摘要

随着人工智能技术的持续进步及多模态感知能力的不断增强，机器对人类行为的深度理解与语义推理正逐步成为构建以人为中心的智能系统的关键基础。行为理解任务正从传统的低层级动作识别向更高层次的语义建模与社会认知推理不断演进。人类行为具有高度的复杂性，既包括物理层面的运动建模，也涉及认知层面的意图识别与心理状态推理，以及社会层面的多主体互动与协同机制。近年来，表征学习为高维数据提供了一种统一的、数据驱动的建模范式。通过构建结构化、可泛化的中间表示，表征学习能够在高维观测信号与高阶语义之间建立紧凑、高效的映射，从而有效提升模型的迁移性与泛化能力。随着自监督学习、对比学习等技术的发展，行为理解与表征学习的融合逐渐成为研究前沿。在这一背景下，如何通过表征学习构建适用于多层次感知与认知任务的结构化、语义化行为表征，已成为人工智能领域一个重要的科学问题。然而，该领域仍面临诸多关键挑战，包括如何学习通用可泛化的行为表征、如何学习时空结构化的行为表征、如何解释隐式行为表征、如何学习行为理解与具身执行的统一表征、如何学习层次化的社交行为表征等等。为了应对上述挑战，本文围绕“基于表征学习的行为理解算法研究”这一主题展开深入探索，主要研究成果如下：

（一）针对行为表征在多任务、多场景下迁移能力不足和时空结构建模能力薄弱的问题，本文提出了一种基于预训练的人体运动表征学习框架 MotionBERT。传统方法通常为特定任务的数据训练特定任务的模型，导致模型在迁移至不同下游任务（如三维姿态估计、动作识别、人体表面重建等）时表现不佳。为提升泛化能力，本文利用大规模异构人体运动数据进行预训练，学习统一的通用运动表征。该方法采用双流时空变换网络结构，同时建模人体骨骼的空间结构与运动的时间动态。在训练过程中引入了二维到三维运动恢复任务，通过部分二维观测数据恢复完整的三维动作序列，从而促使模型学习具有鲁棒性和泛化能力的行为表征。实验结果表明，本文提出的框架在多项任务中均取得了领先性能，特别是在三维姿态估计任务上达到当前最先进水平，并在多个下游任务中实现了小样本条件下的高性能迁移。该框架不仅提升了运动建模的精度，还为多任务行为理解提供了统一的建模范式，具有良好的可扩展性。

（二）针对当前行为理解模型中隐式表征可解释性不足的问题，本文聚焦于机器是否具备对他人心理状态建模与推理的能力，探讨其内部表征是否具有“心智理论”属性。尽管近年来大型语言模型在某些社会推理任务中展现出令人惊讶的表现，但这些能力是否源于对信念、意图等心理状态的真实建模，仍缺乏系统验证。机器是否形成了某种形式的心理表征？如果有，是否可以识别、解析，甚至操控这些表征以提升推

理能力？为此，本文设计了一系列实验，发现语言模型内部确实存在与信念状态相关的神经表征，并可通过线性解码器从神经激活中提取。此外，本文还设计了信念操控实验，验证通过对模型内部信念表征的定向调节，可显著改变其在行为预测、归因等社交推理任务中的表现。这一研究不仅揭示了语言模型在隐式层面具备初步“心智理论”的潜在机制，也为提升行为表征的可解释性、构建具备社会认知能力的人工智能系统提供了新的理论支持与实现路径。

（三）针对当前行为理解与具身执行表征相互割裂的问题，本文提出了一种统一的行为表征学习方法，以实现观察与执行之间的协同建模。镜像神经元研究表明，个体在观察某个动作与执行相同动作时会激活相同的神经元群体，揭示了动作理解与身体执行之间深层次的耦合机制。然而，现有机器学习方法往往将这两类任务视为相互独立，缺乏共享表征的建模思路。本文从表示学习的角度出发，首先观察到模型在无监督条件下学习到的观察表征与执行表征在中间层存在自发对齐的趋势。基于这一观察，进一步设计了一种显式对齐机制：通过两个线性映射器将两类表征投影到共享的潜在空间，并采用对比学习策略最大化对应表征之间的互信息，从而强化两者的一致性。实验结果表明，该方法不仅提升了动作理解与执行任务各自的性能，还促进了它们之间的协同迁移，有效增强了行为表征的表达能力与泛化性能。

（四）针对现有运动预测方法主要聚焦单体行为、忽视多主体间交互与策略建模的问题，本文构建了一个涵盖三维感知与认知建模的联合系统框架，收集了一个新型多主体行为数据集——“五四”篮球训练数据集，涵盖了团队配合、战术策略、社交互动等复杂行为模式，并提出了一种基于认知层级理论的社交行为预测算法。为支持高质量数据采集，本文设计了一套高效多视角动作捕捉系统，通过正交投影方式显著降低计算成本，在保持精度的同时将计算速度提升了十倍，展示了出色的实时性能。在建模方面，本文将多智能体强化学习与生成对抗模仿学习相结合，建立了一个具备认知建模能力的社交预测网络，能够有效捕捉个体之间的互动结构及决策逻辑。实验表明，该方法在多人交互预测任务中显著优于传统序列建模方法，尤其在团队运动中表现出更强的策略理解与预测能力。该研究不仅证明了社交行为预测的可行性，更推动了行为建模从单体分析向多主体认知范式的转变。目前，相关系统已通过北京大学体育教研部与计算中心的合作，实际应用于体育教学与训练分析。

关键词：人工智能，深度学习，行为理解，表征学习

Research on Behavior Understanding Algorithms Based on Representation Learning

Wentao Zhu (Computer Application Technology)
Directed by: Prof. Yizhou Wang

ABSTRACT

With the continuous advancement of artificial intelligence and the increasing power of multimodal perception, a deep understanding of human behavior and semantic reasoning has become a fundamental building block for developing human-centered intelligent systems. The task of behavior understanding is evolving from traditional low-level action recognition to high-level semantic modeling and social-cognitive reasoning. Human behavior is inherently complex, encompassing not only physical motion modeling but also cognitive-level intention recognition, mental state inference, and social-level multi-agent interaction and coordination. In recent years, representation learning has emerged as a unified, data-driven modeling paradigm for high-dimensional data. By constructing structured and generalizable intermediate representations, representation learning enables compact and efficient mappings between high-dimensional observations and high-level semantics, thus significantly improving model transferability and generalization. With the development of techniques such as self-supervised and contrastive learning, the integration of behavior understanding and representation learning has become a research frontier. Against this backdrop, how to construct structured and semantic behavior representations via representation learning—suitable for multi-level perception and cognitive tasks—has become a crucial scientific question in AI. However, the field still faces several challenges, including learning generalizable behavior representations, modeling spatiotemporal structures, interpreting implicit representations, unifying representations for understanding and embodiment, and capturing hierarchical social behavior. To address these challenges, this thesis conducts an in-depth investigation centered on "Behavior Understanding Algorithms Based on Representation Learning", with the following major contributions:

(1) To tackle the limited transferability of behavior representations across tasks and weak modeling of spatiotemporal structures, this work proposes MotionBERT, a pretraining framework for human motion representation learning. Traditional methods often train task-specific

models on task-specific data, which hampers performance in downstream tasks such as 3D pose estimation, action recognition, and human mesh recovery. To enhance generalization, this work leverages large-scale heterogeneous motion data for pretraining, aiming to learn a unified, general-purpose motion representation. MotionBERT employs a dual-stream spatiotemporal transformer architecture to jointly model the spatial structure of human skeletons and temporal motion dynamics. A 2D-to-3D motion reconstruction task is introduced during training, encouraging the model to recover full 3D motion sequences from partial 2D observations. This helps the model learn robust and generalizable representations. Experimental results show that the proposed framework achieves state-of-the-art performance on multiple tasks—especially in 3D pose estimation—and enables high-performance transfer under low-data conditions. It improves motion modeling accuracy and offers a unified modeling paradigm for multi-task behavior understanding with strong scalability.

(2) To address the limited interpretability of implicit representations in current behavior understanding models, this work investigates whether machines can model and reason about others' mental states, exploring whether internal representations exhibit attributes of Theory of Mind (ToM). Although large language models (LLMs) have recently demonstrated surprising capabilities in social reasoning tasks, it remains unclear whether these abilities stem from genuine modeling of mental states such as beliefs and intentions. Do machines form mental representations? If so, can these be identified, interpreted, or even manipulated to enhance reasoning? To this end, a series of experiments reveal the existence of neural representations related to belief states within LLMs, which can be extracted via linear decoders. Additionally, a belief manipulation experiment demonstrates that targeted modulation of these internal belief representations significantly alters model behavior in tasks such as action prediction and social attribution. This research not only unveils a potential mechanism underlying the emergence of ToM-like properties in LLMs but also provides a novel theoretical and technical path toward building explainable and socially intelligent AI systems.

(3) To bridge the gap between behavior understanding and embodied execution, this work proposes a unified behavior representation learning method that models both observation and execution in a coordinated manner. Inspired by mirror neuron studies—which reveal that observing and executing an action activate overlapping neural populations—this work aims to capture the deep coupling between action understanding and physical embodiment. Existing machine learning approaches typically treat these tasks independently, lacking a shared representation space. Starting from the empirical observation that observation and execution

representations tend to align spontaneously in unsupervised learning, this work designs an explicit alignment mechanism: two linear projectors map the two types of representations into a shared latent space, where a contrastive learning objective maximizes mutual information between aligned pairs. Experiments show that this method not only enhances performance in both understanding and execution tasks but also promotes synergistic transfer between them, boosting the expressiveness and generalizability of behavior representations.

(4) To move beyond single-agent behavior prediction and better model multi-agent interactions and strategies, this work develops a unified system that integrates 3D perception and cognitive modeling, along with a novel dataset—the "May Fourth" Basketball Training Dataset—featuring rich behaviors such as teamwork, tactical strategies, and social interactions. A cognitive hierarchy-based social behavior prediction algorithm is proposed. To support high-quality data collection, this work designs an efficient multi-view motion capture system. By adopting an orthographic projection strategy, the system significantly reduces computational cost while maintaining accuracy, achieving a tenfold speedup with real-time performance. On the modeling side, the proposed method combines multi-agent reinforcement learning and generative adversarial imitation learning to build a social prediction network with cognitive reasoning capabilities, effectively capturing interaction structures and decision-making logic among agents. Experimental results show that the proposed method outperforms traditional sequential models in multi-person interaction tasks, demonstrating superior strategy modeling and prediction, especially in team sports scenarios. This study not only validates the feasibility of social behavior prediction but also advances behavior modeling from individual-centric analysis to a multi-agent cognitive paradigm. The system has been practically deployed through collaboration between the Department of Physical Education and the Computing Center at Peking University, supporting teaching and training analytics in sports.

KEY WORDS: Artificial Intelligence, Deep Learning, Representation Learning