VLE: A Vision-Language-Emotion Model for Humanoids in Multiparty Conversation

Peizhen Li

Faculty of Science and Engineering
Macquarie University

Aug 22, 2024

Outline

Motivation

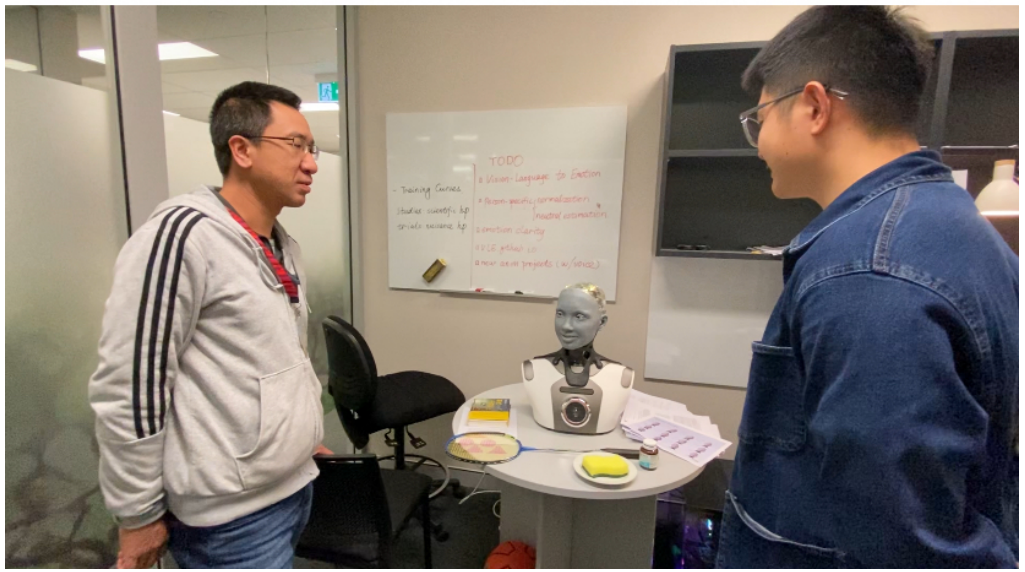
- Enable affective communication between human and humanoid robots
- Equip humanoid robots with the ability to:
  - understand human emotions in multiparty conversation scenarios.
  - deliver emotional responses through facial expressions (why?)
- Real-world applications

Challenges & Solutions

- Challenges
  - active speaker tracking
  - generalize to different persons
  - generalize to different topics
- Solutions
  - track voice activities from microphone
  - person-specific normalization
  - high-capacity architecture

Proposed Method & Results

- Bystander
- Vision-Language-Emotion model
- Demo



Discussion

- Why not incorporate acoustic modality?
  - marginal performance gain
  - limited time
- Audio2Face
  - simulation (flexibility, sim2real gap)
  - interaction with environment
  - MIMO

Limitations & Future Work

Limitations:

- Outputs of the VLE model can not be directly applied to low-level controls/end-effectors of the humanoid robot
- Personalized emotional response
- Parallel empathy only

Future Work:

- An end-to-end model
- Personalized emotional response
- Reactive empathy
- Incorporate acoustic signals

A Sad Story of Hyperparam Tuning for LMM

- 2110.408 MB (parameters + buffers, 3 modalities)
- Experimental results do not always go as expected
- Thousands of trials
  - supervised prototypical contrastive learning (SPCL) for imbalanced classification problem
  - supervised contrastive pretraining using auxiliary image dataset (Aff-Wild2)
  - auxiliary facial expression recognition training task using swin-Transformer
  - facial landmark-aware visual feature extraction and person specific normalization
  - knowledge distillation
- Reference papers

Takeaways

- Tensorboard
  - "Every time you plot something new, you learn something new"
- Hydra
- Git branches
- Tips for H100 server usage
  - Tabby
  - scp or git to synchronize files
- Recommended reading: Deep Learning Tuning Playbook
- Troubleshooting

References

[1] **IJSR 2022** - Facial Emotion Expressions in Human-Robot Interaction: A Survey

[2] **Arxiv 2023** - RT-1: Robotics Transformer for Real-World Control at Scale

[3] **Google DeepMind 2023** - RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

[4] **ACL 2019** - Multimodal Transformer for Unaligned Multimodal Language Sequences

[5] **arXiv 2024** - TelME: Teacher-learning Multimodal Fusion Network for Emotion Reconition in Conversation

[6] **ACL 2023** - A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations

[7] **EMNLP 2022** - Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation

Thank you very much!
Q&A