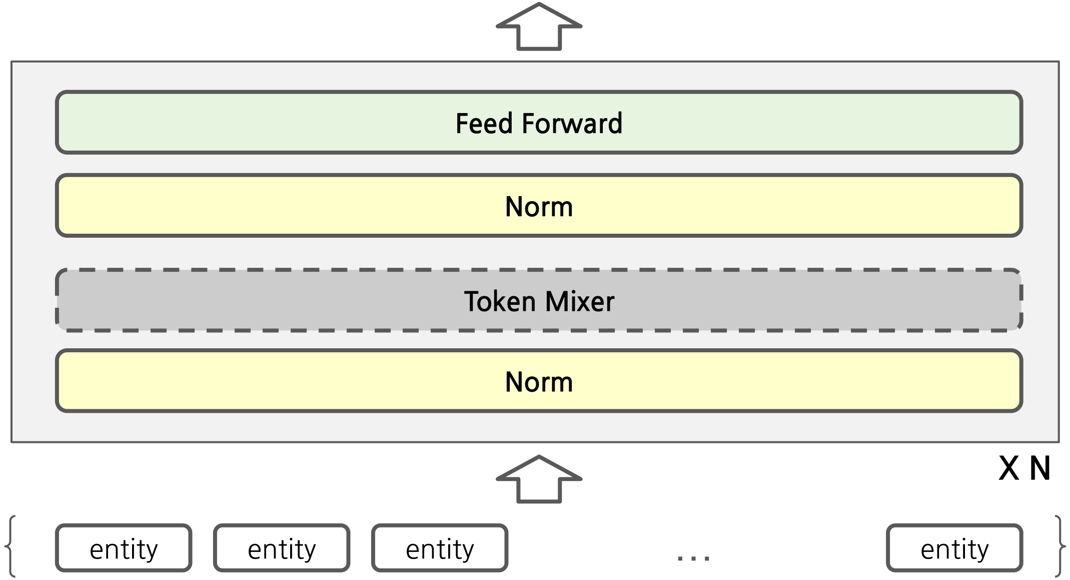
Decision ConvFormer: Local Filtering with MetaFormer Is Sufficient For Decision Making (DC)

ICLR 2024 (Spotlight)

Jeonghye Kim, Suyoung Lee, Woojun Kim[*], Youngchul Sung[*]

KAIST SISReL

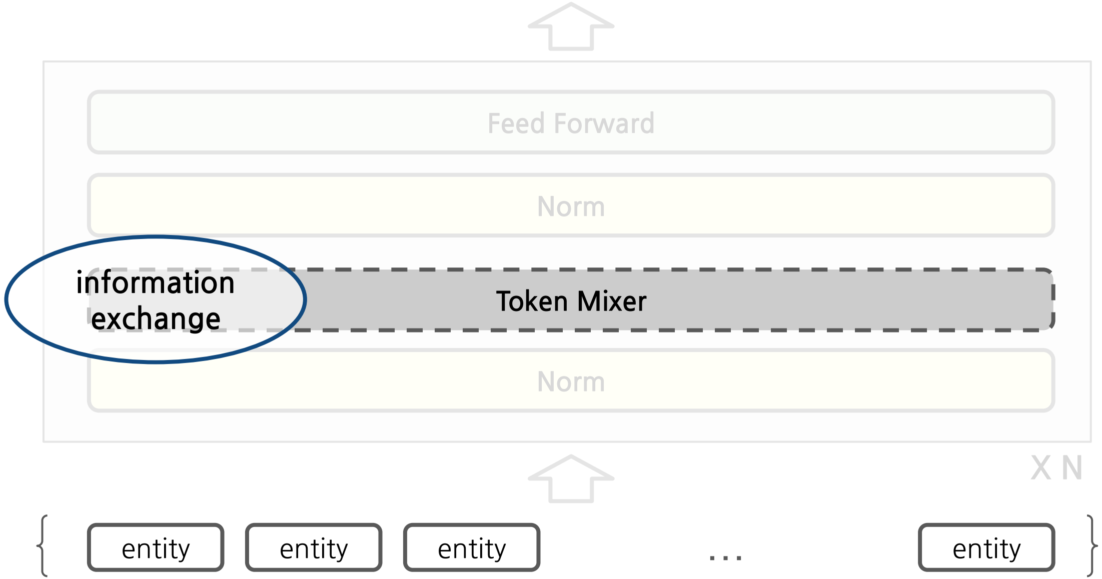
Background - MetaFormer[1]



- As Transformer[2] proved successful in various domains, interest also continued in MetaFormer, a more abstract structure of the Transformer.
- MetaFormer is a general architecture that takes multiple entities in parallel, understands their interrelationship, and extracts important features for addressing specific tasks while minimizing information loss.

[1] Yu, Weihao, et al. "Metaformer is actually what you need for vision." CVPR 2022, [2] Vaswani, Ashish, et al. "Attention is all you need." NeurIPS 2017

Background - MetaFormer



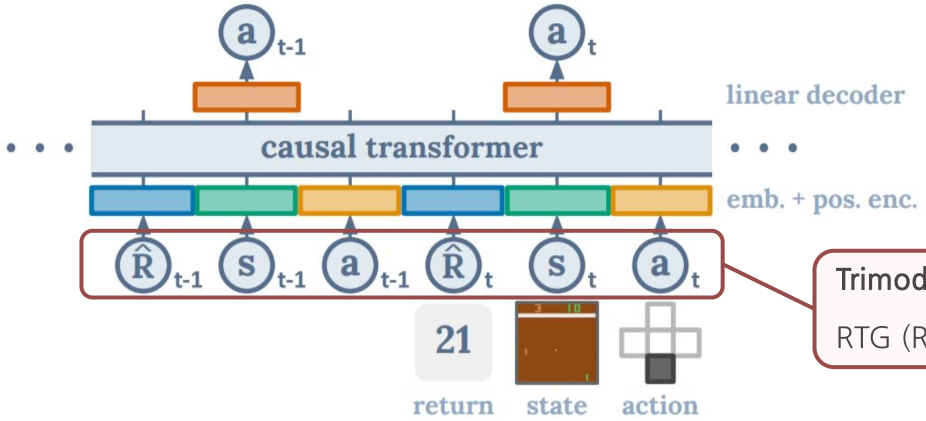
Transformer

Token Mixer = Attention

- Among these components, **the token mixer plays a crucial role in information exchange among multiple input entities.**

Background - Decision Transformer[3]

A representative Transformer model in the Offline RL : Decision Transformer (DT)



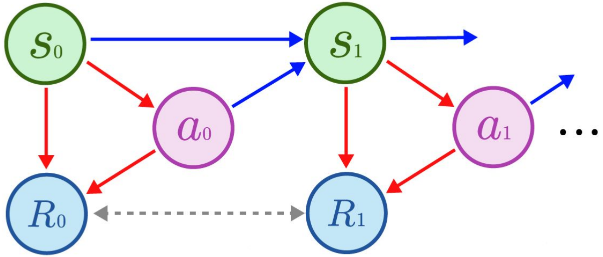
Trimodal token sequence of state, action, and RTG

RTG (Return-to-go) : $\hat{R}_t = \sum_{t'=t}^{T} r_{t'}$

- DT directly leverages history information to predict the next action, resulting in competitive performance compared with existing approaches to offline RL.
- The input trimodal sequence undergoes information exchange through DT's attention module, based on the computed relative importance (weights) between each token and every other token in the sequence.
- **Thus, the way that DT predicts the next action is just like that of GPT-2 in NLP with minimal change.**

[3] Chen, Lili, et al. "Decision transformer: Reinforcement learning via sequence modeling." NeurIPS 2021

Background - The local dependency of offline RL dataset



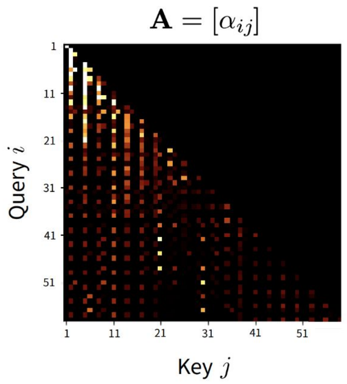
Blue arrows : Markov property
Red arrows : the causal interrelation per a single timestep
Gray dotted line : the correlation of the adjacent returns

- However, unlike data sequences in NLP for which Transformer was originally developed, **offline RL data has an inherent pattern of local association between adjacent timestep tokens due to the Markovian property.**
- This dependence pattern is distinct from that in NLP and is crucial for identifying the underlying transition and reward function of an MDP, which are fundamental for decision-making in turn.

Background - Attention Score Analysis of DT

"Is the attention module initially developed for NLP
still an appropriate local-association identifying structure for data sequences of MDPs?"

Attention scores of Decision Transformer



$$\alpha_{ij} = \text{softmax}(\{\langle q_i, k_{j'} \rangle\}_{j'=1}^{3K-1})_j$$

K : 20, size of attention matrix : 60 x 60 (20 * 3 modal)

*Experimental results in hopper-medium

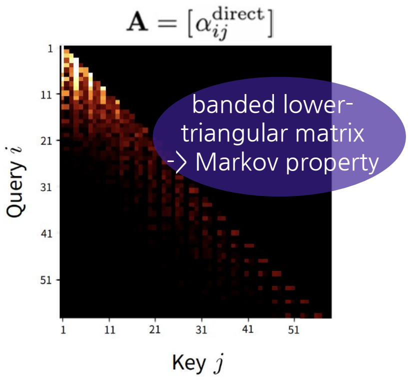
DT considers the history information from the past 20 timesteps to be equally important.

- In reinforcement learning, the state sequence is characterized as a Markov chain, and based on ergodic theory, it is expected to gradually diminish the effects of previous states while ensuring that states that are not consecutive remain independent, contingent on the immediately prior state.
- However, **the results are inconsistent with these theoretical expectations of Markov chains.**

Background - Attention Score Analysis of DT

“Is the attention module initially developed for NLP
still an appropriate local-association identifying structure for data sequences of MDPs?”

Attention scores of direct learning



banded lower-triangular matrix
-> Markov property

$\alpha_{ij}^{\text{direct}}$: Learning parameters

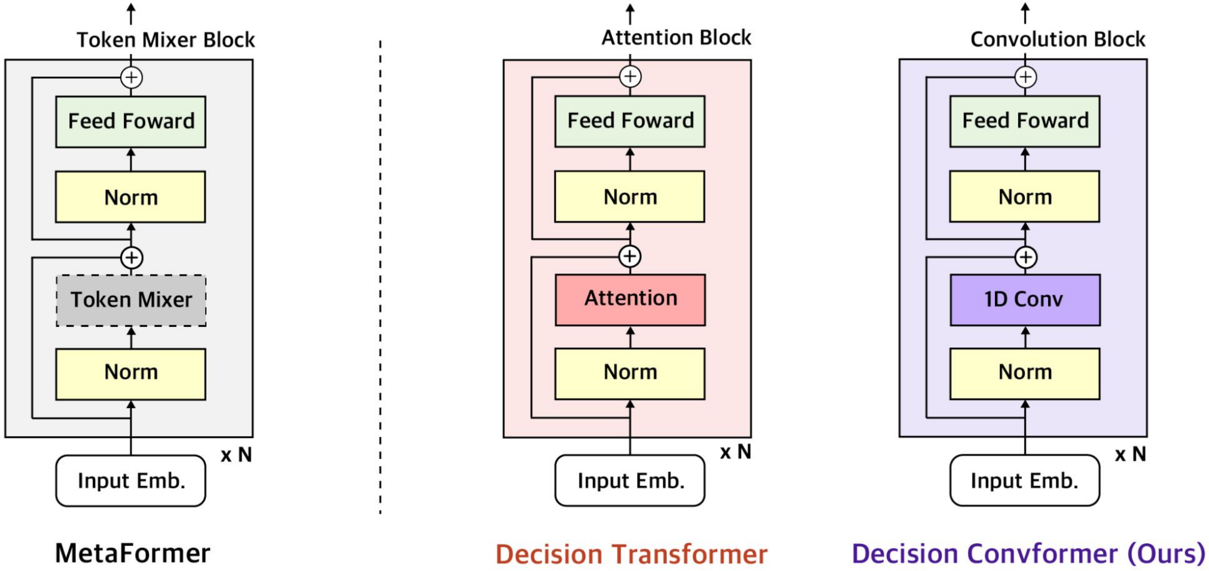
DT	Direct Learning of $\mathbf{A}$
68.4	88.2

- Direct Learning of A - Eech attention score is considered as a single learning parameter.
    - The attention scores of direct learning take the form of a banded lower-triangle matrix - linked to Markov chain theory.
    - Additionally, it demonstrates improved decision-making capabilities compared to DT.
- Thus, the full lower-triangular structure of the attention matrix in DT is an artifact of the method used for parameterizing the attention module and does not truly capture the local associations in the RL dataset.

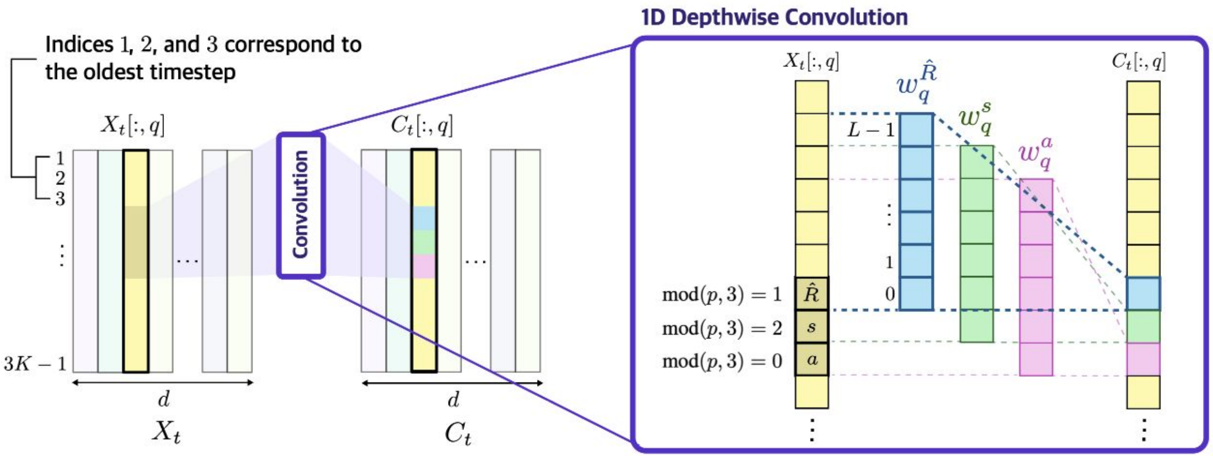
Decision ConvFormer

A new design of a token mixer for MetaFormers as RL action predictors

1. Integrating adjacent past information
2. Input-sequence-independent static linear filtering



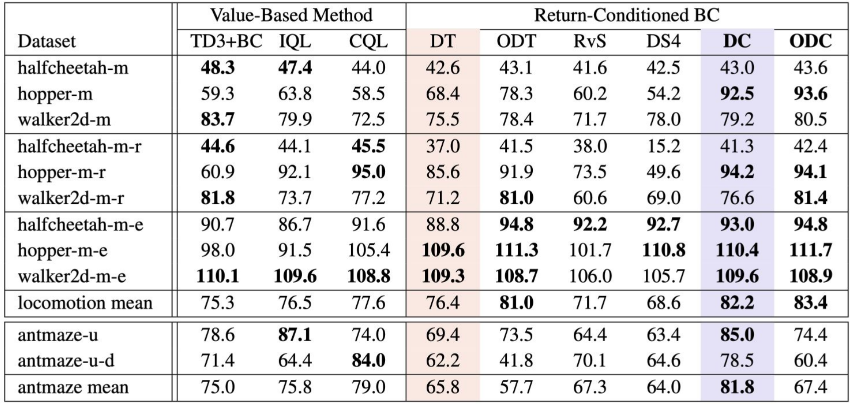
Method - Convolution Module



- Purpose of the convolution module : **integrate the time-domain information among neighboring tokens**.
- For this, 1D depthwise convolution is applied for each hidden dimension across a small window size.
- To effectively extract information associated with each RTG, state, and action token, three different filters are used without sharing parameters; **the RTG filter, state filter, and action filter.**
- The window size is set to 6, considering the current and immediately previous timestep (2 timesteps * 3 token types).

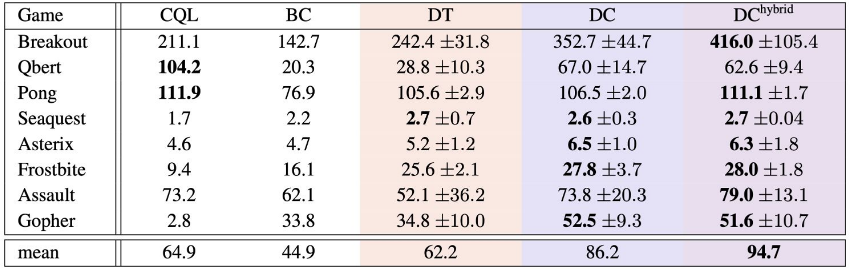
Experiments - Offline Results

Mujoco & Antmaze



	Value-Based Method			Return-Conditioned BC					
Dataset	TD3+BC	IQL	CQL	DT	ODT	RvS	DS4	**DC**	**ODC**
halfcheetah-m	**48.3**	**47.4**	44.0	42.6	43.1	41.6	42.5	43.0	43.6
hopper-m	59.3	63.8	58.5	68.4	78.3	60.2	54.2	**92.5**	**93.6**
walker2d-m	**83.7**	79.9	72.5	75.5	78.4	71.7	78.0	79.2	80.5
halfcheetah-m-r	**44.6**	44.1	**45.5**	37.0	41.5	38.0	15.2	41.3	42.4
hopper-m-r	60.9	92.1	**95.0**	85.6	91.9	73.5	49.6	**94.2**	**94.1**
walker2d-m-r	**81.8**	73.7	77.2	71.2	**81.0**	60.6	69.0	76.6	**81.4**
halfcheetah-m-e	90.7	86.7	91.6	88.8	**94.8**	**92.2**	**92.7**	**93.0**	**94.8**
hopper-m-e	98.0	91.5	105.4	**109.6**	**111.3**	101.7	**110.8**	**110.4**	**111.7**
walker2d-m-e	**110.1**	**109.6**	**108.8**	**109.3**	**108.7**	106.0	105.7	**109.6**	**108.9**
locomotion mean	75.3	76.5	77.6	76.4	**81.0**	71.7	68.6	**82.2**	**83.4**
antmaze-u	78.6	**87.1**	74.0	69.4	73.5	64.4	63.4	**85.0**	74.4
antmaze-u-d	71.4	64.4	**84.0**	62.2	41.8	70.1	64.6	78.5	60.4
antmaze mean	75.0	75.8	79.0	65.8	57.7	67.3	64.0	**81.8**	67.4

Atari

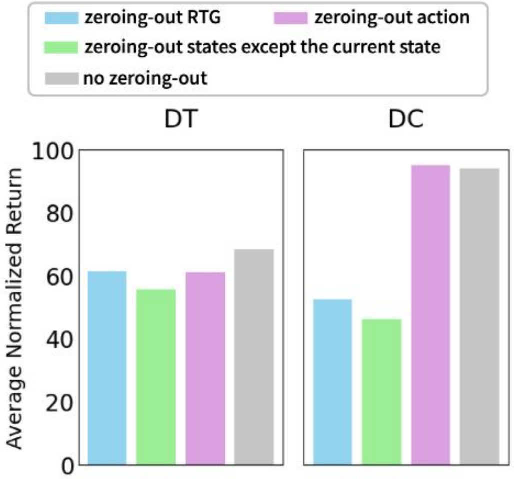


Game	CQL	BC	DT	DC	DC$^{\text{hybrid}}$
Breakout	211.1	142.7	242.4 ±31.8	352.7 ±44.7	**416.0** ±105.4
Qbert	**104.2**	20.3	28.8 ±10.3	67.0 ±14.7	62.6 ±9.4
Pong	**111.9**	76.9	105.6 ±2.9	106.5 ±2.0	**111.1** ±1.7
Seaquest	1.7	2.2	**2.7** ±0.7	**2.6** ±0.3	**2.7** ±0.04
Asterix	4.6	4.7	5.2 ±1.2	**6.5** ±1.0	**6.3** ±1.8
Frostbite	9.4	16.1	25.6 ±2.1	**27.8** ±3.7	**28.0** ±1.8
Assault	73.2	62.1	52.1 ±36.2	73.8 ±20.3	**79.0** ±13.1
Gopher	2.8	33.8	34.8 ±10.0	**52.5** ±9.3	**51.6** ±10.7
mean	64.9	44.9	62.2	86.2	**94.7**

- DC consistently outperforms or closely matches the state-of-the-art performance across all environments.
- DC excels not only in MuJoCo locomotion tasks focused on return maximization but also in goal-reaching AntMaze tasks.

DC effectively combines important information to make optimal decisions specific to each situation, irrespective of whether the context involves high-quality demonstrations, sub-optimal demonstrations, dense rewards, or sparse rewards.

Experiments - Input modal dependency



- To analyze how DT/DC assesses the importance of each modality (RTG, state, action), during the inference stage, we zero out the state, action, and RTG tokens separately.

- In DT, zeroing-out each of the state, action, and RTG tokens results in only a minor performance drop.
- In DC, performance significantly decreases when the RTG and state tokens are zeroed-out, but there is little difference when the action token is zeroed-out.
- **Indeed, DC found out that RTG and state information is more important than action, whereas DT seems not.**

Thank you for listening