Deep Learning vs. Bag of Features in Machine Learning for Image Classification

Sehla Loussaief, Afef Abdelkrim
LA.R.A, Ecole Nationale d'Ingénieurs de Tunis, Université Tunis El Manar. BP 32, le Belvédère 1002.
ENICarthage, Université de Carthage, 35 rue des Entrepreneurs. Charguia II.
Tunis, Tunisie.
sehla.Loussaief@enicarthage.rnu.tn, afef.a.abdelkrim@ieee.org

***Abstract*— The main issue in computer vision and notably image classification problems is image feature extraction and image encoding. Here we show and compare two approaches to solve this problem: the first approach uses the Bag of Features (BoF) paradigm. The second one is based on deep learning and especially Convolutional Neural Networks (CNN). Specifically, we use the "AlexNet" CNN model trained to perform well on the ImageNet dataset. Our results shed light on how the use of CNN is more performant than the BoF in the process of feature extraction in a machine learning framework for image classification. This performance is shown by a series of experimentations that we carried out using the Caltech dataset and many classifier algorithms.**

Keywords*— *computer vision; image classification; feature extraction; machine learning; bag of features; deep learning; convolutional neural network

I. INTRODUCTION

The past decade has seen the rise of different approaches for pattern-recognition task in computer vision. A study has enabled us to identify two methods: The Bag of Features (BoF) and the deep learning technique with the use of convolutional neural network (CNN).

The BoF methods have been used in many computer vision fields such as image classification. The main idea behind BoF approaches is the use of an orderless collection of image features. Even if with this method we lack any structure or spatial information, the image representation would be powerful enough to match or exceed state-of-the-art performance in many of the applications to which it has been applied.

To perform image classification, CNN used in deep learning learn hierarchical layers of representation from input image [1-2]. The use of deep learning has recently proofed impressive performance on many computer vision applications such as classification problems, object detection and tracking, pattern recognition, etc. [3-5].

For image classification, we choose the application of BoF and deep learning technics. As discussed above, to deploy a machine learning framework for image classification we can use either the CNN capabilities or the BoF approach for feature extraction and image encoding. The feature vectors are fed to different classifiers. The purpose of our research is to experiment and compare the performance of classifier algorithms when we use these technics. We held experimentation with the AlexNet CNN which is a well-trained CNN on ImageNet dataset. We also use the Caltech dataset as input to our image classification framework.

The purpose of this paper is to compare the performance of the Bag of Features and deep learning paradigms in image classification task through an experimental evaluation.

The organization of this paper is as follows. Section II exposes a summary of image feature extraction deployed in machine learning framework. Section III presents details on the feature detection and extraction methods commonly used in BoF and deep learning approaches. Section IV looks at the evaluation of BoF and deep learning methods, including datasets, accuracy measures and comparative evaluation. The last section includes our concluding remarks.

II. IMAGE FEATURE EXTRACTION IN MACHINE LEARNING

Feature extraction can be defined as the fact of reducing an algorithm input data when this data is considered too large or redundant for processing. The result of feature extraction process is called feature vector.

The feature vector is supposed to include the most relevant information from the input data which will allow the use of the reduced information instead of the whole initial data without compromising the task process.

Selecting suitable variables is a critical step for successfully implementing an image classification. The use of too many variables in a classification task may decrease the performance of classification. It is imperative to choice only the variables that are most valuable. Therefore, for image classification in machine learning, the input for feature extraction process is a build derived features from an initial set of measured data. These features are intended to be informative, non-redundant, facilitating the subsequent learning steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

To decrease the dimensionality of input data, feature extraction methods have been extensively used. The most widely held method is the classical Principal Component Analysis (PCA) [6]. In the last decade, a multitude of non-linear approaches for reduction of dimensionality have been introduced.

In section III we expose the two investigated methods for image feature extraction which are Bag of Features and deep learning.

978-1-5386-4449-2/18/$31.00 ©2018 IEEE

III. METHODS

This section presents image feature extraction approaches used in this work. The Bag of Features paradigm is exposed in III.A. The based on Convolutional Neural Network method is detailed In III.B.

A. Bag of Features image representation

In Bag of Features method image representation is based on orderless groups of local features. In order to construct a visual vocabulary, features are extracted and vector quantized from each input images. The image features represent local areas of the image. New image's features are assigned to the nearest code in the "codebook". The image is represented as a histogram of codes. The idea of using normalized histogram of codes is inspired from Bag of Words technique used in document classification.

The BoF term vector is a compact representation of an image which discards largescale spatial information and the relative locations, scales, and orientations of the features [7].

Fig. 1 illustrates the process for building a Bag of Features image representation. This process is based on the following steps: (1) Generate visual vocabulary (code book): Features are extracted from all images in a training dataset. A clustering method is then used to cluster these features into a "visual vocabulary". Each generated cluster represents a "visual word" or "term". (2) Assign Terms: Considering a new image, features are extracted and assigned to the closest terms in the vocabulary. (3) Generate Term Vector: The counts of each term that appears in the image is calculated. These counts are then used to generate a normalized histogram representing a "term vector" called the Bag of Features representation of the image.

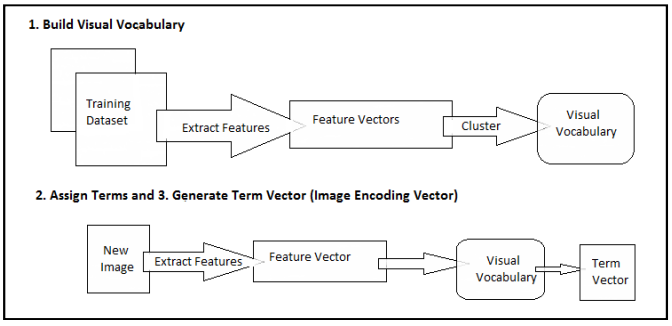

Fig. 1: BoF image representation process

For Bag of Features representation of an image we can use in ways other than simple term frequency. In fact, for the whole BoF process there are a multitude of design choices that we can use at each step.

Feature detection and representation method is one import key decision.

For features detection and extraction, we use the Speed Up Robust Features (SURF) method. It extracts salient features and descriptors from images.

This extractor is preferred over Scale-Invariant Feature Transform (SIFT) due to its concise descriptor length. In SURF, a descriptor vector of length 64 is constructed using a histogram of gradient orientations in the local neighborhood around each key-point [8].

In the clustering step, we use the K-means algorithm. It is selected over Expectation Maximization (EM) to group the descriptors into N visual words and to construct the visual vocabulary as experimental methods have verified the computational efficiency of K-means as opposed to EM [9].

B. Deep Learning: Convolutional Neural Network

Convolutional Neural Networks (CNN) are deep neural networks introduced for image classification purpose. The first CNN role is feature extraction and image vector representation. The second one is image classification. The CNN architecture is based on layers where the output of a layer is the input of the next one.

The global CNN architecture can be divided in two parts. The first one devoted to image vector representation is essentially composed of convolutional layers whereas the second part used for image classification is a fully connected layers.

Each layer delivers image representation level. A set of weights and biases allows layers' connection. The CNN particularity is that for a local image location we will use the weights. Weights shared for the same input location form a filter.

Convolutional part of a CNN is a succession of: (1) a convolution of the input with a set of filters for local feature extraction; (2) a non-linearity function such as the logistic function, for non-linear input data representation learning; and (3) a pooling function, which groups feature statistics at nearby locations and therefore decrease the computational cost.

The last layer output of the convolutional part of a CNN is introduced as input to a fully-connected layer [10].

Fig. 2 illustrates the architecture of the used CNN [11] which is a the pretrained ImageNet Convolutional Neural Network.

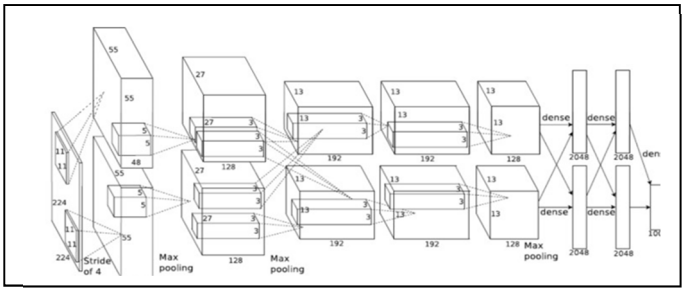

Fig. 2: ImageNet CNN layers

It consists of 5 convolution layers followed by 3 fully connected layers. Convolution layer filter sizes are 11×11, 5×5 and 3×3. The convolution function used in each convolutional layer can be represented by the following equation

$$O^l = pool_P\left(\sigma(O^{l-1} \star W^l + b^l)\right) \qquad (1)$$

Where O^{l-1} is the input feature map to the $l-th$ layer; $\theta^t=\{w^l, b^l\}$ is the set of learnable parameters (weights and biases) of the layer, $\sigma(.)$ is the point-wise non-linearity, $pool$ is a subsampling operation, P represents the pooling zone size (usually $P \times P$ square region) and the operator $\star$ signifies linear convolution function. A multi/hyper-spectral images are used

as input for the first layer, i.e. $O^o = I$, where I ∈ $\mathbb{R}^{R^o x C^o x N_h^o}$ is the input image, R^o and C^o are its width and height and N_h^o is the spectral bands number.

The most relevant hyper-parameters in CNN architectures are: layers number; filters size, and spatial pooling size and type.

The network training method is also an important aspect to deal with. For this purpose we can use a supervised method such as the standard back-propagation [12-13], or an unsupervised method, by the use of greedy layer-wise pre-training [14].

IV. EXPERIMENTAL RESULTS

In this section we expose the capabilities of the presented feature extraction approaches in different scenarios of image classification.

It is devoted to present a series of results of image classification based on the BoF and CNN strategies. Our results are reported on Caltech101 image dataset to which we have added some new images of existing categories.

For all experiments, we run 10 times and the average results were presented. Here we are interested in measuring the classifier accuracy.

A. BoF experimental environment

The deployed BoF environment shown in Fig. 3, includes the use of the SURF technique as a feature extractor and image encoder. This step returns a big number of image feature vectors.

In the clustering step we use the K-means algorithm. It aims to reduce the number of feature vectors. Only cluster centers will be considered as the visual words of vocabulary. BoF encoder step aims to encode each input image of the training dataset into histogram of visual words frequency. Image encoding vectors are than fed into classifiers.

B. CNN experimental environment

To investigate the CNN capabilities in image feature extraction for classification purpose, we choose the well-known AlexNet architecture [15], which is a convolutional neural network trained on the 1.3-million-image ILSVRC 2012 ImageNet dataset [16-17].

This already-trained AlexNet CNN is provided by the Caffe software package [18].

We chose AlexNet because it is widely known and publicly available. We conduct experiments with the Caffe provided model trained on the ImageNet dataset. The Caffe version has a minor difference from the original architecture in [18] that its neural activation functions are rectified linear units (ReLUs) [19] instead of sigmoids.

C. Results and discussion

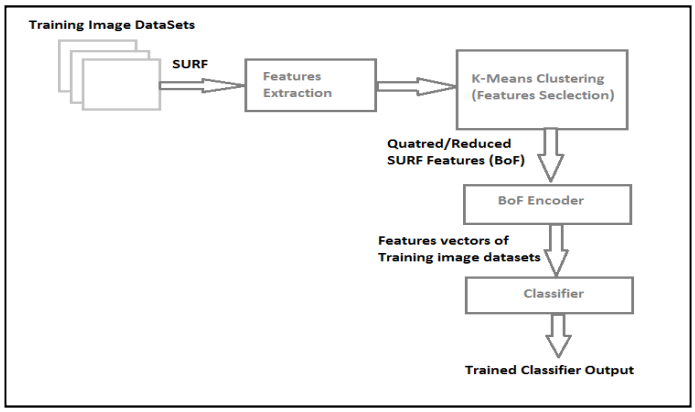

Fig. 3: Image classification process

For the performance measurement of our feature extraction techniques we use the Caltech101[1] dataset to which we add some images.

1) Scenario 1

Here, an approach for classification by a Linear SVM analysis is evaluated with particular regard to the effect of training set size on classification accuracy. The image feature vector size generated by our pre-trained CNN model is equal to 4096. To ensure the same dimension of image encoding vector, we chose the same value as visual dictionary size in the BoF assessment. Thus, in our machine learning framework each image is encoded with a vector of 4096 size.

The performance measures show that deep learning extractor and encoding technique is more robust compared to the increase of the size of the image dataset.

As presented in Fig. 4, experimentations indicate that the SVM classifier's accuracy significantly degrades with the increase of categories number when we use the BoF approach. In contrast the SVM classifier maintain a better accuracy when we use the deep learning extractor and encoding technique (>90%), as shown in Fig. 5.

With 12 categories in image dataset we notice that the SVM accuracy is 30% better when images are encoded based on the CNN technique.

2) Scenario 2

Next, we fix the category number in the dataset and evaluate the accuracy of our machine learning regarding the classifier's

[1] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

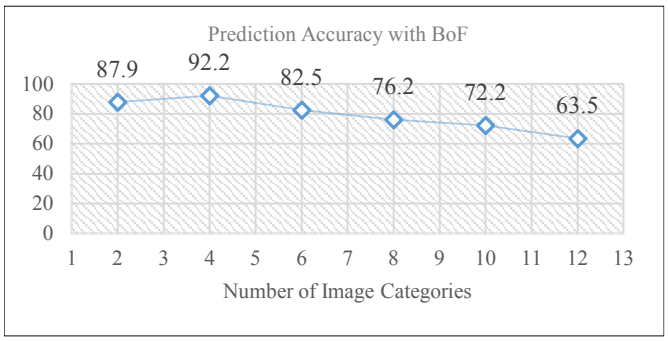

Fig. 4: Classification prediction accuracy with BoF technique

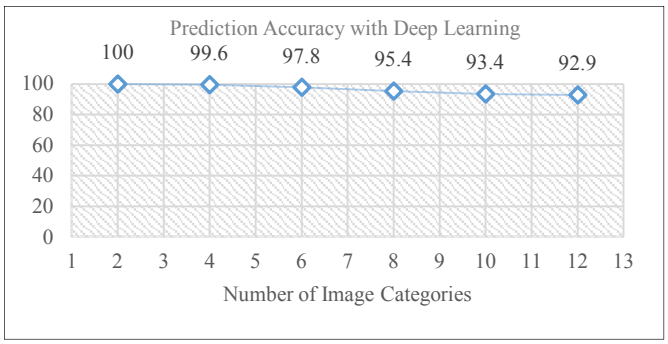

Fig. 5: Classification prediction accuracy with Deep Learning technique

type. In these experimentations we use the SVM, KNN and Decision Trees family.

Support Vector Machine Classification: Here we are interested in evaluating the accuracy of support vector machine (SVM) classifiers based on BoF or Deep Learning technique. The basic idea behind the use of the SVM technique in classification methods is the construction of a hyperplane or a set of hyperplanes in a high or infinite dimension space. Hyperplane which perform a better separation is the one that has the largest distance to the nearest training-data point of any class (functional margin). The accuracy of the classifier is closely related to the value of the margin: the larger the margin, the maller the classifier error. A multiclass classification problem is decomposed to a set of two class classification sub-problems, with one SVM learner for each sub-problem we can use different Kernel function to compute the classifier [20]: Linear kernel, Gaussian kernel, Quadratic and Cubic.

Measurements, presented in Fig. 6, show that the use of deep learning gives a better prediction accuracy. Among the SVM classifier the Quadratic SVM gives a superior result either with the deployment of the BOF technique or that of the CNN based Deep Learning.

a) K-Nearest Neighbors Classification: K-Nearest Neighbors algorithm (k-NN) is a non-parametric function witch can be deployed for classification task. In this method input is the k closest training examples in the feature space and the returned output is a class membership. The classification of a new object is based on the majority vote of its neighbors. The assigned class is the most common one among its k nearest neighbors. KNN-based algorithms differ in the number of

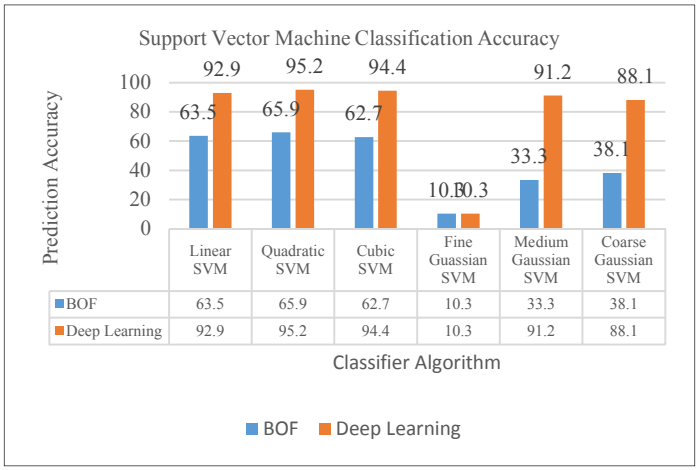

Fig 6: SVM Classifiers accuracies

nearest neighbors to find for classifying each point when predicting and the distance metric used [21].

KNN classifiers investigation, displayed in Fig. 7, confirm that the CNN based deep learning approach provide significantly better results than those based on the BoF mechanism.

It is shown that the Cosine KNN is the most precise during category prediction process.

b) Decision Trees Classification: Decision tree learning method is based on a decision tree. It is considered as a predictive model which maps observations about an item (represented in the branches) to conclusions about the item's target value represented in the leaves [21].

Fig. 8 illustrates the prediction accuracy achieved through the use of likhood decision trees classifiers.

Tests demonstrate that the CNN based Deep Learning technique is more accurate.

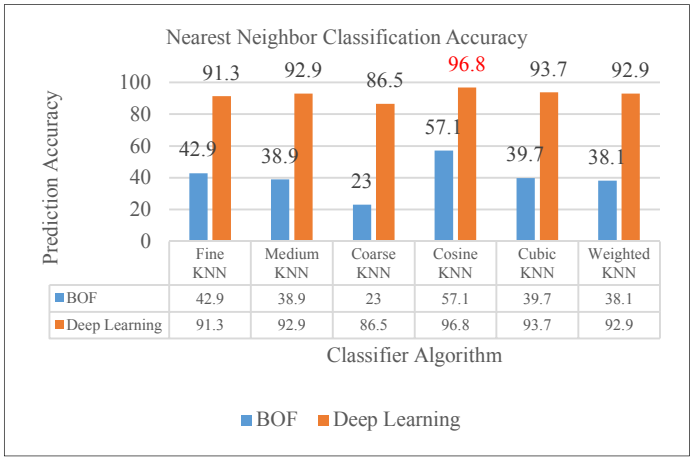

Fig. 7: KNN classifiers accuracies

Measurements show that the Medium Tree which uses medium number of leaves for finer distinctions between classes and number of splits not exceeding a maximum of 20, gives better results than the complex and simple Tree ones.

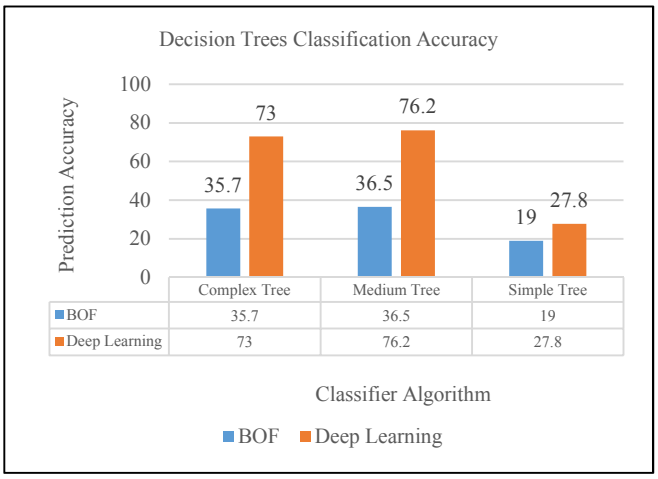

Fig. 8: Decision trees classifiers accuracies

V. CONCLUSION

The main important step in image classification process is feature extraction. This step consists of generating a feature vector from the input data. This reduced information is representative enough to be used instead of the complete input.

In this paper we related two feature extractor and image encoding techniques which are Bag of Features and Convolutional Neural Network based Deep Learning. Convolutional neural networks and BoF methods are investigated with results on Caltech101 Dataset. In BoF analysis, we choose a visual vocabulary size equal to 4096 to ensure the same features vector size as used by the pre-trained ImageNet AlexNet CNN. State-of-the-art image classification algorithms are used to evaluate the two extractor feature approaches. Results show the highest classification accuracy for the Deep Learning approach in comparison with Bag of Features paradigm. Even if for both feature extractor techniques, classification accuracy was related with the size of the training set, the CNN based technique outperform the BoF one. Following the increase in categories number to 12, the Linear SVM classification performance decreased by 20% in the case where BoF was used as feature extracting approach and only by 0.08% with the use of CNN based deep learning. Research carried out in this paper demonstrate that the use of deep learning ImageNet AlexNet CNN for image feature extraction has led to significant gains in accuracy. This result highlights that a trained CNN is very effective at image information encoding. In fact, features extracted by CNN on the image training dataset are highly relevant.

In addition, experiments show that the most accurate classifications are derived from the KNN approach (Cosine KNN, 96.8%) followed by the quadratic SVM with 95.2%. The decision trees derived algorithms don't yield accurate classification (max=76.2%). The large performance gap between these two families of approaches make the BoF image encoding technique useless. That's why Deep Learning based classification has increasingly become important approach for data classification.

REFERENCES

[1] Y. Bengio, "Learning deep architectures for ai". Foundations and trends R in Machine Learning, 2(1): pp. 1-127, 2009.

[2] G. E. Hinton, "Learning multiple layers of representation". Trends in cognitive sciences, 11(10): pp. 428-434, 2007.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks". In Advances in neural information processing systems, pp. 1097-1105, 2012.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Contextdependent pre-trained deep neural networks for largevocabulary speech recognition". Audio, Speech, and Language Processing, IEEE Transactions on, 20(1): pp.30-42, 2012.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification". In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pp. 1701-1708. IEEE, 2014.

[6] I. Jolliffe, "Principal component analysis". Springer, 2002.

[7] S. O'hara and B A. Draper. "Introduction to the bag of features paradigm for image classification and retrieval". arXiv:1101.3354v1 [cs.CV] 17 January 2011.

[8] D.Lowe, "Towards a computational model for object recognition in IT cortex". Proc. Biologically Motivated Computer Vision, p. 2031, 2000.

[9] D. G.Lowe, "Distinctive image features from scale-invariant keypoints". IJCV, 60(2): pp. 91-110, 2004.

[10] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview. Neural Networks", Volume 61, pp. 10-28, January 2015.

[11] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks". Advances in Neural Information Processing Systems 25, 2012.

[12] Y. LeCun, L. Bottou, G. Orr, and K. Muller, "Efficient backprop," in ¨ Neural Networks: Tricks of the Trade. Springer Berlin, 1998, pp. 9-50.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in International Conference on Learning Representations (ICLR2014). CBLS, April 2014.

[14] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527-1554, July 2006.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In Advances in neural information processing systems, pp. 1097-1105, 2012.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei, "Imagenet: A large-scale hierarchical image database". In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 248-255, 2009.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla and M. Bernstein, "Imagenet large scale visual recognition challenge". arXiv preprint arXiv:1409.0575, 2014.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding". arXiv preprint arXiv:1408.5093, 2014.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines". In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807-814, 2010.

[20] K. Crammer, Koby and Y.Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines" (PDF). Journal of Machine Learning Research. 2: pp. 265-292, 2001.

[21] P. Hall, B. Park and RJ. Samworth, "Choice of neighbor order in nearest-neighbor classification". Annals of Statistics. 36 (5): pp. 2135-2152. doi:10.1214/07-AOS537, 2008.

[22] L. Rokach and O. Maimon, "Data mining with decision trees: theory and applications". World Scientific Pub Co Inc. ISBN 978-9812771711, 2008.