Machine Theory of Mind

**نویسنده اول: Neil Rabinowitz**

**سال انتشار: 2018**

**محل چاپ: ICML**

**لینک به صفحه دانلود مقاله: https://arxiv.org/abs/1802.07740**

## در یک جمله بنویسید مقاله چه کار کرده

## چکیده مقاله

Theory of mind (ToM; Premack & Woodruff, 1978) broadly refers to humans’ ability to represent the mental states of others, including their desires, beliefs, and intentions. We propose to train a machine to build such models too. We design a Theory of Mind neural network – a ToM-net – which uses meta-learning to build models of the agents it encounters, from observations of their behaviour alone. Through this process, it acquires a strong prior model for agents’ behaviour, as well as the ability to bootstrap to richer predictions about agents’ characteristics and mental states using only a small number of behavioural observations. We apply the ToMnet to agents behaving in simple grid-world environments, showing that it learns to model random, algorithmic, and deep reinforcement learning agents from varied populations, and that it passes classic ToM tasks such as the “Sally-Anne” test (Wimmer & Perner, 1983; Baron-Cohen et al., 1985) of recognising that others can hold false beliefs about the world. We argue that this system – which autonomously learns how to model other agents in its world – is an important step forward for developing multi-agent AI systems, for building intermediating technology for machine-human interaction, and for advancing the progress on interpretable AI.

## کارای قبلی که مقاله ذکر کرده چی بودند (background literature, previous work)

There have been several previous works related to understanding the behavior and intentions of other agents, but they often faced challenges in terms of complexity, interpretability, and scalability. Here are one of those approaches and its associated problems:

## Hand-Crafted Models:

## Approach: Some works involve manually designing models for agents' behaviors, often based on rule-based systems or predefined decision trees.

## Problem: Hand-crafted models may not generalize well to diverse and complex scenarios. They can also be difficult to scale and adapt as the environment becomes more intricate

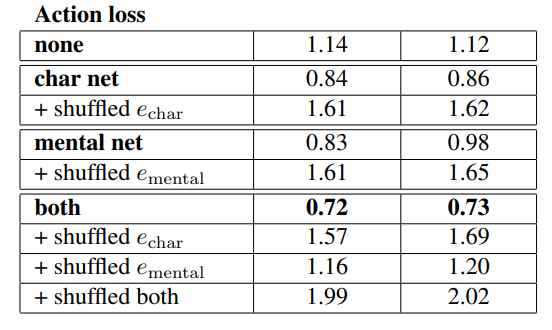
## .به طور خلاصه در یک پاراگراف بگید که مقاله، چالش، ایراد یا ضعف کارهای قبلی رو چگونه ارزیابی کرده

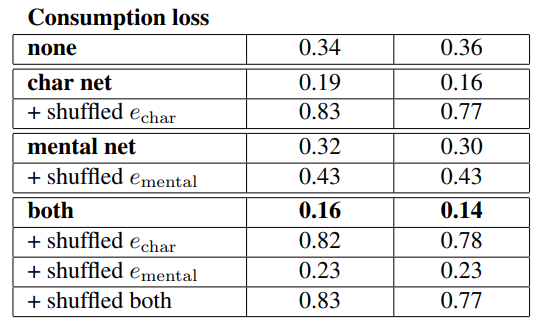
As I said there were problems in terms of complexity, interpretability, and scalability.

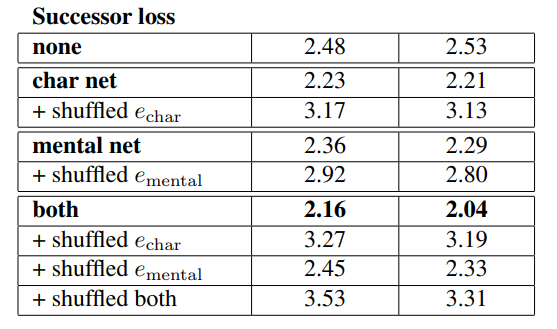
## ایده مقاله برای حل چالشها، ایرادها یا ضعف های پاراگراف قبل چی بوده (خلاصه)

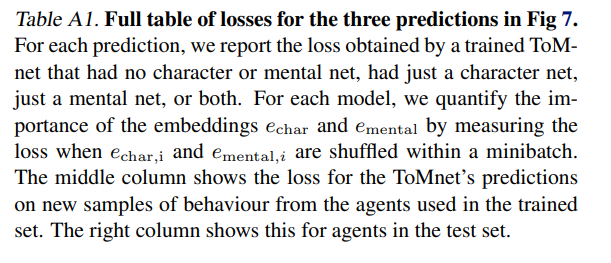
The main idea of the article is to overcome the challenges associated with modeling the behavior and intentions of other agents by using a novel approach called "Machine Theory of Mind."

## مقاله تا چه حد تونسته با روش پیشنهادی خودش به اون چالشها، ایرادها و ضعف ها پاسخ بده (metrics,evaluations,results)









## شما برای حل چالشهای گفته شده چه ایده ای دارین؟ (خلاصه)

No idea ☺

## شما برای بهبود این مقاله چه ایده ای دارین؟ (خلاصه)