Classification of Isolated Volcano-Seismic Events based on Inductive Transfer Learning

Manuel Titos, Angel Bueno, Luz García, Carmen Benítez, J. C. Segura

***Abstract*—Domain-specific problems where data collection is a expensive task, are often represented by scarce or incomplete data. From a Machine Learning perspective, this type of problems have been addressed using models trained in different specific domains as starting point for the final objective-model. The transfer of knowledge between domains, known as Transfer Learning (TL), helps to speed up training and improve the performance of the models in problems with limited amounts of data. In this paper we introduce a Transfer Learning approach to classify isolated volcano-seismic signals at *"Volcán de Fuego"*, Colima (Mexico). Using the well-known convolutional architecture (LeNet) as features extractor and a representative dataset containing regional earthquakes, volcano-tectonic earthquakes, long period events, volcanic tremor, explosions and collapses, our proposal compares the generalization capabilities of the models when we only fine-tune the upper layers and fine-tune overall them. Compared to others state-of-the-art techniques, classification systems based on Transfer Learning approaches provide good generalization capabilities (attaining close 94% of events correctly classified) and decreasing computational time resources.**

***Index Terms*—Transfer Learning, Deep Learning, volcano-seismic signals, classification of isolated events.**

I. Introduction

Seismic signals registered by seismometers in volcanic areas can be classified based on the source mechanisms (seismic events) that originated them [1]. This letter analyzes how Transfer Learning (TL) [2] can be used to speed up training and improve the performance of the models used to classify volcano-seismic signals.

Deployed in vulcanological observatories, Volcano-seismic signals Recognition systems (VSR) provide several advantages: (1) They reflect the nature and underlying physics of the source processes involved. (2) Analyzing these seismic streams, geophysicists can separate rapidly into classes a large number of events, which is important in case of eruptive crisis. (3) They provide consistent catalogues of each type of event improving the knowledge that we have about the state of the volcano. (4) The new knowledge obtained can be used to infer new eruptive crisis studying the temporal evolution of the volcano.

Despite the good performance obtained by the existing classification techniques in terms of time consumption and accuracy rate [3], [4], [5], [6], [7], [8], [9], the accurate recognition of certain kind of events remaining constrained due to the difficulty of creating both, well-labeled and statistically representative datasets [10], [5]. Hence, one of the most challenging objectives in volcano-seismology is the development of robust data pattern extraction mechanisms able to characterize properly each event.

Traditionally, feature engineering approaches based on the knowledge of human experts were used to extract relevant and discriminative information. However, newer approaches are based on deep hierarchical models that do not have to be supplied with such "hand-crafted" features. They can learn representative features from raw data. These new approaches have become the state of the art in many disciplines, improving the traditional ones, at the cost of a much greater demand of training material and computational resources [11].

Given the vast amount of data, computation and time resources required to develop deep hierarchical models, an emerging approach is to exploit what has been learned in one domain (where a lot of labeled training data is available) to improve generalization in another domain where data are scarce. This is what in the related literature is known as Transfer Learning [2], alluding to the fact of the translation of knowledge acquired in a domain to a different one. Instead of starting the learning process from scratch, the basic idea is to use the parameters of a well-trained model in one domain (original domain) as a pre-trained version for a model in a different domain (in which there are much less training data available). After that, pre-trained parameters are fine-tuned using domain-specific available data (in the target domain). Transferring the knowledge acquired in the original domain to the target domain to be used as a starting point for the training of the models.

In this letter, we use LeNet architecture [12], designed for handwritten and machine-printed character recognition, as features extraction algorithm to build a system for automatic classification of volcano-seismic events. Input data, composed by spectrograms, will be processed by LeNet model resulting in a feature vector that will later be used to train several multilayer perceptrons (MLP).

The main contribution of this work is to show the applicability and potential of using hierarchical feature representations obtained by models trained in a different specific problem as efficient information to build a system for automatic classification of volcano-seismic events. Our proposal re-trains the model, keeping the spatial and spectral information extracted by pre-trained model in a different domain as input information.

The rest of the paper is organized as follows: Section II provides a theoretical framework of Transfer Learning approaches, and how it can be used for discriminative feature modelling of volcano-seismic events. Section III describes from the geophysical point of view, the seismic signals registered at *"Volcán de Fuego"*. Section IV describes the experimental setup and presents the results and discussion. Section V concludes the study.

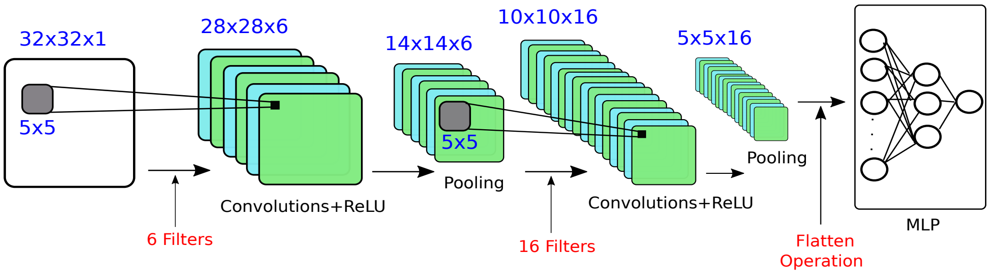

Figure 1: LeNet architecture

II. Theoretical foundations and Related Works

One of the most promising techniques that could someday increases the capacity of generalization of Artificial Intelligence is the transfer of knowledge from an environment (domain) to other environment (domain), widely know as Transfer Learning (TL) [2]. As we mentioned above, TL tries to exploit knowledge that has been learned in one task to improve generalization in a different but related task. Following [2], a domain consists of two components, a feature space and a marginal probability distribution. Given a specific domain, a task (learning problem to be solved) can be defined as a label space and an objective predictive function which will be learned from the training data. Therefore, based on the different relations between domains and tasks, TL can be categorized into inductive transfer learning (ITL), transductive transfer learning (TTL), and unsupervised transfer learning (UTL). Considering the nature of our proposal, we will only describe the inductive approach.

In ITL, target and source task are different. The domains of this two tasks may differ. In this case, as the purpose of both classification tasks differs, some labelled data are required in the target domain to induce its particular predictive model. The parameters of previously trained models (source) can be seen as a starting point of a new developing model, where the later layers of the original model are fine-tuned using available domain-specific data. This approach is based on the idea that low-level features (earlier layers) contain generic information (edge detectors, color regions detectors, etc), while progressively, the middle and the later ones, extract shapes and some task-specific features respectively [13]. Therefore, given computation and time resources required to develop new models from scratch, ITL has become a very useful solution in areas as Computer Vision (CV) [14], Natural Language Processing (NLP) [15] or Automatic Speech Recognition (ASR) [16] in order to speed up training and improve the performance of the models.

Applied to geoscience disciplines, TL has been found to be helpful in domain adaption problem as hyper-spectral images analysis [17], remote sensing data classification [18], wind speed prediction [19] and cyclone tracking [20], among others.

III. Data and Methods

This section describes the dataset used in the study and the proposed architectures used for the experimental setup.

A. *Proposed Architectures.*

Given that convolutional neuronal networks (CNNs) [12] have proven great success in fields such as CV, NLP or ASR, it was decided to incorporate some of the most accurate models as base of our classification system to finally adjust their final layers with our data set.

In this sense, using the spectrogram images as parameterization scheme, we proposed to use LeNet network [12] as base model.

Basically, LeNet network is a CNN with 7 levels of depth trained with MNIST data set to classify handwritten and machine-printed character images of 32x32 pixels in gray scale (Figure 1). The model consist of several convolution layers, each of them followed by a max pooling operation:

- Each convolutional layer can be understood as feature extractor taking as inputs the outputs from its previous layer in the hierarchy. It takes as input a stack of input planes and produces as output some number of output planes know as feature maps. At the same time, each feature map O_k can be understood as a arranged map of responses of a spatially local non-linear operation, applied identically over the whole input planes. The main building block used to construct the non-linear transformation is the convolution operation. Hence, each feature map O_k is associated with one kernel and computed as follows:

$$O_k = \sigma(b_k + \sum_r W_{kr} * X_r) \tag{1}$$

  Being $X_r, W_{kr}, *, b_k, \sigma$ the r-th input channel, the subkernel for that channel, the convolution operation, the bias term, and the element-wise non-linearity (sigmoid, hyperbolic tangent or ReLU) applied to the result of the kernel convolution, respectively.

- The pooling step can be understood as down-sampling operation along the spatial dimensions (width and height). Thus, max pooling operation consists of substituting each sub-window of size pxp by the maximum feature value in it. This procedure can be formalized as follows:

$$H_{k,ij} = \max_p (O_{k,Si+p,Sj+p}) \tag{2}$$

  Where p and S determine the pooling window size and the stride value which corresponds to the horizontal and vertical increments at which pooling sub-windows will be positioned.

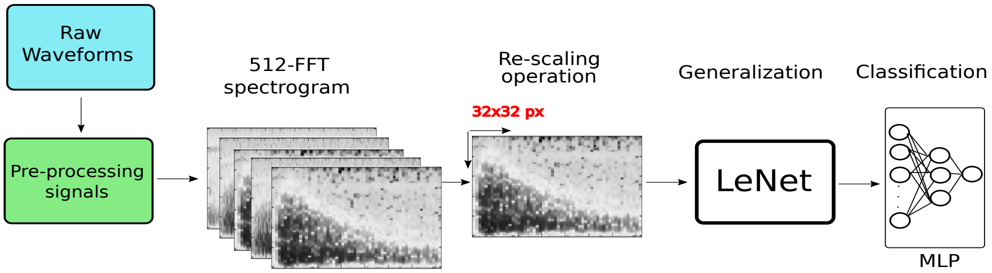

Figure 2: Overview of the data pre-processing pipeline. First, signals are band-pass filtered between 1 and 25 Hz. For each signal, we obtain its spectrogram using a FFT of 512 points. Finally, each spectrogram is resized to 32x32 and transformed to grayscale.

- The final extracted features are flattened and used as input vector in order to feed one (or even two) fully connected layers added in the end.

The basic idea behind LeNet architecture [12] is that the earlier convolutions are able to extract lower features as generic information (edge detectors, color regions detectors, etc) while, later convolutions are specialized on higher-level features as specific shapes.

The topological structure in the first step uses a bank of six 5x5 filters with stride 1. This filter design results in six feature maps of 28x28 pixels. The pooling operation (using a filter width of 2 and stride of 2) reduces the dimension by factor of 2 and ends up with six feature maps of 14x14 pixels. The second step applies another bank of sixteen 5x5 filters, resulting on sixteen feature maps of 10x10 pixels. Again, applying the same pooling operation, sixteen feature maps of 5x5 pixels are obtained.

Once the features have been extracted, they are flattened into a 1-D vector in order to feed one (or even more) fully connected layers with specific number of nodes. Finally, on top, we add a softmax layer to normalize per-class output probabilities corresponding to each of the available events. The final number of parameters is approximately 60K, including the parameters associated to filters design and fully-connected layers.

B. Dataset.

The database used to test TL architectures proposed consists of 9332 volcano-seismic signals distributed per-class as follows: 1738 volcano tectonic earthquakes (VTE), 2699 long period events (LPE), 1170 volcanic tremors (TRE), 455 regional earthquakes (REG), 1406 collapses (COL), 278 explosions (EXP), and 1586 volcanic noise (NOISE). Following [21] and [5] each type of event can be described according to its properties (source mechanism, length, frequency content):

Volcano tectonic earthquakes (VTE)): VTE events are originated by seismic stress when a solid fracture takes place producing a seismic wave; it is possible to identify the P waves (pressure) and S waves (shear) arrivals. Spectral content could reach up to 30Hz.

Regional Earthquakes (REG): These earthquakes occur outside the volcanic structure and are related to tectonic stresses and to fault fractures. They can have larger duration and magnitude than VTE, but similar spectral content. P and S-waves arrivals are generally clear.

Long period events (LPE)): Their sources models are generally associated to the resonance of fluid-filled cavities such as cracks or magmatic conduits, in the shallow part of the volcano. Their spectra usually present one or several dominant peaks below 5 Hz.

Volcanic tremor (TRE)): Its spectral content is below 5Hz, and the duration is highly variable, lasting from a few minutes to months. Volcanic tremor is a sign of high activity inside the volcano. Some theories suggest that it is caused by the movement of magma or gas, being almost identical to long-periods events, except for the duration.

Explosions (EXP)): They are characterized by variable duration (from second to tens of minutes) and a distinctive spectrogram with a narrow energy peak around 20 Hz. Explosions are naturally related to sonic boost waves, produced when the expanding gas is accelerated within the volcano structure.

Lava Flow (COL)): Volcanic debris processes located at the volcano surface exhibiting frequency content above 5 Hz.

Environmental Noise (NOISE): Mainly introduced by nearby populations, human activity will interfere the frequency range where most of the volcanic spectral content is located.

IV. EXPERIMENT AND DISCUSSION

This section illustrates the performance of the proposed method. We compare the results obtained with others methods and parameterization schemes in terms of classification accuracy.

A. MODEL TRAINING

Following [5] the feature extraction process is summarized in Figure 2. The input of the model is the full dataset of 9332 seismic signals (belonging to the station EZ5V4) in the time domain, sampled at 50 Hz and pre-processed using a band-pass filter between 1 Hz and 25 Hz.

After the pre-processing stage, a data-set of 9332 spectrogram images using short-time Fourier Transform (FFT) of 512 points is obtained, with their associated labels. In order to extract representative features using LeNet architecture, we need to adapt the input dimensions of the

Table I: Classification results obtained by different architectures. 1 FC correspond to one fully connected layers. L and A correspond to models where the Last and All layers were fine-tuned. CNN-128 and CNN-512 correspond to models with similar topology to LeNet but without TL stage.

#Model	#Parameterization	#Topology	#Acc(%)	# Speed up(%)
SVM	LPC+Statistical information	RBF Kernel	91.55	-
SVM	LPC+Statistical information	Lineal Kernel	92.32	-
RF	LPC+Statistical information	120 estimators	92.80	-
MLP	LPC+Statistical information	500 hidden units	93.57	-
sDA-2H	LPC+Statistical information	260-385	94.32	-
CNN-LeNet 128 L	**512 FFT Spectrogram**	**LeNet + 1 FC layer (128 units)**	**88.3**	56.5
CNN-LeNet 512 L	**512 FFT Spectrogram**	**LeNet + 1 FC layer (512 units)**	**89.3**	36.4
CNN-LeNet 128 A	**512 FFT Spectrogram**	**LeNet + 1 FC layer (128 units)**	**93.4**	11.3
CNN-LeNet 512 A	**512 FFT Spectrogram**	**LeNet + 1 FC layer (512 units)**	**94.1**	32.3
CNN-128	**512 FFT Spectrogram**	**LeNet + 1 FC layer (128 units)**	**92.5**	-
CNN-512	**512 FFT Spectrogram**	**LeNet + 1 FC layer (512 units)**	**93.0**	-

images at 32x32 px. For that, spectrogram images are rescaled. Finally, the flattened features from LeNet architecture will be used as training data for the classifier. As we are working with different streams of data, a direct consequence of normalizing all images to the same dimensions is that the longer signals lose more information. However, as we shall see later, the large size of the dataset used to train LeNet architecture minimizes the impact of this fact.

Considering the inductive nature of our proposal, we compare the result obtained by the models training all the layers and the later fully connected one. To do that, we use a sigmoid function as activation function and Adam optimizer [22] to optimize the loss function (Negative Log-Likelihood). The over-fitting scenarios are controlled during training using a validation set and early stopping criteria with a patience of 10 iterations. The dataset was divided into training (75%) and test (25%) sets. This yields a training set of 7000 training instances, and 2332 test instances. Furthermore, we used 50% of the test set (1166 instances) as validation data [5]. Other techniques as dropout or batch normalization did not offer any improvement. All the experiments were carried out using cross-validation with four partitions of the original database.

B. CLASSIFICATION OF ISOLATED EVENTS

The basis of the comparative study was derived from [5] where volcano-seismic signals were classified by several classical and deep architectures. As classical approaches, this work used MLP [23], Random Forest (RF) [24] and Support Vector Machines (SVM) [25] with both, linear and radial kernels. Regarding the deep ones, the architectures tested were Deep Belief Networks (DBN) [26] and stacked Denoising AutoEncoders (sDA) [27]. The parameterization scheme was based on linear prediction coefficients (LPC) and statistical information associated to impulsivity of the signals.

The best results obtained for different architectures and parameterization schemes are summarized in Table I. In order to prove the efficiency of the TL-based models against the version trained from scratch, we measure the relation between the runtime of both training algorithms according to :

$$speedup = \frac{TL_{line} - base_{line}}{base_{line}} \quad (3)$$

Where $base_{line}$ is the runtime spent without TL version and TL_{line} is the total runtime achieved with TL version. The reported metrics are based on accuracy. Only the best results obtained after testing several configurations varying different units at first and second hidden layers have been reported.

Compared to TL approach, several conclusions can be drawn from these results:(1) By applying the trained model as feature extraction, we notice that ITL provides useful features for the discriminative stage, outperforming hand crafted ones applied to shallow classical classifiers (SVM, MLP, RF). (2) Although the results obtained do not improve those obtained by deep networks using a specific hand crafted-parameterization [5], they are really promising, most especially considering the vast amount of data, computation and time resources required to develop deep hierarchical models from scratch. (3) Compared to specific features based on signal processing approaches, the ones extracted from spectrogram images have proven to be very useful for the classification of isolated seismo-volcanic events. (4) Given the large number of images of very different quality used to train LeNet architecture, the rescaling operation of spectrogram does not have an undesired effect. The hierarchical features obtained from resized images focus mainly on the contours and shapes of the spectrograms proving to be sufficiently discriminative.

Moreover, it should be pointed out that: (a) the dimension and resolution of the spectrogram images often have a big influence on the performance of the systems. Therefore, changing both, or even the number of input channels, models could obtain better discrimative information, improving the performance in learning and classification tasks. However, the use of higher resolution, even if the models use the same filters design, will result in larger feature maps increasing the number of parameters to be tuned and therefore, affecting the size of the dataset necessary to guide the optimization process. (b) Given the size of the dataset used in this work, we noted that the inclusion of more than one hidden layer between flattened and softmax layers degrades the performance. This degradation may be due to an over-fitting problem. The convolutional features extracted are very representative. Thus, the inclusion of new non-linear transformations lead the system to model noise and memorize rather than to generalize data.

V. Conclusion

In this work we present the use of Inductive Transfer Learning as knowledge base from which to build reliable and efficient volcano-seismic classification systems. Based on the results obtained, we conclude that the use of previously adjusted CNN and, more specifically, the hierarchical learning representation that they implement, can be efficiently exploited in the classification of isolated seismo-volcanic signals, taking advantage of the invariance and locality characteristics that convolution operations offer.

VI. Acknowledgements

The data used in this study have been obtained by the Seismological Network of Colima State (RESCO). They have been provided to the authors with the agreement of Gabriel Reyes Dávila and Raúl Arámbula. This research was funded by TEC2015-68752-R (MINECO/FEDER).

References

[1] B. Chouet. "Volcano seismology". In: *Pure and Applied Geophysics* 160.3-4 (2003), pp. 739–788.

[2] S. Pan et al. "A survey on transfer learning". In: *IEEE T KNOWL DATA EN* 22.10 (2010), pp. 1345–1359.

[3] M. Masotti et al. "Application of Support Vector Machines to the classification of volcanic tremor at Etna, Italy". In: *GEOPHYS RES LETT* 33.20 (2006).

[4] A. Esposito et al. "A neural approach for hybrid events discrimination at Stromboli volcano". In: *Multidisciplinary Approaches to Neural Computing*. Springer, 2018, pp. 11–21.

[5] M. Titos et al. "A Deep Neural Networks Approach to Automatic Recognition Systems for Volcano-Seismic Events". In: *IEEE J-STARS* 11.5 (2018), pp. 1533–1544.

[6] R. Carniel et al. "Detecting dynamical regimes by Self-Organizing Map (SOM) analysis: an example from the March 2006 phreatic eruption at Raoul Island, New Zealand Kermadec Arc." In: *B GEOFIS TEOR APPL* 54.1 (2013).

[7] AD. Jolly et al. "Seismo-acoustic evidence for an avalanche driven phreatic eruption through a beheaded hydrothermal system: an example from the 2012 Tongariro eruption". In: *Journ. Vol. Geoth. Res.* 286 (2014), pp. 331–347.

[8] M. Beyreutherand et al. "Hidden semi-Markov model based earthquake classification system using weighted finite-state transducers". In: *NONLINEAR PROC GEOPH* 18.1 (2011), p. 81.

[9] M. Titos et al. "Detection and classification of continuous volcano-seismic signals with recurrent neural networks". In: *IEEE T GEOSCI REMOTE* 57.4 (2019), pp. 1936–1948.

[10] G. Cortés et al. "Evaluating robustness of a HMM-based classification system of volcano-seismic events at Colima and Popocatepetl volcanoes". In: *Geoscience and Remote Sensing Symposium,2009 IEEE International,IGARSS 2009*. Vol. 2. IEEE. 2009, pp. II–1012.

[11] T. Elsken et al. "Neural Architecture Search: A Survey." In: *J MACH LEARN RES* 20.55 (2019), pp. 1–21.

[12] Y. LeCun et al. "Gradient-based learning applied to document recognition". In: *P IEEE* 86.11 (1998), pp. 2278–2324.

[13] J. Yosinski et al. "How transferable are features in deep neural networks?" In: *ADV NEUR IN*. 2014, pp. 3320–3328.

[14] S.Hoo-Chang et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE T MED IMAGING* 35.5 (2016), p. 1285.

[15] J. Huang et al. "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers". In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, pp. 7304–7308.

[16] J. Deng et al. "Autoencoder-based unsupervised domain adaptation for speech emotion recognition". In: *IEEE Signal Processing Letters* 21.9 (2014), pp. 1068–1072.

[17] H. Yang et al. "Domain adaptation with preservation of manifold geometry for hyperspectral image classification". In: *IEEE J-STARS* 9.2 (2016), pp. 543–555.

[18] D. Tuia et al. "Domain adaptation for the classification of remote sensing data: An overview of recent advances". In: *IEEE geoscience and remote sensing magazine* 4.2 (2016), pp. 41–57.

[19] Q. Hu et al. "Transfer learning for short-term wind speed prediction with deep neural networks". In: *Renewable Energy* 85 (2016), pp. 83–95.

[20] S. Ho et al. "Automated cyclone tracking using multiple remote satellite data via knowledge transfer". In: *Aerospace conference, 2009 IEEE*. IEEE. 2009, pp. 1–7.

[21] R. Arámbula-Mendoza et al. "Seismic activity that accompanied the effusive and explosive eruptions during the 2004–2005 period at Volcán de Colima, Mexico". In: *J VOLCANOL GEOTH RES* 205.1-2 (2011), pp. 30–46.

[22] D.Kingma et al. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[23] I. Goodfellow et al. *Deep learning*. MIT press, 2016.

[24] A. Liaw et al. "Classification and regression by randomForest". In: *R news* 2.3 (2002), pp. 18–22.

[25] C. Burges. "A tutorial on support vector machines for pattern recognition". In: *DATA MIN KNOWL DISC* 2.2 (1998), pp. 121–167.

[26] G. Hinton et al. "A fast learning algorithm for deep belief nets". In: *NEURAL COMPUT* 18 (2006), pp. 1527–1554.

[27] V. Pascal et al. "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion". In: *J MACH LEARN RES* 11.Dec (2010), pp. 3371–3408.