Relational Reasoning via Set Transformers: Provable Efficiency and Applications to MARL

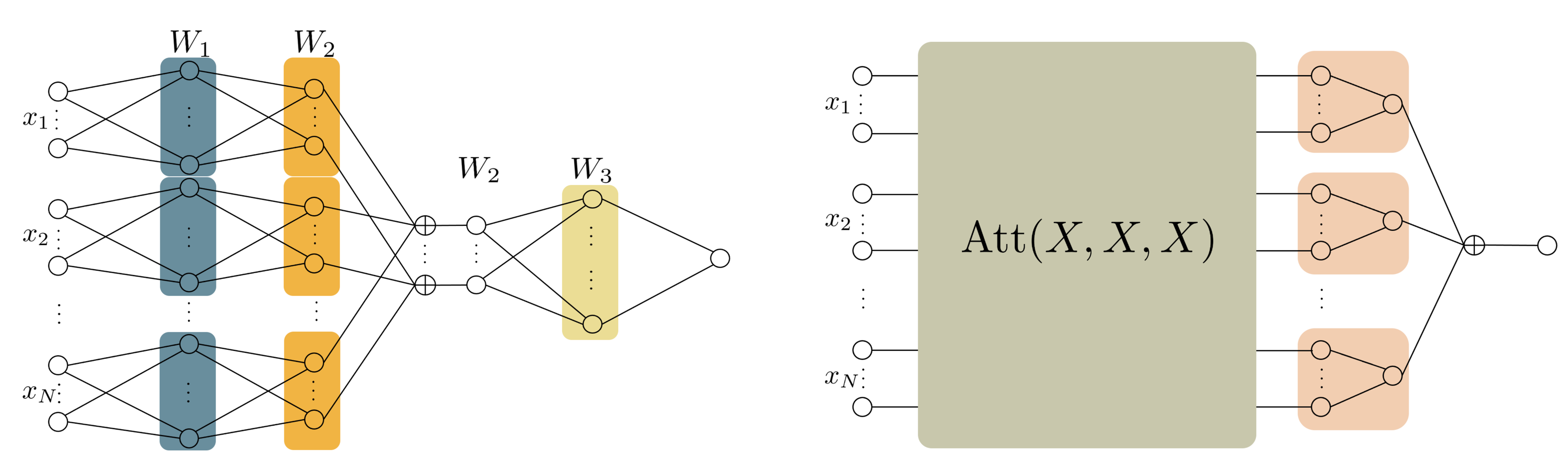
Fengzhuo Zhang†, Boyi Liu*, Kaixin Wang†, Vincent Y. F. Tan†, Zhuoran Yang‡, Zhaoran Wang*

†National University of Singapore, * Northwestern University, ‡ Yale University

Summary of Results

- MultiLayer Perceptrons (MLP) cannot approximate transformer unless the width is **exponential** in the input dimension of each channel.
- The generalization error bound of transformer function class is **independent** of the number of channels.
- The transformer helps to break **"the curse of many agents"**.

Superiority of Transfomer Over the MLP in Terms of Relational Reasoning



(a) $\rho_{\mathrm{ReLU}}(\sum_{i=1}^N \phi_{\mathrm{ReLU}}(x_i))$ with ρ_{ReLU} and ψ_{ReLU} as single-hidden layer neural networks.

(b) Self-attention mechanism $\mathbb{I}_N^\top \mathrm{Att}(X,X,X)w$.

- The function class of the permutation invariant MLP can be defined from deepset

$$\mathcal{N}(W) = \left\{ f : \mathbb{R}^{N\times d} \to \mathbb{R} \;\middle|\; f(X) = \rho_{\mathrm{ReLU}}\Big(\sum_{i=1}^N \phi_{\mathrm{ReLU}}(x_i)\Big) \text{ with } \max_{i\in[3]} W_i \le W \right\},$$

  where ρ_{ReLU} and ϕ_{ReLU} as width-constrained ReLU networks with maximal widths W_1 and W_3.
- The self-attention function class is

$$\mathcal{F} = \big\{ f : \mathbb{R}^{N\times d} \to \mathbb{R} \;\big|\; f(X) = \mathbb{I}_N^\top \mathrm{Att}(X,X,X)w \text{ for some } w \in [0,1]^d \big\}.$$

- Let $W^*(\xi, d, \mathcal{F})$ be the smallest width of the neural network such that

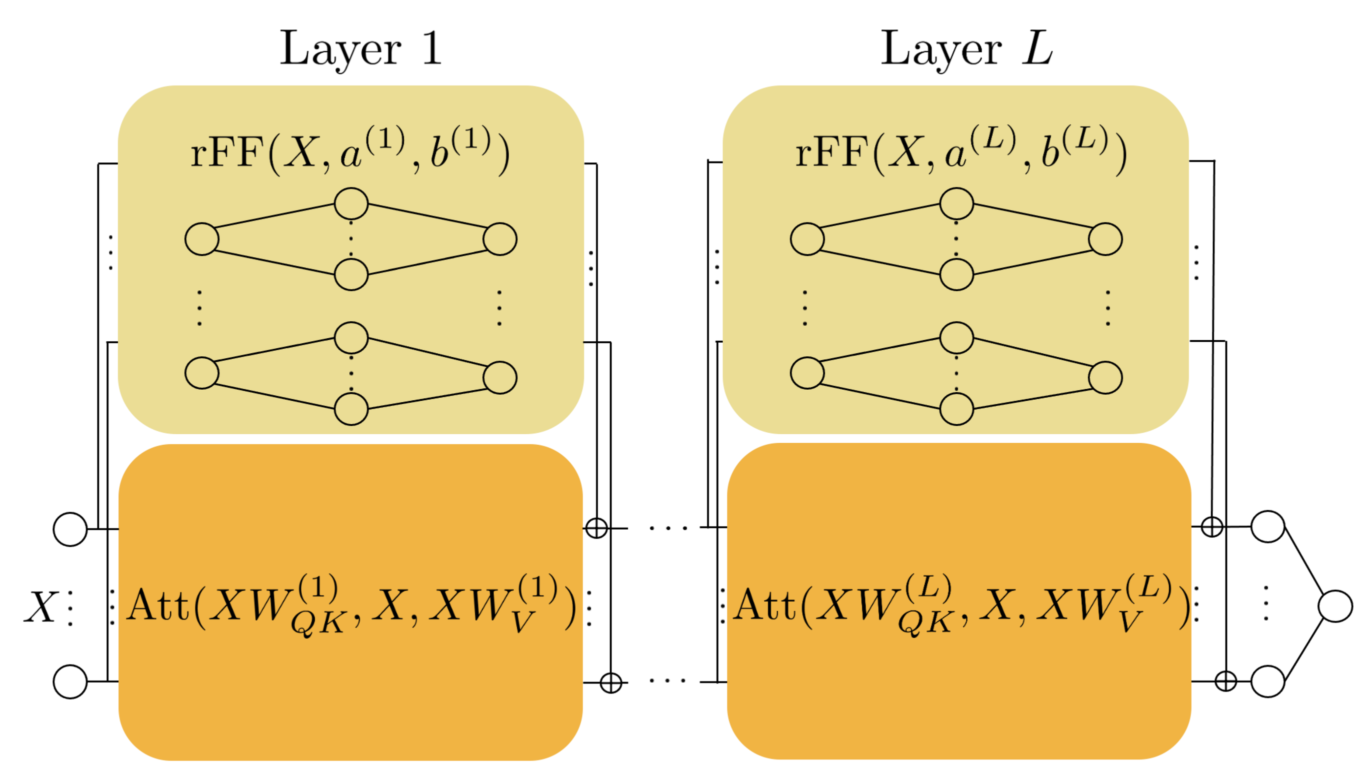
$$\forall f \in \mathcal{F},\ \exists g \in \mathcal{N}(W) \quad \text{s.t.} \quad \sup_{X\in[0,1]^{N\times d}} \big|f(X) - g(X)\big| \le \xi.$$

> With sufficient number of channels N, it holds that $W^*(\xi, d, \mathcal{F}) = \Omega(\exp(cd)\xi^{-1/4})$ for some $c > 0$.

Intuition: The deep set only adopts a single-hidden layer net to reason the relationship among inputs. It is too coarse compared with the self-attention.

Generalization Error is Indep. of the Number of Channels

- We consider the transformer with N channels and L depth.
- The norms of W_{QK} and W_V in attention and weights in fully-connected layer are bounded by a tuple B in our transformer function class $\mathcal{F}_{\mathrm{tf}}(B)$.
- Implement layer-normalization at the output of each layer.



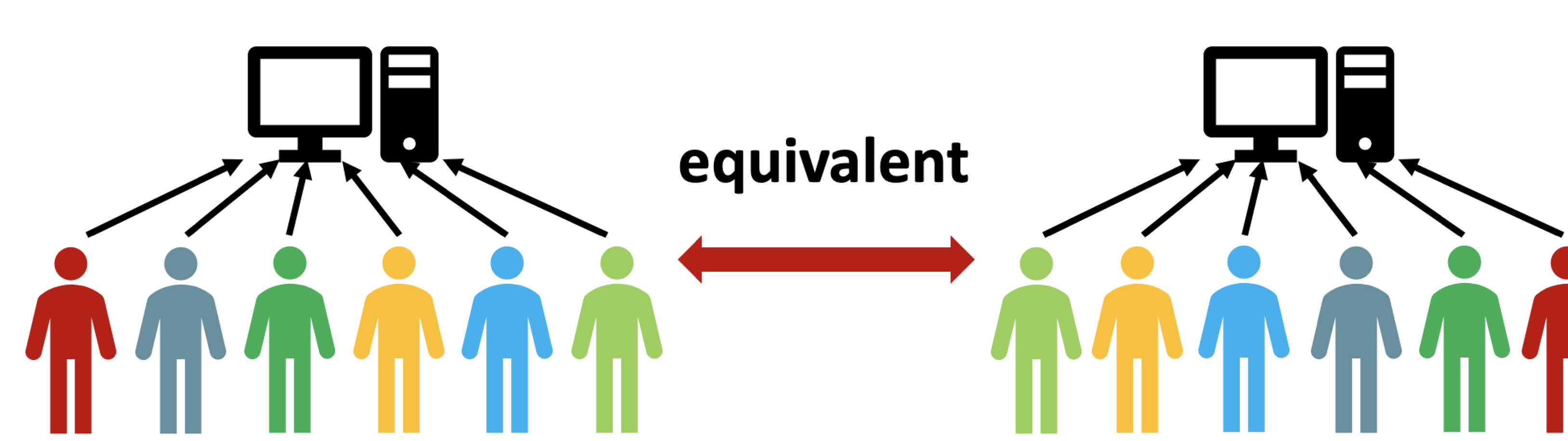
- Aim to predict the value of the response variable $y \in \mathbb{R}$ from the observation matrix $X \in \mathbb{R}^{N\times d}$, where $(X, y) \sim \nu$, and $|y| \le V$.
- Estimate $f : \mathbb{R}^{N\times d} \to \mathbb{R}$ from i.i.d. samples $\mathcal{D}_{\mathrm{reg}} = \{(X_i, y_i)\}_{i=1}^n$.
- The risk of using $f \in \mathcal{F}_{\mathrm{tf}}(B)$ as a regressor on sample (X, y) is defined as $(f(X) - y)^2$.

For all $f \in \mathcal{F}_{\mathrm{tf}}(B)$, with probability at least $1 - \delta$, we have

$$\left|\mathbb{E}_\nu\Big[\big(f(X) - y\big)^2\Big] - \frac{1}{n}\sum_{i=1}^n \big(f(X_i) - y_i\big)^2\right|$$
$$\le \frac{1}{2}\mathbb{E}_\nu\Big[\big(f(X) - y\big)^2\Big] + O\left(\frac{V^2}{n}\left[mL^2d^2 \log\frac{mdL\bar{B}n}{V} + \log\frac{1}{\delta}\right]\right).$$

- Independent of the number of channels N.
- Polynomial in the depth of the network L.

Homogeneous MARL



- We consider the Homogeneous MARL where the reward function and transition kernel are permutation invariant, i.e.,

$$P^*(\bar{S}' \mid \bar{S}, \bar{A}) = P^*\big(\psi(\bar{S}') \,\big|\, \psi(\bar{S}), \psi(\bar{A})\big) \quad \text{and} \quad r(\bar{S}, \bar{A}) = r\big(\psi(\bar{S}), \psi(\bar{A})\big).$$

- There exists an optimal policy that is permutation invariant.
- For any permutation invariant policy π, the corresponding value function V^π and action-value function Q^π are permutation invariant.
- We only consider the class of permutation invariant policies Π, where $\pi(\bar{A} \mid \bar{S}) = \pi(\psi(\bar{A}) \mid \psi(\bar{S}))$ for all permutations ψ.

Pessimistic Model-Free Offline Reinforcement Learning

- Learn from the i.i.d. dataset $\mathcal{D} = \{(\bar{S}_i, \bar{A}_i, r_i, \bar{S}_i')\}_{i=1}^n$.
- Bellman error of $f \in \mathcal{F}_{\mathrm{tf}}(B)$ for π on $\mathcal{D}$ is denoted as $\mathcal{E}(f, \pi; \mathcal{D})$.

Algorithm:

$$\hat{\pi} = \operatorname*{argmax}_{\pi\in\Pi} \min_{f\in\mathcal{F}(\pi,\varepsilon)} f(\bar{S}_0, \pi), \text{ where } \mathcal{F}(\pi, \varepsilon) = \big\{ f \in \mathcal{F}_{\mathrm{tf}}(B) \,\big|\, \mathcal{E}(f, \pi; \mathcal{D}) \le \varepsilon \big\}.$$

The coefficient $C_{\mathcal{F}_{\mathrm{tf}}}$ measures the coverage of the dataset

$$C_{\mathcal{F}_{\mathrm{tf}}} = \max_{f\in\mathcal{F}_{\mathrm{tf}}} \mathbb{E}_{d_{P^*}^{\pi^*}}\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} f(\bar{S}, \bar{A})\big)^2\Big] \Big/ \mathbb{E}_\nu\Big[\big(f(\bar{S}, \bar{A}) - \mathcal{T}^{\pi^*} f(\bar{S}, \bar{A})\big)^2\Big].$$

Assumptions:

- Realizability: $\inf_{f\in\mathcal{F}_{\mathrm{tf}}} \sup_{\mu\in d_\Pi} \mathbb{E}_\mu[(f(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2] \le \varepsilon_{\mathcal{F}} < \infty$
- Completeness: $\sup_{f\in\mathcal{F}_{\mathrm{tf}}} \inf_{\tilde{f}\in\mathcal{F}_{\mathrm{tf}}} \mathbb{E}_\nu[(\tilde{f}(\bar{S}, \bar{A}) - \mathcal{T}^\pi f(\bar{S}, \bar{A}))^2] \le \varepsilon_{\mathcal{F},\mathcal{F}} < \infty$
- The coefficient $C_{\mathcal{F}_{\mathrm{tf}}}$ is finite for the sampling distribution ν.

With probability at least $1 - \delta$, the suboptimality gap of the policy derived in the model-free algorithm is bounded as

$$V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \le O(\text{ Independent of the number of agents } N)$$

- Broken "the curse of many agents".

Pessimistic Model-Based Offline Reinforcement Learning

- System dynamics: $\bar{S}' = F^*(\bar{S}, \bar{A}) + \bar{\varepsilon}$, where F^* is a nonlinear function, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{d\times d})$ for $i \in [N]$ are independent random vectors.
- Learn the system dynamics with transformer function class.

Algorithm:

$$\hat{F}_{\mathrm{MLE}} = \operatorname*{argmin}_{F\in\mathcal{M}_{\mathrm{tf}}} \frac{1}{n}\sum_{i=1}^n \big\|\bar{S}_i' - F(\bar{S}_i, \bar{A}_i)\big\|_{\mathrm{F}}^2 \quad \text{and} \quad \hat{\pi} = \operatorname*{argmax}_{\pi\in\Pi} \min_{F\in\mathcal{M}_{\mathrm{MLE}}(\zeta)} V_{P_F}^\pi(\bar{S}_0).$$

$\mathcal{M}_{\mathrm{MLE}}(\zeta) = \{F \in \mathcal{M}_{\mathrm{tf}}(B') \mid \frac{1}{n}\sum_{i=1}^n \mathrm{TV}(P_F(\cdot|\bar{S}_i, \bar{A}_i), \hat{P}_{\mathrm{MLE}}(\cdot|\bar{S}_i, \bar{A}_i))^2 \le \zeta\}$ is the confidence region.

Assumption

- Realizability: $F^* \in \mathcal{M}_{\mathrm{tf}}(B')$.
- The coefficient $C_{\mathcal{M}_{\mathrm{tf}}}$ is finite for the sampling distribution ν.

With probability at least $1 - \delta$, the suboptimality gap of the policy learned in the model-based algorithm is upper bounded as

$$V_{P^*}^{\pi^*}(\bar{S}_0) - V_{P^*}^{\hat{\pi}}(\bar{S}_0) \le O(\text{ Logarithmic in the number of agents } N)$$

Full Paper is Available at:

https://arxiv.org/abs/2209.09845