Semantic Classification for Big Data Analysis

Amina Taouli, Djamel Amar Bensaber, Nabil Keskes, and Khayra Bencherif

LabRI Laboratoty, Ecole Supérieure Informatique, Sidi Bel Abbes, 22000, Algeria
{ a.taouli, d.amarbensaber, n.keskes, k.bencherif }@esi-sba.dz

Abstract. Big Data can be defined as a dataset that contains a huge volume of information which can be analyzed to discover unknown and useful patterns. In fact, Deep Learning is a technique that is used mainly to facilitate the Big Data Analysis by extracting complex abstractions of high level. Nevertheless, data heterogeneity represents a key challenge in the context of Big Data Analysis. Usually, data providers use different techniques to represent the same real-world object. Moreover, they have a lack of methods for the automation of the classification of the input images. Therefore, we have to add the semantic aspect in the classification for enhancing the Big Data Analysis process. In order to meet this requirement, we put forward an approach that allows semantizing the classification using Convolutional Neural Network and Semantic Memory. We apply the pooling operation to reduce the size of the input data. After a filter succession, the convolution maps are concatenated into a 1-D vector and then we use Semantic Memory to classify the input data.

1 Introduction

Nowadays, the total amount of data consumed and offered on the Web has increased. For this purpose, the researchers expected that in 2020 there will be more than 16 zettabytes of useful data Turner et al. (2014). In connection with this explosion, the term Big Data has emerged to refer to any collection of massive, high-speed and heterogeneous datasets. In this context, we face the main challenges of acquiring, integrating, analyzing and visualizing large amounts of data Curry (2014). This amount of data is captured in various forms containing unstructured data that account for 90% of all data.

In order to make the raw data acquired usable in the recommendation, decision and prediction systems, it is recommended to use Big Data Analysis (BDA). In fact, BDA is the process of adding structures to the data to find facts, relationships, patterns, and extracting hidden information.

Currently, the Deep Learning (DL) has a huge success in several areas such as speech recognition, image processing and natural language Mohammadi et al. (2017). In addition to the data growth, the DL plays an important role in providing predictive solutions for BDA that consist of having a better result, understanding and detecting relationships between data and predicting future instances.

However, many traditional systems are not able to interpret all types of information contained in data sources. On the one hand, we can find different concepts representing the same meaning of a real-world object. On the other hand, the applicability of many existing approaches for BDA is limited to a lack of the automation of the classification of the input images. Thus, the addition of a semantic context to manage these problems brings more effective solutions and makes it possible to better respond to the challenges of Big Data Analysis.

In this paper, we propose an approach to semantize the classification of data using DL algorithms for solving the problem of the volume which consists of Big Data domains and the variety for obtaining quality results in a real time. One of the main reasons that allow the world to recognize the power of DL is the Convolutional Neural Network algorithm that has many advantages in image recognition.

The CNN algorithm consists of several consecutive layers, starting with loading, storing and processing the raw data of the images in the network. Then, the data pass through the convolutional layers, which act as a feature extractor when learning the representations of the characteristics of their input images Rawat and Wang (2017). Generally, we can find pooling and ReLU layers between these convolutional layers. The first layer allows simplifying the information in the output of the convolutional layer Nielsen (2017). The last layer allows applying an activation function on the input data Patterson and Gibson (2017). Finally, the fully connected layer is placed after the determined amount of convolutional layers, ReLU layers and pooling layers that are connected to all the maps activation of the previous layer. The input of the fully connected layer will be smaller than the original inputs due to the reductions of the images specified in the previous operations. We scan the reduced images, which correspond to each features map, and transform each of them into a list of values to arrive at the classification of its images.

In order to answer the problem of semantics, we introduce the semantic aspect via the Semantic Memory statistical method Lund et al. (1995). The Semantic Memory creates a semantic space from the co-occurrences of the input data. We form a matrix where each element represents the force of association between the data represented by the line and the column. The closest neighbors are considered to reflect the semantics of the target data, and therefore they have the higher weight.

The rest of this paper is structured as follows: We review the state of the art approaches in section 2. Section 3 presents an overview of the proposed approach. In section 4, we conclude and suggest directions for future research.

2 Related work

Naturally, large data sets contain mostly unstructured data. Thus, the large-scale processing of such semi-structured or unstructured data sets presents a significant challenge for the BDA Kaisler et al. (2013). In this section, we address the different approaches that introduce the semantic dimension (semantic web technology or statistical methods) to analyze the Big Data in two ways, either using or not DL techniques.

2.1 Semantic for Big Data Analysis without the use of Deep Learning algorithms

The authors of Nural et al. (2015) propose an approach to choose an appropriate model for analyzing large datasets using semantic technologies by developing an ontology analysis for predictive analysis. In Kumara et al. (2015), an ontology-based workflow generation approach is proposed to automatically generate the workflow. This approach uses ASC (Automatic Service Composition) to automate the ontology-based Cross-Industry Standard Process for Data Mining (CRISP-DM) method. In Dayyani (2016), the authors implement and develop a software platform using intelligent components that can analyze a large number of financial data to help humans make decisions. This platform consists of seven layers that organize the data in a semantic layer containing three software components. Paik (2016) is another approach that provides a framework for collecting information from the web, remote devices or sensors by analyzing collected data and transforming them into domain ontology instances. Then, an automation model of Big Data Analysis with CRISP-DM using ASC is used to get a better understanding of the situation. In Yao et al. (2016), a new framework is developed to combine semantic methods and Big Data processing for security analysis in Big Data. The semantic analysis is based on the use of an ontology and HCI analysis that deploys semantic and data analysis.

2.2 Semantic for Big Data Analysis with the use of Deep Learning algorithms

The authors of Huang et al. (2013) developed a new model through the combination of the Latent Semantic Analysis (LSA) method with a Deep Neural Network based structure to classify web documents using a hashing techniques. In Salakhutdinov and Hinton (2009), the authors described a semantic hashing method that is used to find binary codes. In order to fastly retreive the documents, they use a Deep Autoencoder Network and the retropropagation algorithm for finding semantically similar documents regardless of the document size. In Liu et al. (2015), the authors presented a multi-stain Deep Neural Network method for a multi-domain classification. The authors of Shen et al. (2014) developed a latent semantic model based on a Convolutional Neural Network for searching the queries and the web documents to perform a semantic correspondence between them. Zhang et al. (2015) is another work that proposes an empirical study on character-convolutive networks for the text classification. The authors designed ConVnets using an English thesaurus that was obtained from the WordNet where each synonym for a sentence or a word is classified by semantic proximity in the most commonly understood sense Feinleib (2014).

However, the main challenges of these approaches are the lack of semantics in the context of data classification. The first part of the state-of-art approaches introduced the semantic aspect via semantic web techniques working only on structured data. They provide, organize, filter and analyze the data acquired. In addition, they improve the current analysis techniques to meet performance requirements for data volume including compute performance, real-time big data and batch processing, data mining, data association, etc. Unfortunately, these approaches often require knowledge of the experiment design and the choice of a modeling technique which often depends on the purpose of the analysis. Furthermore, the user needs an expertise in ontologies and rule bases. The second part of the state-of-art approaches proposed methods

for combining the semantic aspect with Deep Learning algorithms working with large amounts of heterogeneous high-velocity data. For example, the LSA method is used to statistically discover hidden and underlying semantics. However, it implies too many calculations due to the decomposition of the matrix into two orthogonal matrices and into a diagonal matrix of singular values. In addition, the Semantic Hashing method implies that documents with similar addresses have similar content, but the reverse is not necessarily true.

In order to fill these gaps, we propose an approach that uses semantic memory to facilitate the semantic classification of data involving the characteristic of value and quality.

3 Proposed Approach

In fact, the biological inspiration for CNN is the visual cortex in animals. The goal of a CNN is to learn the high-level features in data via convolutions Ravi et al. (2017). They are recognized for the analysis of images which are the major cause for which the world admits the power of DL Sugomori (2016).

In the health department, a large volume of unstructured patient data is generated from clinical reports and medical images. In recent years, CNN has quickly been adopted by the medical imaging research community because of its proven performance in computer vision.

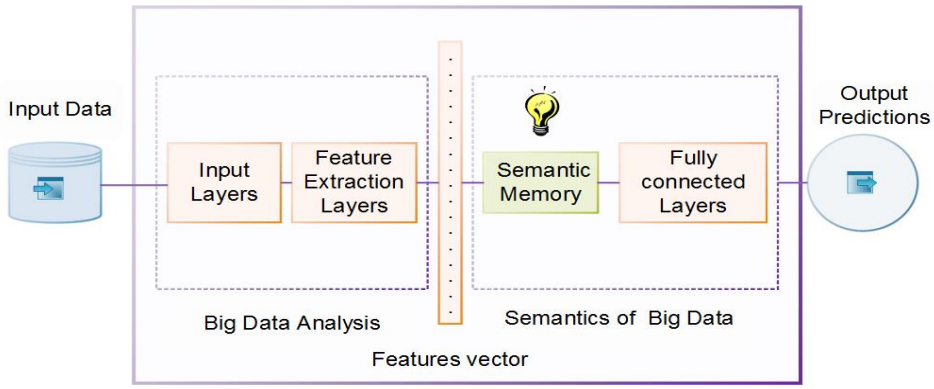

FIG. 1 – *Semantic Memory For Big Data Analysis Architecture*

Figure 1 shows the architecture of our approach which will be detailed in the following sub-sections. Our approach consists of three main parts: the analysis of Big Data, the construction of the characteristics vector and the semantics of Big Data. It takes as input a large number of raw medical images that will be loaded and stored for processing in the network and will produce as output a semantic classification of the input data.

3.1 Big Data Analysis

The input layer provides an image in the form of a pixel array. It allows specifying the width, the height and the number of channels. This number refers to the RGB values for

each pixel representing the fundamental colors (Red, Green, and Blue) Patterson and Gibson (2017). The feature extraction layers include onvolutional and sub-sampling or pooling layers

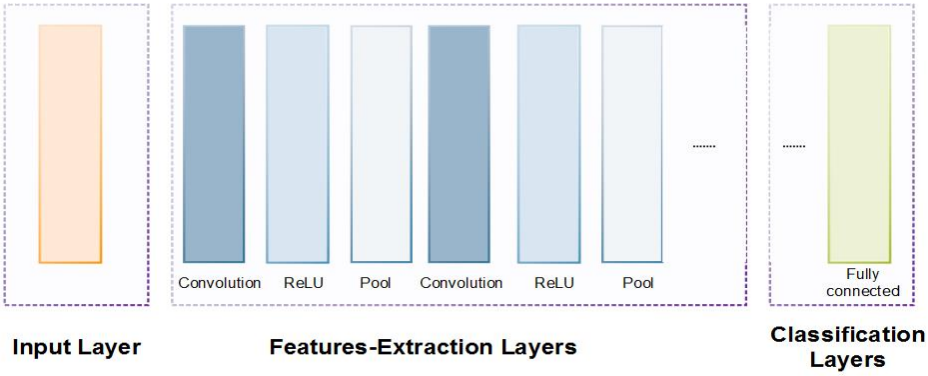

FIG. 2 – *High-level general CNN architecture Patterson and Gibson (2017)*

as already shown in Figure 2.

3.1.1 The convolutive layer

It contains the basic elements of the Convolutional Neural Network. It functions as an image feature extractor that locates the presence of a set of features in input images Rawat and Wang (2017). Thus, we perform a convolutional filtering to drag a window representing the entity on the image and calculate the convolution of each image with each filter which called "kernel" Sugomori (2016).

3.1.2 The ReLU layer

This layer applies an activation function to the input data. Thus, the execution of this function on the input volume modifies the values of the pixels but does not modify the spatial dimensions of the input data Patterson and Gibson (2017). Therefore, it replaces all negative values received as input with zeros.

3.1.3 The pooling layer

It receives several input feature mappings and applies the pooling operation to each of them. The pooling operation reduces the size of the images, but preserves their important characteristics Nielsen (2017). In addition, it makes the network less sensitive to the position of the entities.

3.2 Features vector construction

The convolution maps are laid flat and concatenated into a 1D feature vector. As the images are reduced with the previous operations, the input of this layer will be smaller than the original

inputs. We scan the reduced images, which correspond to each feature map, and transform each of them into a list of values Beysolow II (2017).

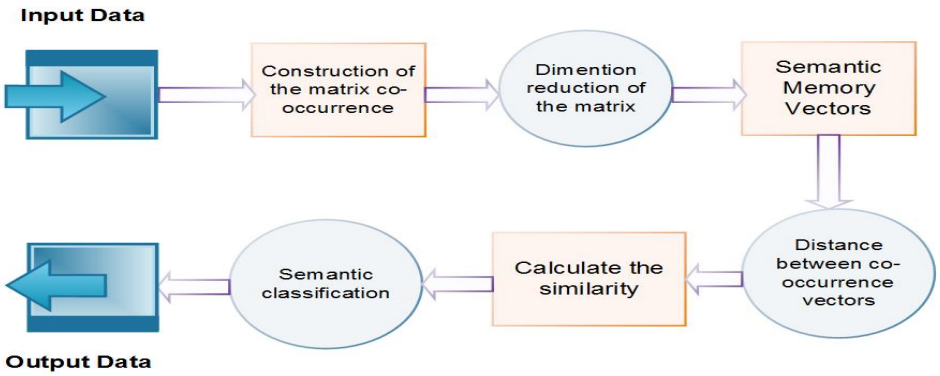

FIG. 3 – *Semantic Memory Workflow*

3.3 Semantics for Big Data

At this stage, we introduce a statistical method called Semantic Memory also known as Hyperspace Analog Language illustrated in Figure 3. The Semantic Memory uses a large-scale semantic space constructed from a co-occurrence matrix by defining a window size. Concretely, if two images are neighborhoods of the window, they are counted as co-occurrents. In order to obtain a set of vectors, we extract columns and rows from the constructed matrix and we group the images by calculating a distance between them. Therefore, the co-occurring images have similar lines. As consequence, the similar images will appear close to each other. The result of this method is then connected to the input of fully connected layers that consist of a set of vectors.

4 Conclusion and future work

In this paper, we presented a Deep Learning approach for Big Data Analysis with the introduction of the semantic aspect. In order to model complex relationships between data, our approach analyzes the data with the techniques of DL which is based on the CNN algorithm that is very effective in the analysis, the classification and the recognition of medical images. Furtheremore, we used the Semantic Memory to exploit each octet of relevant data, provide a better prediction, make smarter decisions and improve our results.

In our futur work, we will implement this architecture in a real-time data analysis platform such as SPARK.

References

Beysolow II, T. (2017). *Introduction to Deep Learning Using R*.

Curry, E. (2014). *In New Horizons for a Data-Driven Economy A Roadmap for Usage and Exploitation of Big Data in Europe.* Springer.

Dayyani, B. (2016). Software architecture design and development of multi-layer highly modular platform using intelligent components for dynamic big data analytics. In *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*, pp. 45–53.

Feinleib, D. (2014). *Big Data Bootcamp: What Managers Need to Know to Profit from the Big Data Revolution*. Apress.

Huang, P.-S., X. He, J. Gao, L. Deng, A. Acero, and L. Heck (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information; knowledge management*, CIKM '13, New York, NY, USA, pp. 2333–2338. ACM.

Kaisler, S., F. Armour, J. A. Espinosa, and W. Money (2013). Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences(HICSS)*, Volume 00, pp. 995–1004.

Kumara, B. T. G. S., I. Paik, J. Zhang, T. H. A. S. Siriweera, and K. R. C. Koswatte (2015). Ontology-based workflow generation for intelligent big data analytics. In *2015 IEEE International Conference on Web Services*, pp. 495–502.

Liu, X., J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval. NAACL.

Lund, K., C. Burgess, and R. Atchley (1995). Semantic and associative priming in high-dimensional semantic space.

Mohammadi, M., A. I. Al-Fuqaha, S. Sorour, and M. Guizani (2017). Deep learning for iot big data and streaming analytics: A survey. *CoRR abs/1712.04301*.

Nielsen, M. (2017). Neural networks and deep learning. `http://neuralnetworksanddeeplearning.com` Last access: 06/05/2018.

Nural, M. V., M. E. Cotterell, and J. A. Miller (2015). Using semantics in predictive big data analytics. In *2015 IEEE International Congress on Big Data (BigData Congress)(BIGDATA CONGRESS)*, Volume 00, pp. 254–261.

Paik, I. (2016). Situation awareness based on big data analysis. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, Volume 2, pp. 911–916.

Patterson, J. and A. Gibson (2017). *Deep Learning A Practitioner's Approach*. O'Reilly Media.

Ravi, D., C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang (2017). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics 21*(1), 4–21.

Rawat, W. and Z. Wang (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation 29*(9), 2352–2449. PMID: 28599112.

Salakhutdinov, R. and G. Hinton (2009). Semantic hashing. *International Journal of Approx-*

imate Reasoning 50(7), 969 – 978. Special Section on Graphical Models and Information Retrieval.

Shen, Y., X. He, J. Gao, L. Deng, and G. Mesnil (2014). Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, New York, NY, USA, pp. 373–374. ACM.

Sugomori, Y. (2016). *Java Deep Learning Essentials*. Packt Publishing Ltd.

Turner, Gantz, Reinsel, and Minton (2014). The digital universe of opportunities : rich data and the increasing value of the internet of things. Technical report, IDC EMC.

Yao, Y., L. Zhang, J. Yi, Y. Peng, W. Hu, and L. Shi (2016). A framework for big data security analysis and the semantic technology. In *2016 6th International Conference on IT Convergence and Security (ICITCS)*, pp. 1–4.

Zhang, X., J. Zhao, and Y. LeCun (2015). Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, Cambridge, MA, USA, pp. 649–657. MIT Press.

Résumé

Les Big Data peuvent être définies comme un ensemble de données contenant un énorme volume d'informations pouvant être analysées pour découvrir des modèles inconnus et utiles. En effet, Deep Learning est une technique principalement utilisée pour faciliter l'analyse des Big Data en extrayant des abstractions complexes de haut niveau. Néanmoins, l'hétérogénéité des données représente un défi majeur dans le contexte de l'analyse des Big Data. Généralement, les fournisseurs de données utilisent différentes techniques pour représenter le même objet réel. De plus, ils manquent des méthodes pour l'automatisation de la classification des images d'entrée. Par conséquent, nous devons ajouter l'aspect sémantique dans la classification pour améliorer le processus d'analyse des Big Data. Afin de répondre à cette exigence, nous proposons une approche qui permet de sémantiser la classification en utilisant le réseau de neurones convolutifs et la mémoire sémantique. Nous appliquons l'opération de pooling pour réduire la taille des données d'entrée. Après une succession de filtres, les cartes de convolution sont concaténées dans un vecteur 1-D, puis nous utilisons la mémoire sémantique pour classer ces données d'entrée.