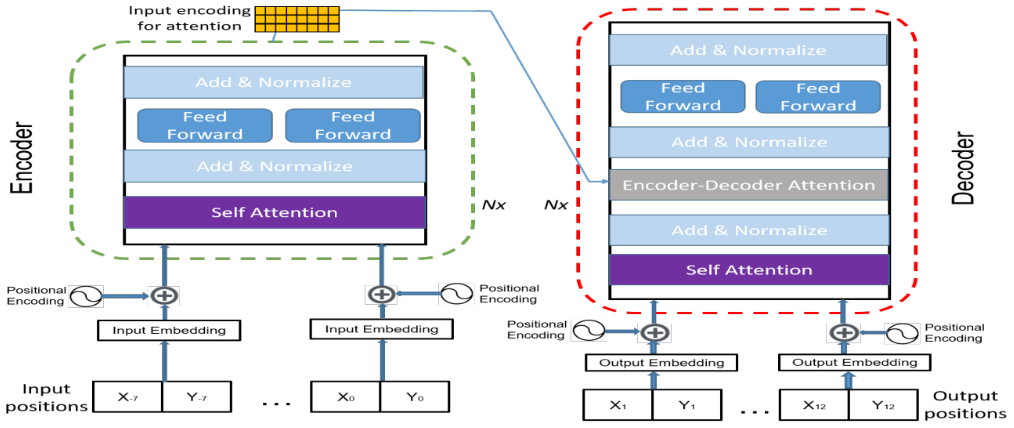
TF has a consistent advantage at every horizon

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(2) Ablation Study Missing and noisy data	TrajNet dataset	Transformer

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(2/3)
Missing and noisy data





Transformer Network (TF)

LSTM can't learn with missing data

Transformer can learn even with missing and noisy data

To replace missing data, it can use simple linear interpolation to improve the results

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(2) Ablation Study Missing and noisy data	TrajNet dataset	Transformer

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(2/3)
Missing and noisy data

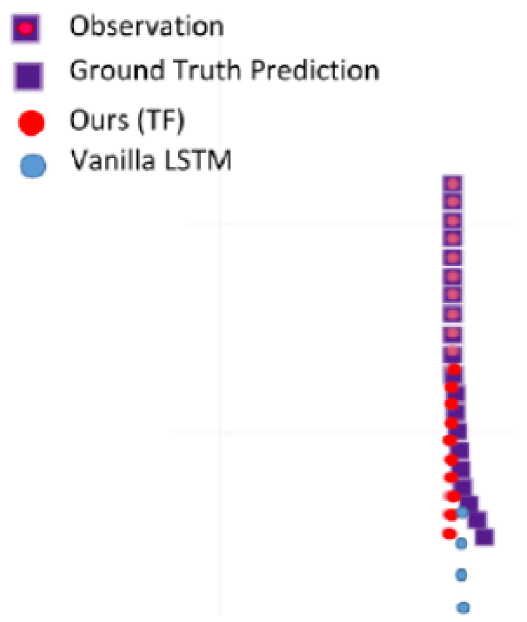
Results of the Experiments

# most recent frames dropped	Drop most recent obs. *including* current frame (FAD/MAD)	Drop most recent obs. *excluding* current frame (FAD/MAD)
0	1.197 / 0.356	1.197 / 0.356
1	1.305/ 0.389	1.267 / 0.373
2	1.409 / 0.429	1.29 / 0.38
3	1.602 / 0.495	1.303 / 0.384
4	1.787 / 0.557	1.313 / 0.387
5	1.897/ 0.593	1.327 / 0.329
6	2.128 / 0.669	1.377 / 0.406
	from 0.356 to 0.669 (91% degrades)	from 0.356 to 0.406 (16% degrades)

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(3) Ablation Study Qualitative results	TrajNet dataset	Transformer, LSTM, TF_q

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(3/3) Qualitative results



a)

(a) The subject going south, with a minimal acceleration

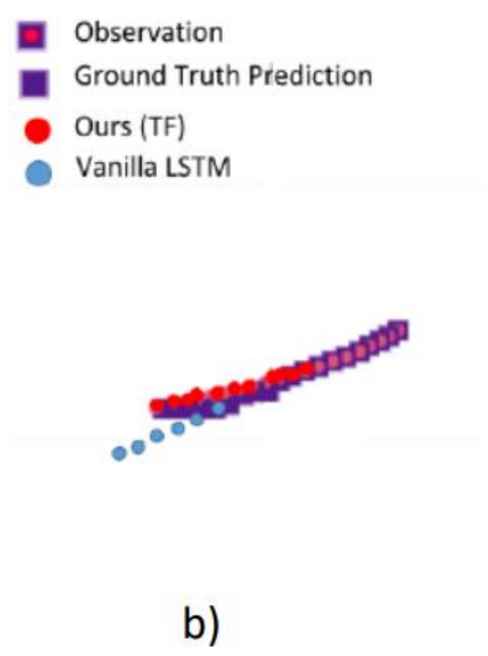
LSTM → Predicts a uniform acceleration toward south

TF → Captures better the dynamics (final direction is not correct)

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(3) Ablation Study Qualitative results	TrajNet dataset	Transformer, LSTM, TF_q

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(3/3)
Qualitative results



(b) The subject going south, with a minimal acceleration

LSTM → Predicts a faster straight trajectory

TF → Followed the bending of the GT more precisely

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(3) Ablation Study Qualitative results	TrajNet dataset	Transformer, LSTM, TF_q

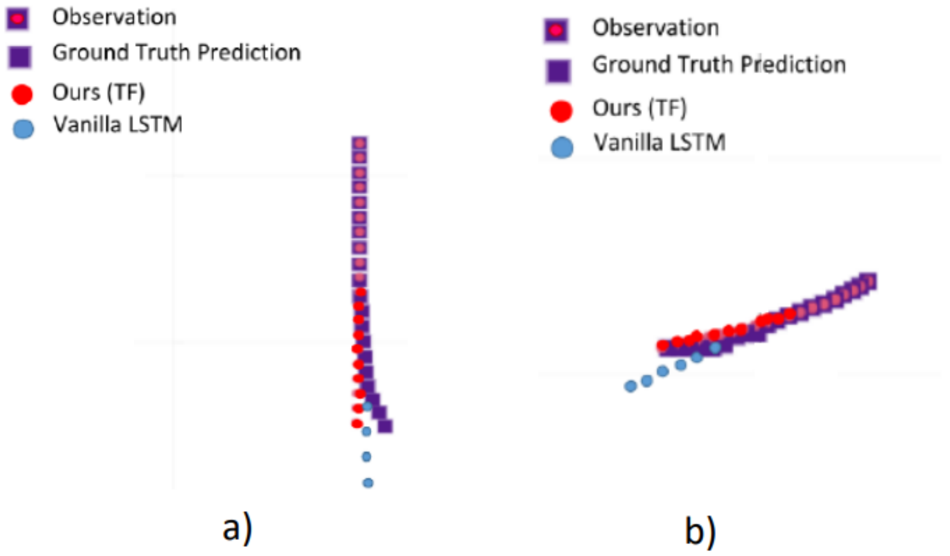
1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(3/3) Qualitative results

Results of the Experiments

- In general, we observed that LSTM generates trajectories way more regular than those predicted by TF

(This is certainly motivated by its unrolling, opposed to the encoder+decoder architecture of TF)

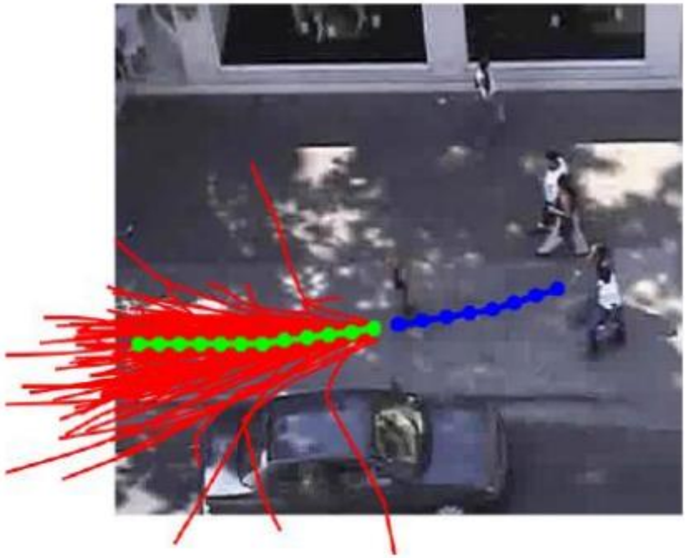
- LSTM is effective on straight trajectories (Zara1), but scarce on bending trajectories (Hotel)



IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(3) Ablation Study Qualitative results	TrajNet dataset	Transformer, LSTM, TF_q

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(3/3)
Qualitative results



c)

(c) Monomodal distribution of TF_q

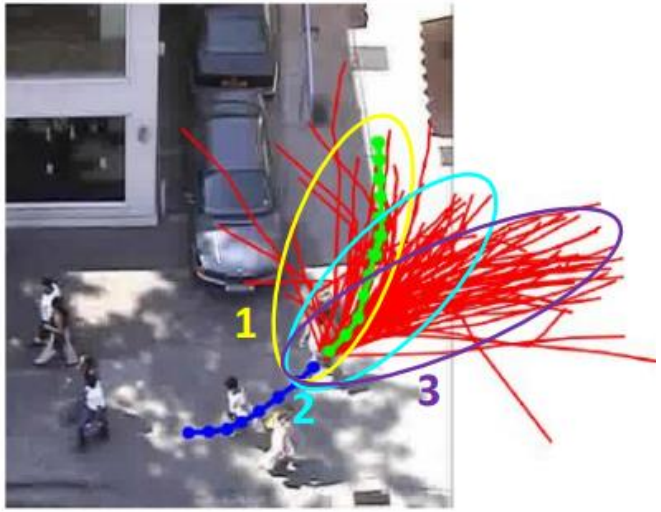
TF_q

Samples concentrated around the GT
They have low probability but are still plausible

IV. Experimental Evaluation

Experiment	Dataset used	Test Model
Exp3-(3) Ablation Study Qualitative results	TrajNet dataset	Transformer, LSTM, TF_q

1. TrajNet Challenge Dataset
2. ETH + UCY Dataset
3. Ablation Study(3/3) Qualitative results



d)

(d) shows that TF has learnt a multimodal distribution

TF_q → one turning north
the other going diagonal
the third(with larger number of them) going east