proceedings

Proceedings

Feature Selection in Big Image Datasets [†]

J. Guzmán Figueira-Domínguez [1,*], Verónica Bolón-Canedo [2] and Beatriz Remeseiro [3]

[1] University of A Coruña, 15071 A Coruña, Spain
[2] CITIC, University of A Coruña, 15071 A Coruña, Spain; veronica.bolon@udc.es
[3] University of Oviedo, 33003 Oviedo, Asturias, Spain; bremeseiro@uniovi.es
[*] Correspondence: j.guzman.figueira@udc.es

[†] Presented at the 3rd XoveTIC Conference, A Coruña, Spain, 8–9 October 2020.

Published: 24 August 2020

Abstract: In computer vision, current feature extraction techniques generate high dimensional data. Both convolutional neural networks and traditional approaches like keypoint detectors are used as extractors of high-level features. However, the resulting datasets have grown in the number of features, leading into long training times due to the curse of dimensionality. In this research, some feature selection methods were applied to these image features through big data technologies. Additionally, we analyzed how image resolutions may affect to extracted features and the impact of applying a selection of the most relevant features. Experimental results show that making an important reduction of the extracted features provides classification results similar to those obtained with the full set of features and, in some cases, outperforms the results achieved using broad feature vectors.

Keywords: feature selection; image feature extraction; big data; computer vision

1. Introduction

Image datasets have grown not only in the number of samples, but also in the number of features that describe them. At this point, it could be reasonable to expect that having more features would provide more information and better results. However, this does not happen, due to the so-called *curse of dimensionality* [1]. In this context, feature selection [2] contributes to the scalability of the machine learning algorithms by finding the most relevant properties of the images and decreasing train and prediction times. However, their efficiency drastically diminishes when dataset dimension grows. Hence, applying big data technologies may ease to use larger datasets. This article addresses the impact of feature selection on image classification using different feature extraction methods. Particularly, this research focuses on the use of filter methods for feature selection with big data technologies.

2. Materials and Methods

This work proposes a pipeline for image classification composed of three main steps: image feature extraction, feature selection and classification. On the one hand, the first step has been implemented in a *Python* package using Keras, OpenCV and scikit-image libraries. On the other hand, the next steps were developed in an *Apache Spark* application that contains independent jobs for both steps. Additionally, features extracted have been stored in *Kaggle datasets*.

1. Feature extraction: In this work, image feature extraction was performed in order to transform image datasets into columnar feature datasets. The techniques applied here are—*bag of features* methods based on feature detection algorithms like SIFT [3], SURF [4] and KAZE [5]; *linear binary pattern* (LBP) methods [6]; and *convolutional neural networks* (ConvNets) used as feature extractors through architectures like VGG, ResNet and DenseNet.

2. Feature selection: Feature selection includes a broad family of dimensionality reduction techniques that achieve reduction by removing the irrelevant and redundant features while keeping the original relevant ones. Particularly, filter methods select a subset of the original feature set independently of the induction model used. Accordingly, these filter methods are more likely to be applied in a big data scenario due to advantages related to computational costs [7]. In such framework, this research has driven the feature selection stage using the big data platform *Apache Spark* and some implementations of such filter methods: Spark's *MLlib* [8] implementation of the χ^2 filter selector [9]; Spark's implementation of the *Relief-F* method [10]; and *ITFS* framework [11] implementation for Spark [12].
3. Classification: Not every available classifier in *Spark MLlib* has a multi-class nature. So, the suitable models in *Spark* for this problem are *Decision Trees*, *Random Forests*, *Naive Bayes* and *Multilayer Perceptron* classifiers. Given the results obtained in the experiments, these two last classifiers were used in the results presented in this manuscript.

In order to carry out the experiments of this research, two datasets were employed—the *ImageNet* dataset, currently hosted by the *Kaggle* platform, which contains 1,281,167 hand-labeled images belonging up to 1000 object categories; and the *Tiny Imagenet* dataset, released as a subset of the original ImageNet, containing very low-resolution images from only a 200-class subset.

3. Results

Regarding results from *Tiny Imagenet*, we noticed that accuracy values provided by features extracted using *bag of features* and *LBP* were quite poor. However, results supplied by features extracted using the ConvNets and applying up to 50% of dimensionality reduction with *Relief-F* (0.6451 top-5 accuracy), χ^2 (0.6422) or *mRMR* (0.6382), outperformed results without feature selection (0.6241).

With respect to experiments carried out with *Imagenet* dataset, features extracted through traditional approaches showed better results with these higher resolution images. Experiments from features extracted using *bag of features*, over the KAZE *keypoints detector*, and applying up to 66% of dimensionality reduction with methods like *mRMR* (0.7674 top-5 accuracy), χ^2 (0.7528) or *ReliefF* (0.7442) showed better results than the ones performed without the selection step (0.7425).

Finally, the accuracy results using features pulled out with a ConvNet like *VGG-19* and feature selection methods were presented quite tight compared to the ones achieved by the own VGG-19 (0.7158 top-1 accuracy and 0.8996 top-5 accuracy). Applying a reduction of a 50% with χ^2 (0.6715 top-1 accuracy and 0.8450 top-5 accuracy) or a reduction of 90% through *mRMR* (0.6554 top-1 accuracy and 0.8143 top-5 accuracy), we notice how results, using a multi-layer perceptron as the classifier model, are below the baseline. However, if we compare the results achieved with a naive Bayes classifier, the baseline (0.6143 top-5 accuracy) is eventually outperformed: 0.6482 top-5 accuracy when applying a reduction up to a 66% with the χ^2 method.

4. Discussion and Conclusions

Contrasting differences on experiments done with all the feature extractors, we can observe some clear tendencies. When feature selection is applied to features extracted with classical techniques, results outperform the baseline collected without making dimensionality reduction. On these techniques, salient information about images is shaped into vectors of a chosen size. As shown in results, this representation may be improved through feature selection techniques. However, when feature selection is applied to *deep features* (i.e., features extracted by pre-trained ConvNets), results are slightly below the baseline without feature selection. This may be explained due to the successive *dropout* layers included in *ConvNets*, which help to remove meaningless information over the layers and represent the best high-order features.

In main terms, results show a clear evidence that feature selection performs a positive impact over features extracted from both datasets. Accuracy values collected in most feature subsets are very close to the ones observed without applying dimensionality reduction. And, in some cases, dimensionality reduction techniques help to outperform classification results using all the features provided by *ConvNets* or *bag of features* extractors. Also, we remark that different feature selection methods stand out depending on the required percentage of feature reduction, so *the best feature selection method* simply does not exist.

Funding: This research has been financially supported in part by European Union FEDER funds, by the Spanish Ministerio de Economía y Competitividad (research project PID2019-109238GB), by the Consellería de Industria of the Xunta de Galicia (research project GRC2014/035), and by the Principado de Asturias Regional Government (research project IDI-2018-000176). CITIC as a Research Centre of the Galician University System is financed by the Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) through the ERDF (80%), Operational Programme ERDF Galicia 2014–2020, and the remaining 20% by the Secretaria Xeral de Universidades (ref. ED431G 2019/01).

Acknowledgments: The authors would like to express their gratitude to the CESGA for the provided resources that allowed this research, and the support of NVIDIA Corporation with the donation of Titan Xp GPUs.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bellman, R.E. *Dynamic Programming*; Dover Publications, Inc.: New York, NY, USA, 2003.
2. Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L.A. *Feature Extraction: Foundations and Applications*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 207.
3. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
4. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. SURF: Speeded up robust features. *Comput. Vis. Image Underst.* **2008**, *110*, 346–359.
5. Alcantarilla, P.F.; Bartoli, A.; Davison, A.J. KAZE features. In Proceedings of the European Conference on Computer Vision, Firenze, Italy, 7–13 October 2012; pp. 214–227.
6. Ojala, T.; Pietikainen, M.; Harwood, D. A Comparative Study of Texture Measures with Classification Based on Feature Distributions. *Pattern Recognit.* **1996**, *29*, 51–59.
7. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Feature selection for high-dimensional data. *Prog. Artif. Intell.* **2016**, *5*, 65–75.
8. Meng, X.; Bradley, J.; Yavuz, B.; Sparks, E.; Venkataraman, S.; Liu, D.; Freeman, J.; Tsai, D.; Amde, M.; Owen, S.; et al. Mllib: Machine learning in apache spark. *J. Mach. Learn. Res.* **2016**, *17*, 1235–1241.
9. Barnard, G. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In *Breakthroughs in Statistics*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 1–10.
10. Ramirez, S. RELIEF-F Feature Selection for Apache Spark. Available online: https://github.com/sramirez/spark-RELIEFFC-fselection (accessed on 6 May 2019).
11. Brown, G.; Pocock, A.; Zhao, M.J.; Luján, M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
12. Ramírez-Gallego, S.; Mouriño-Talín, H.; Martínez-Rego, D.; Bolón-Canedo, V.; Benítez, J.M.; Alonso-Betanzos, A.; Herrera, F. An Information Theory-Based Feature Selection Framework for Big Data Under Apache Spark. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *48*, 1441–1453.