Home Assignment 2: Deep Learning and Embedding

Essay

Almost every part of our world is changing in cycles, in other words – periodically. Therefore, learning such cycles and periodic dependencies is key to understanding our world. Furthermore, correctly extrapolating periodic data would help to predict future events and get ready for them. For example, we could predict global economic growth or decline, and the amount of infected people during epidemic.

Main task consists of creating such activation function for feedforward neural network that would not only decently learn periodic functions but would work with similar quality on no-periodic tasks. Before the inspected article the research field didn't really succeed in learning periodic functions maintaining the quality on the other task. Using different than usual activation functions, such as sin(x), cos(x), or their linear combinations, is not beneficial in comparison to ReLU-based activation functions. This happens because sin(x) and cos(x) produce a lot of local minima, thus making it hard to optimize.

The first step in this research paper is to prove that standard activation functions, such as ReLU and tanh(x) are incapable of correct periodic functions extrapolation. Using asymptotic properties of activation functions and induction, Ziyin *et al.* were able to do so (in reality they didn't even need induction for tanh(x), since it converges to a constant and not to some linear combination like ReLU). Similar to theorems for ReLU and tanh(x), theorems for activation functions based on two above can be proved, since such activation functions share the same asymptotic properties with ReLU and tanh(x).

The second step is proposing the valid activation function, $x + (\sin x)^2$, which is named Snake. The Snake really looks like $x + \frac{(\sin ax)^2}{a}$ or $x - \frac{\cos(2ax)}{2a} + \frac{1}{2a}$, factor of a is present in function to control the frequency of periodicity of the function. In this article, sin(x), cos(x), x + sin(x), and x + cos(x) are discussed as the competitors for $x + (\sin x)^2$, but it was found that not only periodicity, but also monotonicity matters for decent activation function for learning periodic functions. Therefore, sin(x) and cos(x), which lack monotonicity, are not optimal for this task. As for the x + sin(x) and x + cos(x), these functions have $-\frac{x^3}{6}$ as their first non-linear term, while $x + (\sin x)^2$ has x^2. In Ziyin *et al.* opinion, this helps Snake to approximate periodic functions better.

As per results of Snake in different tasks, it shows the fastest loss decrease among sin(x), ReLU, and x + sin(x) on MNIST dataset, and, at the same time, Snake makes possible regressing a sin(x) function in between the training data and outside of it. Blue dots on the picture below are the points of training data, blue line – ground truth, red line – neural network approximation using Snake.

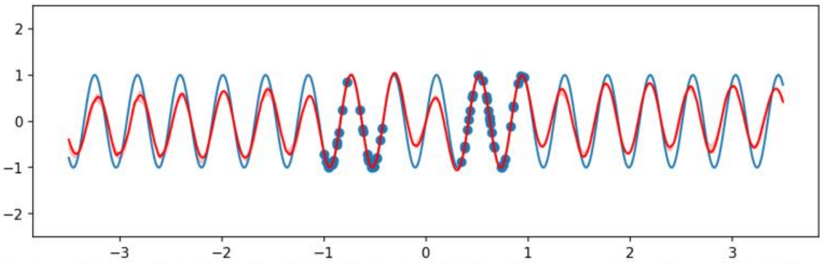


Figure 4: Regressing a simple sin function with Snake as activation functions for $a = 10$.

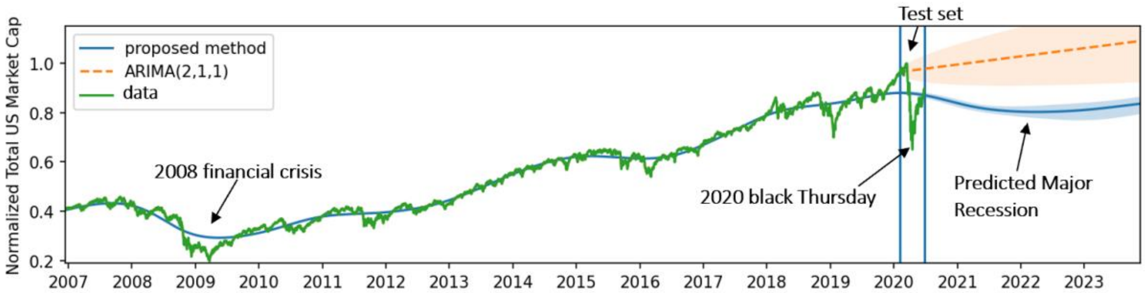
The generalization of Snake's ability to interpolate and extrapolate is proved by "Universal Extrapolation Theorem" and its first corollary in Appendix C. But even though Snake works in theory and on simple tasks, testing Snake on large datasets from industry and verifying its usefulness on real-word experiments is required.

This brings us to the third and final step – showcase of Snake's application on for different tasks. But before we dive into the experiments, we have one thing to discuss – initialization for Snake. Basically, the sufficient initialization is sampling form uniform distribution from range $\left(-\sqrt{\frac{3}{d}}, \sqrt{\frac{3}{d}}\right)$, where d is the dimensionality of the layer. For higher order correction Ziyin *et al.* came up with the following proposition:

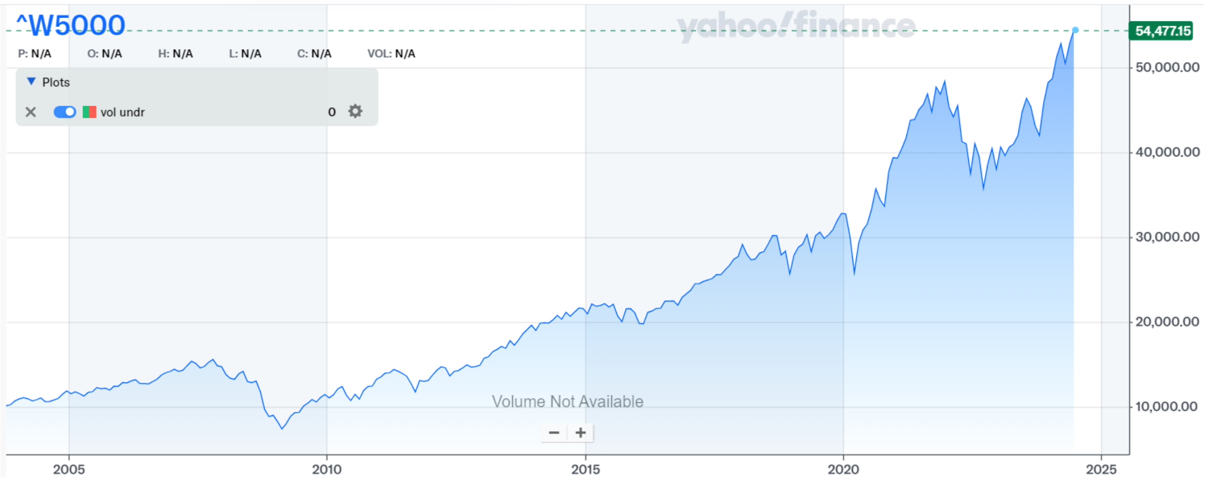
Proposition 1. *The variance of expected value of* $x + \frac{\sin^2(ax)}{a}$ *under a standard normal distribution is* $\sigma_a^2 = 1 + \frac{1+e^{-8a^2}-2e^{-4a^2}}{8a^2}$, *which is maximized at* $a_{max} \approx 0.56045$.

Such correction is more valuable in deeper neural networks and provides an increase in training speed and accuracy. Ziyin *et al.* suggests $0.2 < a < a_{max}$ for standard tasks like image classification and a in range from 5 to 50 for tasks with expected periodicity.

I don't want to spend a lot of time and words on experiments, since they basically proved all expectations Ziyin *et al.* described before section 6 called Applications. For Minamitorishima mean weekly temperature regression, only Snake was able to approximate training data, while ReLU and tanh(x) failed completely. On different scales (ResNet-18 and ResNet-101) of image classification on CIFAR-10 Snake showed similar performance to the best of tested activation functions. For human body temperature modeling Snake was the only function to extrapolate data similarly to training set, in other words extrapolated data laid in the same range as training one. Furthermore, for this task Snake not only captured the long-term periodicity correctly, but the short-term one too, in the form of daily oscillations similar to normal human ones. And all this was achieved with only 25 instances of data! The last experiment was about finances and economy. The total US market capitalization measured by the Wilshire 5000 Total Market Full Cap Index was chosen as a value to predict. The models were trained on the data from 1995 to 31st January of 2020, and tested on data from 2nd February of 2020 to 31st May of 2020. Below I leave a figure with a plot from this experiment.



The standard and traditional ARIMA method for prediction in economics and stock price predicted steady growth after training data, and Snake predicted major recession. Below I leave not normalized, but plot of Wilshire 5000 Total Market Full Cap Index till now.



It seems like none of the models could really predict the future of economics in our constantly changing world.

Before the conclusion, I want to mention one more thing. In comparison to RNN in time series regressing, Snake is incredibly faster and gets relatively faster the longer the time-series are. Moreover, in contrast to RNN, Snake don't have an increase in generalization loss while increasing the white noise variance.

To conclude the essay, this study examines extrapolation capabilities of neural networks beyond the training data, shows that feedforward neural networks with standard activation functions fail to learn periodic functions, and proposes elegant and efficient solution for stated problem that works on toy and real data.

Practice

Look at Jupiter Notebook for that part.