TECHNISCHE UNIVERSITÄT DRESDEN

Computergraphik und Visualisierung

Fakultät Informatik – Institut SMT – Professur Computergraphik und Visualisierung

Review for the Master Thesis by José Hilario:

PRIOR MODELS FOR ROBUST ADVERSARIAL DEEP LEARNING

The aim of the work is to enhance neural networks by introducing a special so-called CRF-layers, i.e. lateral interactions between neurons within the layer. Special attention is paid to robustness against adversarial examples.

The work combines several modern approaches from different fields, like general optimization, graphical models, neural networks, variational inference etc. They are carefully investigated and combined with each other in a theoretically sound manner. Results are illustrative enough, they are properly documented and interpreted. There are many observations and ideas that can be useful for further work.

Unfortunately, the work is written in a somewhat inaccurate manner. First of all, it would be very useful to clearly state from the very beginning not only the goal (i.e. cope with adversarial examples) but also to give a way to approach this goal, as well as to enumerate subtasks to be solved, like approximating marginals, computing averages etc. Otherwise, the overall idea of the developed model becomes more or less clear first in chapter 6.

Literature overview is almost completely devoted to works that consider adversarial problems. There is only two works about the combination CNN+CRF (Dense CRF by Arnab et al.) and no works at all about Bayesian networks, computing marginals, variational inference etc. This makes the literature overview somewhat unbalanced and not really corresponding to the rest of the work, since the main contribution of the master thesis is the development of a novel model, but not a novel learning scheme.

A separate chapter for definitions (ch. 4) seems to be not reasonable, as there are also many further definitions introduced later.

There are some inaccuracies in formulae, like in “Definition 5.1” (integral over variables that are $\{0,1\}$ in context of this work), eq. (5.4) seems to be incorrect, eq. (5.6) is very hard to understand, there is probably a typo in the second equation on page 32, etc.

Adversarial training and robust optimization are not sufficiently explained. They are briefly mentioned in sec. 5.3 on a very general level and later particular experimental setups are given in sec. 7.6. Moreover, in sec. 3 it is stated: “We consider the development of energy-based regularization through the construction of uncertainty sets for adversarial training, using energy-based models“. Unfortunately, this idea (i.e. construction of “meaningful” uncertainty sets) is not further discussed.

Postadresse (Briefe)
Technische Universität Dresden
01062 Dresden

Postadresse (Pakete u.ä.)
Technische Universität Dresden
Helmholtzstraße 10
01069 Dresden

Besucheradresse
Sekretariat:
Nöthnitzer Str. 46
Zi. 2105

Internet
http://www.inf.tu-dresden.de/cgv

Concerning experiments. Was the Ising strength α fixed or learned? If learned, what are the learned values? If fixed, were all experiments performed with the same value of α or each one with its own value? In both cases, how α-s were chosen? Note that setting $\alpha = 0$ turns the approach to the original base-line. Hence, the developed model can not be worse in principle, when tuning α carefully.

Below are some further comments / questions / suggestions:

The assumption that neuron activations should be spatially continuous is in question. Consider e.g. edge features or corner detectors, the activations of which obviously do not need to be spatially compact. It might be however true that the neighbouring activations are highly correlated (instead of to be simply continuous). That would however need a non-supermodular CRF, which does not fit into the proposed framework due to the L-Field optimization.

The final approach is a long chain of approximations. First, all probability distributions are substituted by the corresponding factorized ones. Second, these distributions are replaced by means. Finally, computing marginals is not tractable, hence approximations are employed again. So it might be well asked, what remains from the original idea to use CRF as a part of a CNN.

The notation “prior knowledge” is basically reduced to just a single parameter, i.e. the Ising strength α. This is somewhat too weak and contradicts to the promise (see the abstract) to “… analyze different types of prior knowledge”.

At the end, the developed model is a Feed-Forward neural network, i.e. everything is computed deterministically, both at the inference stage and during the learning by Error Back-Propagation. The starting point of the work is however a Bayesian network, for which the developed model is an approximation. Hence, it would be interesting to investigate the behaviour of the initial Bayesian network (learned approximately), when doing inference e.g. by sampling.

The developed approach seem so to be hard to generalize to other situations, like CRFs with continuous variables or multi-label models. An acknowledgment of this and may be a discussions would be highly appreciated at least in “Future work”.

Despite the above criticism, it should be noted that an essential amount of work was carried out during the thesis. Quite complex aspects are considered and used in a theoretically sound manner. To summarize, I rate the work with **B (very good)**.

Datum

Dr. Dmitrij Schlesinger