Multi-modal Intent Classification for Assistive Robots with Large-scale Naturalistic Datasets

Karun Mathew *karunmatthew@live.in*
Venkata S Aditya Tarigoppula *aditya.tarigoppula@gmail.com*
Lea Frermann *lfrermann@unimelb.edu.au*

THE UNIVERSITY OF MELBOURNE

Motivation

- **Assistive robotic arms** for patients with loss of upper limb control
- Assist with simple **pick-and-place tasks**
- Flexible **voice control**

Challenge

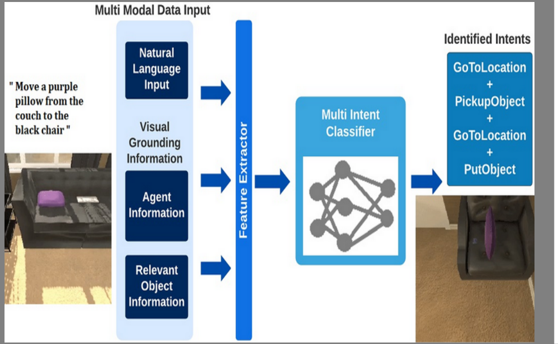
- Map command → action sequence (**intent**)
- **Multi-modal** context (vision and language)
- Language **ambiguity**

Our Contribution

- Multi-modal intent prediction from **diverse inputs**
- **Adapt large-scale ML data sets** to support flexible and robust model development
- **Multi-modal intent classifier**

Simplifying Assumption

- Object recognition and cross-modal entity linking are solved



The ALFRED Dataset (Shridar et al., 2020)

Action **L**earning **F**rom **R**ealistic **E**nvironments and **D**irectives

- Visually grounded language commands
- Crowdsourced -> diverse
- 8,000 indoor scenes

Agent Information	
Agent	{x: -2.50, y: 0.92, z: 2.50, rotation=0}
Scene Information	
FloorPlan:	FloorPlan214
Plate,	{x: -0.31, y: 0.27, z: 5.99}
WateringCan,	{x: -2.28, y: 0.45, z: 4.27}
KeyChain,	{x: -4.31, y: 0.45, z: 6.73}
Box,	{x: -2.40, y: 0.57, z: 4.57}
Laptop,	{x: -2.49, y: 0.53, z: 0.79}
Vase,	{x: -0.60, y: 1.46, z: 5.74}
WateringCan,	{x: -2.40, y: 0.44, z: 3.83}
Language Information	
High Level Task	"Move the purple pillow from the couch to the black chair."
Low Level Subtask 1	"Turn right and walk up to the couch."
Low Level Subtask 2	"Pick up the purple pillow off of the couch."
Low Level Subtask 3	"Turn around and walk across the room, then hand a left and walk over to the black chair."
Low Level Subtask 4	"Put the purple pillow on the black chair."

Augmenting Alfred for Intent Classification

1. Derive commands of varying complexity

> *"Turn right and walk up to the couch."*
> *→ {GoToLocation}*
>
> *"Turn around and walk to the chair. Put the red pillow onto the chair."*
> *→ {GoToLocation, PutObject}*
>
> *"Turn right and walk up to the couch. Pick up the red pillow. Turn around and walk […]. Put the red pillow onto the chair"*
> *→ {GoToLocation, PickObject, GoToLocation, PutObject}*

2. Impose physical constraints: reach angle and maximum reach distance

Final dataset: **150K instances** (70/15/15 train/dev/test)

The DIET classifier (Bunk et al., 2020)

- **D**ual **I**ntent and **E**ntity **T**ransformer
- Text based; embedding and symbolic features
- Maximize similarity between embedded true intent and predicted intent
- Very fast at test time

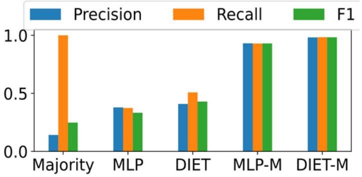
DIET-M: Multi-modal Intent Classification

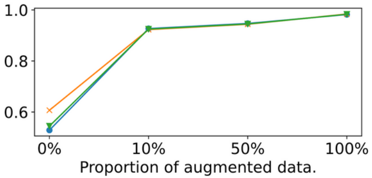
- Concatenate the DIET text encoding with visual features (coordinates)
- Pass through a feed-forward layer+dropout
- Same objective as DIET

Comparison Models

(1) **Majority** vote baseline; (2) **DIET**: Transformer, text-only; (3) **DIET-M**: Transformer, text + visual; (4) **MLP**: Multi-layer perceptron, text-only' (5) **MLP-M**: Multi-modal multi-layer perceptron, text + visual

Results & Take aways





- Powerful language **encoders** enhance intent classification performance (*Left: DIET>MLP*)
- Grounding language in **visual context** boosts performance (*Left: advantage of -M models*)
- Data **augmentation** improved performance ofDIET-M. (*Right*)

References

- Shridhar, Mohit, et al. "Alfred: A benchmark for interpreting grounded instructions for everyday tasks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2020.
- Bunk, Tanja, et al. "Diet: Lightweight language understanding for dialogue systems." *arXiv preprint arXiv:2004.09936* (2020).