

Spectral Metric for Dataset Complexity Assessment

Frédéric Branchaud-Charron¹, Andrew Achkar², Pierre-Marc Jodoin¹

¹ Université de Sherbrooke, ² Miovision Inc.

¹{frederic.branchaud-charron, pierre-marc.jodoin}@usherbrooke.ca, ²aachkar@miovision.com *

Abstract

In this paper, we propose a new measure to gauge the complexity of image classification problems. Given an annotated image dataset, our method computes a complexity measure called the cumulative spectral gradient (CSG) which strongly correlates with the test accuracy of convolutional neural networks (CNN). The CSG measure is derived from the probabilistic divergence between classes in a spectral clustering framework. We show that this metric correlates with the overall separability of the dataset and thus its inherent complexity. As will be shown, our metric can be used for dataset reduction, to assess which classes are more difficult to disentangle, and approximate the accuracy one could expect to get with a CNN. Results obtained on 11 datasets and three CNN models reveal that our method is more accurate and faster than previous complexity measures.

1. Introduction

The number of image-based datasets designed to train deep convolutional neural networks (CNN) have been on the rise in the past few years [8, 11, 23, 25, 29, 35, 44]. One reason for this is the indisputable efficiency of CNNs at classifying image data [9, 17, 23, 39].

A common challenge that arises when building a new image dataset for training a CNN is to identify how challenging the classification problem is, which classes are the most difficult to disentangle, and correspondingly what is the minimum dataset size required to train a CNN. As of today, there is no standard framework to make such determinations. The common way to assess the complexity of an image dataset is by training, finetuning and comparing results from several CNNs, the test accuracy being the usual measure for complexity. However, this procedure is time consuming and, most importantly, requires a fully-annotated dataset which is not available when in the process of building it.

Unfortunately, one cannot predict the accuracy of a CNN by only looking at its architecture. As mentioned by Zhang *et al.* [45] in their attempt to understand why deep neural nets generalize well, deep neural networks can easily fit with zero training error on any input data, including pure random noise. This underlines the sole ability of CNNs to project any input data into a linearly separable space (and thus have a zero training error) while sometimes having poor generalization abilities. Their conclusion is that the structure of a neural net, its hyperparameters, its depth, and its optimizer cannot be used alone to predict its generalization capabilities.

Assessing the complexity of a classification problem may instead start from the analysis of the data at hand with the goal of deriving useful complexity measures (*c-measures*) [2, 3, 14, 18, 41]. The goal of *c-measures* is to assess how entangled classes are assuming that datasets with overlapping classes are more difficult to analyze than those with well separated classes. *C-measures* have been shown effective for a number of applications such as classifier selection [7], automatic noise-filtering adjustment [38], dataset reduction [26], and hyperparameter tuning [31].

Unfortunately, existing *c-measures* have not been designed for large image datasets used to train deep neural networks. While some *c-measures* assume that classes are linearly separable in their original feature space [18], others work only for two-class problems [2, 3, 19]. Also, some *c-measures* are prohibitively slow and memory expensive as they require the analysis of matrices whose size is in the order to the number of training samples and/or the feature dimension size [3, 14].

Another important limitation with existing *c-measures* is the fact that they process raw input data. While this was shown valid for some classification problems [7], it is ill suited for deep neural nets since their learning procedure allows them to project input data onto a different and more easily separable space.

In this paper, we present a novel *c-measure* adapted to modern image classification problems. Instead of processing raw input data like previous approaches, our method first projects the input images onto a lower-dimensional la-

*This work is supported by the FQRNT B1 208594 and Mitacs acceleration IT08995.

tent space. This allows to analyze data whose features are better adapted to what CNNs learn. Our method then estimates pairwise class overlap with a Monte-Carlo method which leads to an inter-class similarity matrix. Following the spectral clustering theory, we compute a $K \times K$ Laplacian matrix where K is the number of classes. Finally, the spectrum of this matrix is used to derive our *cumulative spectral gradient* (CSG) c-measure.

The main advantages of our proposed c-measure are as follows :

1. It naturally scales with the number of classes and the number of images in the dataset;
2. Our metric is fast to compute and does not require the computation of prohibitively large matrices;
3. It has no prior assumption on the distribution of the data;
4. It gives a strong insight on which classes are easily separable and those that are entangled;
5. The metric is highly correlated with CNN generalization capabilities;
6. It can be easily used for dataset reduction.

2. Previous works

The goal of a c-measure is to characterize the difficulty of a classification problem. While several c-measures have been proposed in the past, those by Ho and Basu are by far the most widely used [18]. They proposed 12 different descriptors called $F1$, $F2$, $F3$, $L1$, $L2$, $L3$, $N1$, $N2$, $N3$, $N4$, $T1$, and $T2$. $F1$ is a Fisher’s Discriminant Ratio, $F2$ measures the inter-class overlap, and $F3$ is the largest fraction of points one can correctly classify with a stump decision function. $L1$, $L2$ and $L3$ measures the linear separability of the data, while $N1$, $N2$, $N3$ and $N4$ are nearest neighbor measures which estimate the inter-class overlap. As for $T1$, it measures the total number of hyperspheres one can fit into the feature space of a class and $T2$ is the ratio between the total number of training samples N divided by the dimensionality of the data d .

While the c-measures by Ho and Basu have been shown effective for small non-image datasets [7], those metrics are less suited to analyze large and complex image datasets. For example, $F1$, $F2$, $F3$, $L1$, $L2$ and $L3$ assumes the data is linearly separable which is an over-simplistic assumption when considering modern image datasets. $F1$ requires the computation of $d \times d$ matrices which is problematic memory wise for large d (i.e. for medium to large images) and $F3$ measures the linear separability of each class by accounting for each feature independently which is prohibitively slow when both N and d are large. $T1$ is also prohibitively slow as N gets large since it requires to grow an hypersphere around each data point and $T2$ is not a good complexity predictor as will be shown in the results section.

Although Ho and Basu’s metrics were designed for two-class problems, some researchers generalize it to more than two classes by averaging measures obtained between all possible pair of classes [37, 41]. Also, although recent generalizations of the Ho-Basu c-measures have been proposed [2, 12, 41], none addresses explicitly the problem of classifying large image datasets.

Other c-measures have been proposed. For example, Baumgartner and Somorjai [3] proposed a metric adapted to small biomedical datasets with high dimensionality data. Unfortunately, their c-measures are for two-class problems, assume that the data is linearly separable and require the decomposition of $N \times d$ matrices which is only tractable when N and d are small. Duin and Pekalska [14] quantify the complexity of a dataset with metrics derived from a dissimilarity matrix of size $N \times M$ where N is the training set size and M is the number of “representation” vectors randomly sampled from the training set. They report results on several datasets including two image datasets which contains 2000 or less black and white digits. The authors used the Euclidean distance to measure the similarity between two images, a metric that does not generalize well to real-world images [43].

Like we do, some methods build a graph from the dataset to characterize the intra and inter-class relationships [15, 34]. This type of method requires building a $N \times N$ distance matrix which is problematic memory wise for large datasets. For example, the Hub score by [28] requires to compute $A^T A$ where A is a $N \times N$ adjacency matrix.

To our knowledge, Li *et al.* [27] are the only ones who proposed a c-measure applied specifically to modern image datasets and deep neural networks. They called their measure the *Intrinsic Dimension* which is the minimum number of neurons a model needs to reach its best performances. They show that adding more neurons past the Intrinsic Dimension does not improve test accuracy. Unfortunately, as opposed to what we seek to do, their measure requires multiple training of image classification CNNs through a grid-search approach which is slow and tedious. More details on c-measures can be found in the recent survey paper by [28].

3. Proposed Method

3.1. Class overlap

At the core of our c-measure is the notion of class overlap. Let x be an input image and $\phi(x) \in \mathbb{R}^d$ an embedding for that image. As will be discussed later, ϕ can be any function that projects x to a new dimensional space where images with similar content are close together and the other ones further away. The overlap between two classes C_i and C_j refers to the overall area in the feature space for which $P(\phi(x_k)|C_i) > P(\phi(x_k)|C_j)$ when $\phi(x_k)$ is a member of

class C_j . Class overlap can thus be formulated as [36]:

$$\int_{\mathbb{R}^d} \min(P(\phi(x)|C_i), P(\phi(x)|C_j)) d\phi(x). \quad (1)$$

Unfortunately, the direct calculation of this integral is prohibitively complicated for non-parametric distributions and when d (the dimensionality of the embedded space) is large. Since the overlap between two classes is related to the similarity of their distributions, one may instead use a probability distribution distance function such as the Kullback-Leibler divergence or the Kolmogorov-Smirnov test as a surrogate for Eq.(1). One such function is the probability product kernel of Jebara *et al.* [22] :

$$\int_{\mathbb{R}^d} P(\phi(x)|C_i)^\rho P(\phi(x)|C_j)^\rho d\phi(x) \quad (2)$$

which is a generalization of the Bhattacharyya kernel (and the Hellinger distance) when $\rho = 1/2$. While computing Eq.(2) is as complex as computing Eq.(1) for an arbitrary value of ρ , simplification occurs when $\rho = 1$. In that case, the kernel becomes the inner-product between the two distributions $\int_{\mathbb{R}^d} P(\phi(x)|C_i)P(\phi(x)|C_j)d\phi(x)$ which is the expectation of one distribution under the other : $E_{P(\phi(x)|C_i)}[P(\phi(x)|C_j)]$ or $E_{P(\phi(x)|C_j)}[P(\phi(x)|C_i)]$.

Formulating the inter-class divergence as an expectation function allows one to use Monte-Carlo to approximate it:

$$E_{P(\phi(x)|C_i)}[P(\phi(x)|C_j)] \approx \frac{1}{M} \sum_{m=1}^M P(\phi(x_m)|C_j) \quad (3)$$

where $\{\phi(x_1), \dots, \phi(x_M)\}$ are M samples *i.i.d.* from $P(\phi(x)|C_i)$. One can thus approximate the divergences between two class distributions by averaging the probability of M samples of class C_i to be in class C_j or vice versa. Computing inter-class divergences leads to a $K \times K$ similarity matrix \mathcal{S} where K is the total number of classes and \mathcal{S}_{ij} is the Monte-Carlo approximation of the divergence between C_i and C_j .

Since the underlying model of $P(\phi(x_m)|C_j)$ is *a priori* unknown, we approximate it with a K-nearest estimator:

$$p(\phi(x) | C_j) = \frac{K_{C_j}}{MV} \quad (4)$$

where V is the volume of the hypercube surrounding the k closest samples to $\phi(x)$ in class C_j , M is the total number of samples selected in class C_j and K_{C_j} is the number of neighbors around $\phi(x)$ of class C_j .

3.2. Spectral Clustering

The $K \times K$ similarity matrix \mathcal{S} embodies the overall complexity of a dataset by means of class overlap. Our goal is to extract a measure from \mathcal{S} that would summarize the

complexity of that dataset. For that, we rely on the spectral clustering theory [42] that we briefly review in this section.

Let G be an undirected similarity graph $G = (V, E)$ where V is a set of nodes connected by edges E . An edge E_{ij} is an arc connecting two nodes i and j and whose weight $w_{ij} \geq 0$ encodes how close these two nodes are. A weight of 0 implies no connection between i and j whereas a large weight implies strong similarity (in our case, a weight of 1 implies that i and j are identical). The weight of all edges are put in a $n \times n$ adjacency matrix W where n is the total number of nodes. Note that W is symmetric and positive semi-definite due to the undirected nature of the graph which implies that $w_{ij} = w_{ji}$.

The goal of spectral clustering is to partition G into subgraphs such that the edges between the subgraphs have minimum weight. A set of subgraphs $\{G_1, \dots, G_l\}$ is valid when $G_i \cap G_j = \emptyset, \forall i \neq j$ and $G_1 \cup \dots \cup G_l = G$. An optimal partition of G is one for which the cut has minimum cost : $costCut(G_1, \dots, G_l) = \sum w_{ij}$ for i and j in different subgraphs.

Spectral clustering provides an elegant framework to recover the subgraphs with minimum cut. It starts with a Laplacian matrix whose simplest form is $L = D - W$ where D is a degree matrix $D_i = \sum_j w_{i,j}$. L is symmetric and positive semi-definite, it contains n eigenvalues $\{\lambda_0, \dots, \lambda_{n-1}\}$ that are real and non negative with $\lambda_0 = 0$ and $\lambda_{i+1} \geq \lambda_i$. This set of eigenvalues is called the *spectrum* of L . Interestingly, the n eigenvectors associated to the eigenvalues can be seen as indicator vectors that one can use to cut the graph. Also, the magnitude of their associated eigenvalues is related to the cost of their cut [33]. As such, the eigenvectors associated to the lowest eigenvalues are those associated to the partitions of minimum cost.

3.3. Inter-class adjacency matrix

We formulate our c-measure within the spectral clustering framework for which each node is a class index. In our case, W and L are $K \times K$ matrices where K is the total number of classes. As such, the weight $w_{i,j}$ is the distance between the likelihood distributions of classes C_i and C_j . Thus, a simple dataset for which each pair of classes has little overlap would produce a sparse Laplacian matrix L whose spectrum contains small eigenvalues. On the other hand, a more complex dataset with stronger class overlap would lead to a spectrum with larger eigenvalues.

Since the similarity matrix \mathcal{S} was obtained with a Monte-Carlo approximation of the Jebara kernel, it is not symmetric and thus cannot be used as an adjacency matrix W . Instead, we consider each column \mathcal{S}_i as a signature vector of each class i so two classes with similar likelihood distributions would also have a similar signature vector \mathcal{S}_i and vice versa. We then compute W following the Bray-Curtis

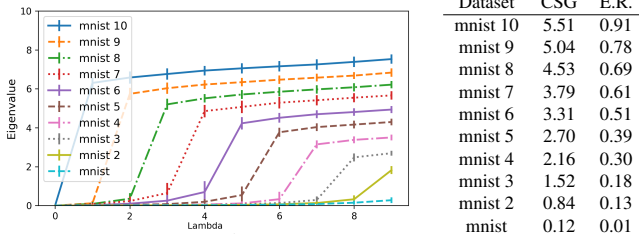


Figure 1: [Left] Spectrum of ten noisy versions of MNIST and [right] our CSG c-measure with the error rate (E.R.) of an AlexNet CNN (figure best viewed in color).

distance function [16]:

$$w_{ij} = 1 - \frac{\sum_k^K |\mathcal{S}_{ik} - \mathcal{S}_{jk}|}{\sum_k^K |\mathcal{S}_{ik} + \mathcal{S}_{jk}|}. \quad (5)$$

This equation implies that $w_{ij} = 0$ when the distributions of classes i and j do not overlap and $w_{ij} = 1$ when the distributions are identical.

3.4. Runtime improvement

Computing the adjacency matrix W with the Bray-Curtis function as well as the Monte-Carlo method (Eq.(3)) is 40 times faster than with a naive implementation (Eq.(2)) for a $K=10$ class problem. This explains why our method is fast and gets good results even with a small number of samples. We came to that number as follows.

First, let us mention that the most computationally intensive operation is the point-wise estimation of a probabilistic distribution function (pdf) $P(\phi(x)|C)$. Since computing Eq.(2) requires M estimations of $P(\phi(x)|C_j)$, the $K \times K$ similarity matrix S requires a total of $K^2 \times M$ pdf estimations, where K is the number of classes and M the number of samples. Also, since Eq.(5) requires no additional pdf estimation, our method requires a total of $K^2 \times M$ pdf estimations to compute the adjacency matrix W .

However, since the Bray-Curtis distance function combines two R^K vectors S_i and S_j , it incorporates the statistics of $2 \times K \times M$ samples at each entry w_{ij} of W . If the same number of samples were to be used by a naive implementation, i.e. that w_{ij} was to be computed with $2 \times K \times M$ samples and Eq.(2), the computation of W would require a total of $4 \times K^3 \times M$ pdf estimations, i.e. $4 \times K$ more pdf estimations than for our method. From there we conclude that our method is 40 times faster than a naive implementation when $K = 10$. Please note that these findings are in line with empirical results.

3.5. The CSG complexity measure

As mentioned before, a dataset with a low eigenvalue spectrum indicates a low inter-class overlap and thus easily separable classes. To illustrate this, we put in Fig. 1

the spectrum of the MNIST dataset (the bottommost cyan curve) which we obtained by processing raw images. Since MNIST contains 10 classes, its spectrum contains 10 eigenvalues. Being a simple dataset, MNIST's spectrum contains mostly near-zero values. We then randomly swap elements between classes to force their distribution to strongly overlap, making this noisy version of MNIST more complex. We first swap elements between two classes (MNIST 2), then between three classes (MNIST 3) all the way to 10 classes (MNIST 10). As one can see, these noisy versions of MNIST lead to a gentle progression of the spectrum profiles. The more entangled classes are, the larger the eigenvalues are. Also, the sooner a strong spectrum gradient occurs ($\lambda_{i+1} - \lambda_i$) the more difficult the dataset is (this gradient discontinuity is also called the *eigengap* in the spectral clustering literature [42]).

The overall complexity of the dataset is thus related to the area under the spectrum curve as well as the position of the eigengap. To account for both observations, we first normalize the eigengap by its horizontal position:

$$\Delta \tilde{\lambda}_i = \frac{\lambda_{i+1} - \lambda_i}{K - i}. \quad (6)$$

The normalization by $K - i$ is at the core of our metric. Depending on where the largest eigengap occurs, its maximum value can only be of $K - i$. The difficulty of cutting the graph is thus related to the position of the largest eigengap. Our c-measure (the *cumulative spectral gradient* (CSG)) is the cumulative maximum (cummax) of the normalized eigengaps :

$$CSG = \sum_i \text{cummax}(\Delta \tilde{\lambda})_i. \quad (7)$$

With a cummax, between two spectrums with the same area under the curve, our CSG measure will be larger for the one with the left-most eigengap. The CSG values for the noisy MNIST datasets are shown on the right of Fig. 1 along side with the test error rate obtained with an AlexNet CNN [24]. As can be seen, our CSG c-measure is heavily correlated to the complexity of the datasets.

Our method is summarized in Algorithm 1.

4. Results

4.1. Embeddings

As mentioned before, the input images x are projected to an embedding space with a function $\phi(x)$. In this paper, we tested four projection functions :

1. *Raw*; the identity function $\phi(x) = x$;
2. *t-SNE*; the t-SNE function [30] which projects the raw input images down to a 2D space;

Algorithm 1: The CSG c-measure algorithm.**Data:** Dataset= $\{(\phi(x_1), t_1), \dots, (\phi(x_N), t_N)\}$ **Args:** M, k**Result:** CSG scoreCompute inter-class similarity matrix \mathcal{S} with Eq.(3)and (4) \forall pair of classes C_i, C_j Compute W (Eq.(5)) $L \leftarrow D - W$ $\{\lambda_1, \dots, \lambda_K\} \leftarrow \text{EigenValues}(L)$

Compute CSG (Eq.(7))

return CSG

3. CNN_{AE} ; the embedding of a 9-layer CNN-Autoencoder trained for 100 epochs;
4. CNN_{AE} t -SNE; the t-SNE function applied to the embedding of the CNN-autoencoder.

4.2. Datasets

In order to gauge performance of our method, we used several image classification datasets of various difficulty levels. Of those datasets, six contain 10 classes, one contains 11 classes and three contain two classes. These datasets are summarized in Table 1 and sorted according to the test error rate (E.R.) obtained with an AlexNet CNN [24]. Note that we replaced the AlexNet local response norm with a batch-norm [20], trained it for 500 epochs on each dataset with a batch size of 32 and the SGD optimizer with the same parameters than in the original paper but without data augmentation. We used Keras [10], Tensorflow [1] and an Nvidia Titan X GPU.

The datasets are the well-known MNIST [25] and CIFAR10 [23]. There is also notMNIST [8], a synthetic dataset of 18,724 letters made of unconventional fonts, and the Street View House Numbers (SVHN) dataset [35], one of the most challenging digit classification datasets with 73,257 images of low resolution street numbers. We also use MioTCD [29], a large dataset of 648,959 vehicles pictured by traffic cameras with varying orientation angles, resolution, time of the day and weather conditions. STL-10 [11] is a 10-class dataset similar to CIFAR-10 but with larger images (96×96 instead of 32×32) and fewer training samples (5,000 instead of 50,000). SeeFood [4] is a two-class dataset (*Hot-dog* vs *No Hot-dog*) with 498 samples derived from the Food-101 dataset [6]. We also use the well-known Inria pedestrian dataset [13] containing 38,634 RGB images of pedestrians or not, and Pulmo-X [21], a two-class pulmonary chest X-Ray dataset for tuberculosis detection containing 662 images. Finally, CompCars [44] is a dataset containing 1,716 car categories of different makes and models. For our experiments, we selected the 10 makes with the highest count and resized the images to 128×128 ,

Datasets	E.R.	K	N	Content
MNIST	0.01	10	50k	Hand written digits
MIO-TCD	0.03	11	649k	Traffic images
notMNIST	0.05	10	18.7k	Printed digits
SVHN	0.08	10	73.3k	Printed digits
Inria	0.10	2	3.6k	Pedestrians
CIFAR10	0.12	10	50k	Various real images
Pulmo-X	0.23	2	662	Pulmonary X-Rays
SeeFood	0.38	2	500	Images of food
STL-10	0.68	10	5k	Various real images
CompCars	0.70	10	6k	Pictures of cars

Table 1: Datasets used to validate our method with the test error rate (E.R.) of an AlexNet CNN [24], the number of classes K , the training set size N and a short summary.

giving us 500 samples per class.

We followed the evaluation methodology specific to each dataset, i.e. we trained and tested the methods on the training and testing set provided with the datasets. For the two datasets without pre-determined train/test split (notMNIST and Pulmo-X) we made a 80-20 Train/Test split and kept the same class proportion.

4.2.1 Hyper-parameters

Our algorithm has two main hyper-parameters: M the number of samples per class used by the Monte Carlo method in Eq.(3) and k the number of neighbors to compute the likelihood distribution of each class in Eq.(4). In Table 2, we show the Pearson correlation score between our c-measure with the CNN_{AE} t -SNE embedding and the error rate of AlexNet on the six 10-class datasets (upper table) as well as the average processing time of our Algo 1 (lower table). As one can see, the choice for k and M has little impact on the quality of the results (except for when M is very small). Also, while the runtime scales almost linearly with M , our method is still fast with timings below 3 seconds, even with $M = 400$ samples per class. This shows that our method does not require a careful adjustment of its hyper-parameters. We found this as well for the other embeddings we tested. As such, we will use $M = 100$ and $k = 3$ for the remainder of this section.

4.3. Experimental results

4.3.1 Comparison with other c-measures

We compared our method to the most widely implemented c-measures, i.e. those by Ho and Basu [18]. We used the C++ DCol library provided by the authors [37] and processed the six 10-class datasets. We thus followed the original methodology provided by the authors which implies no embedding. In addition, we tested two other metrics derived from the spectral theory: the maximum eigenvalue ($\max \Lambda$) and the area under the curve (AUC). These methods are

		k						
		1	3	5	7	9	11	
M	2	0.81	0.79	0.80	0.75	0.76	0.73	Pearson Corr.
	50	0.97	0.97	0.97	0.96	0.96	0.97	
	100	0.97	0.97	0.98	0.98	0.97	0.97	
	200	0.98	0.98	0.98	0.98	0.98	0.98	
	300	0.97	0.97	0.97	0.98	0.97	0.98	
	400	0.97	0.97	0.97	0.97	0.97	0.97	
M	2	0.02	0.02	0.02	0.02	0.02	0.02	Timing (s)
	50	0.30	0.30	0.29	0.30	0.29	0.27	
	100	0.60	0.61	0.60	0.60	0.60	0.60	
	200	1.22	1.21	1.23	1.19	1.20	1.22	
	300	1.82	1.82	1.82	1.78	1.83	1.79	
	400	2.42	2.38	2.42	2.41	2.39	2.39	

Table 2: Correlation values [upper table] and average processing times of Algo 1 in seconds [lower table] for various combinations of hyperparameters M and k .

c-measure	Corr.	p-value	Time (s)
N4	0.069	0.896	3,744
F3	0.267	0.610	3,924
F1	0.501	0.311	72
F2	0.422	0.405	72
T1	0.357	0.487	36,108
T2	0.636	0.175	72
N2	0.652	0.161	36,180
F4	0.725	0.103	3,644
N1	0.741	0.092	17,748
N3	0.773	0.072	36,216
$\max_i \lambda_i$ CNN _{AE} t-SNE	0.88	0.02	0.3 (18,900)
$\sum_i \lambda_i$ CNN _{AE} t-SNE	0.94	≤ 0.01	0.3 (18,900)
CSG Raw	0.696	.125	50 (NA)
CSG CNN _{AE}	0.823	.044	3.6 (13,300)
CSG t-SNE	0.903	.014	0.7 (6,084)
CSG CNN _{AE} t-SNE	0.968	≤ 0.01	0.3 (18,900)

Table 3: Correlation between the accuracy of AlexNet on 6 datasets and 10 c-measures by Ho-Basu [18] and ours methods with four embeddings, the associated p-value and processing time alongside the time to train the autoencoder in the parentheses (measured on CIFAR10).

known in the literature as being related to the similarity between nodes [40]. These turn out to perform worse than our CSG metric. Results are reported in Table 3 together with our method with four embeddings.

The first column contains the Pearson correlation score between the error rate by an AlexNet CNN and each c-measure. As one can see, our method with the CNN_{AE}, t-SNE and CNN_{AE} t-SNE embeddings have a better correlation than any of the existing c-measures with a p-value below the 0.05 bar. The best embedding is CNN_{AE} t-SNE with a significance p-value below 0.01. To illustrate how this embedding correlates with the dataset complexity, we put in

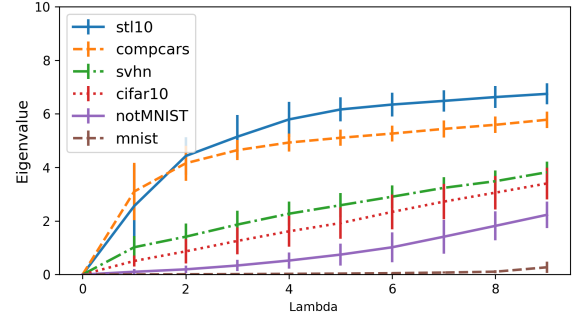


Figure 2: Laplacian spectrum for the 10-class datasets.

Fig. 2 its Laplacian spectrum for the six 10-class datasets. As can be seen, the spectrum plots grow smoothly from the simplest dataset (MNIST), to slightly more complex datasets (notMNIST, CIFAR10 and SVHN) all the way to the most complex datasets (STL-10 and CompCars). Note that we will use the CNN_{AE} t-SNE embedding for the remaining of this section.

As for processing time, our method is faster than the best c-measures F4, N1 and N3. Note the value on the left is the time to execute Algo. 1 whereas the value in parenthesis is the processing time to train a CNN_{AE} and/or run t-SNE. Although that processing time is large (more than one hour) it is much faster than the previous best method N3.¹ The performance of the t-SNE embedding is due to the fact that while t-SNE does not change the nearest neighbours, it does not preserve long-range distances which results in a less convoluted low-dimensional representation. In consequence, the approximation of the volume is better in this representation.

In Table 4, we provide our CSG c-measure with the test error rate of three CNN models as well as their Pearson correlation and p-value. As can be seen, our c-measure correlates well not only with AlexNet, but also with more recent ResNet-50 [17] and XceptionNet [9]. Also, our correlation and p-value with CNN error rates is significantly better than the best existing c-measures [18] even when using our CNN t-SNE embedding. In fact, using embeddings seem detrimental to the overall performance of the existing c-measures. Results on all the existing c-measures with all the embeddings are available in the supplementary material.

We also tested our method on two-class image classification problems. We used the Inria, SeeFood, and PulmoX datasets as well as the deer-dog subset of CIFAR10. Results reported in Table 5 show that our method correlates well with the CNN models, especially AlexNet. Our correlation scores are also better than those of the best c-measure of Ho-Basu (although by a slight margin) although it was specifically designed for two-class problems.

¹The timings were computed on CIFAR10 using a Intel® Xeon® CPU E5-1620 and a NVIDIA TITAN X.

		Error rate		
Datasets	CSG	AlexNet	ResNet-50	Xception
CompCars	2.93	0.70	0.88	0.86
STL-10	3.07	0.69	0.63	0.69
CIFAR10	1.00	0.18	0.19	0.06
SVHN	1.15	0.08	0.07	0.03
notMNIST	0.72	0.05	0.04	0.03
MNIST	0.11	0.01	0.05	0.01
Method		Pearson correlation		
N3/CNN Corr		0.773	0.727	0.681
N3/CNN p-val		0.054	0.102	0.136
N3 CNN t-SNE/CNN Corr		0.837	0.765	0.837
N3 CNN t-SNE/CNN p-val		0.063	0.124	0.144
CSG CNN t-SNE/CNN Corr		0.968	0.935	0.951
CSG CNN t-SNE/CNN p-val		0.01	0.006	0.003

Table 4: [Top] CSG c-measure alongside with test error rates for 3 CNN models on six datasets.[Bottom] Pearson correlation and p-value between the CNN error rates and the N2 and N3 Ho-Basu our CSG c-measure [18] and CNN.

Datasets	CSG	AlexNet	ResNet-50	Xception
SeeFood	0.95	0.38	0.34	0.21
PulmoX	0.55	0.23	0.16	0.11
deer-dog	0.39	0.20	0.02	0.02
Inria	0.32	0.10	0.07	0.03
N3 Ho-Basu/CNN Corr		0.976	0.852	0.862
N3 Ho-Basu/CNN p-val		0.01	0.148	0.138
CSG/CNN Corr		0.995	0.860	0.887
CSG/CNN p-val		0.006	0.130	0.113

Table 5: [Top] CSG c-measure alongside with test error rates for 3 CNN models on four 10-class datasets.[Bottom] Pearson correlation and p-value between ours methods and CNN and between the N3 c-measure [18] and CNN.

4.3.2 Dataset reduction

Dataset reduction (also known as *instance selection* [26]) consists in reducing as much as possible the number of elements in a dataset without losing trained CNN accuracy. One way of doing so is by iteratively removing elements from the dataset up to a point where the CSG measure increases sharply.

We first tested our method on the MIO-TCD dataset [29], a large dataset used for a 2017 CVPR challenge and for which CNN methods got accuracies of up to 98%. Such high accuracies suggest that the dataset is overcomplete and could be reduced without affecting much the CNN accuracy. Results for various reduction ratios are shown in Figure 3. As one can see, the CSG (red dots) stays roughly unchanged for reduction ratios below 80% but then increases sharply after that. This is inline with the AlexNet test error

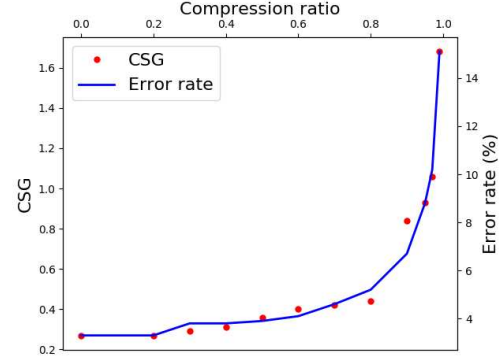


Figure 3: Our c-measure and AlexNet accuracy obtained while reducing the size of the MioTCD dataset.

rate (blue line) although it took less than 5 minutes to produce the CSG measures and 5 days the AlexNet results. We used the same CNN_{AE} embedding for all ratios. We got a Pearson correlation of 0.956 between our CSG dots and the error rate values shown in Figure 3.

Dataset reduction can also be used to measure the similarity between two datasets with very different sizes like CIFAR10 (5,000 training samples per class) and STL-10 (500 training samples per class). While these datasets have visually similar content, the CNN error rates on it are very different (see Table 1). To measure the true distance between those datasets, we progressively reduced the number of samples for each CIFAR10 class to reach that of STL-10. The results in Table 6 show the close bound between our metric and the number of samples in the dataset. With only 500 samples, CIFAR10 gets a CSG score and a CNN accuracy similar but not identical to that of STL-10. This shows that the datasets are similar but not identical, probably due to the fact that the CIFAR10 *Frog* class has been replaced by a *Monkey* class in STL-10 (supplementary material). Here again, it took roughly one minute to produce the CSG scores (after having trained the embedding) and 4 days for the CNN error rates.

4.4. Confusion matrix

While our c-measure can gauge the overall complexity of a dataset with a single measure that correlates with CNN accuracies, we can also use the similarity matrix W (Eq.(5)) to analyze the inter-class distances. As such, one can use a dissimilarity matrix $S = 1 - W$ to visualize the dataset in 2D via an algorithm such as multidimensional scaling (MDS) [5]. This results into plots such as those in Fig. 4. While the classes of MNIST are all well separated, the CIFAR10 plots show that the *cat* and *dog* classes as well as *deer* and *bird* are close to each other, probably due to similar contexts. As for MioTCD, the *bicycle*, *motorcycle* and *pedestrian* classes are in the same vicinity, mainly because

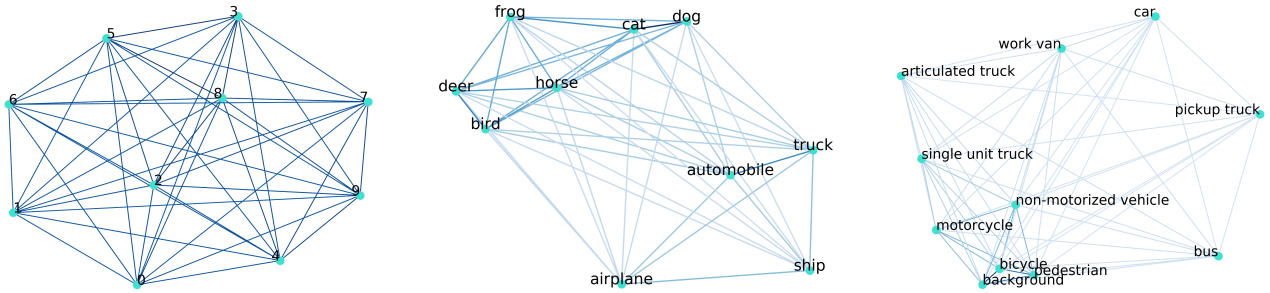


Figure 4: 2D plots of our W matrix for MNIST, CIFAR10 and MioTCD.

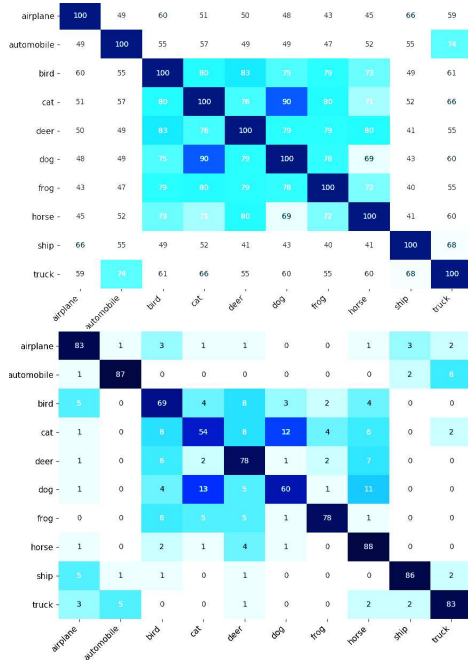


Figure 5: [Top] our W matrix and [Bottom] AlexNet's confusion matrix for CIFAR10.

of their small image resolution, they often contain more compression artifacts and hence be less feature-rich than other classes, making them confusing with featureless background.

As shown in Fig. 5, the W matrix strongly correlates to a real confusion matrix (here AlexNet). Here again, *cat* and *dog* as well as *deer* and *bird* are easily confused.

5. Conclusion

In this work, we proposed a novel complexity measure designed for image classification problems called the cumulative spectral gradient (CSG) which is more accurate and faster than previous methods. We showed that our metric has many uses such as instance selection and class disentanglement. We also showed that the CSG closely matches the accuracy achievable by standard CNN architectures, an important feature when assessing an image dataset.

Dataset	CSG	Error rate
CIFAR10	1.10	0.18
CIFAR10 reduced=4500	1.10	0.19
CIFAR10 reduced=3500	1.26	0.20
CIFAR10 reduced=2500	1.44	0.24
CIFAR10 reduced=1500	2.16	0.28
CIFAR10 reduced=500	2.59	0.42
STL-10	3.16	0.68

Table 6: Effect of reducing the number of samples per class for CIFAR10 on our CSG metric and the AlexNet test error rate.

A future direction of our research would be to determine a procedure to compare the relative complexity of classification problems with different number of classes. The analysis of random subsets of classes could be used as a common representation. Another important direction would be to generalize our method to segmentation and localization problems. As of now, it is not clear how these problems can be described by spectral clustering.

Another future work would be to incorporate our similarity matrix W in the optimization process of a neural network to minimize the interclass divergence. It is our intuition that the *a priori* knowledge of the interclass overlap could be used to force the optimizer to further separate entangled classes, a bit like the triplet loss does. Finally, our metric is not restricted to image classification datasets and could be used in other areas of machine learning such as speech recognition and natural language processing (NLP). These fields already use state-of-the-art embeddings such as Word2Vec [32] and would thus naturally fall into our CSG framework.

References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.
- [2] N. Anwar, G. Jones, and S. Ganesh. Measurement of data complexity for classification problems with unbalanced data. *Journal Stat. Anal. Data Mining*, 7(3):196–211, 2014.

- [3] R. Baumgartner and R. Somorjai. Data complexity assessment in undersampled classification of high-dimensional biomedical data. *Pat. Rec. Letters*, 27(12), 2006.
- [4] D. Becker. Hot dog - not hot dog. [Online]. Available: [urlhttps://www.kaggle.com/dansbecker/hot-dog-not-hot-dog](https://www.kaggle.com/dansbecker/hot-dog-not-hot-dog), 2017.
- [5] I. Borg and P. Groenen. *Modern multidimensional scaling: theory and applications*, volume 40. Wiley Online Library, 2003.
- [6] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *Proc. ECCV*, 2014.
- [7] A. L. Brun, A. S. Britto Jr, L. S. Oliveira, F. Enembreck, and R. Sabourin. A framework for dynamic classifier selection oriented by the classification problem difficulty. *Pat. Rec.*, 76:175–190, 2018.
- [8] Y. Bulatov. notmnist. [Online]. Available: www.kaggle.com/lubaroli/notmnist, 2017.
- [9] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proc. CVPR*, pages 1800–1807, 2017.
- [10] F. Chollet et al. Keras. [Online] Available: <https://github.com/fchollet/keras>, 2015.
- [11] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proc. AIS-TATS*, pages 215–223, 2011.
- [12] L. Cummins and D. Bridge. Choosing a case base maintenance algorithm using a meta-case base. In *Research and Development in Intelligent Systems XXVIII*, pages 167–180. Springer, 2011.
- [13] N. Dalal and B. Triggs. Inria person dataset. [Online]. Available: <http://pascal.inrialpes.fr/data/human>, 2005.
- [14] R. Duin and E. Pekalska. *Object representation, sample size and dataset complexity*. Springer, 2006.
- [15] L. Garcia, A. de Carvalho, and A. Lorena. Effect of label noise in the complexity of classification problems. *Neuro-computing*, 160:108–119, 2015.
- [16] M. Greenacre and R. Primicerio. Measures of distance between samples: non-euclidean. *Multivariate analysis of ecological data*, pages 5–1, 2013.
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.
- [18] T. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE trans on PAMI*, 24(3):289–300, 2002.
- [19] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. Springer, 2012.
- [20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, pages 448–456, 2015.
- [21] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, and G. Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.
- [22] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *JMLR*, 5(7):819–844, 2004.
- [23] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Tech.Report*, 2009.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, 2012.
- [25] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *AT&T Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- [26] E. Leyva, A. González, and R. Pérez. A set of complexity measures designed for applying meta-learning to instance selection. *IEEE Trans on Knowledge and Data Eng.*, 27(2):354–367, 2015.
- [27] C. Li, H. Farkhoor, R. Liu, and J. Yosinski. Measuring the intrinsic dimension of objective landscapes. In *Proc. ICLR*, 2018.
- [28] A. Lorena, L. Garcia, J. Lehmann, M. Souto, and T. Ho. How Complex is your classification problem? A survey on measuring classification complexity. *ArXiv –1808.03591v1*, 2018.
- [29] Z. Luo, F. B. Charron, C. Lemaire, J. Konrad, S. Li, A. Mishra, A. Achkar, J. Eichel, and P.-M. Jodoin. Miotcd: A new benchmark dataset for vehicle classification and localization. *In press at IEEE TIP*, 2018.
- [30] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9(11):2579–2605, 2008.
- [31] R. G. Mantovani, A. L. Rossi, J. Vanschoren, B. Bischl, and A. C. Carvalho. To tune or not to tune: recommending when to adjust svm hyper-parameters via meta-learning. In *Proc. IJCNN*, 2015.
- [32] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, 2013.
- [33] B. Mohar. *Some applications of Laplace eigenvalues of graphs*, pages 225–275. Springer, 1997.
- [34] G. Morais and R. C. Prati. Complex network measures for data set characterization. In *Proc. BRACIS*, pages 12–18. IEEE, 2013.
- [35] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. In *Proc. NIPS*, page 5, 2011.
- [36] E. Nowakowska, J. Koronacki, and S. Lipovetsky. Tractable Measure of Component Overlap for Gaussian Mixture Models. *ArXiv – 1407.7172*, 2014.
- [37] A. Orriols-Puig, N. Macia, and T. K. Ho. Documentation for the data complexity library in c++. *Universitat Ramon Llull, La Salle*, 196, 2010.
- [38] J. Saez, J. Luengo, and F. Herrera. Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification. *Pat. Rec.*, 46(1):355–364, 2013.
- [39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ACPR*, 2015.
- [40] S. Sorgun and S. Buyukkose. Bounds for the largest laplacian eigenvalues of weighted graphs. *International Journal of Combinatorics*, 2013, 2013.
- [41] J. Sotoca, R. Mollineda, and J. Sanchez. A meta-learning framework for pattern classification by means of data complexity measures. *Int. art. rev. iber de IA*, 10(29):31–38, 2006.

- [42] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- [43] L. Wang, Y. Zhang, and J. Feng. On the euclidean distance of images. *IEEE Trans on PAMI*, 27(8), 2005.
- [44] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proc. CVPR*, pages 3973–3981, 2015.
- [45] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires re-thinking generalization. In *Proc. ICLR*, 2017.