The Choice of Feature Representation in Small-Scale MobileNet-Based Imbalanced Image Recognition

Michał Koziarski[1,2], Bogusław Cyganek[1,2] and Kazimierz Wiatr[1,2]
[1]*AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland*
[2]*Academic Computer Center Cyfronet AGH, Ul. Nawojki 11, 30-950 Kraków, Poland*

Keywords: Imbalanced Data Classification, Small-Scale Image Recognition, Convolutional Neural Networks, Feature Representation, MobileNet.

Abstract: Data imbalance remains one of the most wide-spread challenges in the contemporary machine learning. Presence of imbalanced data can affect the learning possibility of most traditional classification algorithms. One of the the strategies for handling data imbalance are data-level algorithms that modify the original data distribution. However, despite the amount of existing methods, most are ill-suited for handling image data. One of the possible solutions to this problem is using alternative feature representations, such as high-level features extracted from convolutional layers of a neural network. In this paper we experimentally evaluate the possibility of using both the high-level features, as well as the original image representation, on several popular benchmark datasets with artificially introduced data imbalance. We examine the impact of different data-level algorithms on both strategies, and base the classification on MobileNet neural architecture. Achieved results indicate that despite their theoretical advantages, high-level features extracted from a pretrained neural network result in a worse performance than end-to-end image classification.

1 INTRODUCTION

The problem of data imbalance remains one of the most wide-spread challenges in the contemporary machine learning (Krawczyk, 2016). It occurs whenever the number of observations in one of the classes (*majority class*) is significantly higher than the number of observations in one of the other classes (*minority class*). Traditional learning algorithms are ill-equipped for handling significant data imbalance, displaying bias towards recognizing objects as belonging to the majority class, at the expense of performance on the minority classes. Despite its prevalence in traditional machine learning, the impact of data imbalance on the problem of image recognition only recently started gaining attention of the research community (Buda et al., 2018). At the same time, inherent characteristics of the image data, such as its high dimensionality and spacial properties, pose a challenge for the traditional methods of dealing with data imbalance (Lusa et al., 2012). This problem becomes further pronounced in the small-scale image recognition task, where the amount of data is relatively small (Japkowicz and Stephen, 2002), at least compared to the amount required to train contemporary convolutional neural networks.

In this paper we examine the possibility of using different feature representations in the small-scale imbalanced image recognition problem. We focus on two most prevalent representations: original image data, used directly to train convolutional neural networks, and high-level features extracted from the top layers of a pretrained network, which can be further used to train a traditional classification algorithm. We experimentally evaluate the performance of both feature representations on several popular image recognition benchmarks with artificially introduced data imbalance, and combine them with a number of state-of-the-art data-level strategies of handling data imbalance, not designed with the image data in mind. In the remainder of this paper we discuss the problem of imbalanced image recognition, presents details of the conducted experimental study, and discuss its outcomes.

Koziarski, M., Cyganek, B. and Wiatr, K.
The Choice of Feature Representation in Small-Scale MobileNet-Based Imbalanced Image Recognition.
DOI: 10.5220/0009357206330638
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 633-638
ISBN: 978-989-758-402-2; ISSN: 2184-4321

2 IMBALANCED IMAGE RECOGNITION PROBLEM

Methods for handling data imbalance can be divided into techniques that modify the original data distribution (*data-level approaches*), either by removing existing observations (*undersampling*) or generating new observations (*oversampling*), and modifications to the existing learning algorithms that account for disproportion between the number of observations between the classes (*algorithm-level approaches*). In the context of image recognition both data-level approaches pose unique challenges that have to be addressed to successfully create a convolutional neural network. In the case of undersampling the main issue lies in the fact that convolutional neural networks tend to require large quantities of data during training, and by performing undersampling we artificially decrease the amount of available data. In particular in the case of highly imbalanced, small-scale datasets, for which the amount of data is limited to begin with, performing undersampling might not be feasible. On the other hand, in the case of oversampling the main problem is that of handling novel data creation. The most straightforward approach is to simply duplicate the existing observations from the minority class. This, however, can lead to overfitting of the classification algorithm, which was previously demonstrated in the context of decision trees (Chawla et al., 2002). The alternative is to generate synthetic observations based on the existing data: by far the most prevalent strategy of that type is the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). SMOTE is based on the idea of interpolation between a given minority object and one of its nearest minority neighbors. However, this process was designed with a traditional data in mind, and is ill-suited for image data, which was demonstrated in Figure 1. As can be seen, the interpolated image loses the spatial properties of its prototypes, producing impossible in practice data that can potentially negatively affect the process of convolutional neural networks training. Since SMOTE is the cornerstone for most of the contemporary methods for handling the data imbalance, the majority of the existing oversampling algorithms will display negative characteristics when applied to the image data.

An alternative to applying the data-level strategies for handling data imbalance directly to the images is using different feature representation. Since in the recent years convolutional neural networks emerged as a de facto standard in the image recognition domain, an obvious choice is extraction of high-level image representations directly from a convolutional neural network. This is usually done by capturing the outputs of one of the last convolutional layers of a network. Produced features, depending on the input image size and the architecture of a particular network, might also have smaller dimensionality than the input images. Furthermore, such extracted features are more suitable for data interpolation, since their geometrical properties differ from the original image data. Finally, particularly in the case of small-scale image recognition problem, extracting the features from an existing network opens up the possibility of using pretraining: training the network that will be used during the feature extraction on another dataset, possibly consisting of a significantly larger amount of data. However, despite all of the aforementioned possible advantages, in practice it is not clear if, and if so to what extent, using high-level features extracted from a convolutional neural network is feasible in practice, especially in presence of data imbalance. Even more so, it is not clear what effect using different feature representations will have on traditional data-level approaches to handling data imbalance.

3 EXPERIMENTAL STUDY

To evaluate the feasibility of using different feature representations in the imbalanced image recognition task we conducted an experimental analysis on several image recognition benchmarks. Specifically, we considered the possibility of using features extracted from pretrained convolutional neural network as an alternative to the original image representation. Furthermore, we examined the feasibility of different data resampling techniques in both settings. In the remainder of this section we give a detailed description of the set-up of the conducted experiments, and we discuss the achieved results.

Datasets. Despite the prevalence of data imbalance in real-life image recognition problems, most contemporary benchmark datasets provide data with a balanced class distribution. To the best of our knowledge, as of yet there exist no dedicated benchmark for imbalanced image recognition. The alleviate this issue we artificially introduced data imbalance to three popular benchmark datasets: MNIST (LeCun et al., 1998), consisting of 60000 train and 10000 test grayscale images of handwritten digits with dimensionality of 28×28 pixels, CIFAR-10 (Krizhevsky et al., 2009), consisting of 50000 train and 10000 test 32×32 color images of objects such as airplanes, birds and trucks, and STL-10 (Coates et al., 2011),

Figure 1: An example of SMOTE interpolation applied to the image data.

consisting of 5000 train and 8000 test 96×96 color images, with classes similar to CIFAR-10. Each of the datasets consists of 10 classes, which is crucial for the problem of imbalanced data resampling, where the multi-class nature of the data further complicates the relations between the classes. During the described experiments we consolidated the image format across the datasets by resizing them to the dimensionality equal to 64×64 and, in case of MNIST, replicating the grayscale images to obtain 3 color channels.

To introduce data imbalance we performed a following procedure: first of all, we randomly ordered all of the classes. Secondly, for a specified imbalance ratio (IR), the total number of observations per class equal to n, and the number of classes $M = 10$, we calculated the ratio of observations for i-th class as

$$r_i = 1 + \frac{\mathrm{IR} - 1}{M - 1} \cdot (i - 1), \tag{1}$$

and the desired number of observations for i-th class as

$$n_i = \frac{1}{r_i} \cdot n, \tag{2}$$

with $i \in \{1, 2, ..., 10\}$. Finally, for each of the classes we undersampled the observations up to the point of achieving the specified number of observations. Both train and test partitions used the same ratios of observations for individual classes, and the number of observations was scaled depending on the size of respective partition. This procedure was repeated 10 times for each of the datasets, creating folds in which different classes were designated as either the majority or the minority. Final results were averaged across the folds.

Feature Representations. In the conducted experiments we considered two different feature representations commonly used in the image recognition task. First of all, original image representation, more specifically 64×64 color images. Convolutional neural networks take advantage of the spatial properties of the image data and do not require any feature extraction. However, image data is ill-suited for advanced oversampling techniques, such as SMOTE, which natively supports only one-dimensional feature representation. To mitigate this problem, when applying resampling on the image data it was vectorized prior to the resampling, and reverted to the original multi-dimensional format afterwards.

Secondly, we considered the possibility of using high-level features extracted from a pretrained convolutional neural network. Extracting features from a previously trained model is a common practice that enables the possibility of using different classification algorithms. It is especially useful when relatively small amount of data is available, since convolutional neural networks typically require large quantities of data to achieve a satisfactory performance. Finally, in theory higher level feature representation can be useful for data resampling. For instance, SMOTE algorithm uses data interpolation between two nearby observations to generate new synthetic instances. This can produce unrealistic observations when dealing with image data, posing an additional difficulty during training of convolutional neural networks, which were designed for dealing with spatially plausible data. In this paper we used the features extracted from the last convolutional layer of MobileNet (Howard et al., 2017) network, flattened into a 4096-dimensional vectors. Used network was previously trained on the ImageNet (Deng et al., 2009) dataset, and was not finetuned to the specific benchmark datasets.

Classification Algorithms. Depending on the feature representation, we considered two methods of data classification. First of all, for the original image representation we trained a MobileNet. The networks weights were once again initialized with the ones obtained during training on ImageNet dataset. However, during classification the network was afterwards finetuned to each of the considered benchmark datasets. To achieve this, the fully-connected layer of the original model was replaced with a global average

pooling layer, followed by a 1024-dimensional fully-connected layer with a ReLU activation function and an another 10-dimensional fully-connected layer with a softmax activation function, all with a randomly initialized weights. The final model was than trained on the target dataset for 50 epochs, using RMSprop algorithm with learning rate equal to 0.0001 and batch size equal to 32. On the other hand, when using the high-level features extracted from the pretrained MobileNet, the SVM classifier with a RBF kernel and a regularization constant equal to 1.0 was used. The implementations of MobileNet and SVM were taken from, respectively, Keras (Chollet et al., 2015) and scikit-learn (Pedregosa et al., 2011) machine learning libraries.

Resampling Techniques. To reduce the negative impact of data imbalance we considered several data-level algorithms. We investigated the possibility of using both the oversampling and the undersampling algorithms. Specifically, we used the random oversampling (ROS), random undersampling (RUS), SMOTE, SMOTE combined with Edited Nearest Neighbours (S+ENN), and undersampling based on instance hardness threshold (IHT) (Smith et al., 2014). Both the oversampling and the undersampling was performed up to the point of achieving balanced class distributions. The implementations of the algorithms provided in the imbalanced-learn (Lemaître et al., 2017) were used.

Evaluation Metrics. Traditionally, the metric most commonly used to evaluate the performance of image recognition algorithms is classification accuracy. However, it is not a proper choice for evaluating the performance on the data with skewed class distribution, since it assigns weight of the miss-classification of individual classes as proportional to the number of observations that they consist of. Instead, we evaluate the performance of classifiers on multi-class imbalanced data using three dedicated skew-insensitive metrics (Branco et al., 2017): Average Accuracy (AvAcc), Class Balance Accuracy (CBA), and Geometric Average of Recall (MAvG). They are expressed as follows:

$$AvAcc = \frac{\sum_{i=1}^{M} TPR_i}{M}, \tag{3}$$

$$CBA = \frac{\sum_{i=1}^{M} \frac{mat_{i,i}}{max\left(\sum_{i=1}^{M} mat_{i,j}, \sum_{i=1}^{M} mat_{j,i}\right)}}{M}, \tag{4}$$

$$MAvG = \sqrt[M]{\prod_{i=1}^{M} recall_i}, \tag{5}$$

where M is the number of classes, and $mat_{i,j}$ stands for the number of instances of the true class i that were predicted as class j.

Results. In the conducted experiments we considered three benchmark datasets: CIFAR-10, MNIST and STL-10, and three different levels of imbalance: small (IR = 2.0), medium (IR = 5.0) and high (IR = 10.0). Furthermore, for both of the considered feature representations we evaluated the baseline case in which no data resampling was applied, as well as resampling with SMOTE, S+ENN and IHT algorithms. Additionally, for the original image representation we considered ROS and RUS algorithms: we did not use them in combination with the SVM classification of convolutional neural networks features since, contrary to the previously described algorithms, random approaches produce the same results regardless of the choice of feature representation.

We present the observed results in Table 1 and Table 2, with the former containing the results for the individual datasets, and the later the results for the individual imbalanced ratios. Several important conclusions can be made based on the observed results. First of all, using the original image representation to train a convolutional neural network led to achieving a significantly better performance than using the high-level features extracted from a pretrained network, regardless of the choice of dataset, imbalance ratio or performance metric. It is not clear what caused such a discrepancy, but several possible factors can be discussed. To begin with, using a pretrained model can lead to a failure when the training data varies drastically between the pretrained model and the target domain. However, while a possible reason in case of MNIST, both CIFAR-10 and STL-10 share the same classes with the ImageNet: in fact, STL-10 is a subset of ImageNet dataset. Second, more plausible explanation is that pretrained models do not achieve satisfactory performance without additional finetuning. It is not obvious whether that is the case for MobileNets: while it is true that fully-connected layers are not usable when the spatial dimensionality of the input images changes, used feature extraction procedure discarded all of the fully-connected layers are relied solely on the output of the convolutional layers. Still, it is possible that due to the increased size images are interpreted by the network as having low resolution, which was shown to degrade the performance of several neural architectures (Koziarski and Cyganek, 2018). Final, most interesting explanation is that convolutional neural networks are, by default, more resilient to the presence of class imbalance than traditional learning algorithms, such as SVM. This

Table 1: Results for the individual datasets, averaged over the considered imbalanced ratios.

		Original image representation, CNN classifier						CNN features, SVM classifier			
Metric	Dataset	Base	ROS	RUS	SMOTE	S+ENN	IHT	Base	SMOTE	S+ENN	IHT
AvAcc	CIFAR	0.8718	**0.8744**	0.8296	0.8617	0.7320	0.7779	0.7007	0.7164	0.6642	0.6089
	MNIST	0.9924	0.9931	0.9903	**0.9933**	0.9918	0.9694	0.9622	0.9672	0.9619	0.8732
	STL-10	0.7461	**0.7466**	0.6885	0.7397	0.5073	0.6737	0.5103	0.5326	0.5199	0.5139
CBA	CIFAR	0.8544	**0.8584**	0.7693	0.8450	0.6555	0.6868	0.6735	0.6957	0.5406	0.5125
	MNIST	0.9894	0.9906	0.9837	**0.9908**	0.9886	0.9316	0.9549	0.9590	0.9399	0.7776
	STL-10	**0.7139**	0.7130	0.6104	0.7041	0.3696	0.5872	0.4437	0.4729	0.3830	0.4170
MAvG	CIFAR	0.8661	**0.8691**	0.8249	0.8542	0.7210	0.7720	0.6410	0.6946	0.5919	0.5996
	MNIST	0.9923	0.9931	0.9903	**0.9933**	0.9918	0.9690	0.9617	0.9669	0.9616	0.8692
	STL-10	0.7288	**0.7296**	0.6751	0.7187	0.3107	0.6644	0.2156	0.3175	0.1577	0.4782

Table 2: Results for the individual imbalanced ratios (IR), averaged over the considered datasets.

		Original image representation, CNN classifier						CNN features, SVM classifier			
Metric	IR	Base	ROS	RUS	SMOTE	S+ENN	IHT	Base	SMOTE	S+ENN	IHT
AvAcc	2.0	0.8853	**0.8864**	0.8721	0.8823	0.7280	0.8586	0.7622	0.7714	0.7208	0.7325
	5.0	0.8696	**0.8705**	0.8336	0.8651	0.7501	0.8009	0.7167	0.7323	0.7188	0.6607
	10.0	0.8552	**0.8572**	0.8027	0.8474	0.7531	0.7616	0.6943	0.7125	0.7064	0.6029
CBA	2.0	0.8682	**0.8684**	0.8488	0.8644	0.6641	0.8259	0.7300	0.7426	0.6349	0.6740
	5.0	0.8527	**0.8538**	0.7834	0.8463	0.6769	0.7219	0.6814	0.7018	0.6203	0.5516
	10.0	0.8369	**0.8398**	0.7311	0.8292	0.6727	0.6577	0.6607	0.6832	0.6082	0.4813
MAvG	2.0	0.8822	**0.8827**	0.8684	0.8788	0.6425	0.8554	0.7316	0.7509	0.5703	0.7245
	5.0	0.8629	**0.8635**	0.8276	0.8573	0.6861	0.7959	0.5881	0.6530	0.6032	0.6492
	10.0	0.8421	**0.8456**	0.7943	0.8301	0.6948	0.7542	0.4986	0.5751	0.5377	0.5733

hypothesis is partially supported by the results presented in Table 2, where the relative drop of the performance with the increase of imbalance ratio is, in the baseline case, higher for the features extracted from the convolutional neural network than the original image representation. Still, more research would be required to reliably confirm this hypothesis.

Secondly, it is worth noting that the choice of data resampling depends on both the chosen feature representation and the type of objects present in images. Generally, the only technique that improved the performance of classification with original image representation for all of the datasets was random oversampling. Both random undersampling, as well as the more advanced techniques, that is SMOTE, S+ENN and IHT, actually resulted in a deteriorated performance, in some cases significantly. The only exception was MNIST dataset, for which SMOTE actually produced a marginally better results than both the baseline and the random oversampling. This leads to a conclusion that unless the considered data is relatively simple, applying any advanced resampling techniques directly to the images is not advisable. Random oversampling, while improving the performance in all of the cases, did so only marginally, even for high imbalance levels. On the other hand, using SMOTE was a feasible strategy when operating on a high-level features extracted from the convolutional neural network, leading to a greater improvement in performance. However, due to the significantly worse baseline performance, simply using the original image representation without any resampling was still a preferable strategy.

4 CONCLUSIONS

The aim of this paper was examining the possibility of using different feature representations in small-scale imbalanced image recognition task. To this end we experimentally evaluated the performance of MobileNet on three popular image recognition benchmarks: CIFAR-10, MNIST and STL-10, with different levels of introduced data imbalance. We considered two feature representations: original image rep-

resentation and high-level features extracted from a pretrained convolutional neural network, as well as several popular data-level techniques for alleviating the negative impact of data imbalance. Presented results indicate that using the original image representation with simple random oversampling leads to the best results on the considered benchmark datasets. Contrary to the results that could be expected for tabular data, using other resampling techniques usually led to a deteriorated performance. Furthermore, while using SMOTE improved performance for the features extracted from the pretrained network, the overall performance of that approach was still significantly worse than simply using the original image representation. Observed results suggest feasibility of several further research directions: first of all, since applying SMOTE actually produced a better performance for the features extracted from the convolutional neural network, it is possible that proposing a better feature representation would preserve this effect, while improving the overall performance. Such additional feature representations could include: features extracted from a finetuned neural network, earlier layers of the network, different neural architectures, or autoencoders. Secondly, the observed results suggest that convolutional neural networks may be more resilient to the presence of data imbalance than traditional learning algorithms, such as SVM. The observed improvement in performance due to using dedicated data preprocessing algorithm was also relatively smaller than for the SVM. This, if confirmed with further studies, could indicate that either dealing with data imbalance is less pressing problem in the image recognition task, or the data-level strategies are not a suitable approach for solving it.

ACKNOWLEDGEMENTS

This work was supported by the Polish National Science Center under the grant no. 2017/27/N/ST6/01705.

REFERENCES

Branco, P., Torgo, L., and Ribeiro, R. P. (2017). Relevance-based evaluation metrics for multi-class imbalanced domains. In *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part I*, pages 698–710.

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chollet, F. et al. (2015). Keras. https://keras.io.

Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Koziarski, M. and Cyganek, B. (2018). Impact of low resolution on image recognition with deep neural networks: An experimental study. *International Journal of Applied Mathematics and Computer Science*, 28(4).

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563.

Lusa, L. et al. (2012). Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 89–94. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Smith, M. R., Martinez, T., and Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine learning*, 95(2):225–256.