Thesis and post-doc days, DILS
June,17th, 2019

VERIFICATION AND VALIDATION OF DEEP LEARNING ALGORITHMS

Girard-Satabin Julien (10-2018 – 10-2021)
Chihani Zakaria (CEA)
Charpiat Guillaume, Schoenauer Marc (INRIA TAU)

PRESENTATION

WHY DOING A PHD ON THIS TOPIC ?

Classical programs

```
let [illegible]
    List.fold_right (fun x l -> (x.Node_proto.input)::l) ns []
et fold_nodes_outputs ns =
    List.fold_right (fun x l -> (x.Node_proto.output)::l) ns []
(*Dict containing initializer tensors dimensions and raw data*)
let dict_tensors ts =
    let t_name x = (match x.Tensor_proto.name with
                    | Some n -> n
                    | None -> "NO_NAME") in
    let t_dim x = x.Tensor_proto.dims in
    let t_raw x = (match x.Tensor_proto.raw_data with
                    | Some rd -> rd
                    | None -> "NO_RAW") in
    List.fold_left (fun m x -> MapStr.add (t_name x) (t_dim x,t_raw x) m)
                    MapStr.empty ts

let parse_model_from_file fp =
    let ch = open_in_bin fp in
    let buf = Piqirun.init_from_channel ch in
    parse_model_proto buf

let node_graph (g:Graph_proto.t) =
    let nodes = g.node
    and inputs = g.input
    and outputs = g.output
    and initi = g.initializer_
    in
    let i_nodes = fold_value_info_names inputs
    and o_nodes = fold_value_info_names outputs
    and c_nodes = single_value_list "C_NODE" (List.length nodes) []
    in
    let c_ops = fold_nodes_ops nodes
    and i_ops = no_op_list (List.length i_nodes)
    and o_ops = no_op_list (List.length o_nodes)
```

Explicit control flow

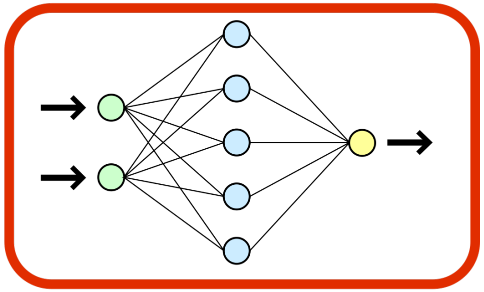
Explicit specifications

Abstractions and well known concepts

Needs to be robust

Already here

Deep learning



Generated control flow

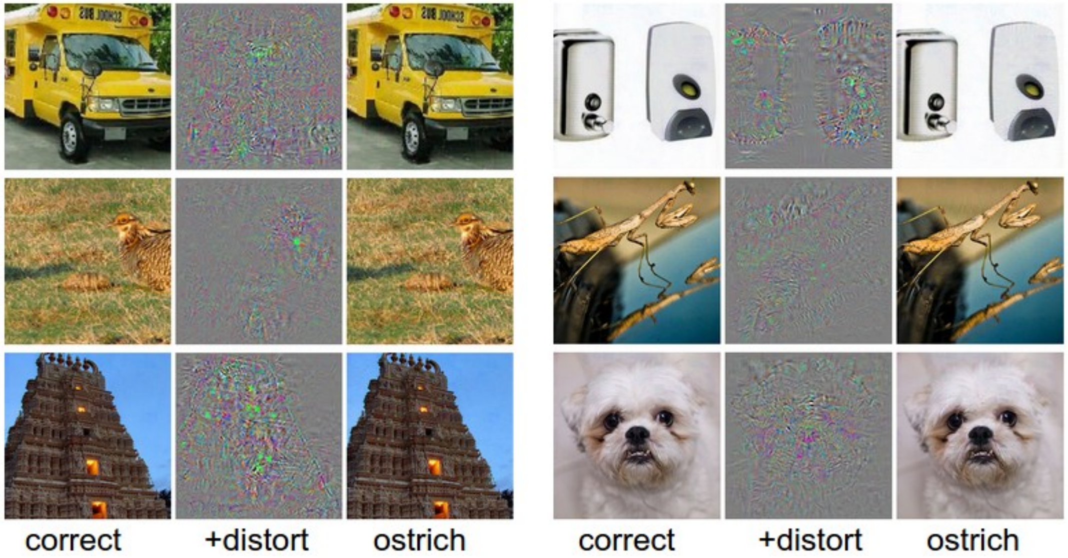
Implicit specifications

Very few abstractions and reusability

Needs to be robust

My thesis' goal

Adversarial examples



- Transferable between programs (Papernot et al.)
- No systematic way to spot them
- Easily synthetized
- Work in the physical world on image, video, sound

STATE OF THE ART (2/2)

Vulnerabilities

- Widely focused on adversarial examples (Carlini et al., Madry et al., Goodfellow et al.,)
- Research on information leakage (Papernot et al., Tramèr et al., Abadi et al.,)
- Very few formal properties

Formally prove robustness

- Exact methods (Katz et al. for SMT-based solving, Tjeng et al. for MILP-based solutions)
- Overapproximation methods (Vechev’s team with abstract interpretation, Wong et al., Tsui-wei et al., for piecewise linear overapproximations)

STATE OF THE ART (2/2)

Vulnerabilities

- Widely focused on adversarial examples (Carlini et al., Madry et al., Goodfellow et al.,)
- Research on information leakage (Papernot et al., Tramèr et al., Abadi et al.,)
- Very few formal properties

Formally prove robustness

- Exact methods (Katz et al. for SMT-based solving, Tjeng et al. for MILP-based solutions)
- Overapproximation methods (Vechev's team with abstract interpretation, Wong et al., Tsui-wei et al., for piecewise linear overapproximations)

More can be done

- Robustness study of a new deep neural network basic block
- Bibliography
- Scientific comitteee (ForMaL spring school)
- Reproduction of state of the art results on adversarial robustness

SCIENTIFIC APPROACH

- ONNX2SMT : from a neural network to a SMT formula
- Constraint-programming approaches
- Formalize privacy properties, especially differential privacy on dataset/algorithm couples

CONCLUSIONS & PERSPECTIVES

- Extend and enhance existing tools using CP

- Formalize privacy properties to check

DISSEMINATION ACTIVITIES

- Course on formal methods, INRIA TAU
- Seminar on exact verification of neural networks, CEA (soon)
- Scientific comittee, tutorial and presentation, ForMaL (spring school funded by DigiCOSME) : https://formal-paris-saclay.fr/
- *A security study of ODE Nets, Girard, Charpiat, Chihani, Schoenauer*

Commissariat à l’énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

Établissement public à caractère industriel et commercial | RCS Paris B 775 685 019