Probabilistic properties and AI safety:

AI is now embedded into a number of everyday life applications. More and more, we depend on neural networks to make even critical situations, such as control and motion planning for autonomous cars, and it is of primary importance to be able to verify their correct behavior. But for some applications, e.g. perception in autonomous systems, through classifiers, we can only hope for probabilistic safety. The context of this internship is to find qualitative measures of how likely a neural net classifier, or a neural net used in a control system will exhibit a certain behavior.

A classical property for such classifiers is some form of robustness: if an input image is slightly perturbed, the classifiers should classify this perturbed image the same way as the original one. But the perturbations on images are generally not just independent perturbations on pixels, but rather effects (blurring, luminosity, contrast changes…) that make the perturbation correlated among pixels. In the context of this internship, we are thus looking for methods to define and propagate joint probability distributions, or rather, sets of joint probability distributions. This can be done using work on imprecise probabilities (PBoxes and Demster-Shafer structures, see e.g. [AAOB+13], polynomial forms, see e.g. [SSYC+20]), combining it with the notion of copula, as in e.g. [Sch23], see also the work for probabilistically verifying model-based control systems [GFS+24] and the work for verifying non-deterministically neural-network based control systems [EGSP22].
Finally, if time permits, we will look at properties beyond robustness and in particular explainability of neural networks. This generally relies on finding some form of correlation between features (and absence thereof) within inputs, and outputs, as done in e.g. LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanation) and others, see e.g. [RSBY+22].

References:

[AAOB+13] Assalé Adjé, Olivier Bouissou, Jean Goubault-Larrecq, Eric Goubault, Sylvie Putot. “Static Analysis of Programs with Imprecise Probabilistic Inputs”. VSTTE 2013.

[EGSP22] Eric Goubault, Sylvie Putot, “RINO: Robust INner and Outer Approximated Reachability of Neural Networks Controlled Systems”. CAV 2022.

[GFS+24] Ander Gray, Marcelo Forets, Christian Schilling, Scott Ferson, and Luis Benet.
“Verified propagation of imprecise probabilities in non-linear odes”. International Journal of Approximate Reasoning 2024.

[MLAD23] Matthew Landers, Afsaneh Doryab, “Deep Reinforcement Learning Verification: a survey”. ACM Computing Reviews, 2023.

[MMPV15] Ignacio Montes, Enrique Miranda, Renato Pelessoni, and Paolo Vicig. Sklar’s theorem in an imprecise setting.Fuzzy Sets and Systems. 2015..

[OS20] Matjaz Omladic and Nik Stopar. “A full scale sklar’s theorem in the imprecise setting”. Fuzzy Sets and Systems. 2020.

[RSBY+22] Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, Lu Liu. “Explaining deep neural networks: A survey on the global interpretation methods”. Neurocomputing, 2022.

[Sch23] Bernhard Schmelzer. “Random sets, copulas and related sets of probability measures”. International Journal of Approximate Reasoning. 2023.

[SSYC+20] Sriram Sankaranarayanan, Yi Chou, Eric Goubault, Sylvie Putot. “Reasoning about uncertainties in Discrete-Time Dynamical Systems using Polynomial Forms”. Neurips 2020.