Towards a Continuous Hyperparameter Representation for Neural Networks

William Guss* Other Author† Other Author‡

June 26, 2017

Abstract

*Email: wguss@berkeley.edu
†Email: other@berkeley.edu
‡Email: other@berkeley.edu

Contents

1 Planning & Unorganized Results (*)

1.1 Motivation & Goal

1.2 Questions/Hypothesis

1.3 Theory

1.3.1 Desired Results

1.3.2 Some Exposition

1.4 Experiments

The following are a set of desired experiments to verify the newly proposed hyperparameter representation.

1.5 Reading List

1.6 Related Notes

- Continuous Hidden Dimension
- **Some Thoughts on Local Search on Hidden Units.** Let $\mathcal{N}$ be the $\mathfrak{n}$-discrete instatntiation of the following DFM

$$\mathcal{O} : \boxed{\mathbb{R}^n} \xrightarrow{\mathfrak{d}} \boxed{L^1(E(\gamma))} \xrightarrow{\mathfrak{f}} \boxed{\mathbb{R}}$$

  where $E : \mathbb{R} \to \mathcal{L}(\mathbb{R})$ is a function which parameterizes the domain over which the $\mathfrak{f}$-functional integrates.

  It was concluded in the last note that if $E(\gamma) = [0, \gamma] \in \mathcal{L}(\mathbb{R})$ then we have the following problem for the piecewise constant parameterization of weights on $\mathfrak{f}, \mathfrak{d}$. Let $F : \mathbb{R} \to \mathbb{R}$ be some loss function, and then computation of the local gradient ascent path gives

$$\begin{aligned}
\frac{\partial F}{\partial \gamma} &= \frac{dF}{dy^2}\frac{\partial y^2}{\partial \gamma} \\
&= \frac{dF}{dy^2} \cdot \left[\frac{\partial}{\partial \gamma} \int_{[0,\gamma]} \sum_{k=1}^{\infty} [\sigma \circ \mathfrak{d}(x)](u) \chi_{k \cdot [0,1]}(u) W_k^1 \, d\mu(u) \right]_{\mathfrak{n}} \\
&= \frac{dF}{dy^2} \cdot \left[\sum_{k=1}^{\infty} [\sigma \circ \mathfrak{d}(x)](\gamma) \chi_{k \cdot [0,1]}(\gamma) W_k^1 \right]_{\mathfrak{n}} \\
&= \frac{dF}{dy^2} \cdot y^1_{\lfloor \gamma \rfloor} W^1_{\lfloor \gamma \rfloor}.
\end{aligned}$$

In otherwords, gradient ascent on F with respect to γ will increase γ if the error will decrease when the contribution of the last output neuron is increased (in magnitude); that is, if $\gamma' > \gamma$ then $(\gamma - \lfloor\gamma\rfloor)$ increases, and thus E decreases by virtue of the term

$$\int_{\lfloor\gamma\rfloor\cdot[0,1]} y^1(u)W^1_{\lfloor\gamma\rfloor}\, d\mu(u) = (\gamma - \lfloor\gamma\rfloor)y^1_{\lfloor\gamma\rfloor}W^1_{\lfloor\gamma\rfloor}$$

increasing. Searching over γ is effecitvely the same as spending extra time changing the weight $W^1_{\lfloor\gamma\rfloor}$ using two linearly dependent parameters, $(\gamma - \lfloor\gamma\rfloor)$ and $W^1_{\lfloor\gamma\rfloor}$, itself[1].

Thus we are led to the question: *Is hyperparameter search a matter of immediate model accuracy or expected capacity for change, and in that distinction, does optimizing hyperparameters with respect to model accuracy coorespond to optimization on model capactiy and visa versa?* Let us examine this question in two contexts.

Above, we noted that a local search on γ decreased error in exactly the same fashion as standard gradient descent, but a step in γ of more than integral amount can increase error. To see this let $k = \lfloor\gamma\rfloor$. When $\Delta\gamma > 1$ then the $(k+1)$th neuron is then "enabled" so-to-speak. However, this $(k+1)$th neuron may perform a computation that increases error and so in the next step of gradient descent $\Delta\gamma$ would be negative, retreating away from the added model capacity of a randomly intiialized $(k+1)$th neuron. That is not to say that γ might not increase again, repeating the process, or in the limit of such oscilations the update $W^1_{k+1} - \alpha\partial E/\partial W^1_{k+1} \to W^1_{k+1}$, will eventually contribute to model accuracy, but relying on these dynamics as a result with no guarentees of convergence is questionable. Despite the fact that $\mathcal{N}$ may need additional model capacity[2], local search on capacity with respect to accuracy may not yield the required capacity to increase accuracy in the limit.

TODO: Include brief analysis of richard's paper.

1.7 Timeline

T

[1] An additonal conclusion is, at least by analogy, that local search on $E(\gamma)$ at any one place assumes that adjacent neurons have similar values

[2] There are functions which are unlearnable without a sufficient number of neurons for example.