Machines are Among Us

Zach Minot
Georgia Institute of Technology
2nd Year Undergraduate, CS
zjminot@gatech.edu

Charles Gunn
Georgia Institute of Technology
2nd Year Undergraduate, CS & Math
cgunn30@gatech.edu

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

To what extent can neural network models communicate with each other and discover each other's identity? How would they use this information in a competitive setting? For example, in a social deduction game, players attempt to uncover each other's hidden allegiance—typically with one "good" team and one "bad" team. Players must utilize deductive reasoning to find the truth or instead lie to keep their role hidden. In this paper, we explore if neural networks can be successfully trained to compete in a scenario such as this, and how would the opposing parties interact during the period of debate.

1.1. Among Us

Among Us is a currently popular social deduction game, where the "imposters" attempt to sabotage and kill all of the "crewmates". Crewmmates have to complete tasks and figure out who the imposters are and eliminate them before the imposters win. At certain points in the game, after periods of no direct communication, players debate the roles of each individual based on information previously acquired through their personal experience. At the end of this discussion, every player votes on a single player to be eliminated. The player with the most votes is eliminated, and if there is a tie, no one is voted out. We chose to emulate this game based on the overall simplicity of the two roles and the requirement of communication for either party to succeed. If the crew do not exchange information and all vote the same person, the vote could result in a tie or a crew being eliminated. If the imposters do not bluff, the crew can easily spot the liars among the group. This provides ample room to explore and experiment with the communication between the two opposing parties.

1.2. Adverserial Networks

Within this design space, there are adversarial parties working against each other. In the deep learning realm, adversarial situations appear in adversarial examples [5] and within GANs (generative adversarial networks) [4] In particular, the latter often designs a contest between two neural networks, in the form of a zero-sum game. We build upon these concepts and foundations in our work.

1.3. Multi-agent Communication

Inherently, a social deduction game requires multiple agents to be trained and contested. This has been explored within the deep learning problem space with multi-agent subproblems. Both cooperative [2] [3] and adversarial [1] communication has been experimented with, showing that models can effective share and also selectively protect information. We reference these approaches we generate active adversarial communication between neural network models.

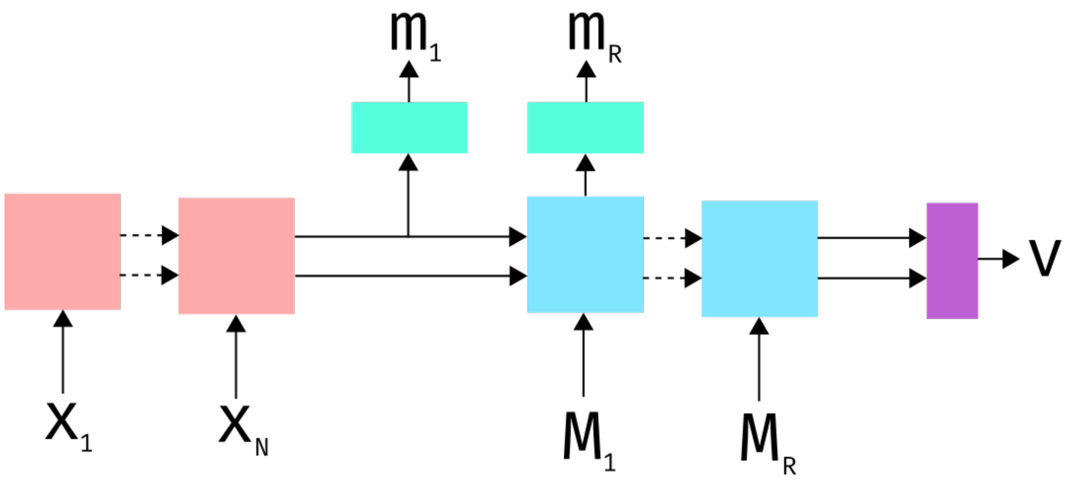

Figure 1. Diagram of the Agent model. The red section is the perception LSTM, which takes in a sequence of events. The blue section is the communication LSTM, which recieves messages, and generates messages using the green MLP. Finally, the purple section is the voting MLP, which produces the agent's vote vector.

2. Approach

2.1. Deductive Situation

2.2. Modeling Interpretation and Communication

2.3. Zero-sum Target

2.4. Training Scheme

2.5. Challenges

3. Results

3.1. Oscillating Scores

3.2. Situation Hyperparameters

3.3. Model Evolution

3.4. Interpreting Communcation Vectors

3.5. Future Directions

References

[1] Martín Abadi and David G. Andersen. Learning to protect communications with adversarial neural cryptography. *CoRR*, abs/1610.06918, 2016.

[2] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate to solve riddles with deep distributed recurrent q-networks. *CoRR*, abs/1602.02672, 2016.

[3] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *CoRR*, abs/1605.06676, 2016.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adverserial examples, 2015. Published as a conference paper at ICLR 2015.