Self-Constraining Deep Neural Networks

Stefan Leijnen (s.leijnen@hva.nl)
Amsterdam University of Applied Sciences &
The Asimov Institute for Artificial Creativity and Constraint

The multi-layered architecture of deep neural networks provides a model for learning abstractions of features, which may in turn act as features for a higher layer (Bengio, LeCun, & Hinton, 2015). This leads to a hierarchical representation of the input data, where connections are trained to project the activations in a lower layer to a particular activation pattern (for deterministic networks) or particular set of activation patterns (for nondeterministic networks) in a higher layer. This (set of) pattern(s) is determined by a constraint on the complete set of all possible activation patterns in this layer with respect to its input layer.

Constraint, a psychological concept referring to structure (Garner, 1962), is used here as the amount of interrelatedness between a system's variables measured in informational terms, which can be expressed in terms of uncertainty (i.e. orthogonal variables yield maximum uncertainty). Applying this measure of constraint to a pair of neural network layers gives maximum constraint (i.e. minimal uncertainty) if each unique activation pattern in the lower layer systematically projects to a unique activation pattern in the higher layer. The usefulness of a training algorithm is then determined by its rate of sampling and finding constraints that allow the model to yield sufficient predictions (Leijnen, 2012).

The types of constraint that can emerge from training are limited by a network's particular topology. Typically, the numbers of units and layers need to be determined pre-training. A sub-optimal topology may lead to intractable computation times, e.g. when the network consists of more layers or units than strictly necessary. Technological advances in GPU processing and data storage have mitigated computational intractability but may not provide a lasting solution when the demand for computational resources continues to increase.

The network architecture presented here, called Self-Constraining Deep Neural Networks (SCDNN), is a model where the set of possible constraints that can emerge from training are not restricted by the network's topology. Instead, constraints are generated, eliminated, preserved, and selected dynamically through self-organization (Leijnen, Heskes, & Deacon, 2016) of activation signals.

Self-organization may occur when recurrent connections cause activation signals to become amplified over time. When only a constrained subset of the possible activation distributions becomes sampled due to the biases introduced by these amplifications, this constraint manifests itself as a pattern – self-organization, in this sense, can also be understood as a kind of self-simplification. The types of patterns that may emerge are bound by the size of the network rather than by its topology. External conditions may be favorable for the preservation of some constraints; other constraints are eliminated spontaneously as the necessary conditions for signal amplification fail to persist. Together, these processes create a learning process where a particular (set of) constraint(s) is selected for.

The self-constraining dynamics of such a neural network may provide a biologically plausible explanation for learning, as selection is neither centralized nor superimposed. Moreover, it allows for a less restrictive search space compared to deep networks initialized with a fixed topology.

References

Bengio, Y., LeCun, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Garner, W. (1962). *Uncertainty and structure as psychological concepts*. Wiley.

Leijnen, S. (2012). Emerging symbols. In T. Schilhab, F. Stjernfeld, & T. Deacon (Eds.), *The symbolic species evolved.* Dordrecht: Springer.

Leijnen, S., Heskes, T., & Deacon, T. (2016). Exploring constraint: Simulating self-organization and autogenesis in the autogenic automaton. In *Proceedings of the 15th international conference on artificial life.*