Hebbian Learning and Preference Upgrade

Introduction

The Basics of Neural Network Semantics

- I'd like to include the related work here as "A Brief History of Neural Network Semantics" (it's really a historical account). Maybe I should start with this. Also discuss the "holy grail" of this theory
- Semantic Encodings (make this a subsection, discuss) – how this work relates. We're really using two different words for the same thing, with a slight difference in focus

Neural Network Models for Preference Upgrade

For both, focus on showing that the two are semantically equivalent, *and only then*, at the end, mention axioms & completeness by reduction

Making Neurons Wire Together; Radical Upgrade

If They Fired Together; Conservative Upgrade

Iterated Hebbian Learning; Knowledge-Preserving Upgrade

- Focus on showing that the two are semantically equivalent, *and only then*, at the end, mention axioms & completeness by reduction

Single-Step Hebbian Learning; Bubble Upgrade

- Focus on showing that the two are semantically equivalent, *and only then*, at the end, mention axioms & completeness by reduction

Stabilized Hebbian Learning; TODO

Why Bother with Completeness?

- tldr; completeness of any of these update systems is *equivalent to* having neural network model building, i.e. being able to do AI alignment (never unlearn)
- Here's where I should do a demonstration of this!!! Have something that I want a Hebbian learner to never unlearn, and then actually build the network that does this! (Even if I end up leaving the stabilized variant as an open problem)

Conclusions and Future Directions

References