4/3/17
Morgan Bryant

A Model of Skill Transfer using Modular Neural Networks: Project Proposal

Contents:
-- Modular neural network shortcomings compared to simple neural networks
-- A specific ability that modular networks have over standard networks
-- Cognitive science literature: justifications and connections
-- Specific project proposal
-- Extensions: prospective

I propose using a modular neural network that I've designed to model the phenomenon of cognitive skill transfer, specifically as a consolidation of multiple skills for the sake of a new, more complicated task. This project will test its capability for skill transfer compared to other neural architectures.

1) Modular Neural Network Shortcomings Modular neural networks are similar to simple ones; the key differences are that simple ones suffer crosstalk [reference] but are more expressive. Let's consider what other advantages modular networks might have. Among the various tasks that a _modular_ NN can do, most tasks that might seem like initially like improvements turn out not. For example, take the following transfer task: train a modular neural network to do a task involving skills A and B, then test whether it has improved ability to learn a task only involving skill A. However, transfer is already better done by the standard neural network strategy of taking a network trained on a relevant task on supercomputers and then replacing the top layer and training the resulting network to the new task. Additionally, there's little research to support that a top-layer-replacing scheme is what human minds do when they transfer learn. And beyond transfer, there are numerous tasks that may seem promising but actually don't work out. Specifically: * The ability to change its train-time and test-time efficiency is not a critical point, since it is already a bloated model. * The distributed modules have no reason to be easier to examine critically than normal standard neural network layers. * Similarly, incorporating hand-selected operations as linear units within the modules for the sake of testing bottlenecks has little practical utility. See [10] for more about the tradeoffs between modular and standard networks.

2) Modular Neural Network Advantages There is one objective, however, that standard network procedures have been insufficient for: an ability to smoothly *combine* skills learned elsewhere. That is, if two standard networks learn tasks that develop skills A and B respectively, there is no clear way to train a composite network that can use both the skills A and B naturally.

Modular networks, however, have the potential to do this. Modular networks, and the exemplar Modular Internal Attention Network (ModIAN), have a bipartite focus on skill development and skill application, represented by applying the module to X_theta (parameter half-layer) and X_B (selection half-layer) in ModIAN. Comparatively, standard

networks only care about skill development. The skill application aspect of modular networks gives them the ability to deliberately relearn how to apply skills that are already developed, avoiding redeveloping those skills.

A high-level specification given two ModIANs N1 and N2 that learned distinct modules M1 and M2 and how to apply them to their respective tasks, a third ModIAN N3 could be developed with its module initialized as a concatenated M3 = [M1, M2]; N3 would then only need to learn how to *use* the pre-developed skills in its M3 for its new task, avoiding retraining M3 and only training X_theta and X_B.

If this hypothesis is true, then a collection of module networks can be trained independently to learn specific skills and later be combined in a system that can use those modules effectively, in a way that current networks are incapable of doing naturally. Additionally, this model may be signal a more general procedure that allows machine learning and deep learning components to be combined in a deep learning framework, recursively and arbitrarily: the ability for the network to 'think' about what it is doing suggests a simple meta-cognition. From another lens, this can be thought of a neural-network-guided ensemble of neural networks.

3) Cognitive science literature: justifications and connections Beyond the machine learning implications, this model is psychologically grounded, and its effectiveness would support some existing theories of cognitive skill learning and transfer.

- ModIAN can model how a person combines learned skills to accomplish a more complicated task, supporting the schema hypothesis that knowledge is used, stored, and reused in 'chunks' for repeated arbitrary access, in contrast to relearning skills for every novel event. Accordingly, [3] claims that skill acquisition uses a complex interaction of elementary components.
- Preparation for Future Learning (PFL) interpretations of transfer supported [8]
- [2] claims that human skill acquisition requires integration and reuse, so a neural model claiming to do the same should have a response to this. Integration is solved by the neural aspect of the network: the new network would aggregate the various module's outputs into overall output by training a new fully-connected layer. Reuse is maintained by using previously-trained networks.
- [3],[5],[6] and [7] support that there are three fuzzy phases of learning by practice: phase 1 takes place when target task is first encountered, and the cognitive strategy is to understand the rules and use analogical reasoning to apply previous knowledge; in phase 2, learning proceeds by feedback of results; and in phase 3, performance is increased by finding new strategies, shortcuts, and 'inlining'. The model below would support such a framework.

4) Specific Project Proposal The specific proposed project will compare the ability for a network to mimic the human process of *skill synthesis* as a keynote example of transfer learning. The first baseline experiment is to simply see if a new network can transfer effectively, given a new task that can use old skills.

As a note, cognitive (composite) skill learning that does not involve transfer is being explicitly ignored for this experiment. For example, VanLehn [7] identifies three questions that human learning events can be analyzed by: What provoked the person to switch from

the task to learning? How did they reason? Did / how did they retrieve principles? This current experiment avoids these questions as well.

Consider two networks N1 and N2 with trained modules M1 and M2. Initialize a new N3 with the module M3 set to M3=[M1 M2], and the rest of the weights are random. The specific learning scheme for the network N3

1. The first phase of learning, which corresponds to internalizing instructions, would most generally correspond to a search procedure over numerous possible component modules to use in the method. We assume that this determination of module composition is spontaneous [9] as a transfer task; so, we simply provide N3 with its initial module for this experiment.
2. Second, we *fix* M3 from being updated according to backpropagation. Train the rest of the layer – that is, X_theta, X_B, and FC. This corresponds to the cognitive strategy of attempting to apply previous skills.
3. Finally, unfix M3 and let it be trained as well. This improves the overall performance as a specialized system. This also correlates to the optimization/generalization phase of human learning, as the person begins to consider this a task on its own, subject to individual

This procedure will be compared to these five alternatives: (1) Two standard neural networks are trained for the two tasks, and a third network takes the two but replaces the two top layers with a single large layer; (2) N3 never does step 2; (3) N3 never does step 3; (4) N3 learns from scratch; and (5) a standard network learns from scratch. We hypothesize the above procedure outperforms alternate procedures: (1) and (2) explicitly do not undergo attempts to apply previous strategies, and (3) does not demonstrate an *intentional focus* on applying previous strategies. (4) and (5) are comparable in terms of cumulative learning speed in contrast with the hypothesis, as well as the tradeoffs between retroactive interference [10] and expressibility. It would also befit comparing against other similar, state-of-art modular networks that are capable of doing a similar task.

If results are sparse, we may try ignoring the N2 network and M2 module, and instead make M3 = [M1 Noise].

The specific task is undecided. A first possibility is: task for N1 is standard MNIST recognition, task for N2 is addition, and task for N3 is addition of two MNIST. However, it would be more ideal to identify a task where there is some direct cognitive science literature about – perhaps the frog papers? Even if we hand-create the tasks, ideally they would be of roughly similar complexity.

The FC layer might be better represented by mapping to the next layer I components of all N outputting modules, using at least B/I such mappings: else, spatial crosstalk concerns [10]. Until the network has issues, I probably won't make this change.

5) Extensions

Papers of particular relevance, sourced from [7] pg '533', Bolivar90, Singley&Anderson89, Frensh&Geary93, indicate that a small amount of practice on subset tasks (ie, N1 and N2) help N3, but more practice (ie training) will result in highly diminishing returns. A simple, reasonable task would be to vary the amount of transferrable material is by halting N1 and N2 training early.

In implementation, perhaps anneal the amount that the M3's weights are affected. In a further attempt to mimic the human ability to also identify errors, an extension

involving adding new modules is outlined below. The extension effectively builds up the new module M3 with additional module units as extra computational units to assist consolidation. It seems psychologically grounded but would require legwork.

1. Perform the fixed-M step as outlined above; corresponds to the cognitive strategy of attempting previous skill application first. If the task error is low enough, exit.
2. Otherwise, switch off between the following two operations:
    a. Append 1 new module m to the system of modules M as M[:-1]. Train only this new module and X and FC, as in step one. This corresponds to an attempt to learn an additional skill, to combine with previous skills, and also determine how to apply them in concert. The network designer could also permit adding more than 1 new module, if there is some prediction about *how* many new ones are expected to be useful. If the task is sufficiently solved, exit: a new skill-schema was learned and the individual applied "knowing with" [8] growth of the overall skill.
    b. Else, if (a) was insufficient, retrain the module system as a whole: train back through all M, X, and FC. This involves an entire reevaluation of the skillset. If the task is sufficiently solved, exit: corresponds roughly to a person having to 'let go' and reevaluated a faulty prior epistemological belief. Otherwise, the model needs to loop starting back at (2) and the individual underestimated the amount of new skills they would need to acquire to solve this. If no looping is performed, it suggests the task was too difficult to use previous skills on, and instead, an approximate network was needed corresponding to replicative (not applicative) knowledge.

Another task: see Catrambone (1994a,b), (Catrambone & Holyoak 1990, 1987) from [7, pg 10 "522"]: transfer significantly increases when tasks are modified to highlight transfer opportunities. This could be studied as: Does M4 = [M3 Noise] learn to use M3 components? A further extension here concerns the how the amount of similarity between tasks 1/2/3 and task 4 can affect transfer, as there are results that have the human equivalents (see [7] pg '533', Bolivar90, Singley&Anderson89).

A next significant task could be how an agent can decide what skills from previous learned modules should be applied to a new task: this may be as simple as making a small Decider network (a standard NN) choose how to initialize X_B, and this might benefit from installing the X_B/X_theta optimization switch capability.

It may be interesting to test how well the ModIAN can shut off certain inputs when training certain kinds of components; selective specialization has the added bonus of shielding.

Gating mechanism: not inspired by but similar to LSTM gates; theoretically justified by [11].

References

[1] Cattell, Raymond B. Theory of fluid and crystallized intelligence: A critical experiment. Journal of Educational Psychology, Vol 54(1), Feb 1963, 1-22.

[2] Salvucci, Dario D. "Integration and reuse in cognitive skill acquisition." *Cognitive Science* 37.5 (2013): 829-860.

[3] 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.
[4] Fitts, P. M. Perceptual-motor skill learning. In A, W. Melton (Ed.), Categories of human learning. New York: Academic Press, 1964.
[5] Schwartz, M. S. & Fischer, K. W. (2004) Building general knowledge and skill: cognition and microdevelopment during science learning. In A. Demetriou & A. Raftopoulos (Eds.), Cognitive developmental change: Theories, models, and measurement. Cambridge, U.K.: Cambridge University Press.
[6] Cognitive Skill Acquisition. Encyclopedia of the Sciences and Learning, Springer US. 2012.
[7] VanLehn, Kurt. Cognitive Skill Acquisition. Annual Reviews of Psychology 47: 513-39. 1996.
[8] 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.
[9] VanLehn, Kurt. "Rule-learning events in the acquisition of a complex skill: An evaluation of CASCADE." *The Journal of the Learning Sciences* 8.1 (1999): 71-125.
[10] Auda, Gasser, & Kamel, Mohamed, Modular Neural Networks: A Survey, International Journal of Neural Systems, Vol. 9, No. 2, 129-151, April 1999.
[11] McLachlan, G.J. & Basford, K.E. (1988) Mixture Models: Inference and Applications to Clustering. New York: Marcel Dekker.