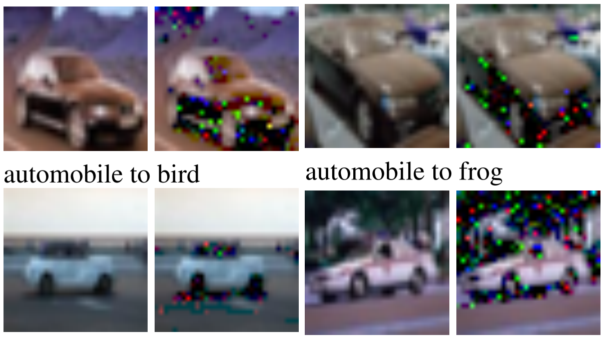
Safety Verification of Deep Neural Networks [*]

Xiaowei Huang
University of Liverpool, UK

Deep neural networks have achieved impressive experimental results in image classification, matching the cognitive ability of humans in complex tasks with thousands of classes. Many applications are envisaged, including their use as perception modules and end-to-end controllers for self-driving cars. Let $\mathbb{R}^n$ be a vector space of images (points) that we wish to classify and assume that $f : \mathbb{R}^n \rightarrow C$, where C is a (finite) set of class labels, models the human perception capability, then a neural network classifier is a function $\hat{f}(x)$ which approximates $f(x)$ from M training examples $\{(x^i, c^i)\}_{i=1,..,M}$. For example, a perception module of a self-driving car may input an image from a camera and must correctly classify the type of object in its view, irrespective of aspects such as the angle of its vision and image imperfections. Therefore, though they clearly include imperfections, all four pairs of images in Figure 1 should arguably be classified as automobiles, since they appear so to a human eye.

Classifiers employed in vision tasks are typically multilayer networks, which propagate the input image through a series of linear and non-linear operators. They are high-dimensional, often with millions of dimensions, nonlinear and potentially discontinuous: even a small network, such as that trained to classify hand-written images of digits 0-9, has over 60,000 real-valued parameters and 21,632 neurons (dimensions) in its first layer. At the same time, the networks are trained on a finite data set and expected to generalise to previously unseen images. To increase the probability of correctly classifying such an image, regularisation techniques such as dropout are typically used, which improves the smoothness of the classifiers, in the sense that images that are close (within ϵ distance) to a training point are assigned the same class label.


automobile to bird

automobile to frog

automobile to airplane

automobile to horse

Figure 1: Automobile images (classified correctly) and their perturbed images (classified wrongly)

Unfortunately, it has been observed that deep neural networks, including highly trained and smooth networks optimised for vision tasks, are unstable with respect to so called *adversarial perturbations*. Such adversarial perturbations are (minimal) changes to the input image, often imperceptible to the human eye, that cause the network to misclassify the image. Examples include not only artificially generated random perturbations, but also (more worryingly) modifications of camera images that correspond to resizing, cropping or change in lighting conditions. They can be devised without access to the training set and are transferable, in the sense that an example misclassified by one network is also misclassified by a network with a different architecture, even if it is trained on different data. Figure 1 gives adversarial perturbations of automobile images that are misclassified as a bird, frog, airplane or horse by a highly trained state-of-the-art network. This obviously raises potential safety concerns for applications such as autonomous driving and calls for automated verification techniques that can verify the correctness of their decisions.

[*] This talk is based on [1], which is published in CAV2017.

Safety of AI systems is receiving increasing attention in view of their potential to cause harm in safety-critical situations such as autonomous driving. Typically, decision making in such systems is either solely based on machine learning, through end-to-end controllers, or involves some combination of logic-based reasoning and machine learning components, where an image classifier produces a classification, say speed limit or a stop sign, that serves as input to a controller. A recent trend towards "explainable AI" has led to approaches that learn not only how to assign the classification labels, but also additional explanations of the model, which can take the form of a justification explanation (why this decision has been reached, for example identifying the features that supported the decision) In all these cases, the safety of a decision can be reduced to ensuring the correct behaviour of a machine learning component. However, safety assurance and verification methodologies for machine learning are little studied.

The main difficulty with image classification tasks, which play a critical role in perception modules of autonomous driving controllers, is that they do not have a formal specification in the usual sense: ideally, the performance of a classifier should match the perception ability and class labels assigned by a human. Traditionally, the correctness of a neural network classifier is expressed in terms of *risk*, defined as the probability of misclassification of a given image, weighted with respect to the input distribution μ of images. Similar (statistical) robustness properties of deep neural network classifiers, which compute the average minimum distance to a misclassification and are independent of the data point, have been studied and can be estimated using tools such as DeepFool and cleverhans. However, we are interested in the safety of an *individual decision*, and to this end focus on the key property of the classifier being *invariant* to perturbations *at a given point*. This notion is also known as pointwise robustness or local adversarial robustness.

Contributions. In this paper we propose a general framework for automated verification of safety of classification decisions made by feed-forward deep neural networks. Although we work concretely with image classifiers, the techniques can be generalised to other settings. For a given image x (a point in a vector space), we assume that there is a (possibly infinite) region η around that point that incontrovertibly supports the decision, in the sense that all points in this region must have the same class. This region is specified by the user and can be given as a small diameter, or the set of all points whose salient features are of the same type. We next assume that there is a family of operations Δ, which we call manipulations, that specify modifications to the image under which the classification decision should remain invariant in the region η. Such manipulations can represent, for example, camera imprecisions, change of camera angle, or replacement of a feature. We define a network decision to be *safe* for input x and region η with respect to the set of manipulations Δ if applying the manipulations on x will not result in a class change for η. We employ discretisation to enable a *finite exhaustive* search of the high-dimensional region η for adversarial misclassifications. The discretisation approach is justified in the case of image classifiers since they are typically represented as vectors of discrete pixels (vectors of 8 bit RGB colours). To achieve scalability, we propagate the analysis *layer by layer*, mapping the region and manipulations to the deeper layers. We show that this propagation is sound, and is complete under the additional assumption of minimality of manipulations, which holds in discretised settings. In contrast to existing approaches, our framework can guarantee that a misclassification is found if it exists. Since we reduce verification to a search for adversarial examples, we can achieve safety *verification* (if no misclassifications are found for all layers) or *falsification* (in which case the adversarial examples can be used to fine-tune the network or shown to a human tester).

We implement the techniques using Z3 in a tool called DLV (Deep Learning Verification) and evaluate them on state-of-the-art networks, including regularised and deep learning networks. This includes image classification networks trained for classifying hand-written images of digits 0-9 (MNIST), 10 classes of small colour images (CIFAR10), 43 classes of the German Traffic Sign Recognition Benchmark (GTSRB) and 1000 classes of colour images used for the well-known imageNet large-scale visual recognition challenge (ILSVRC). The perturbed images in Figure 1 are found automatically using our tool for the network trained on the CIFAR10 dataset.

References

[1] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. *https://arxiv.org/abs/1610.06940*, 2016.