Morgan Bryant, mrbryant@stanford.edu

Overview This proposed project test the ability for a network to mimic the human process of *skill synthesis* as a keynote example of transfer learning. The key experiment will compare various networks at their ability to transfer effectively given a new task that can use old skills.

The experiment question is: what neural network architecture best mimics human transfer learning? Up to date, networks have shown extensive ability to learn tasks, but they often take large amounts of time to do so. Due to effects such as catastrophic interference [1,2], networks are also notoriously bad at learning new skills while retaining ability to perform old ones. In contrast, natural minds can both learn new tasks quickly if they are similar to previously known tasks [3] and reuse (or transfer) previous skills as composite skill acquisition [4,5].

This project will attempt to mimic human patterns of transfer learning using neural networks, with the assumption that a network with similar patterns may be more cognitively realistic. To do so, we will compare six artificial networks to human results at their ability to transfer via combining previously learned skills into new, complex skills.

The task is a composite classification task. Task A requires certain skills, Task B requires some different skills from Task A, and Task C requires only skills that are the used for either Task A or B. Specifically, humans will be tested on the question A, "Is this creature a *fep*?" where the discrimination of species *fep* is given by a logical formula of features, such as whether the creature has antennae or whether the creature has either spots or stripes. Task B will be a question, "Is this creature a *bar*?", where a different species *foo* is given by a different combination of formulae. Finally, Task C will be "Is this creature a *bep*?" where *bep* is different from both *fep*s and *bar*s but that the features that determine *bep* are drawn from the sets of features that determine *fep* or *bar*.

Human trials will be conducted over Amazon Mechanical Turk. The results of human trials will be used as a comparison against the neural models. The specific query sought is: How quickly to humans learn task C *relative* to learning tasks A or B? Machine trials will consider the following six networks:

1. Two standard but independent neural networks are trained for the two tasks A and B, and a third network takes the two nets A and B but replaces the two top layers with a single large layer for task C
2. Same as (1), but the two nets A and B are frozen when they are conjoined for C.
3. A standard network is tasked to learn all three tasks A, B, and C, in smoothed sequence: at first, all the tasks are from A, but gradually B tasks are added in, then C tasks.
4. A modular network as in [6] is used, except that the third options in C are only provided once A and B are learned.
5. A modular attention network where two *modules* are learned for A and B each, and the third module for C is initialized as [A B].
6. A modular attention network as in (5), but with modules for A and B frozen once combined into C.

The artificial network with the best machine trials will be compared to human results to see if it indeed can transfer learned skills at a similar relative rate as humans do.

References

1. McCloskey, Michael, and Neal J. Cohen. "Catastrophic interference in connectionist networks: The sequential learning problem." *Psychology of learning and motivation* 24 (1989): 109-165.
2. Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the National Academy of Sciences* (2017): 201611835.
3. Salvucci, Dario D. "Integration and reuse in cognitive skill acquisition." *Cognitive Science* 37.5 (2013): 829-860.
4. Anderson, J. R. Acquisition of cognitive skill. 1982. *Psychological Review, 89,* 369-406. Source from *Learning and Memory: From Brain to Behavior*, Gluck, Mark, Mercado, Eduardo, and Myers, Catherine, Worth Publishers, New York, 2008.
5. Bransford, John D., & Schwartz, Daniel L. Rethinking Transfer: A Simple Proposal With Multiple Implications. Review of Research in Education, Chapter 3, Vol. 24, p61-100. 2001.
6. Michael I. Jordan and Robert A. Jacobs, “A Competitive Modular Con- nectionist Architecture”, MIT, MA, USA, https://papers.nips.cc/paper/430- a-competitive-modular-connectionist-architecture.pdf