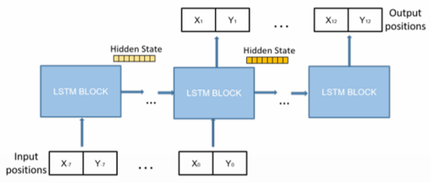
Transformer Networks for Trajectory Forecasting

Giuliari Francesco, Cristani Marco - University of Verona. {name.surname}@univr.it
Hasan Irtiza - Inception Institute of Artificial Intelligence. irtiza.hasan@inceptioniai.org
Galasso Fabio - Sapienza University of Rome. galasso@di.uniroma1.it

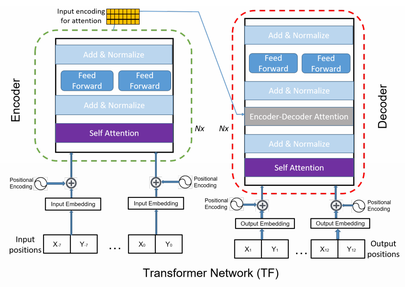
LSTM vs Transformer

LSTM



- Makes use of recurrent layers to deal with sequential data
- History encoded in a hidden state
  - Limited memory
  - Risk of forgetting initial observations
- No distinction between observed values and previous predictions
  - Prediction errors are fed back into the network and amplify over time

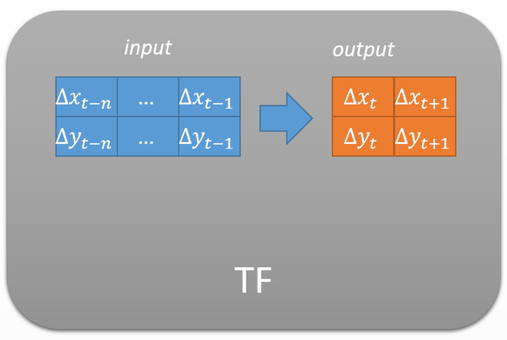
The Transformer[1]



- Use attention mechanisms to deal with sequential data
- All the input elements are available during inference
  - No loss of information
- Different treatment to observed values and previous predictions
  - Prediction errors can be kept in check

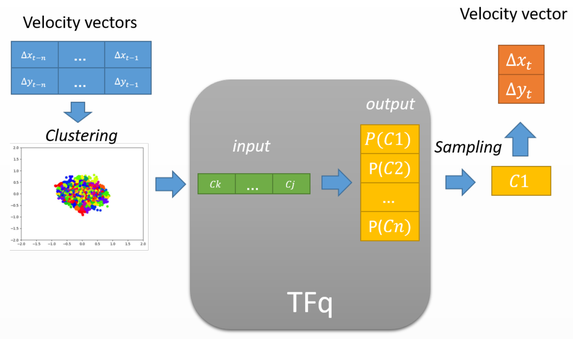
Models

We propose two models based on The Transformer Network[1]



TF for Accurate Single trajectory prediction

- Regression model
- *Input:* Sequence of Velocity vectors
- *Output:* Sequence of predicted velocity vectors



TF_q for Multi modal predictions

- Quantized version of TF
- *Input:* Sequence of velocity clusters ids
- *Output:* Probabilities over the velocity clusters

Multi Trajectory prediction: ETH+UCY

Comparison against SoA models following the best-of-20 protocol. Results are reported as MAD/FAD, with the standard protocol of 8 observation and 12 predictions.

Architecture	LSTM-based				TF-based
Used Features	Individual	Social		Soc.+ map	Ind.
Method name	S-GAN-ind	S-GAN	Trajectron++	Soc-BIGAT	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**

Only S-GAN-ind uses the same amount of information as our TFq, the other methods are reported for the sake of comparison against more complex methods.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, u., Polosukhin, I. (2017). Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000–6010).

Code

Code is available on Github:

Trajnet:Single prediction

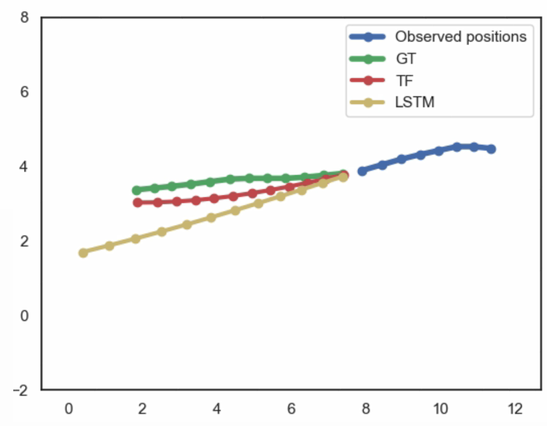
TrajNet Challenge results (world plane Human-Human).

Method	**FAD**	**MAD**	**Needs social cues**
TF	***1.197***	***0.356***	***no***
REDv3	1.201	0.360	no
SR-LSTM	1.261	0.37	yes
S.Forces (ewap)	1.266	0.371	yes
TF_q	*1.300*	*0.416*	*no*
BERT	*1.354*	*0.440*	*no*
BERT_NLP_pt.	*1.357*	*0.447*	*no*
MX-LSTM	1.374	0.399	yes
S.Forces (attr)	1.395	0.412	yes
LSTM	1.793	0.491	no
S-GAN	2.107	0.561	yes

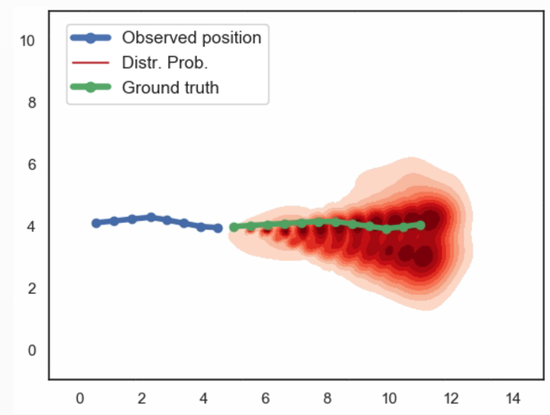
Blue italic indicates approaches proposed in this work.

Qualitative results

Our **TF** is able to predict the motion with much higher accuracy than standard LSTM based models.



Our **TF_q** can predict true multi-modal trajectories in a true data-driven manner, with no information about the underlying distribution.



Acknowledgements

This work is partially supported by the Italian MIUR through PRIN 2017 - Project Grant 20172BH297: I-MALL - improving the customer experience in stores by intelligent computer vision, and by the project of the Italian Ministry of Education, Universities and Research (MIUR) ”Dipartimenti di Eccellenza 2018-2022”.