EVALUATION OF THE DIPLOMA THESIS "PRIOR MODELS FOR ROBUST ADVERSARIAL DEEP LEARNING" BY JOSE ANANIAS HILARIO REYES

The thesis considers the question whether introducing prior knowledge and models into deep networks can robustify them w.r.t. adversarial attacks. In particular, it proposes to introduce lateral interactions into convolutional layers of deep networks for image classification. This requires to consider stochastic convolutional networks in which each layer is modeled as a conditional random field (CRF). Since the forward propagation for such models is intractable, it becomes necessary to use some approximation algorithm for computing the marginal probabilities of the nodes and propagating them through the network. The author considers a particular approach proposed by Djolonga and Krause (2014), which can be applied for supermodular CRFs and has the attractive property of being differentiable w.r.t. to its inputs.

Chapters 1-3 of the thesis introduce the problem of susceptibility of deep networks to adversarial attacks, give an overview on robust learning and an extensive review of related work.

Chapters 4-5 give all the necessary technical concepts needed for the proposed approach. This includes stochastic neural networks, conditional random fields and the approximation approach used for computing their marginal probabilities. It is shown how to combine these concepts and methods for convolutional networks with CRF layers. The last technical chapter summarises the approach and shows its end-to-end differentiability.

Chapters 7-8 present the experiments and conclude them. The main analysed question is whether CRF-CNNs are more robust w.r.t. to adversarial attacks than their standard counterparts. For this both variants of learning, standard and adversarial, where used for each of the two model variants. The experiments where carried on MNIST and MNIST fashion datasets.

Thesis strengths. The thesis is well structured and well written. It introduces and covers all necessary concepts, approaches and algorithms with sufficient detail. The review of the state of the art is exemplary. The experiments and their outcomes are reproducible.

Thesis weaknesses. The thesis is in my view too long. The derivation of certain well known facts could be omitted. Examples are the basic properties of the KL-divergence and the Hammersley-Clifford theorem establishing the relation between Markov random fields and Gibbs random fields. This also holds for the redundant repetition of the network architecture description in the experimental part. On the other hand, it would be desirable if some implementation details of the approximation approach and the Frank-Wolfe algorithm where given in the thesis.

The technical part as a few small imprecisions, as e.g. the formulation of the Hammersley-Clifford theorem, which should be given after introducing MRFs and GRFs.

Overall evaluation. In my opinion Jose Hilario approached the task in a highly motivated and diligent way. His breadth and eagerness for knowledge are exemplary. On the other hand, the theoretical penetration and in depth understanding of this admittedly challenging

Date: June 19, 2019.

and complicated approach was not easy for him and required his constant effort. I recommend to accept the thesis and evaluate it with the grade B.

BORIS FLACH, CZECH TECHNICAL UNIVERSITY IN PRAGUE, DEPT. OF CYBERNETICS
Email address: flachbor@cmp.felk.cvut.cz