Review of Bachelor Thesis: "Application of artificial neural networks for malware detection in HTTPS traffic", by Jan Bodnár, 2017

Reviewer: RNDr. Petr Somol, Ph.D.

I read the thesis with great interest as the topic is well chosen to address a problem of crucial importance in computer network security. Neural networks are notoriously popular in many fields these days, yet the problem of HTTPS classification remains largely unresolved by any means, including NNs.

The thesis is built around experiments on preprocessed industrial data, hence the thesis could focus entirely on the modeling problem without distraction. The thesis has two principal parts: 1) design and evaluate NN as classifier, 2) utilize NN formalism for similarity search. The two parts complement each other well.

The first part is a delight to read, I have rarely seen such clear, straightforward and compact explanation of NN formalism. The only thing that hinders this part is inferior formatting of formulas. The content is technically sound, experiments provided in sufficient extent and scientific discussion led correctly. I particularly appreciate attention to technical detail and completeness of discussion, that covers all facts precisely regardless how favourable they may seem w.r.t. to the performance of the proposed solution. The results thus have practical value and can be fully trusted.

The second part is in fact more scientifically interesting, while at the same time suffers from much less thorough treatment. The presented material and proposed solution based on Siamese NNs is extremely interesting. The reasoning about choices made (for defining this particular modification of Siamese NNs) is again well thought through and correct. In this sense the quality of presented ideas is beyond Bachelor Thesis level and can easily get extended to Master level. The problem is, however, the incompleteness of descriptions, missing details, and generally much lower quality of presentation in this part when compared to the first part. The key problems as I see them are: 1) I particularly miss a diagram and possibly textual description of how the proposed Siamese network should be designed in detail and how it should work step by step, all this is currently introduced so briefly that without study of external sources it is difficult to understand the concept; even then some details remain unclear, 2) I agree that it is difficult to evaluate the quality of learned classification-clustering models, however, a discussion could have been be provided how to approach the problem. Here it is only assumed that 2D plots of data are self-explanatory, 3) even in experiments that seem compatible between two thesis parts the author uses confusingly different measures. If I understand it right, e.g., "precision" has a different meaning in first and second thesis parts. The second part does look like having been written in rush. Having pointed out the key problems, I should nevertheless emphasise that the second part provides relevant and novel scientific results that deserve appreciation.

List of other issues:

- Equations would be better numbered and referred to by number, see e.g. the default LaTeX style
- Occasional typos in formulas, e.g., page 7, G(m|... should probably be G(m_i|... etc. By the way, how to find the value of d mentioned six lines before? Next page first formula – is not there brackets missing? Bottom of page 20 – C is a set, so in place of denominator it should be as |C| ?
- When referring tables or figures from text, don't just drop a number in text flow as in "...first two columns of 4.2 the networks...". I would also recommend to call tables Table and not Figure as is the case now.
- In some tables you list complete details about NN configuration, in some not. E.g., in 4.2 there is no mention of ReLU although in 4.3 there is; am I right it belongs to both?
- Having read section 4.2.1 I would be interested to see results of 2 or 3-hidden-layer networks with Hinge Loss. But this is not a flaw, just my interest.
- Page 28, more details about random forest and knn setup of concurrent experiments would be helpful for the reader
- Page 32 why do you describe the principle of your Siamese NN using programmatic notation only? 1) It differs from previous text, and 2) without more details makes understanding of the proposed model extremely hard. Also, even in this crucial part there seems to be a typo: autoencoder_clean should be defined as (decode(encode(c))-c)^2, shouldn't it? The same for autoencoder_malware
- Page 33 before referring table 5.12 and 5.13 I would refer 5.10 too

For future scientific work I recommend the author to improve in two things: formal presentation quality esp. of mathematical matter, and global consistency and completeness of topic coverage – keep the same notation across whole work, provide enough detail about crucial topics, and try to address all hard relevant questions that can come to your (or anyone's) mind. The thesis shows that you have a great scientific mind; the impact of your future work will now partly depend on you improving in areas I have mentioned.

To summarize, in view of the high quality of scientific ideas and material versus the non-negligible flaws in presentation, I recommend this thesis to be accepted as Bachelor Thesis, and recommend to grade it as "Very Good".

RNDr. Petr Somol, Ph.D.

Research Fellow
ÚTIA AV ČR, Dept. Of Pattern Recognition
Pod vodárenskou věží 4, Praha 8
and
Head of Research
Cognitive group at Cisco Systems
Karlovo náměstí 10, Praha 2

29 Aug 2017