Searching for Activation Functions

https://arxiv.org/pdf/1710.05941.pdf

Paper Summary By:

Ankit Dhankhar

Thursday 3rd January, 2019

INTRODUCTION

- Author of the paper used automated search techniques to discover novel activation functions.
- The best found activation function, which they call Swish, is

$$f(x) = x.sigmoid(\beta x)$$

. where β is a constant or trainable parameter.

SUMMARY

- Experiments shows that Swish consistently matches or outperforms ReLU on deep networks applied to variety of challenging tasks.
- Choice of search algorithm depends on the size of search space; it is possible to enumerate entire search space but when search space is large it call lead to extremely large search space (i.e. in order of 10^12), making exhaustive search infeasible.
- For large search space they used an RNN controller. At each time-step controller predicts a single component of activation function, which is fed back to controller in next time-step, and this process is continued until every component of activation function is not predicted. And constructed activation function fro string is used to train a 'child network' with candidate activation function. Child networks validation accuracy is used to update search algorithm.
- In exhaustive search, a list of top performing activation function order by validation accuracy is maintained where as in case of RNN controller is trained with reinforcement learning to maximize validation accuracy(which act as reward).

- Complicated activation functions under-performed simpler activation function potentially due to an increases difficulty in optimization.
- A common structure shared by top activation function is the use id raw preactivation function in final binary function.
- Functions that use division tend to perform poorly because the output explodes when denominator is near 0.
- They focused evaluating the activation function $f(x) = x.\sigma(\beta x)$, which they call Swish, where $\sigma(z) = 1 + exp(-z))^{-}1$ is sigmoid function.
- Unlike ReLU, Swish is smooth and non-monotonic.
- The derivative of Swish is:

$$\begin{aligned} f'(x) &= \sigma(\beta x) + \beta x.\sigma(\beta x)(1 - \sigma(\beta x)) \\ &= \sigma(\beta x) + \beta x.\sigma(\beta x) - \beta x.\sigma(\beta x)^{x} \\ &= \beta x.\sigma(x) + \sigma(\beta x)(1 - \beta x.\sigma(\beta x)) \\ &= \beta f(x) + \sigma(\beta x)(1 - \beta f(x)) \end{aligned}$$

- If $\beta = 0$, Swish becomes the scaled linear function $f(x) = x/2$, and as $\beta \to \infty$, the sigmoid component approached 0-1 function, so it becomes ReLU function.