Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Václav Volhejn
Thesis title	Smoothness of Functions Learned by Neural Networks
Year submitted	2020
Study program	Computer Science
Study branch	General Computer Science

Review author	RNDr. Milan Straka, Ph.D.	Reviewer
Department	Ústav formální a aplikované lingvistiky	

Overall

	good	OK	poor	insufficient
Assignment difficulty	X			
Assignment fulfilled	X			
Total size *... text and code, overall workload*	X			

In past years, deep neural networks have achieved great success in many areas. Surprisingly, the reason why neural networks trained by SGD do generalize so well is not yet fully understood. Recently, researchers came with a proposition that SGD may be biased towards solutions with low complexity, which could act as a built-in regularization.

The thesis tries to quantify this phenomenon. Notably, several measures of smoothness are defined, and two experiments, determining whether these measures are being implicitly minimized, are performed.

Overall, I consider this thesis to highly exceed the standard level of a bachelor thesis. It performs up-to-date world-level research on interesting and not yet explained phenomenon, and as far as I can tell, the experiments are valid.

I include two comments regarding the performed experiments. These are not meant as criticism, but instead as areas of possible improvement if the work should be published.

- I was missing other baselines than just polynomial interpolation in Table 4.1 – spline interpolation of different degrees would be a natural extension.
  Out of curiosity, I evaluated spline interpolation of degree 1 (piecewise linear function), 2 and 3 using `scipy.interpolate.splrep`, arriving at

	NNs	Polynomials	Spline 1	Spline 2	Spline 3
gradient_norm	0.74 ± 0.13	0.36 ± 0.20	0.75 ± 0.19	0.73 ± 0.24	0.48 ± 0.28
path_length_f	0.65 ± 0.16	0.44 ± 0.22	0.63 ± 0.20	0.75 ± 0.20	0.63 ± 0.26
path_length_d	0.75 ± 0.11	0.51 ± 0.14	0.94 ± 0.05	0.66 ± 0.21	0.76 ± 0.21
weights_product	0.80 ± 0.13	0.38 ± 0.15	0.97 ± 0.03	0.53 ± 0.28	0.34 ± 0.23

- In the explicit regularization experiments, I was wondering whether there is a correlation between L2 and L1 norms and the smoothness measures. Therefore, I would appreciate if L2/L1 norm could be used both as an evaluation measure (while regularizing using some smoothness) and as a regularization (while evaluating smoothness measures).

Thesis Text

	good	OK	poor	insufficient
Form *... language, typography, references*	X			
Structure *... context, goals, analysis, design, evaluation, level of detail*	X			
Problem analysis	X			
Developer documentation		X		
User Documentation		X		

The thesis is written in English, it reads well and I found only a minimal number of grammatical errors.

Some comments regarding the text follows:

- In Definition 1, if gradient descent should find a local minimum, the learning rates must converge to zero (and not be fixed).
- I believe the derivative of ReLU is often defined as 0 at 0, because it allows to compute the derivative of ReLU using only its output value as $[\mathit{output\ value} > 0]$.
- The $\boldsymbol{w}_i^{(1)}$ in Equation (3.8) should be $\boldsymbol{W}_i^{(1)}$.
- If the initialization presented in Equation (4.1) should be equal to Glorot initializer, either α must be 2 or the ℓ should be $\sqrt{\frac{6\alpha}{n_{in}+n_{out}}}$.
- In Section 4.3.2, “Additionally, we balance the classes by truncating the dataset to 10000 examples of each digit” seems incorrect, because there are only 6 000 examples of each digit in the training portion of MNIST (and 1 000 in the test set). Perhaps the author meant 10 000 elements altogether, i.e., 5 000 examples of each digit? (Also, it would be better to typeset the number with a thousands separator.)
- In Section 4.3.2, there is a missing *by* in “... what is observed Maennel et al. [2018].”

Thesis Code

	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*		X		
Implementation *... naming conventions, formatting, comments, testing*	X			
Stability	X			

The source code for all the experiments and analyses is attached to the thesis. I was able to run it and retrain some of the thesis experiments, so the most important goal of a research source code – replicability – is achieved.

However, there are minor areas where the code could be improved:

- The list of required packages is incomplete and does not contain any information about required versions.
- The polynomial baseline of increasing training set size in one dimension (presented in Table 1) cannot be evaluated (the `smooth.analysis.get_interpolation_measures` is even marked as *unused*); but I was able to execute it after a simple modification, reaching very similar results.
- I would find it more straightforward if the training scripts produced the results in the `CSV` format used by the reporting scripts (or if the reporting script used the produced `feather` files).

Overall grade Excellent
Award level thesis No

Date Signature