DEEP LEARNING IN VERY HIGH RESOLUTION REMOTE SENSING IMAGE INFORMATION MINING. COMMUNICATION CONCEPT

Corina Vaduva (1), Inge Gavat (1), Mihai Datcu (1)(2)

University Politehnica of Bucharest (1), German Aerospace Center (2)

ABSTRACT

This paper presents the image information mining based on a communication channel concept. The feature extraction algorithms encode the image, while an analysis of topic discovery will decode and send its content to the user in the shape of a semantic map. We consider this approach for a real meaning based semantic annotation of very high resolution remote sensing images. The scene content is described using a multi-level hierarchical information representation. Feature hierarchies are discovered considering that higher levels are formed by combining features from lower level. Such a level to level mapping defines our methodology as a deep learning process. The whole analysis can be divided in two major learning steps. The first one regards the Bayesian inference to extract objects and assign basic semantic to the image. The second step models the spatial interactions between the scene objects based on Latent Dirichlet Allocation, performing a high level semantic annotation. We used a WorldView2 image to exemplify the processing results.

***Index Terms*—** Information theory, deep learning, semantic annotation.

1. INTRODUCTION

The communication theory developed by Shannon [1] is the foundation for the information transmission and data compression which are vital in a variety of applications in nowadays. The main idea was statistically expressed as a specific signal to be transmitted through a channel that is characterized by a probability distribution over the set of all likely outputs given each input. The source of the communication channel randomly generates messages that will be first encoded. That is to each message a certain signal class will be assigned. When the transmission is completed, the receiver needs to decode the output and decide what particular signal was the actual input.

The information theory had an important role on remote sensing (RS) image construction, allowing detailed studies based on the analogy between the communication channel and the image formation model [2]. It also has a major contribution to algorithms development for information mining from the collected RS data.

Along with the technology development and the improvement of images spatial resolution, the information mining requires a deeper analysis due to the increased amount of observable details. Several levels for image analysis are mandatory now for a complete scene understanding. For each level, models must be developed according to the type of information desired, namely pixel based, object based or structure based. A set of algorithms for representing and organizing multiple levels in order to express complex relationships among data defines a deep learning process. Feature hierarchies are discovered considering that higher levels are formed by combining features from lower level. One can resume that the deep learning consists of creating functions mapping between different levels of abstraction [3]. This theory is currently an emergent subject for speech and language processing [11]. Other deep learning approaches and applications are overviewed in [12], presenting the process as a new challenge in the artificial intelligence research field.

In this article, the authors present a deep learning algorithm for semantic annotation of very high resolution (VHR) RS images by means of the communication theory. The information mining process is considered similar to data transmission through a communication channel. An image is modeled in order to encode scene content. The aim is to learn a set of rules and to discover an accurate dictionary of symbols, such that the map received by the user to express a real meaning of the scene.

The presented processing chain is following a deep hierarchical representation of the image content on different levels, from primitive features to pairs of objects. Two major learning steps are considered. Through the first one, objects are extracted using Bayesian inference [4]. In the second step, an analysis of objects' spatial interactions is performed by applying Latent Dirichlet Allocation (LDA) model [5]. The results illustrate maps containing spatially alike pairs of objects with semantic meaning.

The article begins with the description of a communication channel concept for deep learning in Chapter 2. In order to sustain the idea of this concept, Chapter 3 presents a deep hierarchical information representation of for VHR RS images. The objects extraction process is depicted in Chapter 4, while the analysis of objects' spatial interactions is described in Chapter 5. The next chapter, number 6, emphasizes some characteristics of the proposed concept and presents the results for a VHR image portraying the city of Bucharest. The conclusions of our work are given in the last section.

2. COMMUNICATION CHANNEL CONCEPT FOR DEEP LEARNING

The main purpose of this section is to explain the deep learning process for semantic annotation of RS images using the communication channel concept. By definition, the standard communication model consists of three elements: a source, a channel and a receiver, connected by an encoder and decoder (Figure 1). The idea of information transmission consists of decoding the received message with a small probability of error, which is finding those code-symbols mutually completely separable.

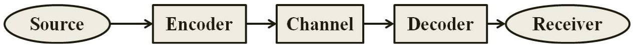

Figure 1. Standard communication model.

The problem of semantic annotation resembles to information transmission. The source is the image that we need to code and send to the user in the shape of a thematic map where the assigned labels are the mutually separable code-symbols required to define the semantic meaning of land structures without significant losses. The process of data analysis is similar to the channel modulation. The scheme in Figure **2** illustrates a communication system model for image semantic annotation.

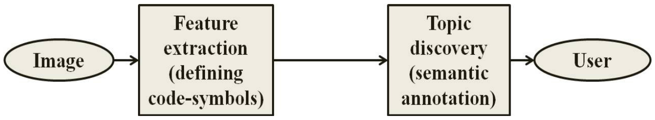
Image
Feature extraction (defining code-symbols)
Topic discovery (semantic annotation)
User

Figure 2. Semantic information mining model.

The construction of the communication system describing the information mining from RS images is becoming more complex with the increasing of the image spatial resolution. This is explained by the fact that each type of detail observable in the scene introduces a new level in the image processing chain. For instance, a low resolution image analysis is only pixel-based, while a VHR image requires a three levels modeling, as it will be presented in the next chapter. The hierarchical architecture of the new system confirms the necessity of a deep learning process.

3. DEEP HIERARCHICAL INFORMATION REPRESENTATION FOR VHR RS IMAGES

The content of the VHR RS optical images is remarkably complex, providing details about objects like small houses, trees, or even cars on parking area. In order to be able to completely describe such an extent of information, a deep hierarchical representation such in Figure 3 is recommended.

The lowest level is depicted by the primitive feature vectors color, texture and shape. At the next level of representation, single objects like building, highway, forest, boat or lake are described by unique combinations of the primitive feature vectors. Basic semantics is thus defined.

But individually considered, these objects cannot describe the scene, they cannot capture the global meaning of the scene. According to their neighboring and spatial positioning, they give different interpretations to the scene. Therefore, in order to extract real useful meaning from the content of an image, we group the objects two by two and compute descriptors for the objects spatial interactions at the next level of abstract representation.

Their modeling leads to the discovery of semantic rules that define the last level of information representation, the semantic classes. Residential area, commercial area, harbor are high-level semantic features for describing the content of remote sensed Earth's surface.

Following this hierarchical representation, we apply the learning algorithms, for object extraction and spatial interaction modeling.

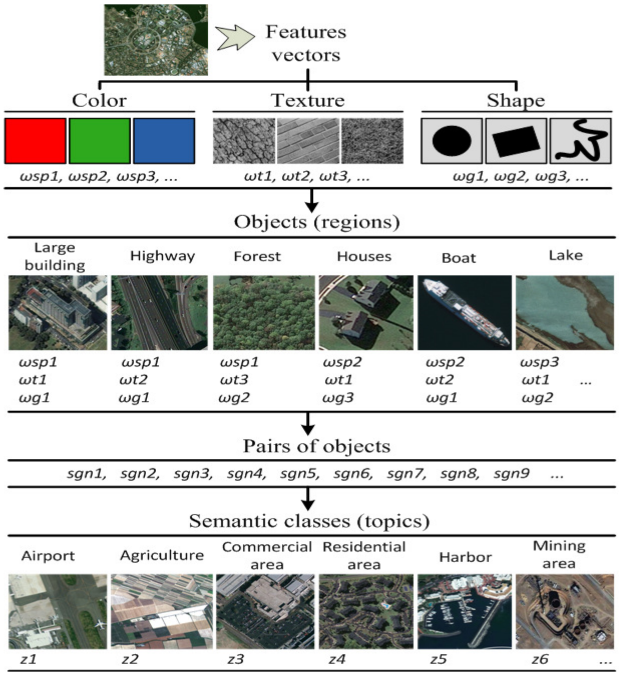

Figure 3. Deep hierarchical information representation of the image content. Each level of abstract representation is characterized by a specific feature.

4. BAYESIAN INFERENCE FOR OBJECT EXTRACTION: BASIC SEMANTIC ANNOTATION

In order to extract objects, we use a human-centered concept for interactive learning and probabilistic retrieval of user-specific cover labels. This is the Knowledge Based

Information Mining (KIM) System [4], [6], [7] a system able to explore large collections of data, performing a very good local classification. The human computer interaction is made through a graphical interface where positive and negative example can be given to train the system. When the learning phase is over, KIM is able to assign meaning to the primitive features.

The KIM system implementation is also based on a hierarchical information representation that models the image content in a Bayesian formulation.

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)} \tag{1}$$

The Bayes rule relates the posterior probability $p(H|D)$ to the likelihood $p(D|H)$. $p(H)$ and $p(D)$ are the prior of belief in the hypothesis H and respectively the prior predictive distribution. The concept of probability as the frequency of an event is here understood as the degree of certainty of a particular hypothesis [4]. Using the strong stochastic models available for image processing, the decision considering the H need to be made based in the posterior odds:

$$\Lambda = \frac{p(H|D)}{p(-H|D)} = \frac{p(D|H)}{p(D|-H)}\frac{p(H)}{p(-H)} \tag{2}$$

If H is the hypothesis of D as a realization of a certain cover type defined by the user, then the posterior odds allow the user to decide that the result obtained is indeed the label he was looking for. Once the decision is made, the posterior odds may be computed in order to complete the image content modeling. The procedure can be repeated for several cover types.

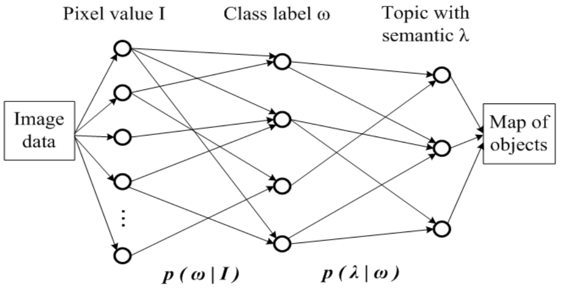

Figure 4. A simplified version of information representation for objects extraction. The learning process is similar to a standard transmission procedure.

The formalism implemented in KIM may be divided in two main parts, as illustrated in Figure 4. At first, the image classification based on the primitive features is performed in an unsupervised, application-free manner. The second part represents the user driven learning method to discover specific cover types.

Each of the two learning parts can be seen as a communication system. While the classification in the first part encodes the data using class labels as code-symbols $p(\omega|I)$, the user based modeling in the second part uses the class labels as inputs to be encoded by means of topics λ containing basic semantics $p(\lambda|\omega)$. Nevertheless, the object extraction process may be resumed to a single communication system $p(\lambda|I)$.

5. LATENT DIRICHLET ALLOCATION ANALYSIS FOR SPATIAL INTERACTION MODELING: HIGH LEVEL SEMANTIC ANNOTATION

The labels assigned to the withdrawn objects represent a basic label of semantics. Nevertheless, a proper information mining from VHR imagery requires further analysis. In order to perform a high level semantic annotation and offer to the user a thematic map with real meaning encapsulated, we need to compute descriptors for the spatial interaction of objects.

The spatial information represents an important element of VHR image understanding and scene description. Therefore, we group the objects two by two and for each pair we compute a spatial signature [8] based on the histogram of forces exerted between the two objects and tending to move them one to another on a certain direction [9]. The study of spatial interaction is a solution to bridge different levels of information representation, extending the pixel and object interpretation to scene understanding. This is considered an intermediate level between two learning steps in a deep learning process.

We apply further the Latent Dirichlet Allocation (LDA) to analyze the spatial information. Even if this algorithm was developed for text retrieval, LDA is a generative probabilistic model based on the bag of words assumption that can be used for any collection of discrete data [10], including RS imagery. It is founded on a three-level hierarchy, in which documents of a corpus are represented as random mixtures over latent topics and each topic is characterized by a distribution over words Figure **5**.

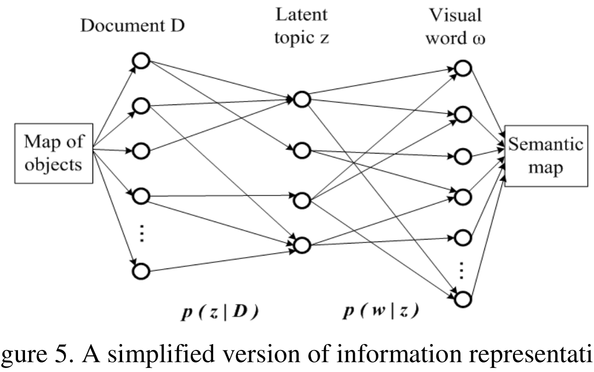

Figure 5. A simplified version of information representation for semantic topic discovery. The learning process is similar to a standard transmission procedure.

The LDA modeling of spatial interaction implies image text analogy like in [5]. A dictionary of “visual words” is defined over the whole set of spatial signature computed for the image using a classification process. The configurations spatially alike in the scene will receive the same label. Thereby, acquired classes of signatures will represent the visual words, and their number, the size of the dictionary.

Once the vocabulary defined, each tile of the image is represented by a sequence of N visual words. By counting the visual words occurrences, we get a word frequency vector for each sub-image. LDA models each visual word in a document as a sample from a mixture model, where the mixture components can be viewed as representations of topics. As a consequence, LDA leaves flexibility to assign a different topic to every observed word in a document. Each document is represented as a mixture of topic proportions, reducing the size of the dictionary to the number of topics. The topic structure is learned without any usage of background knowledge. The retrieved data presents, besides similarities with the observable structures in the scene, an amount of latent information.

The generative process within the LDA model [10] requires a training set given by a user consisting of a few documents containing useful information for the application. Rules are discovered based on words statistics and the global parameters α and β of the model are estimated over the entire collection of data. Considering these estimations, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words ω is given by

$$p(D|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(\omega_n|z_n,\beta)\right)d\theta \qquad (3)$$

The representation of the LDA formalism given in Figure 5 relies on a hierarchy that confirms the assumption of a deep learning process. The similarity with the communication system is still perceptible. We can treat the LDA model as a single communication system $p(\omega|D)$, or it can be regarded as a series of two such systems encoding words based on topics $p(\omega|z)$ and topics based on documents $p(z|D)$.

6. DISSCUSSIONS AND RESULTS

The deep learning processing described in this article has powerful abilities to learn dependencies and distributions. Its understanding is simplified by a comparison with the communication system. Moreover, the modeling can be treated as a concatenation between several communication channels. The RS images are not visual signals, but recordings of an instrument about the Earth surface. The communication channel model is a unitary representation which highlights their component elements and helps in formalizing hypotheses about image content. This approach was used for objects extraction as well as for semantic topics discovery.

There are two important aspects to be discussed, namely the objects extraction (KIM system learning) and the objects spatial interactions analysis (LDA modeling). They both model the link between class labels and semantic topic, but in the opposite direction. In KIM pixels label are modeled such that they form objects (topics with basic semantic), while LDA assumes that latent topics are modeled in order to compute the probability of a visual word (the spatial signature corresponding to a pair of objects) to receive the semantic label corresponding to the topic. The objects extraction is a supervised method, where the training dataset must be strongly relevant for the cover type that a user desire to define. The LDA analysis is also supervised, but the information included in the training data set is not definitive for the obtained semantic topics, because LDA is able to latently discover hidden information from the data provided. One can say that the training set is randomly chosen and its content is not significant for the semantic labeling of topics. Another difference regards the validation of the semantic annotation performed. The ground truth is not available for VHR images; it can only confirm large areas of forest, city, water or agriculture. Still, the objects labeling may be identified through a visual inspection, while the semantic topics containing pairs of objects with similar positioning is difficult to assess.

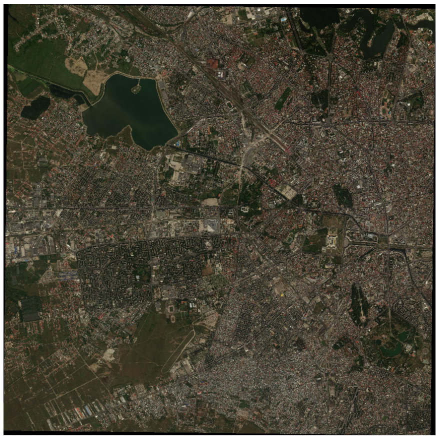

Figure 6. WorldView2 image. Bucharest, Romania. [13]

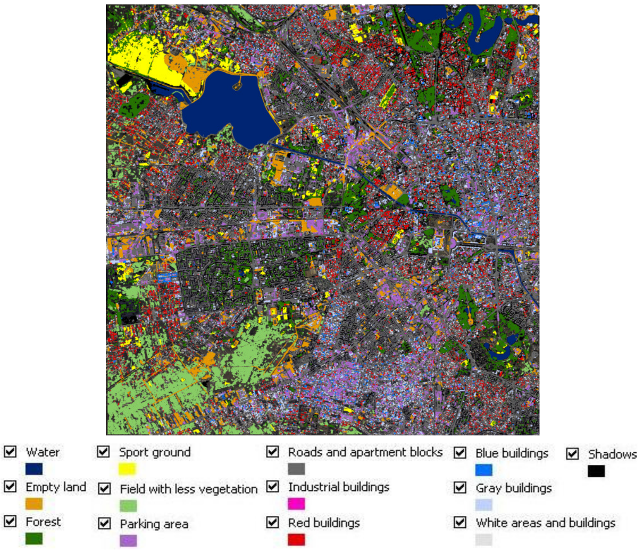

Figure 7. Scene land cover. The result of objects extraction. Thematic map with basic semantic included. These objects will be grouped 2 by 2 and the pairs modeled using LDA.

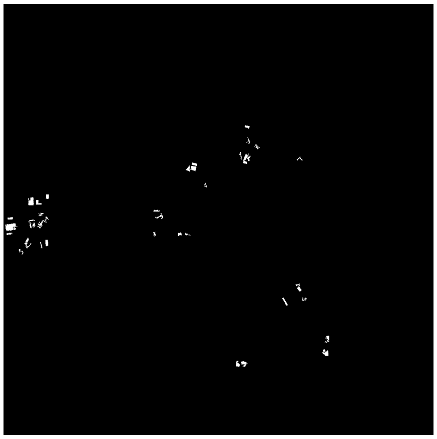

Figure 8. An example of high level semantic topic expressing real meaning: industrial areas. The topic contains groupings of "parking areas (in gray) with industrial buildings (in white)" with the same spatial arrangement.

In order to exemplify the proposed methodology, we used a WorldView2 image covering an area of approximately 100 square kilometers over the city of Bucharest, Romania (Figure 6).

As we can see, the semantic topics obtained contain a high level of information that usual feature extraction methods are not able to distinguish. It can supersede a human visual inspection due to its characteristic of identifying configurations not easy to be observed by a user.

The proposed methodology allows further quantitative evaluations for the learning process. Their significance is emphasized by the objective nature of the image signal.

7. CONCLUSIONS

The deep learning process has an emergent impact especially on recent research related to artificial intelligence. It was regarded as a method to supersede human learning with machine learning. The study described on the present paper focuses on this type of analysis for a high level semantic annotation of VHR RS images. An abstract hierarchical representation was provided for the image content. The information contained was encoded level by level through an analogy with the standard communication channel: the image is the source, the feature extraction analysis encodes the information, the semantic topic discovery decodes the image content and finally the user receives the message.

The proposed methodology is based on two main procedures for semantic learning. The first uses Bayesian inference to extract objects and add basic semantic label to the image, while the second applies a latent LDA analysis to model the spatial interaction between objects within the scene and to annotate the image with real meaning semantic labels. The user contribution in the whole process is considerable, but differs greatly in the two learning steps. For objects extraction, the training set must be consistent with the desired cover type. The LDA modeling instead is not a user driven process and is able to discover information that is not observed by the user in the training data set.

A WorldView2 image illustrating the city of Bucharest was used for an urban assessment and semantic annotation based on the proposed methodology. Classes expressing real meaning such as for instance: industrial area containing "parking areas nearby industrial buildings".

8. ACKNOWLEDGEMENTS

This research has been developed in the frame of the ESA-PECS project ROKEO – Romanian Centered Knowledge Earth Observation.

9. REFERENCES

[1] C.E. Shannon, "A Mathematical Theory of Communication", *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623-656, 1948.

[2] J.A. O'Sullivan, R.E. Blahut and D.L. Snyder, "Information - Theoretic Image Formation", *IEEE Transaction on Information Theory*, vol. 44, no. 6, pp. 2094-2123, October 1998.

[3] Y. Bengio, "Learning Deep Arhitectures for AI", *Foundation and Trends in Machine Learning,* vol. 2, no. 1, pp. 1-127, 2009.

[4] M. Schroder, H. Rehrauer, K. Seidel and M. Datcu, "Interactive Learning and Probabilistic Retrieval in Remote Sensing Image Archives", *IEEE Transactions on Geosciences and Remote Sensing*, vol. 38, no. 5, pp.100-119, September 2000.

[5] D. Bratasanu, C. Vaduva, I. Gavat and M. Datcu, "Latent knowledge discovery in satellite images", *ESA-EUSC 2011: Image Information Mining: Geospatial Intelligence from Earth Observation*, pp 29-32, March 2011.

[6] H. Daschiel and M. Datcu, "Design and evaluation of human machine communication for image information mining" *IEEE Transactions on Multimedia*, vol. 7, no. 6, December 2005..

[7] H. Daschiel and M. Datcu, "Image information mining system evaluation using information-theoretic measures", *EURASIP Journal on Applied Signal Processing, v*ol.14, pp.2153–2163, 2005

[8] C. Vaduva, D. Faur, I. Gavat and M. Datcu, "Data mining and spatial reasoning for satellite image characterization", *The 8th Communications International Conference*, Bucharest, Romania, COMM 2010, ISBN 978-1-4244-6361-9, pp. 173-176.

[9] P. Matsakis and L. Wendling, "A new way to represent the relative position between areal objects", *IEEE Trans. on Pattern Analysis and Machine Intel.*, vol. 21, no. 7, pp. 634-643, July 1999.

[10] D.M. Blei, A.Y. Ng and M.I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research 3, pp. 993-1022, March 2003.

[11] D. Yu, G. Hinton, N. Morgan, J.T. Chien and S. Sagayama, "Introduction to the Special Section on Deep Learning for Speech and Language Processing", *IEEE Transaction on Audio, Speech and Language Processing*, vol. 20, no.1, January 2012.

[12] I. Arel, D.C. Rose and T. Karnowski, "Deep Machine Learning – A New Frontier in Artificial Intelligence Research", *IEEE Comp.Intel. Magazine*, vol. 5, no. 4, pp. 13-18, Nov. 2010.

[13] www.digitalglobe.com.