Trajformer: Trajectory Prediction with Local Self-Attentive Contexts for Autonomous Driving

Manoj Bhat[1] **Jonathan Francis**[1,2*] **Jean Oh**[1]
[1]Carnegie Mellon University [2]Bosch Research Pittsburgh
{mbhat, jmf1, hyaejino}@andrew.cmu.edu

Abstract

Effective feature-extraction is critical to models' contextual understanding, particularly for applications to robotics and autonomous driving, such as multimodal trajectory prediction. However, state-of-the-art generative methods face limitations in representing the scene context, leading to predictions of inadmissible futures. We alleviate these limitations through the use of self-attention, which enables better control over representing the agent's social context; we propose a local feature-extraction pipeline that produces more salient information downstream, with improved parameter efficiency. We show improvements on standard metrics (minADE, minFDE, DAO, DAC) over various baselines on the Argoverse dataset. We release our code at: https://github.com/Manojbhat09/Trajformer.

1 Introduction

Precise contextual understanding and feature-extraction is critical in developing reliable models for, e.g., trajectory prediction in autonomous driving scenarios. Typically, generative model classes can be seen as a composition of encoder and decoder structures, where the encoder is responsible for mapping observations to some intermediate feature representation and the decoder is responsible for leveraging this representation for generating a set of future trajectory predictions. Because the trajectory predictions are conditioned on this intermediate representation, we can also say that the quality of the predicted trajectories is a function of the comprehensiveness of the feature representation itself, in terms of its ability to encode all of the salient events in the world in which the agent operates.

Park et al. [2020] highlight a distinction between two major types of environmental context, described with respect to the ego-agent: scene-to-agent context and agent-to-agent context. The former characterises the static properties of the environment, such as admissible driving regions and infrastructural road elements (e.g., traffic lights), and the latter characterises more dynamic elements of the environment, such as pedestrians, other vehicles, and other moving obstacles. Whereas various works have addressed the use of scene-to-agent context as in differentiable driveable-area maps, efficient encoding of agent-to-agent context remains an open challenge for autonomous driving, despite being widely studied for human trajectory prediction [Alahi et al., 2016, Gupta et al., 2018, Robicquet et al., 2016, Vemula et al., 2018].

Relating to agent-to-agent context, specifically, the ego-agent must model its state, amid the set of time-dependent social relationships between all the other agents in the scene. The ability to model these social interactions is crucial, because it informs the ego-agent about the behaviours that are appropriate, given the scene-to-agent context; furthermore, these interactions also inform the ego-agent about the proper social etiquette for driving alongside other agents on the road. Indeed, both aspects of the agent-to-agent context have direct impacts on the predicted trajectories, downstream. However, the highly multimodal nature of the scene complicates models' ability to extract salient

*Correspondence.

Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

information about agent behaviour and social etiquette, leading to the predictions of inadmissible futures in autonomous driving [Casas et al., 2020, Park et al., 2020]. In this work, we propose the use of self-attention [Vaswani et al., 2017] that serves as a better structure of modeling the important latent factors in dynamical scenes.

We propose `Trajformer`, an end-to-end model that addresses prior limitations in modeling the social contextual relationships, in multimodal trajectory prediction for autonomous driving. We achieve this through the use of a self-attention-based encoding structure, allowing for better characterisation of agent behaviour and social etiquette, by providing focused local features given other objects in the scene (e.g., dynamic obstacles, pedestrians, other vehicles). We validate our approach on the Argoverse dataset [Chang et al., 2019] and show substantial performance improvement over baselines and ablations, across standard metrics (minADE, minFDE, DAO, and DAC). Additionally, we show significantly improved model parameter-efficiency, relative to the state-of-the-art. We publish our tools and code for replicating our experiments: `https://github.com/Manojbhat09/Trajformer`.

2 Related Work

Encoder structures for multimodal trajectory prediction. Social contextual modeling has been widely studied in the area of human trajectory prediction [Alahi et al., 2016, Gupta et al., 2018, Robicquet et al., 2016, Vemula et al., 2018, Giuliari et al., 2020]. Giuliari et al. [2020] introduced a method for utilizing transformer models [Vaswani et al., 2017] to produce pedestrian trajectory predictions with multiple mode support. We propose a similar encoder with spatial priors, in the form of projected embeddings from map croppings [Dosovitskiy et al., 2020], with specific application to modeling the social context in autonomous driving scenarios. With our performance improvements, we illustrate the contribution of transformer models in the crucial feature-extraction step – linking annotation-free diverse prediction with self-attention, for long-range dependency modeling [Vaswani et al., 2017, Li et al., 2017].

Another promising encoding structure for forecasting models is in the use of graph-based encoders. Liang et al. [2020] represented the scene context (i.e., lane center-lines) as a discrete graph and, given past poses therein, they used attention to extract salient features for predicting trajectory futures. Similarly, Messaoud et al. [2020] and Monti et al. [2020] highlight the components necessary for predicting futures, such as: graph-based feature extractors, image feature extractors, network classes to perform fusion of these information modalities, and the prediction heads themselves. We propose a more integrated approach – with a single self-attention-based backbone, for learning social behavior+etiquette and generating diverse+admissible futures. Further, while these approaches focused on the spatial context, we pursue performance gains through modelling the social context.

Importance of model size for trajectory prediction. Carion et al. [2020] discuss the importance of understanding the trade-off between model capacity and performance, in the context of object detection and semantic segmentation tasks; here, image features are extracted from a transformer-based backbone network and joined with a positional encoding representation. Dosovitskiy et al. [2020] proposed how image croppings can be directly utilized for feature-extraction with self-attention. In this work, we are inspired by these discussion and make specific application to multimodal trajectory prediction for autonomous driving.

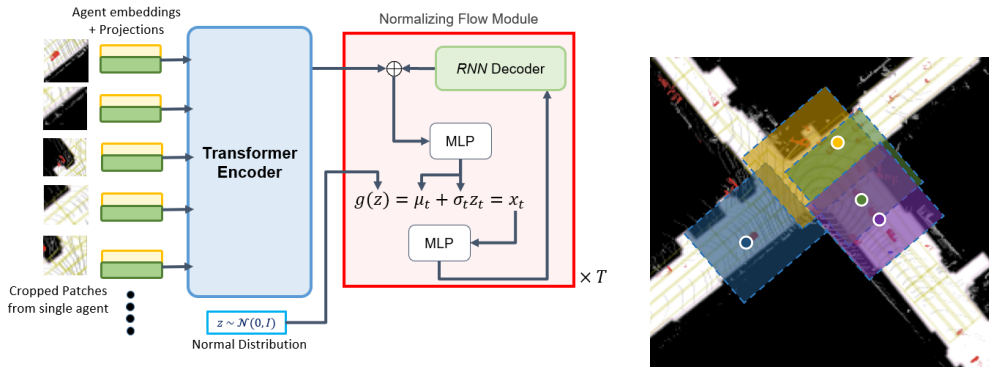
3 Local Self-attention for Trajectory Prediction

In this work, we unify a self-attention-based [Vaswani et al., 2017] encoder structure with a normalising flow-based decoder structure [Park et al., 2020], for multimodal trajectory prediction in autonomous driving. Our method is illustrated in figure 1a and is further described below.

Model structure. To encode the social factor between A agents, we combine their individual past-trajectory encodings, through consecutive additive and multiplicative fusion [Liu et al., 2018], to generate A embeddings. For each of these agent embeddings, t past time-step poses are raised to N-dimensional vectors and are combined with patch and positional embeddings, for each time-step. In this way, we generate an $N \quad t$-dimensional input for the transformer model.

Let $\boldsymbol{S} \quad \{S^1, S^2, \ldots, S^A\}$ denote the set of agent trajectories for A agents in a given scene, with S^a being the concatenation of past trajectory segment S^a_{past} and ground-truth future trajectory segment S^a_{future}. Here, a single step is indicated by $S^a_t \; 2 \; \mathsf{R}^2$, for agent a and time-step t. Thus, $\boldsymbol{S}_{\text{past}}$ is the collection of past trajectory segments for all agents, and the observation set of all 3-second past trajectories is denoted by $O \quad f\boldsymbol{S}_{\text{past}}, \Phi, \phi g$, where is a scene embedding [Park et al., 2020] and ϕ is the positional embedding [Vaswani et al., 2017]. We want to model the posterior distribution over future trajectories, for all agents in the scene snapshot, $q(\boldsymbol{S}_{\text{pred}}|O)$.

Past agent trajectories are projected to a (higher) d-dimensional space, in preparation for input to the transformer, i.e., $\mathbf{e}^a_{obs} = \mathtt{MLP}_{proj}(\mathsf{S}^a)$. For encoding the contextual information, we extract an m^2 pixel neighbourhood



(a) (b)

Figure 1: (a) Model architecture with Agent pose embeddings, cropped image and positional embeddings fused for input to the transformer encoder and Flow based decoding for producing T future poses. This reduced architecture can be useful for Trajectory prediction for embedded platforms in Robotic applications. (b) Depiction of Patch croppings produced from the BEV image. The different colors indicate different Agents in the scene. And the colored area is a fixed K K pixel size for each such crops. These patches are then again cropped into 16 16 patches and linearly projected to produce projections at the input of the transformer model.

image patch, centered around each vehicle, from the *birds-eye-view* (BEV) map of the scene: $\mathbf{e}^a_{patch}$. The BEV map contains coloured objects and superimposed LiDAR points. Figure 1b illustrates the agent-wise pixel neighbourhood, which capture local contextual information.

We perform sine-distance positional encoding of the map representation, which is then added to each agent's flattened sequence of patch vectors. We then calculate a fused representation, combining the local environment information and state history, as a *Hadamard* product between each agent's past trajectory embedding and its corresponding scene context [Liu et al., 2018]:

$$\mathbf{e}^a_{fused} = \mathbf{e}^a_{obs} \quad [\texttt{ENC}_{pos}(\mathbf{e}^a_{map}) + \texttt{FLATTEN}(\mathbf{e}^a_{patch})]$$

These fused representations are fed to a standard transformer encoder, which contains alternating layers of multi-headed self-attention and MLP blocks. The output is a set of latent codes – one for each agent: $\mathbf{c}^a_{latent} = \texttt{ENC}_{tr}(\mathbf{e}^a_{fused})$, with $\mathbf{c}^a_{latent}$ 2 $\mathsf{R}^{D \times A}$ with hyperparameter D.

The normalizing-flow-based generative decoder features an implicit auto-regressive design and performs a differentiable and bijective mapping, from the latent codes to the set of agent-wise trajectory predictions [Park et al., 2020, Rhinehart et al., 2019] (see figure 1a for illustration): $g_\theta(z_t; \mu_t, \sigma_t) = \sigma_t \ z_t + \mu_t = S^a_{pred,t}$, with z_t $N(\mathbf{0}, I)$ 2 R^2. Here, θ is the set of model parameters and μ_t 2 R^2 and σ_t 2 $\mathsf{R}^{2\times2}$ are projected parameters. Iterating through time, we get the predictive trajectory S^a_{pred} for each agent. By sampling multiple instances of z_{pred} and mapping them to trajectories, we get various hypotheses of future.

Following Park et al. [2020], we train our model according to the symmetric cross-entropy objective [Rhinehart et al., 2018], with the annotation-free discrete grid map as the (estimated) prior distribution $\hat{p}$ and with a model degradation coefficient of $\alpha = 0.5$.

4 Experiments

We benchmark two instances of our approach, `Trajformer-12` and `Trajformer-24`, with respectively 12 and 24 layers in the transformer encoder. We set the size of the trajectory encoder projection d to be 1024 ($\texttt{MLP}_{proj}$ is a single-layer projection), pixel neighbourhood width/height m to be 16, and the dimension of the latent code D to be 256. We choose a batch size of 128 and train with Adam optimizer. We use linear learning rate warm-up and decay. It takes 3 days to train each model on a NVIDIA 1080 Ti GPU device, with batch data processed as in Park et al. [2020], from the *Tracking* split of the Argoverse dataset [Chang et al., 2019].

5 Results

Quantitative results are summarized in Table 1, and some qualitative results are shown in 2. We observe a new state-of-the-art performance in our model, compared to Park et al. [2020], in both qualitative and quantitative results. Most of the maneuvers have been refined by the model. An interesting observation in figure 2b suggests

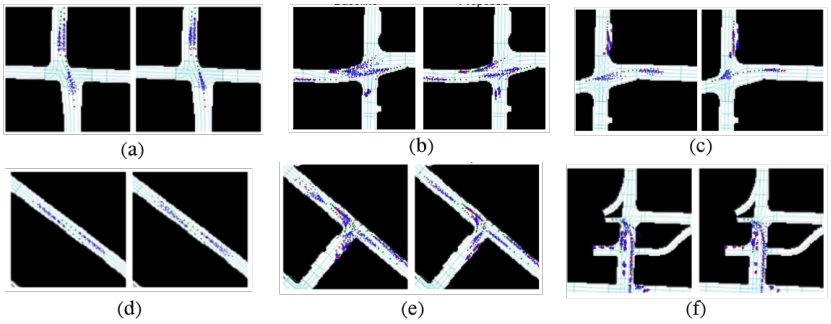


Figure 2: Qualitative illustration of model performance [right], compared to best-performing baseline [Park et al., 2020][left]. Observations: (a) more precise & confident on straight in-lane trajectories; (b) more confidence in the maneuver, indicated by a cluster of trajectories; (c) more confident and diverse alternative maneuvers, with attention to standing vehicles, due to simple intersection map lane-start/end prior; (d) equivalent lane-change maneuver on empty roads, due to attention on immediate local activities; (e) more conservative turning maneuver with more agent-activity; (f) reduced confidence in strong curve lane-change maneuvers.

Table 1: Comparison of improvements over baseline models on Argoverse. The metrics are abbreviated as follows: MINADE(**A**), MINFDE(**B**), RF(**C**), DAO(**D**), DAC(**E**). Improvements indicated by arrows. : larger is better, as long as **A** and **B** are small.

	A (↓)	B (↓)	C (↑)*	D (↑)*	E (↑)*
LSTM	1.441	2.780	1.000	3.435	0.959
CSP [PARK ET AL., 2020]	1.385	2.567	1.000	3.453	0.963
MATF-D [ZHAO ET AL., 2019]	1.344	2.484	1.000	1.372	0.965
DESIRE [LEE ET AL., 2017]	0.896	1.453	3.188	15.17	0.457
MATF-GAN [PARK ET AL., 2020]	1.261	2.313	1.175	11.47	0.960
R2P2-MA [RHINEHART ET AL., 2018]	1.108	1.270	2.190	37.18	0.955
DATF [PARK ET AL., 2020]	0.730	1.124	3.282	28.64	0.968
TRAJFORMER-12 (ours)	0.684	0.885	3.359	27.71	0.972
TRAJFORMER-24 (ours)	**0.621**	**0.719**	**3.868**	**28.21**	**0.973**

Table 2: Model size comparison (with optimizer state), in megabytes (MB) and number of parameters.

Model-layers	SIZE (.TAR)	#PARAMS
DATF [PARK ET AL., 2020]	4.7 MB	462K
TRAJFORMER-12 (ours)	2.1 MB	164K
TRAJFORMER-24 (ours)	2.9 MB	192K

that rule-based maneuvers, such as the right-of-way in intersections, are learned and followed by model (the left vehicle gains the right-of-way). We also note a significant failure mode: where, if the velocity of a particular agent in a particular frame is high, the target 6 trajectory points are spaced uniformly and with a much greater *intra*-point distance (2x), compared to the average spacing between each data-point in a trajectory from the Argoverse dataset. Here, the model fails to predict the appropriate spacing, despite producing the right modes of trajectories. The reason for this phenomenon is hypothesised to be the size of the local neighbourhood context that was chosen while training the model. Compared to DATF [Park et al., 2020], the model is significantly lighter in time-complexity and memory-intensity, due to the social attention and scene attention blocks provided by the transformer encoder. The `Transformer-12` and `Transformer-24` model instances did not vary significantly in the qualitative and quantitative results: the aforementioned observations remain true for both models.

6 Conclusion

In this paper, we proposed `Trajformer`, an end-to-end model that addresses prior limitations in modeling the social contextual relationships, in multimodal trajectory prediction for autonomous driving. We introduced the use of self-attention-based encoding for applications to autonomous driving, allowing for better characterisation of agent behaviour and social etiquette, given other objects in the scene (e.g., dynamic obstacles, pedestrians, other vehicles). Our approach uses a single self-attention-based backbone, for learning social behavior+etiquette and generating diverse+admissible futures. We validated our approach on the Argoverse dataset [Chang et al., 2019] and showed substantial performance improvement over baselines and ablations, across standard metrics (minADE, minFDE, DAO, and DAC). Additionally, we show significantly improved model parameter-efficiency, relative to the state-of-the-art.

References

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *European Conference on Computer Vision*, 2020.

Sergio Casas, Cole Gulino, Simon Suo, and Raquel Urtasun. The importance of prior knowledge in precise multimodal prediction. *arXiv preprint arXiv:2006.02636*, 2020.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting, 2020.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.

J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2017. doi: 10.1109/TMM.2016.2642789.

Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. *arXiv preprint arXiv:2007.13732*, 2020.

Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018.

Kaouther Messaoud, Nachiket Deo, Mohan M. Trivedi, and Fawzi Nashashibi. Trajectory prediction for autonomous driving based on multi-head attention with joint agent-map representation, 2020.

Alessio Monti, Alessia Bertugli, Simone Calderara, and Rita Cucchiara. Dag-net: Double attentive graph neural network for trajectory forecasting. *arXiv preprint arXiv:2005.12661*, 2020.

Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R Jadhav, Paul Pu Liang, and Louis-Philippe Morency. Diverse and admissible trajectory forecasting through multimodal context understanding. *European Conference on Computer Vision*, 2020.

Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 772–788, 2018.

Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2821–2830, 2019.

Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 549–565, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12126–12134, 2019.