Manitest: Are classifiers really invariant?

Alhussein Fawzi
alhussein.fawzi@epfl.ch
Pascal Frossard
pascal.frossard@epfl.ch

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

Abstract

Invariance to geometric transformations is a highly desirable property of automatic classifiers in many image recognition tasks. Nevertheless, it is unclear to which extent state-of-the-art classifiers are invariant to basic transformations such as rotations and translations. This is mainly due to the lack of general methods that properly measure such an invariance. In this paper, we propose a rigorous and systematic approach for quantifying the invariance to geometric transformations of any classifier. Our key idea is to cast the problem of assessing a classifier's invariance as the computation of geodesics along the manifold of transformed images. We propose the *Manitest* method, built on the efficient Fast Marching algorithm to compute the invariance of classifiers. Our new method quantifies in particular the importance of data augmentation for learning invariance from data, and the increased invariance of convolutional neural networks with depth. We foresee that the proposed generic tool for measuring invariance to a large class of geometric transformations and arbitrary classifiers will have many applications for evaluating and comparing classifiers based on their invariance, and help improving the invariance of existing classifiers.

1 Introduction

Due to the huge research efforts that have been recently deployed in computer vision and machine learning, the state-of-the-art image classification systems are now reaching performances that are close to those of the human visual system in terms of accuracy on some datasets [18, 33]. Questions emerge to what differences remain between human visual system and state-of-the-art classifiers. We focus here on one key difference, namely the problem of *invariance to geometric transformations*. While the human visual system is invariant to some extent to geometric transformations, it is unclear whether automatic classifiers enjoy the same invariance properties. The importance of invariance in classifiers has been outlined in recent works [22, 30], and effective solutions for transformation-invariant classifications have been proposed by either adapting the classification rules with proper distance metrics [11, 16, 29, 36], or by improving the features used for classification [1, 4, 24]. To validate such new design choices and to understand how to further improve classifiers' invariance, it becomes however primordial to develop general methods to properly measure the robustness of classifiers to geometric transformations of data samples. Previous works have proposed methods to evaluate the invariance of classifiers, either by controlled changes in simple images [3], or by specific tests for features of popular neural network architectures [13]. These previous studies are however limited, as they are restricted to one-dimensional transformations (e.g., rotations only), to particular types of classifiers (e.g., neural networks) or to simple images (e.g., sinusoidal images), and are based on heuristically-driven quantities. Another approach for measuring invariance consists in generating datasets with transformed images, and measuring the accuracy of classifiers on these datasets [19, 21, 31]. This is however laborious and involves building a novel well-designed dataset to compare all classifiers on a common ground.

In this paper, we propose a principled and systematic method to measure the robustness of arbitrary image classifiers to geometric transformations. In particular, we design a new framework that can be applied to any Lie group $\mathcal{T}$ and to any classifier f regardless of the particular nature of the classifier. For a given image, we define the invariance measure as the minimal distance between the identity transformation and a transformation in $\mathcal{T}$ that is sufficient to change the decision of the classifier f on that image. In order to define the transformation metric, our novel key idea is to represent the set of transformed versions of an image as a manifold; the transformation metric is then naturally captured by the geodesic distance on the manifold. Hence, for a given image, our invariance measure essentially corresponds to the minimal geodesic distance on the manifold that leads to a point where the classifier's decision is changed. A global invariance measure is then derived by averaging over a sufficiently large sample set. Equipped with our generic definition of invariance, we leverage the techniques used in the analysis of manifolds of transformed visual patterns [9, 14, 38] and design the Manitest method built on the efficient Fast Marching algorithm [15, 35] to compute the invariance of classifiers.

Using Manitest, we quantitatively show the following results: (i) The invariance of convolutional neural networks and scattering transforms largely outperform SVM classifiers, (ii) Two classifiers can have a similar accuracy, but have different invariance scores, (iii) The invariance of convolutional neural networks improves with network depth, (iv) On natural images classification task, baseline convolutional networks are not invariant to slight combinations of translations, rotations, and dilations (v) Data augmentation can dramatically increase the invariance of a classifier. The latter result is particularly surprising, as an SVM with RBF kernel trained on augmented samples can outperform the invariance of convolutional neural networks (without data augmentation) on a handwritten digits dataset. Besides these results, we showcase examples illustrating the introduced invariance scores. By providing a systematic tool to assess the classifiers in terms of their robustness to geometric transformations, we bridge a gap towards understanding the invariance properties of different families of classifiers, which will hopefully lead to building new classifiers that perform closer to the human visual system. The code of Manitest is available on the project website[1].

2 Problem formulation

2.1 Definitions

We consider a mathematical model where images are represented as functions $I : \mathbb{R}^2 \to \mathbb{R}$, and we denote by L^2 the space of square integrable images. Let $\mathcal{T}$ be a Lie group consisting of geometric transformations on $\mathbb{R}^2$, and we denote by p the dimension of $\mathcal{T}$ (i.e., number of free parameters). For any transformation τ that belongs to $\mathcal{T}$, we denote by I_τ the image I transformed by τ. That is, $I_\tau(x,y) = I(\tau^{-1}(x,y))$. Examples of Lie groups include the rotation group SO(2) ($p = 1$, described by one angle) and the similarity group ($p = 4$, described

[1] http://sites.google.com/site/invmanitest/

by a 2D translation vector, a dilation and an angle).

Consider an image classification task, where the images are assigned discrete labels in $\mathcal{L} = \{1,\dots,L\}$, and let f be an arbitrary image classifier. Formally, f is a function defined on the space of square integrable images L^2, and takes values in the set $\mathcal{L}$. Our goal in this paper is to evaluate the invariance of f with respect to $\mathcal{T}$. Given an image I, we define the invariance score of f relative to I, $\Delta_{\mathcal{T}}(I;f)$, to be the *minimal normalized distance* from the identity transformation to a transformation τ that changes the classification label, i.e.,

$$\Delta_{\mathcal{T}}(I;f) = \min_{\tau \in \mathcal{T}} \frac{d(e,\tau)}{\|I\|_{L^2}} \text{ subject to } f(I_\tau) \neq f(I), \tag{1}$$

where e is the identity element of the group $\mathcal{T}$ and $d : \mathcal{T} \times \mathcal{T} \to \mathbb{R}^+$ is a metric on $\mathcal{T}$ that we define later (Section 2.2). The invariance score quantifies the resilience of f to transformations in $\mathcal{T}$, namely larger values of $\Delta_{\mathcal{T}}(I;f)$ indicate a larger invariance. It is worth noting that our definition of $\Delta_{\mathcal{T}}$ is related to the recent work in [32] that defined adversarial noise as the minimal perturbation (in the Euclidean sense) required to misclassify the datapoint. However, instead of considering generic adversarial perturbations, we focus on minimal *geometric transformations*, with a metric borrowed from the group $\mathcal{T}$.

For a given a distribution of datapoints μ, the global invariance score of f to transformations in $\mathcal{T}$ is defined by

$$\rho_{\mathcal{T}}(f) = \mathbb{E}_{I \sim \mu} \Delta_{\mathcal{T}}(I;f). \tag{2}$$

The quantity $\rho_{\mathcal{T}}(f)$ depends on f as well as the distribution of datapoints μ. However, to simplify notations, we have omitted the dependence on μ, assuming the distribution is clear from the context. In practical classification tasks, the true underlying distribution μ is generally unknown. In that case, we estimate the global resilience by taking the empirical average[2] over training points: $\hat{\rho}_{\mathcal{T}}(f) = \frac{1}{m}\sum_{j=1}^{m} \Delta_{\mathcal{T}}(I_j;f)$.

2.2 Transformation metric

We discuss and introduce the distance used for the invariance score $\Delta_{\mathcal{T}}(I;f)$. It should be noted that $\mathcal{T}$ is possibly a multi-dimensional group (i.e., the transformations in $\mathcal{T}$ are described by many parameters of different nature such as translation, rotation, scale, ...); hence, defining a trivial metric that measures the absolute distance between transformation parameters is of limited interest, as it combines parameters possibly of different nature. Instead, a more relevant notion of distance is one that *depends on the underlying image I*. In that case, $d(\tau_1,\tau_2)$ quantifies the change in *appearance* between images I_{τ_1} and I_{τ_2}, rather than an absolute distance between the two transformations. Consider for example the *image distance* $d_I(\tau_1,\tau_2) = \|I_{\tau_1} - I_{\tau_2}\|_{L^2}$. While d_I explicitly depends on the underlying image I, it fails to capture the intrinsic geometry of the family of transformed images. To illustrate this point, we consider a simple example of images in Fig. 1 with two transformed versions I_{τ_1} and I_{τ_2} of a reference image I_{τ_0}. Note that $d_I(\tau_0,\tau_1) = d_I(\tau_0,\tau_2)$, as both transformed objects have no intersection with the reference object. However, it is clear that I_{τ_2} incurred a large rotation and translation, while I_{τ_1} underwent a slight vertical translation. Hence, the distance metric should naturally satisfy $d(\tau_0,\tau_1) < d(\tau_0,\tau_2)$, which is not the case for the image distance. This is crucial in our setting, as a classifier that recognizes the similarity of the objects in I_{τ_2}

[2] In practice, it is sufficient to consider an empirical average over a sufficiently large random subset of the training set. The number of samples is chosen to achieve a small enough confidence interval.

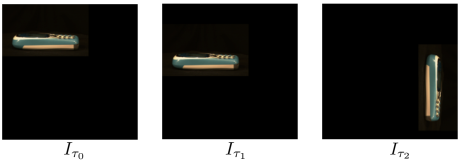


Figure 1: Schematic representation of the problem encountered by using metric the L^2 metric. Black pixels indicate pixels with value 0, and I_{τ_1}, I_{τ_2} are obtained by applying a combination of rotation and translation to I_{τ_0}. Image taken from [12].

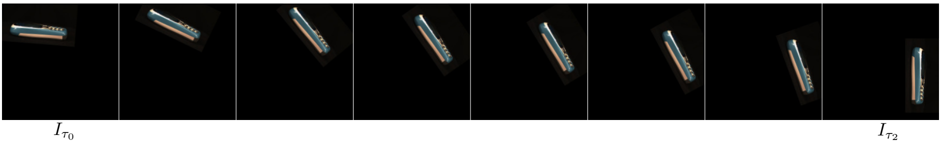


Figure 2: Images along the geodesic path from I_{τ_0} to I_{τ_2}

and I_{τ_0} is certainly more robust to transformations than a classifier that merely recognizes the similarity between I_{τ_1} and I_{τ_0}, and should be given a higher score. This example underlines a well-known fundamental issue with the L^2 distance that fails to capture the intrinsic distance of the curved manifold of transformed images (see e.g., [9, 34]). To correctly capture the intrinsic structure of the manifold, we define d to be the length of the shortest path belonging to the manifold (i.e., the *geodesic distance*). For illustration, we show in Fig. 2 images along the geodesic path from τ_0 to τ_2; the geodesic distance is then essentially the sum of *local* L^2 distances between transformed images over the geodesic path. We formalize these notions as follows.

Let $\mathcal{M}(I)$ be the family of transformed images $\mathcal{M}(I) = \{I_\tau : \tau \in \mathcal{T}\}$. Equipped with the L^2 metric, $\mathcal{M}(I)$ defines a metric space and a continuous submanifold of L^2. Following the works of [14, 38] that considered similar manifolds in different contexts, we call $\mathcal{M}(I)$ an *Image Appearance Manifold* (IAM), and we follow here their approach. Assuming that $\gamma : [0,1] \mapsto \mathcal{T}$ is a C^1 curve in $\mathcal{T}$, and that $I_{\gamma(t)}$ is differentiable with respect to t, we define the *length* $L(\gamma)$ of γ as

$$L(\gamma) = \int_0^1 \left\| \frac{d}{dt} I_{\gamma(t)} \right\|_{L^2} dt. \tag{3}$$

Note that Eq. (3) is expressed in terms of the L^2 metric in the image appearance manifold and corresponds to summing the local L^2 distances between transformed images over the path I_γ. We now show that $L(\gamma)$ can be expressed as a length associated to a Riemannian metric on $\mathcal{T}$ that we now derive. Defining the map

$$F : \mathcal{T} \to \mathcal{M}, \quad \tau \mapsto I_\tau,$$

we have

$$\frac{d}{dt} I_{\gamma(t)} = (F \circ \gamma)'(t) = dF_{\gamma(t)}(\gamma'(t)),$$

where dF_τ denotes the differential of F at τ, and γ' is derivative of γ. It follows that

$$L(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))} dt$$

where g_τ is the *Riemannian metric* (i.e., a positive bilinear form on $T_\tau \mathcal{T}$, the tangent space of $\mathcal{T}$ at τ), given by:

$$g_\tau(v,w) = \langle dF_\tau(v), dF_\tau(w) \rangle_{L^2} \text{ for all } v,w \in T_\tau \mathcal{T}.$$

Note that g can be equivalently seen as the pullback of the L^2 metric on $\mathcal{M}(I)$ along F. By choosing a basis in the tangent space, the length $L(\gamma)$ can be equivalently written

$$L(\gamma) = \int_0^1 \sqrt{\gamma'(t)^T G_{\gamma(t)} \gamma'(t)} dt,$$

where $G_{\gamma(t)}$ is the $p \times p$ positive definite matrix associated to the bilinear form g.

Example 1 (Rotation, $\mathcal{T} = SO(2)$) The transformation group $\mathcal{T}$ is parametrized with a rotation angle θ ($p = 1$). In this case, the matix G_θ is of size 1 by 1, and equal to

$$G_\theta = \left\| \frac{\partial I_\theta}{\partial \theta} \right\|_{L^2}^2. \quad \square$$

Example 2 (Dilation+Rotation). The group $\mathcal{T}$ has 2 degrees of freedom; namely a scale parameter a, and a rotation angle θ. The Riemannian metric reads

$$G_\tau = \begin{bmatrix} \left\langle \frac{\partial I_\tau}{\partial a}, \frac{\partial I_\tau}{\partial a} \right\rangle & \left\langle \frac{\partial I_\tau}{\partial a}, \frac{\partial I_\tau}{\partial \theta} \right\rangle \\ \left\langle \frac{\partial I_\tau}{\partial \theta}, \frac{\partial I_\tau}{\partial a} \right\rangle & \left\langle \frac{\partial I_\tau}{\partial \theta}, \frac{\partial I_\tau}{\partial \theta} \right\rangle \end{bmatrix}. \quad \square$$

Having defined the length of a curve on $\mathcal{T}$, the geodesic distance between two points τ_1, τ_2 is defined as the length of the shortest curve joining the two points:

$$d(\tau_1, \tau_2) = \inf\{L(\gamma) : \gamma \in C^1([0,1]), \gamma(0) = \tau_1, \gamma(1) = \tau_2\}.$$

Finally, our problem therefore consists in computing the global invariance score, or equivalently $\Delta_{\mathcal{T}}(I; f)$ defined in Eq. (1), where d is the geodesic distance. In other words, our problem becomes that of computing the minimal geodesic distance from the identity transformation to a transformation that is sufficient to change the estimated label of f.

3 Invariance score computation

The key to an efficient and accurate approximation of $\Delta_{\mathcal{T}}(I; f)$ lies in the effective computation of geodesics on the manifold $(\mathcal{T}, G)$ that we address as follows.

Let $u(\tau) = d(e, \tau)$ be the *geodesic map* that measures the geodesic distance between the (fixed) identity element and τ. The geodesic map satisfies the following Eikonal equation [26]

$$\|\nabla u(\tau)\|_{G_\tau^{-1}} = 1 \text{ for } \tau \in \mathcal{T} \backslash \{e\}, \text{ and } u(e) = 0, \tag{4}$$

where $\|x\|_A = \sqrt{\langle x, x \rangle_A}$ with $\langle x, y \rangle_A = x^T A y$. Moreover, it was proved in [7] that the geodesic map u is the *unique* viscosity solution of the Eikonal equation, provided that $\tau \to G(\tau)$ is continuous. Many numerical schemes rely on the Eikonal equation characterization to approximate the geodesic map. We use here the popular *Fast Marching (FM) method* [15], a fast front propagation approach that computes the values of the discrete geodesic map in increasing order. We only provide here a brief description of FM due to space constraints, and focus on the case where the manifold $\mathcal{T}$ is two-dimensional (i.e., $p = 2$). The extension to arbitrary dimensions is straightforward, and we refer to [26, 28] for more complete explanations and computations.

Algorithm 1 Manitest method (with $p = 2$) for computing $\Delta_{\mathcal{T}}(I; f)$

```
Initialize $U(e) = 0$, $U = \infty$ otherwise, and tag all nodes as *unknown*.
while termination criterion is not met do
    Select the *unknown* node $\tau_{\min}$ that achieves minimal distance $U$.
    Tag $\tau_{\min}$ as *known*.
    If $f(I_{\tau_{\min}}) \neq f(I)$, set $\Delta_{\mathcal{T}}(I; f) \leftarrow U(\tau_{\min}) / \|I\|_{L^2}$ and terminate.
    for all *unknown* $\tau \in \mathcal{N}(\tau_{\min})$ do
        Update $U(\tau)$ to be the minimum of itself, $U(\tau_{\min}) + \|\tau - \tau_{\min}\|_{G_\tau}$ and the expression in Eq.(5).
    end for
end while
```

We assume that the manifold $\mathcal{T}$ is sampled using a regular grid; let $\mathcal{T}_*$ be the sampling of $\mathcal{T}$, and U be the discrete vector that approximates u at the nodes. The structure of Fast Marching is almost identical to Dijkstra's algorithm for computing shortest paths on graphs [8]. The main difference lies in the update step, which bypasses the constraint of propagation along edges. For a given node τ, define $\mathcal{N}(\tau)$ to be the set of neighbours of τ (see illustration in Fig. 3). In the FM algorithm, each grid point is tagged either as *Known* (nodes for which distance is frozen), or *Unknown* (nodes for which distance can change in subsequent iterations). Initially, the grid points are set to *Unknown*, and U is set to ∞, except $U(e)$ that is set to zero. At each iteration of FM, the unknown node $\tau_{\min}$ with smallest U is selected, and tagged as *Known*. Then, each unknown neighbour $\tau \in \mathcal{N}(\tau_{\min})$ is visited, and $U(\tau)$ is updated as follows: $U(\tau)$ is set to be the minimum of itself, $U(\tau_{\min}) + \|\tau - \tau_{\min}\|_{G_\tau}$ and

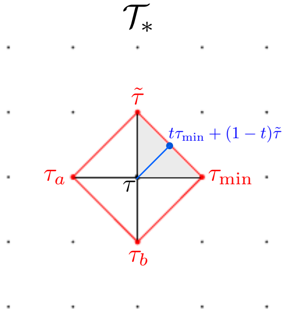


Figure 3: Schematic representation of the discretized manifold $\mathcal{T}_*$, and the Fast Marching update rule. In this figure, we have $\mathcal{N}(\tau) = \{\tilde{\tau}, \tau_{\min}, \tau_a, \tau_b\}$.

$$\min_{t \in [0,1]} tU(\tau_{\min}) + (1-t)U(\tilde{\tau}) + \|t\tau_{\min} + (1-t)\tilde{\tau} - \tau\|_{G_\tau}, \tag{5}$$

for each known $\tilde{\tau}$ such that $(\tau, \tau_{\min}, \tilde{\tau})$ forms a triangle (see Fig. 3). It is worth noting that, unlike Dijkstra, FM seeks the optimal point (possibly outside the set $\mathcal{T}_*$) on the neighbourhood boundary that minimizes the estimated distance at τ, under a linear approximation assumption (Eq. 5). Fortunately, the problem in Eq. (5) can be solved in closed form, as it corresponds to the minimization of a scalar quadratic equation [28].

The Manitest method, which applies FM algorithm to compute $\Delta_{\mathcal{T}}(I; f)$, is given in Algorithm 1 in the two dimensional case. The algorithm is stopped whenever a transformation that changes the classification label is found.[3] The nodes and metrics are generated on-the-fly in order to avoid spending unnecessary ressources on far-away nodes that might be farther than the minimal transformation that satisfies $f(I) \neq f(I_\tau)$ and therefore never visited.

[3] To ensure the termination of the algorithm (even if no successful transformation is found) we limit the number of iterations N to 50,000. However, in all our experiments, this limit was never reached, and the algorithm terminated by successfully finding a transformation that satisfies $f(I_\tau) \neq f(I)$.

The complexity of Manitest is $O(N \log(N))$, where N is the number of visited nodes if a min-heap structure is used [26] (for constant p, and constant cost for evaluation of f). It is important to note however that the complexity of the algorithm has an exponential dependence on the dimension p since our method involves the enumeration of simplices in dimension p; this is however not a big limitation as our main focus goes to low-dimensional transformation groups (e.g., $p \leq 6$ for affine transformations).

Finally, we note that when the metric is isotropic (i.e., G_τ is proportional to the identity matrix for all τ), FM provides a consistent scheme. That is, as the discretization step tends to zero, the solution computed by the algorithm tends towards the viscosity solution of the Eikonal equation. Unfortunately, for arbitrary anisotropic metrics, consistency is however not guaranteed, and the exact computation of the geodesics becomes much more difficult and computationally demanding (see [2, 23, 25, 27]). However, we observed that the anisotropy of the considered metric is generally not very large in the vicinity of e (although it exceeds the theoretical limit of guaranteed consistency). This leads to empirically accurate estimates of the geodesic distance using Manitest, when the discretization step is sufficiently small. Finally, we stress that that all previous methods addressing the metric anisotropy can readily be applied to our setting, and we leave that as future work.

4 Experiments

We propose now a set of experiments to study the invariance of classifiers in different settings. In particular, we consider the following transformation groups:

- $\mathcal{T}_{\text{trans}}$: in-plane translations of the image ($p = 2$),
- $\mathcal{T}_{\text{dil+rot}}$: dilations and rotations around the center of the image ($p = 2$),
- $\mathcal{T}_{\text{sim}}$: similarity transformations that describe combinations of translations, dilations and rotations around the center of the image ($p = 4$).

In all experiments, we used a discretization step of 0.5 pixels for translations, $\pi/20$ radians for rotation, and 0.1 for dilation for Manitest. Finally, the transformed images have the same size as the original image, and we use a zero-padding boundary condition.

4.1 Handwritten digits dataset

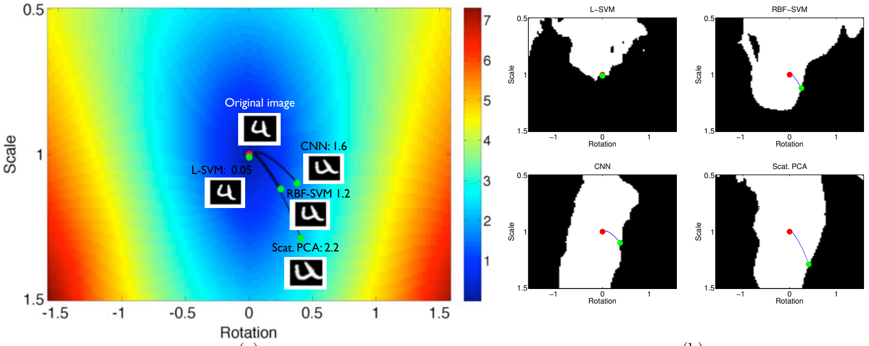
We first compare the invariance of different classifiers on the MNIST handwritten digits dataset [20]. We consider the following classifiers:

1. **Linear SVM** [10],
2. **SVM with RBF kernel** [6],
3. **Convolutional Neural Network** [37]: we employ a baseline architecture with two hidden layers containing each a convolution operation (5×5 filters with 32 feature maps for the first layer and 64 for the second layer), a rectified linear unit nonlinearity, and a max pooling over 2×2 windows followed by a subsampling. The architecture is trained with stochastic gradient descent, with a softmax loss.
4. **Scattering transform followed by a generative PCA classifier**. We used the same settings as in [5], and we refer to that paper for more details.

Table 1 reports the performance of the different classifiers under study, and their invariance scores $\hat{\rho}_{\mathcal{T}}(f)$ using Manitest. As expected, the linear and RBF-SVM classifiers compare poorly to other classifiers in terms of invariance. This is due to the construction of the CNN and Scat. PCA, which explicitly take into account the invariance through pooling operations, while others do not. Moreover, it can be noted that Scat. PCA outperforms CNN in terms of robustness to translations, and global similarity transformations, even if the two

Group	L-SVM	RBF-SVM	CNN	Scat. PCA
Test error (%)	8.4	1.4	**0.7**	0.8
Translations ($\mathcal{T} = \mathcal{T}_{\text{trans}}$)	0.8	1.3	1.7	**2.1**
Dilations + Rotations ($\mathcal{T} = \mathcal{T}_{\text{dil+rot}}$)	0.8	1.5	**1.9**	1.8
Similarity ($\mathcal{T} = \mathcal{T}_{\text{sim}}$)	0.6	1.1	1.5	**1.6**

Table 1: Accuracy and invariance scores of different classifers on the MNIST dataset.



(a) (b)

Figure 4: Distance map with $\mathcal{T}_{\text{dil+rot}}$ group (a), and correctly classified regions (b), for the four tested classifiers on an example image of digit "4". Geodesic paths are also shown.

classifiers have similar test error. This result is in agreement with the theoretical evidence [5, 24] showing that scattering classifiers are invariant to deformations.

To further get an insight on the invariance of the classifiers, we focus on the two-dimensional group $\mathcal{T}_{\text{dil+rot}}$, and show in Fig. 4 (a) the geodesic distance map for an example image of digit "4" computed starting from the identity transformation (shown by a red dot at the center). Moreover, we overlay the minimally transformed images that change the labels of each of the classifiers, along with the corresponding geodesic paths. On this example, the Scat. PCA classifier is the most robust: a large dilation, accompanied with a rotation is required to change the classification label. In contrast, the linear SVM is easily "fooled" with a slight dilation. In Fig. 4 (b) we illustrate in white the region of the Rotation-Scale plane, where the classifier outputs the correct label "4". Interestingly, the CNN and Scat. PCA classifiers are largely invariant to dilations (indicated by the vertical shape of the white region), while being moderately robust to rotations.

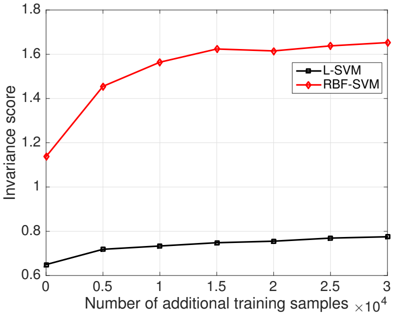
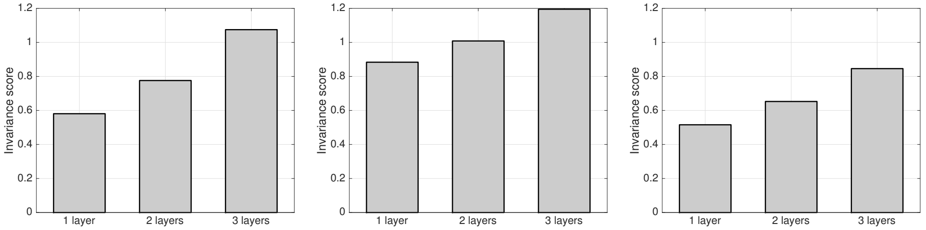


Figure 5: Invariance score versus number of additional training samples, for MNIST, with $\mathcal{T} = \mathcal{T}_{\text{sim}}$.

In vision tasks, it is common practice to augment the training data with artificial examples obtained by slightly distorting the original examples to achieve invariance. Although this practice is known to improve the classification performance of the classifiers on many tasks, its effect on the invariance of the classifier is not quantitatively understood. Fig. 5 illustrates the Manitest invariance scores for L-SVM and RBF-SVM classifiers trained on augmented training sets obtained by randomly generating transformations[4] from the similarity group $\mathcal{T}_{\text{sim}}$, on the MNIST dataset. Both classifiers improve their invariance score as more transformed samples are added to the training set. This result has moreover an element of surprise, as RBF-SVM succeeds in improving its invariance score by around 50% with mere additions of artificial examples in the training set, and outperforms the invariance of CNN (without data augmentation). Moreover, the obtained score is comparable to Scat. PCA classifier, which is carefully designed to satisfy invariance properties. This experiment permits to characterize the actual power of data augmentation for *learning* the invariance from the data.

4.2 Natural images

In this second experimental section, we perform experiments on the CIFAR-10 dataset [17]. We focus on baseline CNN classifiers, and learn architectures with 1, 2 and 3 hidden layers. Specifically, each layer consists of a successive combination of convolutional, rectified linear units and pooling operations. The convolutional layers consist of 5×5 filters with respectively $32, 32$ and 64 feature maps for each layer, and the pooling operations are done on a window of size 3×3 with a stride parameter of 2. We build the three architectures gradually, by successively stacking a new hidden layer on top of the previous architecture (kept fixed). The last hidden layer is then connected to a fully connected layer, and the softmax loss is used. Moreover, the different architectures are trained with stochastic gradient descent. On the test set, the error of the three architectures are respectively 35.6%, 25.0% and 22.7%.

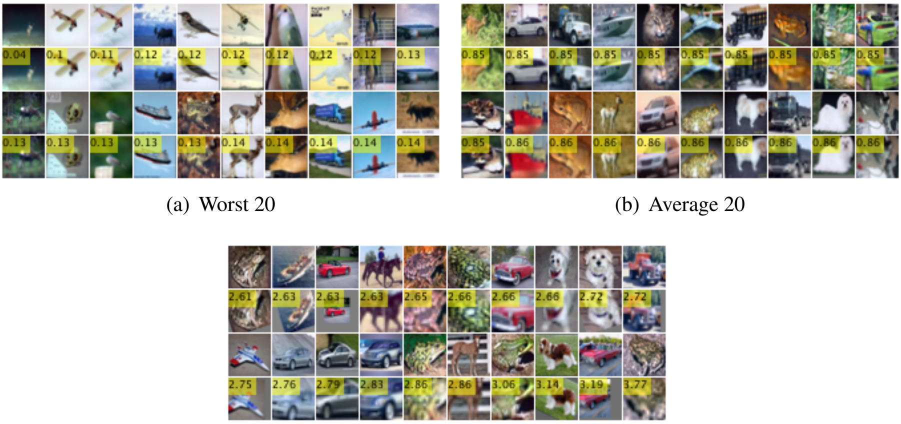


(a) Translations (b) Dilation + Rotation (c) Similarity

Figure 6: Invariance scores of CNNs on $\mathcal{T}_{\text{trans}}$, $\mathcal{T}_{\text{dil+rot}}$ and $\mathcal{T}_{\text{sim}}$, for the CIFAR-10 dataset.

We show in Fig. 6 the Manitest invariance scores of the three architectures. Our approach captures the *increasing* invariance with the number of layers of the network, for the three groups under study. This result is in agreement with empirical studies and previous known belief [1, 13] that invariance increases with the depth of the network. However, while previous results were measuring the invariance with respect to a one dimensional transformation group (e.g., rotation only), Manitest provides a systematic and principled way of verifying the increased invariance of CNNs with depth on more complex Lie groups (e.g., similarity transformations). Interestingly enough, it should be noted that despite the relatively small difference in performance between the two and three layers architectures, the invariance score strongly increases. This highlights again that invariance and performance measures capture two different properties of classifiers.

[4] Random transformations are constrained as follows: translation of at most 3 pixels in each direction, a scaling parameter between 0.7, and 1.3, and a rotation of at most 0.2 radians.



(a) Worst 20

(b) Average 20

(c) Top 20

Figure 7: Illustration of images having (a) worst, (b) average, (c) top invariance to **similarity** transformations (i.e., $\mathcal{T} = \mathcal{T}_{\text{sim}}$), for the three-layer CNN. The odd rows show the original images, and the even rows show the minimally transformed images changing the prediction of the CNN. The Manitest invariance score $\Delta_{\mathcal{T}}(I; f)$ is indicated on each transformed image. All original images are **correctly classified** by the 3-layer CNN.

Compared to the handwritten digits task, note that the Manitest scores obtained on the CIFAR task are generally much smaller, which suggests that it is harder to achieve invariance on this task. To visualize the level of invariance of the 3-layer CNN on the CIFAR-10 dataset, we show in Fig. 7 sorted example images. For images with an average invariance score or less, note that the distinction between the transformed and original images are hardly perceptible. This suggests that the CNN is not robust to combinations of translations, rotation and dilation, even if it achieves a high accuracy. On the other hand, the difference between the original and the minimally transformed images are clearly perceptible for the top-scored images, even though a human observer is likely to correctly recognize the class of the transformed images.

5 Conclusion

In this paper, we proposed a systematic and rigorous approach for measuring the invariance of any classifier to low-dimensional transformation groups. Using a manifold perspective, we were able to convert the problem of assessing the classifier's invariance to that of computing geodesic distances. Using Manitest, we quantified the increasing invariance of CNNs with depth, and highlighted the importance of data augmentation for learning invariance from data. We believe Manitest will be used to perform an in-depth empirical analysis of different classification architectures, in order to have a better understanding of the building blocks that best preserve invariance, and potentially build more robust classifiers.

Acknowledgments. We are grateful for the comments provided by the anonymous reviewers. We also thank Laurent Jacques and Gabriel Peyré for their insights on Fast Marching. We thank Luca Baroffio and Hamza Fawzi for their comments on the paper draft.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[2] Fethallah Benmansour and Laurent D Cohen. Tubular structure segmentation based on minimal path method and anisotropic enhancement. *International Journal of Computer Vision*, 92(2):192–210, 2011.

[3] Pietro Berkes and Laurenz Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5(6):9, 2005.

[4] Joan Bruna. *Scattering representation for recognition*. PhD thesis, 2012.

[5] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

[6] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[7] Michael G Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton–Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

[8] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[9] David Donoho and Carrie Grimes. Image manifolds which are isometric to euclidean space. *Journal of mathematical imaging and vision*, 23(1):5–24, 2005.

[10] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[11] Alhussein Fawzi and Pascal Frossard. Image registration with sparse approximations in parametric dictionaries. *SIAM Journal on Imaging Sciences*, 6(4):2370–2403, 2013.

[12] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1): 103–112, 2005.

[13] Ian Goodfellow, Honglak Lee, Quoc V Le, Andrew Saxe, and Andrew Y Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646–654, 2009.

[14] Laurent Jacques and Christophe De Vleeschouwer. A geometrical study of matching pursuit parametrization. *IEEE Transactions on Signal Processing*, 56(7):2835–2848, 2008.

[15] Ron Kimmel and James A Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences*, 95(15):8431–8435, 1998.

[16] Effrosyni Kokiopoulou and Pascal Frossard. Minimum distance between pattern transformation manifolds: Algorithm and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1225–1238, 2009.

[17] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, 10:1–40, 2009.

[20] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 97–104, 2004.

[22] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] Qingfen Lin. Enhancement, extraction, and visualization of 3d volume data. *PhD thesis*, 2003.

[24] Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

[25] Jean-Marie Mirebeau. Anisotropic fast-marching on cartesian grids using lattice basis reduction. *SIAM Journal on Numerical Analysis*, 52(4):1573–1599, 2014.

[26] Gabriel Peyré, Mickaël Péchaud, Renaud Keriven, and Laurent D Cohen. Geodesic methods in computer vision and graphics. *Foundations and Trends in Computer Graphics and Vision*, 5(3–4):197–397, 2010.

[27] James A Sethian and Alexander Vladimirsky. Fast methods for the eikonal and related Hamilton–Jacobi equations on unstructured meshes. *Proceedings of the National Academy of Sciences*, 97(11):5699–5703, 2000.

[28] James A Sethian and Alexander Vladimirsky. Ordered upwind methods for static Hamilton–Jacobi equations: Theory and algorithms. *SIAM Journal on Numerical Analysis*, 41(1):325–363, 2003.

[29] Patrice Y Simard, Yann A LeCun, John S Denker, and Bernard Victorri. Transformation invariance in pattern recognition–tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 239–274. Springer, 1998.

[30] Stefano Soatto and Alessandro Chiuso. Visual scene representations: Sufficiency, minimality, invariance and deep approximation. In *International Conference on Learning Representations (ICLR) Workshop*. 2015.

[31] Kihyuk Sohn and Honglak Lee. Learning invariant representations with local transformations. In *International Conference on Machine Learning*, 2012.

[32] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[33] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[34] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[35] John N Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE Transactions on Automatic Control*, 40(9):1528–1538, 1995.

[36] Nuno Vasconcelos and Andrew Lippman. Multiresolution tangent distance for affine-invariant classification. *Advances in neural information processing systems*, pages 843–849, 1998.

[37] Andrea Vedaldi and Karel Lenc. Matconvnet-convolutional neural networks for matlab. *arXiv preprint arXiv:1412.4564*, 2014.

[38] Michael B Wakin, David L Donoho, Hyeokho Choi, and Richard G Baraniuk. The multiscale structure of non-differentiable image manifolds. In *Wavelets XI in SPIE International Symposium on Optical Science and Technology*, 2005.