A Trajectory is Worth Three Sentences: Multimodal Transformer for Offline Reinforcement Learning (Supplementary Material)

Yiqi Wang\textsuperscript{1} \quad Mengdi Xu\textsuperscript{2} \quad Laixi Shi\textsuperscript{1} \quad Yuejie Chi\textsuperscript{1}

\textsuperscript{1}Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
\textsuperscript{2}Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

The following information is provided to support discussion in the paper: (A) hyperparameter and training comparisons between the Decision Transformer \cite{Chen21} (DT), DT-large, and our Decision Transducer (DTD); (B) evaluations curves of DT-large and DTD on medium-expert, medium, medium-replay of the D4RL benchmark \cite{Fu20}.

A HYPER-PARAMETERS FOR DT-LARGE AND DTD

In Table 1, we have listed the architecture difference between DT-large and DTD.

Table 1: Hyperparameters. While DT, DT-large, and DTD’s Join net are all temporal transformers \cite{Radford19} taking multimodal input, their input modalities varies according to the design.

<table>
<thead>
<tr>
<th>Models</th>
<th>size dimension</th>
<th>Modality Encoders</th>
<th>Biasing Combiner</th>
<th>Joint Encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>layers heads modality</td>
<td>layers heads modality</td>
<td>layers modality</td>
<td>layers heads modality</td>
</tr>
<tr>
<td>DTd</td>
<td>2.52M 128</td>
<td>3 1 uni-modal</td>
<td>1 1 bi-modal</td>
<td>1 2 bi-modal</td>
</tr>
<tr>
<td>DT</td>
<td>0.7M 128</td>
<td>\</td>
<td>\</td>
<td>3 1 tri-modal</td>
</tr>
<tr>
<td>DT-large</td>
<td>2.41M 213</td>
<td>\</td>
<td>\</td>
<td>4 3 tri-modal</td>
</tr>
</tbody>
</table>

Figure 1: Learning rate (lr) schedule. DTD and DT-large are trained with the same number of gradient steps. However, on small dataset, DTD’s lr (schedule 1) will be exponentially decayed after linear warm-up while DT-large don’t (schedule 2). Since DTD is much more sample efficient than DT, a decaying lr will present it to overfit the data.

DT-large and DTD were both trained with 25k gradients steps of 256 batch size before evaluations. We found DTD learns faster than DT and may overfit the smaller dataset from D4RL benchmark \cite{Fu20}, such as medium and medium-replay.
Therefore, we provide an exponentially decay learning rate schedule for DTd on smaller dataset other than medium-expert. Both DT-large and DTd has a peak learning rate 0.0001. We provide the learning rate curve throughout the training process as in Figure 1. DT-large always use schedule 2 provided by Chen et al. [2021] while DTd use schedule 2 when dataset is large (medium-expert) and choose exponentially decaying schedule 1 when dataset is small to avoid potential overfitting.

B EVALUATIONS CURVES OF DT-LARGE AND DTD

The normalized scores on D4RL benchmark [Fu et al., 2020] across 4 runs with independent training seed per run and 3 different evaluation seeds was plotted across all 3 environments (hopper, walker2d, halfcheetah) and across all dataset (medium-expert, medium, medium-replay). Both DT-large and DTd was trained 25k gradient steps with a batch size of 256 and was evaluated every 250 gradients steps following the protocol we discussed above. Curves are the average result across 4 runs and the shaded area corresponds to the standard deviation. As shown in the Figure 2, DTd is more sample efficient than DT on medium-expert data across all environment but such an advantage is not consistently observed on medium and medium-replay dataset across 3 environments as shown in Figure 3 and Figure 4.

References
