Measuring the effect of nuisance variables on classifiers

Alhussein Fawzi
alhussein.fawzi@epfl.ch

Pascal Frossard
pascal.frossard@epfl.ch

Signal Processing Laboratory (LTS4)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

In real-world classification problems, *nuisance* can cause wild variability in the data. Nuisance corresponds for example to geometric distortions of the image, occlusions, illumination changes or any other deformations that do not alter the ground truth label of the image. It is therefore crucial that designed classifiers are robust to nuisance variables, especially when these are deployed in real and possibly hostile environments. We propose a probabilistic framework for efficiently *estimating* the robustness of state-of-the-art classifiers and *sampling* problematic samples from the nuisance space. This allows us to visualize and understand the regions of the nuisance space that cause misclassification, in the perspective of improving robustness. Our probabilistic framework is applicable to arbitrary classifiers and potentially high-dimensional and complex nuisance spaces.

Ingredients.

Classifier: We consider an arbitrary classifier that is provided through its conditional distribution $p_{\text{cl}}(c|x)$, which represents the probability that an image x is classified as c by the classifier.
Nuisance: Let $\mathcal{T}$ be the set of nuisances, and let $p_{\mathcal{T}}(\theta)$ denote a prior probability distribution on $\mathcal{T}$ that captures our region of interest in the nuisance space. For example, when $\mathcal{T}$ denotes the occlusion nuisance set, $p_{\mathcal{T}}(\theta)$ might take large values for small occlusions (covering small parts of the image), and smaller values for large occlusions.

Measuring the robustness to nuisance.

We define the robustness $\mu_{\mathcal{T}}(x)$ as the average confidence of the classifier on the transformed samples:

$$\mu_{\mathcal{T}}(x) := \mathop{\mathbb{E}}_{\theta \sim p_{\mathcal{T}}} \left[p_{\text{cl}}(\ell(x)|T_\theta x) \right], \tag{1}$$

where $\ell(x)$ is the ground truth label of x, $T_\theta x$ is the image x transformed by θ.
Given a data distribution p_d, we define the *global* robustness to nuisance variables in $\mathcal{T}$ as the average of $\mu_{\mathcal{T}}(x)$, i.e.,

$$\rho_{\mathcal{T}} := \mathop{\mathbb{E}}_{x \sim p_d} \left[\mu_{\mathcal{T}}(x) \right]. \tag{2}$$

Estimation of the average robustness.

Monte Carlo approximation.

$$\hat{\mu}_{\mathcal{T}}(x) = \frac{1}{N} \sum_{i=1}^{N} p_{\text{cl}}(\ell(x)|T_{\theta_i} x), \text{ with } \theta_i \overset{\text{iid}}{\sim} p_{\mathcal{T}}, \tag{3}$$

$$\hat{\rho}_{\mathcal{T}} = \frac{1}{M} \sum_{j=1}^{M} \hat{\mu}_{\mathcal{T}}(x_j), \text{ with } x_j \overset{\text{iid}}{\sim} p_{\text{d}}. \tag{4}$$

How many samples are needed to have a good approximation of $\rho_{\mathcal{T}}$?

$$\text{If } NM \geq \frac{\ln(2/\delta)}{2t^2}, \text{ then } \mathbb{P}(|\hat{\rho}_{\mathcal{T}} - \rho_{\mathcal{T}}| \leq t) \geq 1 - \delta.$$

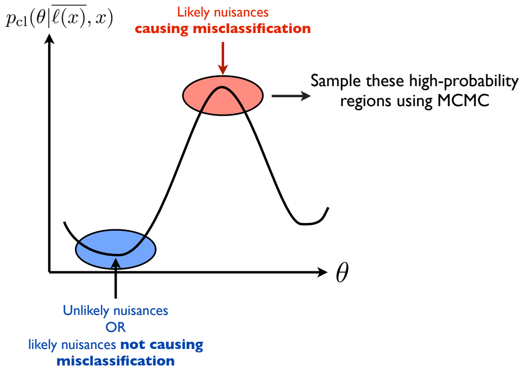
Sampling from the problematic region.

While $\rho_{\mathcal{T}}$ measures the *average* likelihood of the classifier, it is also crucial to visualize and understand the *problematic* regions of the nuisance space where the classifier has low confidence on transformed images.
The problematic regions are mathematically described by the *posterior* distribution

$$p_{\text{cl}}(\theta|\overline{\ell(x)}, x) \propto p_{\text{cl}}(\overline{\ell(x)}|T_\theta x) p_{\mathcal{T}}(\theta),$$

where $p_{\text{cl}}(\overline{\ell(x)}|T_\theta x) = 1 - p_{\text{cl}}(\ell(x)|T_\theta x)$ is the probability that $T_\theta x$ is *not* classified as $\ell(x)$.

Sampling from this posterior distribution allows us to "diagnose" the weak spots of the classifier; that is, nuisance parameters that can cause classification errors.



Likely nuisances
causing misclassification

Sample these high-probability
regions using MCMC

Unlikely nuisances
OR
likely nuisances **not causing misclassification**

Experiments.

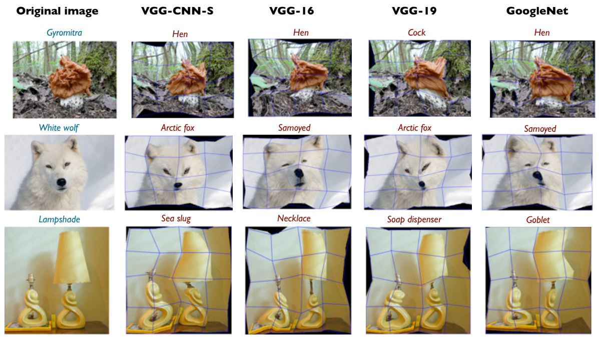
Quantitative evaluation of the robustness to affine transformations.

We evaluate quantitatively the robustness to affine transformations of different CNN architectures on MNIST. We show that

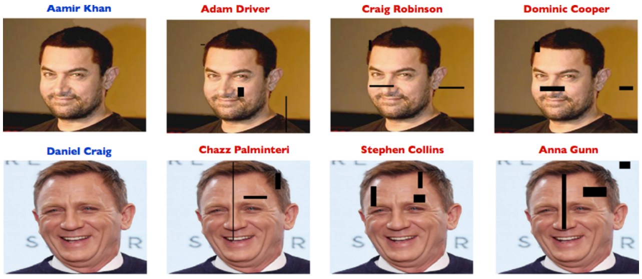
- Deeper networks are more robust to nuisances,
- While dropout leads to significant improvements in test accuracy, it has no effect on the robustness,
- Data augmentation and spatial transformers [1] can lead to a quantitatively significant boost of the robustness.

Robustness of classifiers to piecewise affine transformations.

We show samples drawn from the posterior distribution $p_{\text{cl}}(\theta|\overline{\ell(x)}, x)$ for different state-of-the-art networks trained on ImageNet.



Robustness of face recognition to occlusions.
We show different samples drawn from the posterior distribution $p_{\text{cl}}(\theta|\overline{\ell(x)}, x)$, where the classifier is VGG-Faces [2].



With relatively small occlusions, the classifier can achieve trivial errors (e.g., Daniel Craig → Anna Gunn).

[1] M. Jaderberg, K. Simonyan, and A. Zisserman. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2008–2016, 2015.

[2] O.M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *British Machine Vision (BMVC)*, 1(3):6, 2015.