HPNet: Dynamic Trajectory Forecasting with Historical Prediction Attention

Supplementary Material

1. Training Objective for Joint Prediction

With a simple adjustment of the training objective, HPNet can be used for joint trajectory prediction. In joint prediction, we treat the predictions of all agents in the same mode as a single predicted future, so the k_t-th mode to be optimized is determined based on the minimum joint endpoint displacements between the predicted future $\{L^1_{t,1\sim N,k}\}_{k\in[1,K]}$ and the ground truth $G_{t,1\sim N} = \{g_{t+1,1}, g_{t+1,2}, ..., g_{t+F,N}\}$:

$$k_t = \underset{k\in[1,K]}{\arg\min} \sum_{n=1}^{N} (l^1_{t+F,n,k} - g_{t+F,n}) \tag{1}$$

Then, the regression loss function contains two Huber losses for the trajectory proposals and the refined final trajectories, respectively:

$$\mathcal{L}^t_{reg1} = \sum_{n=1}^{N} \mathcal{L}_{Huber}(L^1_{t,n,k_t}, G_{t,n}), \tag{2}$$

$$\mathcal{L}^t_{reg2} = \sum_{n=1}^{N} \mathcal{L}_{Huber}(L^2_{t,n,k_t}, G_{t,n}). \tag{3}$$

Besides, the future probability scores $\hat{\pi}_{t,k}$ are optimized by using the cross-entropy loss function:

$$\mathcal{L}^t_{cls} = \mathcal{L}_{CE}(\{\hat{\pi}_{t,k}\}_{k\in[1,K]}, k_t). \tag{4}$$

Overall, the total loss function of the whole model is formulated as follows:

$$\mathcal{L} = \frac{1}{T} \sum_{t=-T+1}^{0} (L^t_{reg1} + L^t_{reg2} + L^t_{cls}). \tag{5}$$

2. Implementation Details

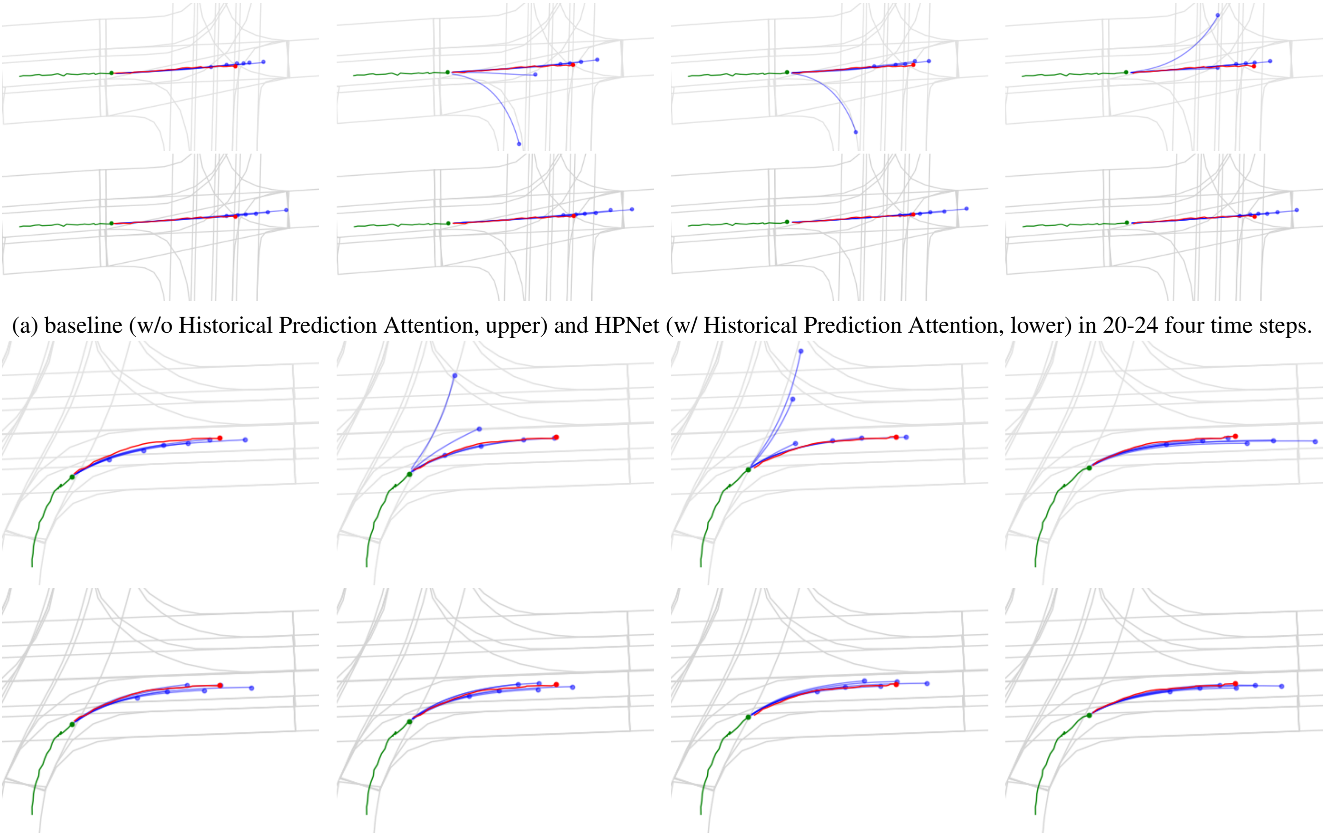
Our model trains on 8 RTX 4090 GPUs for 64 epochs, using the AdamW [1] optimizer with a batch size of 16, dropout rate of 0.1, and weight decay of 1×10^{-4}. Initial learning rates are 5×10^{-4} for Argoverse and 3×10^{-4} for INTERACTION, with a cosine annealing scheduler for rate decay. On Argoverse, we apply a single Spatio-Temporal Attention layer and two Triple Factorized Attention layers, setting a 50 radius for all local areas and a 20-time span for historical frames and predictions. Data augmentation techniques—horizontal flipping, agent occlusion, and lane occlusion—are used with ratios of 0.5, 0.05, and 0.2, respectively. On INTERACTION, the setup involves one Spatio-Temporal Attention layer and three Triple Factorized Attention layers, with an 80 radius for local areas and a 10-time span for frames and predictions. The augmentation techniques of horizontal flipping and lane occlusion remain the same without the use of agent occlusion. The model sizes for Argoverse and INTERACTION are 4.1M and 5.3M,respectively.

3. Inference Latency

We also take a closer look at the inference latency, which is important for practical applications. The inference latency is reported on the Argoverse validation set with an NVIDIA V100 GPU. In the training stage, parallel processing is employed for fast training. In the testing stage, both parallel processing and serial processing can be used depending on the practical requirements. With serial processing, the time required for HPNet to predict all trajectories for all agents in a single time step is 27.62 ms, and the baseline model without Historical Prediction Attention takes 22.76 ms. When parallel processing, the time required for HPNet and the baseline model to predict trajectories for all agents at all 20 time steps is 92.02 ms and 81.08 ms respectively. These results indicate that the inference latency introduced by Historical Prediction Attention is small. In practice, we suggest serial processing. Overall, Historical Prediction Attention will not impede the real-time operational capabilities of autonomous driving systems, while concurrently enhancing prediction performance.

4. Quality Results

In Fig. 1, we show two representative examples. In Fig. 1 (a), the agent is in the middle lane and is about to move forward. The baseline gives three types of possible motion goals (*i.e.*, go straight, turn right and go straight, turn left and go straight) at four consecutive moments, which undoubtedly complicates the subsequent decision-making process. In contrast, HPNet provides a stable, accurate motion goal (*i.e.*, go straight). In Fig. 1 (b), the agent exhibits a clear intention to turn right. The baseline gave two motion goals at the middle two moments (*i.e.*, go straight and turn right), while at each of the other two moments, it gave only one motion goal (*i.e.*, turn right).In contrast, HPNet also gives a stable and accurate motion goal (*i.e.*, turn right). All in all, these examples clearly and intuitively demonstrate the great improvement of Historical Prediction Attention on the accuracy and stability of trajectory prediction, indicating its importance and effectiveness.



(a) baseline (w/o Historical Prediction Attention, upper) and HPNet (w/ Historical Prediction Attention, lower) in 20-24 four time steps.

(b) baseline (w/o Historical Prediction Attention, upper) and HPNet (w/ Historical Prediction Attention, lower) in 20-24 four time steps.

Figure 1. Qualitative results on the Argoverse validation set. The lanes, historical trajectory, ground truth trajectory, and six predicted trajectories are indicated in grey, green, red, and blue, respectively.

References

[1] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1