Motion Question Answering via Modular Motion Programs

Mark Endo*, Joy Hsu*, Jiaman Li, Jiajun Wu
[markendo, joycj, jiamanli, jiajunw]@stanford.edu

ICML International Conference On Machine Learning

Motion Understanding

HumanMotionQA task + **BABEL-QA dataset** + **NSPose method**

HumanMotionQA task
- Evaluates **human activity understanding** through question answering in **long-form** motion capture data

BABEL-QA dataset
- Consists of complex, **multi-step** questions about **real world** human motion sequences
- Requires detection of motor cues, understanding of specific motion attributes, and temporal reasoning

NSPose method
- Is a **neuro-symbolic** approach for motion QA
- Operates on **variable length motion sequences**
- Jointly learns the QA task and **action localization**
- Does not require entity-centric input and leverages **temporal motion programs**

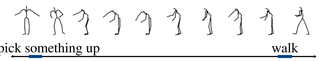
HumanMotionQA Task

Input
- Unannotated **human motion sequences** represented by 3D joint positions $S \in \mathbb{R}^{T \times J \times 3}$; T as number of timesteps and J as number of joints
- **Question** referring to actions in the motion sequence

Output
- **Answer** from a set vocabulary, including concepts that relate to different attributes (e.g., motion or body)

Question : What direction does the person move before they walk and after they use their right foot?
Answer : right

Question : What body part does the person use after they take/pick something up and before they walk?
Answer : right hand



BABEL-QA Dataset

BABEL-QA
- Is built from the BABEL dataset of **real world human motion sequences**
- Has question answer pairs procedurally generated with **logical building blocks** of `filter`, `query`, and `relate`
- Contains questions that query for **action**, **direction**, and **body part** attributes, with temporal relations of *before*, *after*, and *in between* specifying the action of interest

Query Action

Ann: walk | transition | take/pick something up | place something | transition | left | walk
Question: What action does the person do before they move left?
Program: `query_action(relate(before, filter(left)))`
Answer: place something

Ann: transition | kick | jump | walk, backwards
Question: What action does the person do before they move backwards and after they kick?
Program: `query_action(intersection(relate(before, filter(backwards)), relate(after, filter(kick))))`
Answer: jump

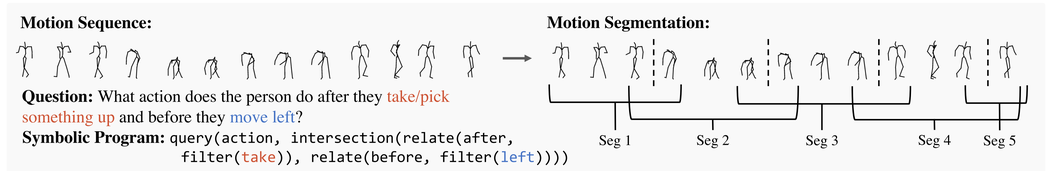
Query Direction

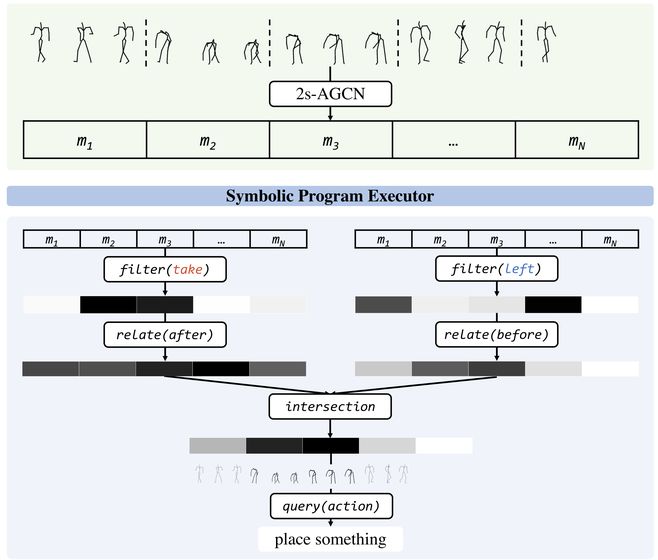
Ann: walk | transition | knock, right hand | left | walk
Question: What direction does the person move after they use their right hand?
Program: `query_direction(relate(after, filter(right hand)))`
Answer: left

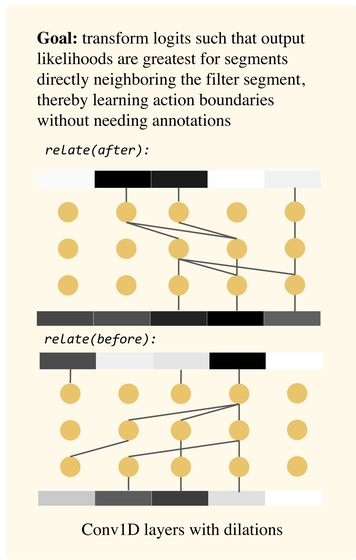
Query Body Part

Ann: right hand | transition | jump, kick, right | right hand
Question: What body part does the person use after they jump?
Program: `query_body_part(relate(after, filter(jump)))`
Answer: right hand

NSPose Method







NSPose combines
- **Modular motion programs**
- Neural networks that identify attributes from motion capture data

NSPose's programs
- Learn **concept embeddings** such as *walk*, *left*, and *right hand*
- Are executed in a hierarchical structure that follows the underlying reasoning process of the question

Temporal motion programs
- Are implemented as **1D convolutional layers** with dilation
- Enable learning of temporal action boundaries

Results

NSPose outperforms a variety of baselines: MotionCLIP models that learn powerful human motion representations & end-to-end 2s-AGCN models that leverage the same feature extractor as NSPose without programs

Model	Overall	Query Action				Query Direction				Query Body Part			
		All	Before	After	Btw	All	Before	After	Btw	All	Before	After	Btw
CLIP	0.417	0.467	0.380	0.452	0.591	0.366	**0.467**	0.292	0.222	0.261	0.261	0.278	**0.333**
2s-AGCN-MLP	0.355	0.384	0.353	0.411	0.273	0.352	0.378	0.250	0.278	0.228	0.261	0.130	**0.333**
2s-AGCN-RNN	0.357	0.396	0.349	0.396	0.409	0.352	0.400	0.396	0.278	0.194	0.261	0.111	0.167
MotionCLIP-MLP	0.430	0.485	0.411	0.470	0.545	0.361	0.400	0.271	0.333	0.272	0.304	0.222	**0.333**
MotionCLIP-RNN	0.420	0.489	0.461	0.441	0.606	0.310	0.400	0.333	0.222	0.250	**0.333**	0.167	**0.333**
NS-Pose (Ours)	**0.578**	**0.627**	**0.618**	**0.620**	**0.639**	**0.598**	0.389	**0.583**	**0.750**	**0.325**	0.296	**0.471**	0.083