Qeios

Peer Review

Review of: "Don’t Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving"

Jean-Bernard Hayet[1]

1. Computer Science, Centro de Investigación en Matemáticas (CIMAT), Guanajuato, Mexico

Overall comments

- This paper describes a planning scheme for autonomous driving based on the principle of planning-by-prediction: The system executes plans corresponding to what it estimates is the most likely outcome of a predictive model. The proposed approach, called MomAD, is not built from scratch, but instead is built on top of SparseDrive, focusing its improvements on the temporal consistency of consecutive predictions, which tends to be an underestimated problem and which is not trivial to handle given the multi-modal nature of these algorithms.
- A frustrating impression from reading the paper is that the method section (Section 3) is too short to understand the details of the approach, and I feel that the reader is obliged to stay at a high level of the method description.

Technical comments

- The related work section is short and very unbalanced: There is a very long list of references appearing at once at the beginning of this section, but with no nuance among the methods. I would recommend just citing a few relevant ones and explaining briefly in what they differ (as you do with UniAD, VAD, VAD2). Also, because you build upon SparseDrive, I would recommend giving a few more comments about it in Section 2: How does it implement multi-modality, for example?
- I understand well the motivation behind the selection in Eq. 3, but it gives me some doubts: All the weight is put on the distance to the historical trajectory, and *none* on the probabilities given by the predictor. Could it be a potential problem when faced with dangerous, sudden situations? If something

doi.org/10.32388/S4HEJ7

new appears on the road, for example, the images could influence the predictor in strongly changing the current trajectory and put low probabilities on the trajectories that are closer to the previous one.

- In Section 3.2, it is not clear to me how Q_t^{p*} (or Q_t^p) is computed from the trajectories. Is it from the 'sparse perception' module from Fig. 2?
- If I understand correctly how TTM and MPI work, TTM selects one (only one, among K) of the currently predicted trajectories, based on its closeness to the previous most likely trajectory; then it uses the previous context and a cross-attention mechanism with this selected trajectory to produce, again, a K-dimensional multimodal query which is used to produce refined predictions. A genuine doubt is: Why collapsing the time-consistent trajectory to just one instance if you come back to having K proposed trajectories anyway? Why not re-weighting the probabilities in the set of currently proposed trajectories to avoid passing through this deterministic (and potentially risky, as commented above) step of selecting just one trajectory, the closest to the historical one?
- After Equation 4, it is commented that the MLP goes from $\mathbb R^{D_q}$, but a few lines above it is also said that Q^p_{t-1} is $\mathbb R^{K*D_q}$; is this because the MLP processes each of the queries (among the K) separately? If so, I think it should be stated clearly.
- In the description of 3.2 (Eq. 5), what is the 'PlanHead' module? I suppose it is one of the modules of SparseDrive?
- Section 3.3 is very, very short; I think it would gain from having a basic mathematical description of this module; I suppose that the denoising module is learned separately? Or is the whole set of modules end-to-end? I see the answer to this question in appendix A.3, but I think it should be commented on in the main text.
- In Appendix A.3: please define clearly the different loss terms.

Experimental results

- In the description of the results, the sentence "We use Benchdrive..." does not seem to introduce a second testing dataset, but rather complement the previous sentence. Maybe you could say "We also use Benchdrive..." to make that point clear.
- I do not understand the two protocols mentioned in Table 1. I think it would be better to stick with one protocol to avoid confusion.
- Again in Table 1, I do not see clearly how the notion of bold/underline is used. Please explain it better.

- The results presented in Table 1 are good but get close to those of SparseDrive. Those in Table 2 (for more complex trajectories within nuScenes) are more interesting, and the difference with SparseDrive is clearer. Same thing for Table 3 (longer horizons). A detail for Table 3: Maybe consider using the same convention as in Table 2 for presenting the relative differences.
- Table 4 is hard to understand: Why are there two versions of MomAD? What is MomAD (Euclidean)? How is the Success rate evaluated? What are the Effi/Comf metrics?
- Table 5: again, what are the underlined figures?
- Table 8: I do not understand well what the TP flag is: What does it mean if you do not use a trajectory predictor?

Presentation details

- In Fig. 1.c, I have some problems understanding the middle image; If I understand it correctly, the upper trajectory is the one that has been predicted from $t-1$ and the lower ones are the ones resulting from the prediction step at t. So, I would not expect to see the yellow trajectory coming from $t-1$ as high in the common reference frame (I suppose the predictions coming from t should be "higher" than those from $t-1$, assuming that the car is going to the upper side). Also, it is not clear (not from the figure, nor from the caption or main text) what the dashed lines represent. It could be related to the Hausdorff distance that you describe elsewhere, but if it were the case, I would expect more of these lines (one for each predicted waypoint). Hence, I would urge the authors to describe this figure better.
- In Fig. 2, I would recommend helping the reader by commenting on its components better; what is the output of the "Sparse Description Module"? Is it the Q^m_t,Q^p_t?
- In the same figure, does the blue box always correspond to the "raw" features extracted from the images? If so, I think it would be better to use the F^{ins}_t all the time.
- There is a problem in Eq. 6 (definition of TPC), as the terms within the sum do not have indices i. Also, I suppose for this definition that the positions at T^{Pred} are moved by one step to make them coincide with the current ones?

Typos:

- Page 10: Planing -> Planning

Declarations

Potential competing interests: No potential competing interests to declare.