Neural Network research 1987-2002 relevant to Deep Learning Research --2019 (Hanson, SJ).

Walking through the recent 2017 and 2018 N√IPS (“NeurIPS”, nee “NIPS”) conference is an intellectual delight, both in terms of the enormous and exciting progress in AI, but for those of us who were part of the last Neural Network revolution, is also a walk down memory lane. This annotated bibliography is offered as one source of (highly biased selection, of course), research we had done that may be relevant or useful to future Deep Learning research and possibly provide some context for potentially new research directions. Queries or comments to jose@rubic.rutgers.edu.

- Hanson S. J. & Kegl J. (1987), *PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences*, ***Ninth Annual Conference of the Cognitive Science Society***, *Seattle, pp. 106-119.* ***This paper describes one of the first “large” Auto-encoders to also train on a “large” data corpus of 1 million running words—the BROWN Corpus. The auto-encoder took 4 days to converge on a VAX11/780 training on 35k sentences. The network consisted of 24,615 weights and ~40 hidden units (one of the largest at the time). PARSNIP learned lexical category pattern completion and limited recursion from merely producing sentences from input to output through a bottleneck hidden layer (yes we tried more layers at the time—but it didn’t learn for well know reasons now-VAX11/780?!).***

- Hanson, S. J. & Burr, D. J. (1988), *Minkowski-r Back-Propagation: Learning in Connectionist Models with Non-Euclidian Error Signals*. IEEE Neural Information Processing Conference. **American Institute of Physics Or Advances in Neural Information Processing-0** pp. 348-358. ***This paper was in the first NIPS conference (sometimes referred to ad NIPS-0, a direct offshoot of the Caltech “HOPFEST”). This paper describes a novel regularization and noise filtering method that explores the backpropagated error signal as a distance metric. Standard squared error –Euclidean error-- is generalized to Minkowski error metrics (e.g. L1, L), resulting in LAD loss (also LASSO penalty) and TUKEY’s 1.5 noise reduction hueristic. The paper also provides a few examples of the distance metrics with backpropagation and the advantage of r values <2, very little work has occurred since this time involving r>2, which allows for heavier tails in the error residual distributions.***

  Hanson, S. J. & Pratt, L. Y. (1989), *Comparisons of Constraints on Minimal Networks with Back-Propagation*, **Advances in Neural Information Processing-1** ,Morgan Kaufmann, pp. 177-186. ***This was the first published account of regularization in Backpropagation. This paper was initiated with a discussion I had with Dave Rummelhart on regularization. He had developed a version of “weight decay” and in this paper we derived his particular version and other generalizations which are presently popular in deep-learning implementations today.***

- Hanson, S. J.(1990)*, A Stochastic Version of the Delta Rule,* ***PHYSICA D****, 42, 265-272.* ***This was one of the early methods to inject adaptive noise into a Neural Network. There were a few others algorithims around this time, but mainly focused on hidden units. The DROPOUT algorithm, popular today in Deep Learning is actually a special case of the Stochastic Delta Rule (SDR) with clustered binomial weight noise (see recent papers below). In general SDR weights are conceptualized as random variables (gaussian) with parameters mean and variance, which are used as to update weights in the network. This results in model averaging, smoothing over local minima, and a local simulated annealing per weight inducing a directed random parameter search. (see recent papers)***

- Hanson, S. J. (1990), *Meiosis Networks****, Advances in Neural Information Processing-2,*** *Tourezsky, (Ed.), Morgan Kaufmann, pp.533-542.* ***This paper was an example of an early neural network construction algorithm along with Scott Falman’s ‘Cascade Correlation’ also developed in 1990. It was based on the stochastic delta rule (SDR) and theoretically could build random lattices dependent on gradient weight***

noise which indexed where in the hidden unit layer and location feature detectors lacked required complexity for the task. Initially, the algorithm was designed to build both within layers and across layers of the network, thus constructing a tree like structure as a function of the complexity of feature development.

- Hanson S. J., Olson, C., (1990), ***Connectionist Modeling and Brain Function*: The Developing Interface.** MIT Press/Bradford, 396 pp. (BOOK). ***This was an early example of compilations of Brain Modeling and Neural Network intersections, which continue in promising ways today with Deep-Learning architectures. Based on a workshop in Princeton, 1988, organized with C. Olson and G. Miller.***
- Olson C. & Hanson, S. J., (1990), *Spatial Representation of the Body*, **In Connectionist Modeling and Brain Function: The Developing Interface**, S. J. Hanson & C. Olson, (Eds.) MIT Press/Bradford Books, pp. 193-254. (Book CHAPTER) ***One of the early examples of mapping complex high dimensional biological motion with neural networks. In this case a geometrically realistic arm model with 13 DOFs, was mapped from eye gaze coordinates to random configurations of arm positions based on shoulder/elbow/wrist/neck rotational values. Recovered and identified hidden unit surfaces, were similar to those measured in Primate cortex (c.f., Georgopoulus) varying over elbow/shoulder continuous values. There were a number of innovations introduced in this paper, one was "residual error neural networks", these type of networks could learn the error from the primary network in order to cancel the error of the primary network. Subsequent networks could learn the error of the secondary residual network and so on. Typically in these motor mapping tasks, 3-5 networks would be able to model the approximation error to some negligible level improving the overall map performance.***
- Hanson S. J. & Burr, D. J., (1990), *What Connectionist Models Learn: Toward a theory of representation in Connectionist Networks*, **Behavioral and Brain Sciences,** 13, 471-518. ***This was one of the first theoretical accounts of how learning and representation interact in Neural Networks. In contrast to AI representational theories, which tended to focus on logical form or theory of mind accounts, this paper provided a framework for thinking about distributed structures and their characterization. Technical innovations that the paper Introduced included hierarchical clustering and multivariate analysis of hidden units and their interpretation.***
- Harnad, S. Hanson, S. J. & Lubin, J. (1991), *Categorical Perception and the Evolution of Supervised Learning in Neural Networks*, In D. W. Powers & L. Reeker (Eds.) **Working Papers of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology** pp. 65-74. ***This paper continued the investigations of auto-encoder properties. In this case we were able to combine an auto-encoder with a classification bias (this was one the first cases to combine auto-encoding and classification in the same neural network) and learn simple perceptual categories and show that the representation of the boundary between the categories became hyper-expanded increasing the between-member distances and reducing the within member distances wrt to actual feature distance. This effect appeared related to the sigmoidal function representing neural activation and the auto-encoding/classification compared to auto-encoding the same stimuli independently.***
- Hanson, S. J. & Gluck, M. A. (1991), *Spherical Units as Dynamic Consequential Regions: Implications for Attention, Competition and Categorization*, **Advances in Neural Information Processing-3**, R. Lippman, J. Moody, & D. Touretzsky, (Eds.), Morgan Kaufmann, pp., 656-665. ***Roger Shepard began to develop neural network models to account for a basic perceptual phenomena involving feature correlations and attention. This work was based on work Shepard had done in 1961 and 1987. Others had built in new parameters or kernels in neural networks which could then fit the pattern of data—but didn't rely on unbiased learning. Gluck and I showed that a specific competition kernel for hidden units could show the same fit to the pattern of data, simply by exposure to the tasks, and merely the standard backpropagation learning given the special HU kernel. Recent work (Hanson, Caglar, & Hanson, 2018) replicates and expands this work to showing DL architectures even without any special kernels also fit the same complex pattern of human data.***

- Hanson, S. J., (1991), *Behavioral Diversity, Search, and Stochastic Connectionist Systems*, In **Neural Network Models of Conditioning and Action**, M. Commons, S. Grossberg & J. Staddon (Eds.), New Jersey: Erlbaum, pp. 295-345. (Book CHAPTER) ***This book chapter showed how SDR could be generalized the to reinforcement learning, in particular Temporal Difference learning. Using animal behavior research showing that behavioral diversity increases immediately after reinforcement. In this paper we show*** **that this increase could be the basis of random behavioral search increasing the probability of response/consequence contingency. SDR would allow this transitory increase in behavioral noise, producing faster and more efficient reinforcement learning algorithms simulated in the paper.**

- Hanson, C. & Hanson, S. J. (1992), *Development of Schemata During Event Parsing: Neisser's Perceptual Cycle as a Recurrent Connectionist Network,* In **Fourteenth Annual Conference of the Cognitive Science Society**, Bloomington, pp 438-449. also in **J. Cognitive Neuroscience** , 1996. ***This paper proposed that an RNN could be used to represent the Perceptual Cycle, a concept first developed by Ulrich Neisser. Neisser proposed the perceptual cycle as theoretically connecting attention, memory and perception as a fundamental cognitive dynamic. The RNN predicted event change points in order to decode continuous movie stimuli. The hidden unit structure recovered the event transition structure of the original movie sequence and was able to model the human event change time series.***

- Hanson, S. J., Petsche, T., Kearns, M. & Rivest, R. (1994), **Computational Learning Theory and Natural Learning Systems**, Vol 2, MIT Press, Bradford, 447pp. (BOOK) ***This was the second volume of a four volume series that provided an intersection of neural networks, machine learning and learning theory, which at the time were all considered to be distinct sub-fields with different goals, data sets and algorithms. This series of workshops brought leaders of all three groups together in a common format for the presentation of new research and common discussion across the learning sciences.***

- Hanson , S. J. & Gurvits, L *The Temporal Triangle, Generalizations of Temporal Difference methods.* (1994) **SCR-LS-024.** ***This technical report was based on an idea of using higher order temporal differences to do sequence credit assignment. L. Gurvits, developed the mathematical framework for the HO temporal differences and proved some relevant theorems showing that HOTD converges. This was never published but it is still referred to in the control/reinforcement literature.***

- Hanson, S. J., (1995), *Some comments and variations on Back-propagation.* In T**he Handbook of Back-propagation**, Y. Chauvin & D. Rummelhart (Eds.), New Jersey: Erlbaum, pp. 292-323. (Book CHAPTER) ***This chapter summarized a cluster of variations on Back-propagation since the 90s, that included SDR, weight decay, Mieosis networks and other variations. The chapter makes the case that Back-propagation as framed by Rummelhart, was in fact unique, productive and transformative of modern Artificial intelligence.***

- Petsche, T., Marcantonio, A., Darken, C., Hanson, S. J., Kuhn, G.K., Santoso, I. (1996), *A neural network autoassociator for induction motor failure prediction.* **Neural Information Processing Systems-4,** pp. 924-931. ***Another exploration of Auto-encoders ("auto-associators") for modeling signal variation of a mechanical device (motor in this case) and capturing a model of normative performance that could be frozen and then thresholded for anomaly prediction. This system was eventually deployed for Siemens windshield motors, used in mercedes and other vehicles.***

- Japkowicz, N., Hanson, S.J. & Gluck, M. (2000). *Nonlinear Autoassociation is not equivalent to PCA.* **Neural Computation**, 12, 531-545. ***Yet another auto-encoder exploration, in this case a series of experiments with autoe-ncoders ("auto- association") that even with a single hidden layer were shown not to be equivalent to principle components analysis especially in highly nonlinear mappings. This was a precursor to the deeplearning demonstration of the non-equivalence with PCA.***

- Hanson S. J. & Negishi M., *(2002) On the Emergence of Rules in Neural Networks,* **Neural Computation***, 14, 1-24.* ***This paper provides one of the first demonstrations of grammar learning transfer using a RNN to learn regular grammars (FSM). Training on the same FSM grammer, the lexicon was switched out to a***

new lexicon once the network was able to master the original FSM+Lexicon-1. After 9 such lexicon switches, generalization was measured for a 10th switch, with no learning. Remarkably, the RRN was able to transfer the FSM to the unseen novel lexicon with at least 60% savings.

RECENT PAPERS relevant to Deep Learning and Neural Networks.

Hanson, C, Caglar, L. R. & Hanson, S. J. (2018) Attentional Bias in Human Category Learning: The Case of Deep Learning, **Frontiers in Psychology https://www.frontiersin.org/article/10.3389/fpsyg.2018.00374, 9, 1664-1078, 10.3389/fpsyg.2018.00374**

Frzaier-Logue, N. and Hanson, S.J. (2018) Dropout is a special case of the stochastic delta rule: faster and more accurate deep learning, **arXiv.org>cs> arXiv:1808.03578**