2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Efficiency Analysis of Multi-Head Attention Models for Social Dynamics Prediction

Ha Q. Ngo, Christoph Henke, Frank Hees

Abstract— **Our research aim is, to investigate the problems of dynamic environment understanding for service robot navigation in social areas. This paper presents an efficiency analysis of multi-head attention models for real-time social dynamics prediction on limited computational hardware.**

Keywords: attention-based learning, social dynamics modeling and prediction, situation understanding, human-robot interaction, autonomous navigation in mixed traffic.

I. Introduction

In the scenario of human-robot crossing spaces, autonomous robots navigate complex social dynamic environments while avoiding collision and following social norms. In this context, a service robot needs to perform not only a high-quality perception of the surrounding static physical scene but also an accurate prediction of the surrounding dynamics. Since social areas are usually dense and highly dynamic, automatic trajectory prediction is an extremely challenging problem, which takes into account the complexity of diverse input data. The challenge is how to efficiently exploit the input data for accurate modeling and prediction of social dynamics to close the gap between perception and situation understanding for autonomous navigation.

II. Terminology and Background

This research challenge has drawn the great interest of the Robotics community, where people try to bring autonomous service robots closer to daily life [1, 2]. In recent years, learning-based methods [3-6] have outperformed traditional methods such as *Social Forces* [7] or *Markov Decision Process–based* methods [8, 9]. Since social dynamics can be represented as time-series data, *Recurrent Neural Networks* such as *Long-Short Term Memory Networks (LSTMs)* are able to predict the joint distribution of future trajectories [10-19]. However, this method has shown its limitation in prediction accuracy due to its recurrent nature and architecture complexity. Therefore, learning of long-term data dependencies in complex time-series processes remains as an ill-posed problem.

As a solution, *Attention-based Models* have shown their abilities to overcome the mentioned above problem [20]. This method has proven great results in *Natural Language Processing (NLP)* domain by replacing recurrence completely by attention mechanisms, which make the network able to learn the relationships between all data tokens (i.e., parts of data). As the successors in the NLP research domain, Transformer [21] and its variants can also learn the recurrent nature of sequential data using Positional Encoding [22-25]. Transformer introduces a *Multi-Head Attention (MHA)* architecture, which is the concatenation of several Single-Head Attention (SHA) modules. This lets Transformer-based networks learn different semantic meanings of attention, which can significantly improve prediction accuracy, while modeling complex contexts. However, all-to-all comparisons are expensive and lead to an increase in model size, which is a big disadvantage of Transformer-based models for real-time applications.

This paper investigates the problems of using attention-mechanisms for social dynamics prediction by adopting the power of the Transformer network. We deliver an experimental efficiency analysis of different MHA architecture configurations for small-size models.

III. Methodology

A. Problem formulation

Imagine that an autonomous service robot is moving among n other traffic agents (e.g. cars, buses, cyclists, pedestrians, etc.). In order to understand the surrounding social dynamics, the robot needs to predict the future states of surrounding agents. Given an observation of state history $S = \{S_i\}_{i=1}^{n}$, where $S_i = \left\{\{s_j^i\}_{j=1}^{m}\right\}_{t=-T}^{0}$ is the state history of i^{th} agent characterized by m features over previous T time steps (**figure 1**). We predict the future states $\widehat{S} = \{\hat{S}_i\}_{i=1}^{n}$,

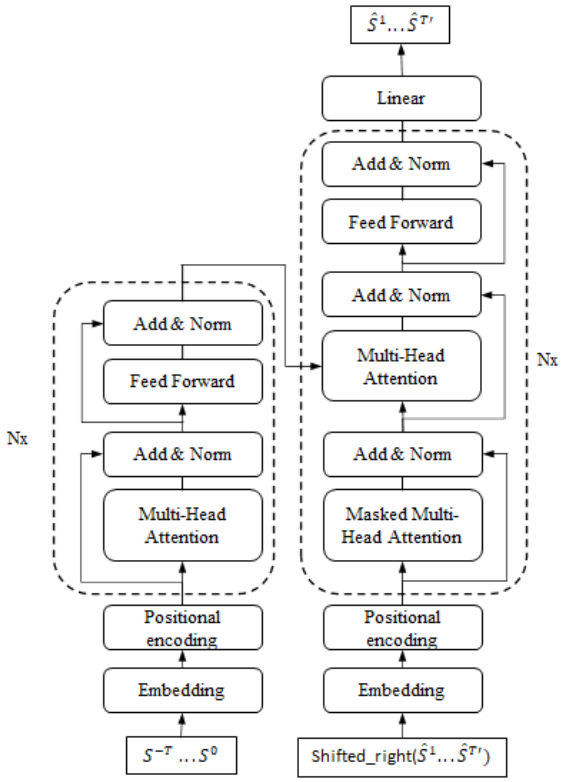

Figure 1 – Transformer-based model for social dynamics prediction

Ha Q. Ngo, Christoph Henke and Frank Hees are with the Institute for Management Cybernetics, 52068 Aachen, Germany. (Corresponding author's e-mail: *quangha.ngo@ ifu.rwth-aachen.de*).

where $\widehat{S}_i = \left\{\{s_j^i\}_{j=1}^m\right\}_{t=1}^{T'}$ are the predicted states of each agent for next T' time steps .

B. Transformer-based model

In this work, we apply a standard Transformer architecture, which consists of the following main components:

The *encoder* is a stack of N encoders to learn an implicit representation of input data. Each encoder is broken down to a MHA layer, a *Feed-Forward (FF)* layer with residual connections [26], and layer-normalization [27] in between. Each MHA is a concatenation of h Scaled Dot-Product Attention modules. This allows the network to learn different meanings of attention to provide a better context representation.

Similarly, the *decoder* is a stack of N decoders to learn the relationships of the previous outputs and generate predicted future states according to the input representation from the encoding component. Each decoder consists of a Masked-MHA layer, an MHA layer, and a FF layer. Each of these layers is also followed by a residual connection and layer-normalization.

The *Embedding and Positional encoding components* convert each observed token $s_i^t = \left(\{s_j\}_{j=1}^m\right)_i^t$ to a vector of dimension d_{model}: $e_i^t = L(s_i^t) + P(t)$, where L is a linear embedding layer and P is a positional encoding function using sine/cosine stamping.

IV. Experiments and Results

A. Datasets

We train the models on the *rounD datasets* [28], which are large datasets of naturalistic road users. Each data point of the datasets consists of 13 different features including 2D coordinates, heading, width, length, velocity, acceleration, etc.

B. Experiment setup and training

The experiments are conducted on a Geforce GTX1080Ti GPU. Small models (size <=7.2 MB) are tested for evaluation in this work since we consider applying attention-based models for real-time prediction on autonomous robots with limited hardware capacity. In order to explore the efficiency of using MHA models, we manipulate the embedding size d_{model}, the number of layers N, and the number of heads h for different network configurations. The training process for each test case continues until the prediction accuracy is not improved by 1% through the last 100 epochs.

C. Results

The experiment results are showed in Table I for six different model sizes by three different architecture configurations. We analyze the configuration efficiency based on the prediction accuracy over the number of attention heads. In this context, the lower average displacement value (Avg.Disp.) indicates the higher prediction accuracy.

The results in Table I show that MHA models (h>=2) always provide better results than (SHA) models (h=1) due to its ability to learn more attention meanings of the context. However, increasing the number of attention heads while keeping the model size does not guarantee for the improvement of prediction accuracy. In some test cases, when the number of attention heads is large (h=8), the prediction accuracy is not improved.

TABLE I. Experiment Results

d_{model}	N	h	Size (MB)	Avg.Disp (m)
32	1	1	2.4	0.8512
32	1	2	2.4	0.8506
32	1	8	2.4	0.8488
32	2	1	3.5	0.8432
32	2	2	3.5	0.8417
32	*2*	*8*	*3.5*	*0.8402*
32	3	1	4.7	0.8434
32	3	2	4.7	0.8416
32	3	8	4.7	0.8419
32	4	1	5.8	0.8423
32	4	2	5.8	0.8415
32	*4*	*8*	*5.8*	*0.8411*
64	1	1	4.9	0.8483
64	1	2	4,9	0.8462
64	1	8	4.9	0.8469
64	2	1	7.2	0.8433
64	*2*	*2*	*7.2*	***0.8422***
64	2	8	7.2	0.8423

d_{model} – embedding size, N – number of layers, h – number of attention heads, *Arg.Disp.* – average displacement (m), Size – model size(MB)

V. Discussion and Future Work

Since the prediction accuracy does not change monotonously over the raise of attention heads, it is still unclear how to archive the best prediction accuracy using MHA architecture. When the size of each attention head is not big enough, the model may not be able to learn the complex context correctly. Furthermore, since the query and key matrices of each attention head are initialized randomly and parallelly updated during the training process, different attention heads may refer to the same data subspaces, while some other data subspaces may be ignored. This may lead to the downgrading of the prediction results. Therefore, it is important to consider the optimal configuration for MHA models based on the model size limit and input data complexity. Unfortunately, experimental searching for optimal configuration is time-costly on large datasets. We propose to solve these problems in further works by developing a novel method for automatic searching of optimal MHA configurations.

VI. Conclusion

This work showed the advantages of MHA models against SHA models for social dynamics prediction. Our experimental analysis argued that MHA architecture is useful for improving prediction accuracy. However, it is expensive to search optimal configuration for MHA models in case of real-time applications with limited computational hardware for complex social dynamics prediction.

Acknowledgment

This work was conducted as part of the interdisciplinary project UrbANT, which is supported by the German Federal Ministry of Education and Research under the funding code 16SV7919.

REFERENCES

1. Kruse, T., et al. *Legible robot navigation in the proximity of moving humans*. in *2012 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*. 2012.
2. Rudenko, A., et al., *Human Motion Trajectory Prediction: A Survey.* arXiv:1905.06113 [cs], 2019.
3. Tai, L., et al., *Socially Compliant Navigation through Raw Depth Inputs with Generative Adversarial Imitation Learning.* arXiv:1710.02543 [cs], 2017.
4. Kretzschmar, H., et al., *Socially compliant mobile robot navigation via inverse reinforcement learning.* The International Journal of Robotics Research, 2016. 35(11): p. 1289-1307.
5. Long, P., W. Liu, and J. Pan, *Deep-Learned Collision Avoidance Policy for Distributed Multi-Agent Navigation.* arXiv:1609.06838 [cs], 2016.
6. Liu, Y., A. Xu, and Z. Chen, *Map-Based Deep Imitation Learning for Obstacle Avoidance.* p. 6.
7. Helbing, D. and P. Molnar, *Social Force Model for Pedestrian Dynamics.* Physical Review E, 1995. 51(5): p. 4282-4286.
8. Bennewitz, M., et al., *Learning Motion Patterns of People for Compliant Robot Motion.* The International Journal of Robotics Research, 2005. 24(1): p. 31-48.
9. Ziebart, B., et al. *Planning-based Prediction for Pedestrians*. in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*. 2009.
10. Alahi, A., et al. *Social LSTM: Human Trajectory Prediction in Crowded Spaces*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. IEEE.
11. Chandra, R., et al., *TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions.* arXiv:1812.04767 [cs], 2018.
12. Fernando, T., et al., *Soft + Hardwired Attention: An LSTM Framework for Human Trajectory Prediction and Abnormal Event Detection.* arXiv:1702.05552 [cs], 2017.
13. Gupta, A., et al., *Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks.* arXiv:1803.10892 [cs], 2018.
14. Vemula, A., K. Muelling, and J. Oh, *Social Attention: Modeling Attention in Human Crowds.* arXiv:1710.04689 [cs], 2017.
15. Sadeghian, A., et al., *SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints.* arXiv:1806.01482 [cs], 2018.
16. Hoshen, Y., *VAIN: Attentional Multi-agent Predictive Modeling.* arXiv:1706.06122 [cs], 2017.
17. Yingfan, H., et al. *STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction*. 2019.
18. Brownlee, J., *Encoder-Decoder Long Short-Term Memory Networks*, in *Machine Learning Mastery*. 2017.
19. Ngo, H.Q., C. Henke, and F. Hees, *An End-to-End Learning Approach for Trajectory Prediction in Pedestrian Zones.* ACM/IEEE International Conference on Human-Robot Interaction. Workshop on The Forgotten in HRI: Incidental Encounters with Robots in Public Spaces, 2020.
20. Luong, M.-T., H. Pham, and C.D. Manning, *Effective Approaches to Attention-based Neural Machine Translation.* 2015.
21. Vaswani, A., et al., *Attention Is All You Need.* arXiv:1706.03762 [cs], 2017.
22. Dehghani, M., et al., *Universal Transformers.* 2018.
23. Devlin, J., et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* 2018.
24. Irie, K., et al., *Language Modeling with Deep Transformers.* 2019.
25. Zeyer, A., et al., *A Comparison of Transformer and LSTM Encoder Decoder Models for ASR.* IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2020.
26. He, K., et al., *Deep Residual Learning for Image Recognition.* 2015.
27. Ba, J.L., J.R. Kiros, and G.E. Hinton, *Layer Normalization.* 2016.
28. Krajewski, R.a.M., Tobias and Bock, Julian and Vater, Lennart and Eckstein, Lutz. *The rounD Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany*. 2020; Available from: https://www.round-dataset.com/.