Overview of the 2017 RedICA Text-Image Matching (RICATIM) Challenge

Luis Pellegrin, Hugo Jair Escalante, Alicia Morales, Eduardo F. Morales, and Carlos A. Reyes-García
Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Tonantzintla, Mexico
{pellegrin,hugojair,a.morales,emorales,kargaxxi}@inaoep.mx

***Abstract*—This paper describes the design and analysis of results of the 2017 RedICA: Text-Image Matching (RICATIM) challenge. This academic competition faces the image labeling problem (assigning words to images) as one binary classification. Motivated by recent success of representation learning, we built a data set for binary classification in which each instance is the learned representation of a pair of an image and a word. Instances are labeled as positive, if the word is relevant for describing the content of the image and negative otherwise. Thus, participants of the challenge had to develop binary classification methods to distinguish between relevant and irrelevant text-image matchings. The challenge attracted 43 participants, that provided quite original and competitive solutions. The performance obtained by the top ranked participants was impressive, improving the performance of the baseline considerably. In this paper we describe the approached problem, the challenge design (including data and evaluation protocol), and provide an overview of the results achieved by participants.**

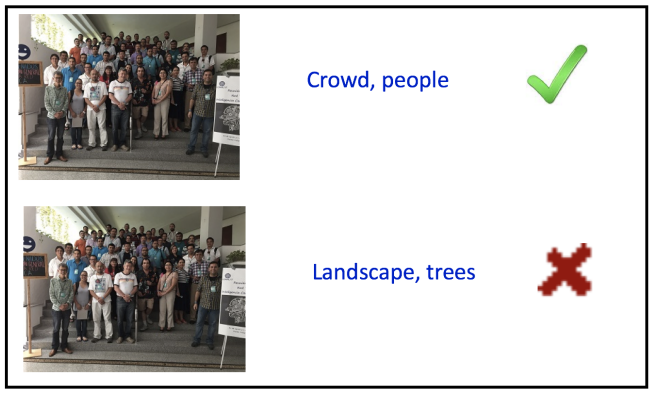

Fig. 1. Sample of a matching correct between image and text.

I. Introduction

The goal of Automatic Image Annotation (AIA) is to assign keywords to images describing their visual content. AIA is an area of constant development due to great applicability in many tasks involving information systems. A prevalent way to address the AIA task is based on supervised learning, defining the task as a general classification problem where words are seen as classes and one or more labels are assigned to each image. However, this kind of scenarios does not offer scalability in the label assignation process, because they only can assign a few labels (those labels associated to the class) from thousands of alternatives, e.g. see ImageNet [1]. Instead, unsupervised AIA (UAIA) rely on text mining methods that process collections of weakly labeled images (e.g., webpages and the images they contain) to assign free-vocabulary labels to images. The main limitation of UAIA being the inherent noise in weakly labeled images. Both approaches have complimentary benefits and limitations (see e.g., [2]), in this paper we focus in a hybrid approach to AIA, where unsupervised learning is used to assign free-vocabulary labels to images.

On the other hand, significant progress has been achieved in machine learning and pattern recognition, leading to a great variety of modeling techniques, such as those based on deep learning [3], and distributional representations [4]. Both approaches, successfully dominate computer vision [5] and natural language processing [6] fields, respectively. Of particular interest are methods for representation learning[1], that allow us to automatically learn representations for data that may show discriminative capabilities, e.g. see [7] for images, and [8] for texts.

Considering the aforementioned strengths of representation learning methods in vision and language, we approach the AIA task as a text-image matching problem based on learned features. In a nutshell, both images and labels are represented with state of the art representation learning methods (using pretrained models). Then, we generate instances for a binary classification task that consists in determining whether a pair of text-image (the concatenation of their learned representations) is a relevant match. Where, by relevant match we mean that the label is relevant for describing the visual content of the associated image, see Figure 1. In this way, virtually any word (for which we can generate a distributed representation) could be used for describing the content of images. Overcoming the limitation of traditional supervised AIA methods. On the other hand, methods for this task rely on pairs of ground truth labellings, which are cleaner than those used in unsupervised AIA. With this formulation, that can be considered hybrid between supervised and unsupervised UAIA, we aim to alleviate the limitations of both approaches. To the best of our knowledge this is the first work approaching the AIA problem in this way.

[1]Representation learning refers to methodologies for automatically discovering data representations for different tasks (e.g., classification) from raw data and without human intervention [3].

978-1-5386-0819-7/17/$31.00

Motivated by this text-image matching problem, we organized an academic challenge around it. Our overall aim was to explore the feasibility of this new image annotation scheme. We generated a challenging data set for the task and asked participants to develop solutions for the proposed problem. In addition to data, we provided an evaluation protocol and prizes to motivate participation. The outcome of the challenge was quite promising: participants developed solutions that achieved quite competitive performance, evidencing the feasibility of the task. This paper describes the challenge design and summarizes its results and main findings. Overall, the challenge was quite competitive and encouraging. We plan to organize future editions of this challenge in the short term.

The rest of this paper is organized as follows. Section II provides an overview of the challenge. Section III presents the evaluation results. Finally, Section IV outlines conclusions derived from this work.

II. The RICATIM Challenge

This section provides an overview of the the RICATIM challenge. The challenge was run on the CodaLab[2] platform, and had a duration of about 45 days. This challenge was organized and sponsored by *RedICA*[3]*: Red Temática CONACyT en Inteligencia Computacional Aplicada*, and it is expected to be the first of a series of periodic challenges organized by this academic networks. Results of the challenge and winners will be announced with ENAIC/SNAIC[4] 2017. The rest of this section provides details on the design and organization of the challenge.

A. Approached problem

The aim of this challenge is to assess the feasibility of approaching the AIA problem as one of binary classification. Participants were provided with a data set for a binary classification problem in which the feature space of each instance encodes a text-image (keyword-image) pair, see Figure 2. Where the class of the instance is 1 when the keyword is relevant for describing the image, and 0 otherwise. The keyword to image relevance was determined with an undisclosed methodology for the participants, that may not be intuitive to them, even if they have access to images and labels (i.e., a keyword may be relevant even if it is not an object visually observable in the image). Images in the data set were represented by Convolutional Neural Network (CNN) based features, whereas keywords were encoded with their word2vec representation, see next section for details. The class of each instance was determined with three variants that we will keep undisclosed to some extend because further rounds of this challenge will be organized. Additionally, participants were provided with the raw images and the actual words, so that participants can take advantage of such information.

[2]https://competitions.codalab.org
[3]http://redica.mx/
[4]ccc.inaoep.mx/SNAIC/2017

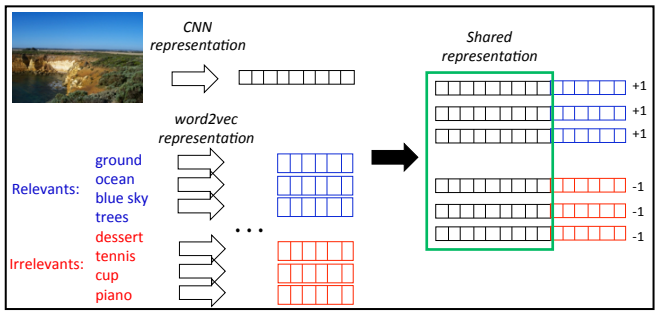

Fig. 2. General framework used for generating positive/negative instances of the approached problem. The representations of images and words are concatenated to generate the instances of the binary classification task. The rectangle shows the visual part of each instance.

B. Data

To create the RICATIM data set, initially a set of 3,300 images was taken from the IAPR TC-12 [9], [10]. IAPR TC-12 consists 20,000 real world images taken from tourism travel agencies, i.e. see (a)-(d) in Figure 3. Every image has a caption/description manually annotated, i.e. see (f) in Figure 3. Moreover, the segmented and annotated IAPR TC-12 (SAIAPR TC-12) benchmark [10] is an extended version where images were manually segmented and annotated at the region level (i.e. see (e) in Figure 3) with 250 keywords approximately (arranged hierarchically). We used labels information from both the IAPR TC12 data set and its extended version.

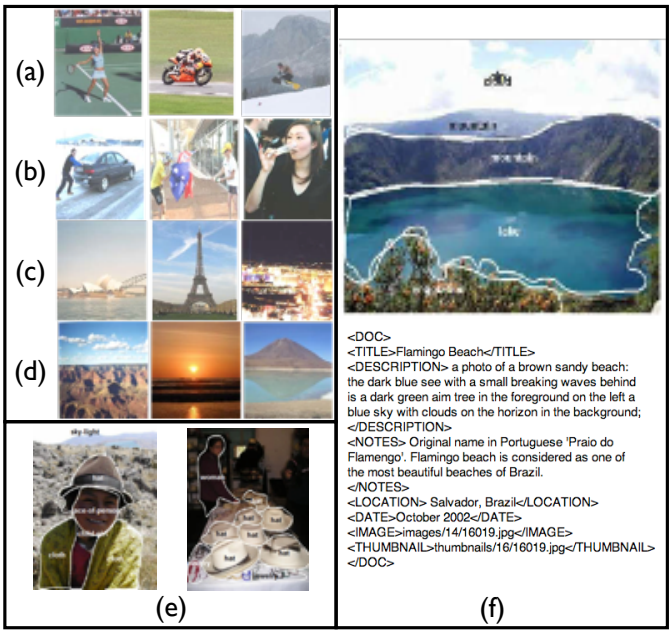

Fig. 3. Sample of images from [9], [10].

As previously mentioned, an instance in our data set consists of a concatenation of a visual and a textual representation (see Figure 2). For representing images, we used a CNN-based representation extracted by using a pretrained deep network:

each image was preprocessed and passed through a pretrained 16-layer CNN-model [11], penultimate layer activations were used as visual representation (a vector of 4096 elements). On the other hand, keywords were represented by their word2vec representation [8] (200-dimensional vectors were considered). Word2vec representations were obtained by using Wikipedia as training collection.

In order to generate the labels of instances for the data set, the initial image set was divided into three sets: $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$. Each set containing the same images, but using a different strategy for generating labels. Accordingly, three different approaches for generating labels were adopted: One manual and two semi-automatic ones. In the following, we describe the three different approaches used to generate positive/negative instances. For each approach, we concatenate the representation of the image and that of the labels.

1) *Region-level labels.* For the set $\mathcal{X}$, we used the labels assigned to images according the the considered data set. As we mentioned before, each image taken from SAIAPR TC-12 is annotated with n labels out of 250 possible (region-labels were associated to the global image they appear in). For a given image $i \in \mathcal{X}$, the generation of positive instances (relevant text-image pairs) was straightforward: manually assigned labels to image i were considered as relevant. The labels for negative instances (non relevant text-image pairs) were produced by taking labels randomly from the semantically-farthest keywords to the manually assigned labels. Where the semantic distance was estimated by the distances among the word2vec representations of labels. For each positive instance a negative one was generated, taking care that labels used as negative were different from the positives. A different number of instances was created from each image, depending of the number of labels assigned, having an average of eight labels (positive and negatives) per image.
2) *Annotated captions.* For the $\mathcal{Y}$ set we used captions assigned to images for generating relevant and irrelevant text-image pairs, see (f) in Figure 3. First, captions were indexed with a bag-of-words (BoW), then a TF-IDF weighting scheme was applied. For generating positive instances, given an image $i \in \mathcal{Y}$ we considered as relevant labels to those words from the caption with higher TF-IDF value. One negative instance was created for each positive one, with a similar strategy as described before. This time, as the vocabulary extracted from the captions is large, i.e. 7,708 different terms, the terms used as negative were not taken from the last positions. Instead, using the word2vec corresponding to whole extracted vocabulary, a matrix of cosine distances among these vectors was calculated. Thus, empirically a distance range was chosen in order to take terms to be used as negative, i.e. the negative label for a given positive label was taken randomly from the nearest 200-400 labels to the positive one (we found that in these range irrelevant but related words can be found). A different number of positive/negative instances was created from each image, in a random range between three and six labels, at the end, this $\mathcal{Y}$ set had an average of 10 labels per image.
3) *Unsupervised Automatic Image Annotation.* In this case, we used an UAIA method proposed in [2] to generate relevant and irrelevant text-image pairs for set $\mathcal{Z}$. As previously mentioned UAIA methods can annotate images with labels using large vocabularies. For generating relevant pairs in set $\mathcal{Z}$, we took as relevant labels to the top 3-6 labels generated with the UAIA method, negative labels were determined according to the same methodology as before. This set had an average of 10 labels per image.

From the three approaches used in the methodology a total of 31,128 instances were generated. Next, the following disjoint partitions were created with 30,000 instances taken equally from the generated instances (from approaches described above):

- Training data (labeled data, can be used to train and develop models). This partition is formed by 20,000 instances, where 40% are positive instances and 60% are negative instances.
- Validation data (unlabeled data, participants can make predictions during phase 1 to get immediate feedback in the leaderboard). This partition is formed by 5,000 instances, where 60% are positive instances and 40% are negative instances.
- Test data set (released in the final phase, performance on test data will be used to determine the winners). This partition is formed by 5,000 instances, but this time 50% are positive instances and 50% are negative instances.

C. Schedule

The challenge comprised two stages: development and final:

- **Development Phase:** Participants had access to labeled development (training) and validation data, with ground-truth labels and raw data (images and words). During this phase, participants could receive immediate feedback on their performance on validation data through the leaderboard in CodaLab.
- **Final Phase:** The unlabeled final (test) data and its raw data corresponding, was provided along with validation labels. Therefore, participants were able to use labeled validation data for training your models, and then submitting predictions for the test data. The participants also had to send their code and fact sheets describing their methods. All participants code was verified and replicated prior announcing the winners.

To be eligible for prizes, the winners had to publicly release their code and worksheets.

The timeline for the challenge was as follows:

- *3rd July, 2017:* Beginning of the challenge, release of development and validation data.

- *14th August, 2017:* Release of test data and validation ground truth labels.
- *16th August, 2017:* Submission deadline for prediction in test set. Release of worksheet template.
- *18th August, 2017:* Submission deadline for code and fact sheets.
- *19-23th August, 2017:* Verification phase (code and fact sheets).
- *24th August, 2017:* Winners notification.
- *4-8th September, 2017:* Presentation of results at the ENIAC-SNAIC and award winners.

D. Baseline Systems

Two baseline methods were implemented, the first one was simple enough to give the participants a wide margin for improvement. This baseline was a basic classification using LIBLINEAR [12] without performing any preprocessing in the training data set. In order to establish another point of comparison, a second baseline was defined using the random forest implementation from the CLOP[5] toolbox.

E. Metrics and Evaluation

For the challenge evaluation, we used the classic metrics of accuracy and F1, the former being the measure that was used to officially rank participants. Accuracy is a well known statistical measure widely used in a binary classification. On the other hand, F1 is the harmonic mean of the precision p and the recall r of the binary classification problem.

III. Analysis of Results

This section is divided into three subsections that together provide a comprehensive analysis of the challenge.

A. Participation

The participation was numerous, especially considering that this was the first edition of the task and that the challenge lasted slightly more than one month. In total, 43 participants took part, making more than 220 submissions to the leaderboard. The following teams participated in the RICATIM challenge:

- **I3GO+.** A collaboration among the INFOTEC-UMSNH, INFOTEC-UP, CONACYT-CentroGEO, CONACYT-INFOTEC, all institutions in Mexico, it was formed by Jose Ortiz Bejar, Claudia Sanchez, Daniela Moctezuma, Sabino Miranda, Eric Sadit Tellez and Mario Graff.
- **MIGUE, TAVO & ANDRES.** A collaboration between the *Instituto Nacional de Astrofísica, Óptica y Electrónica* and the *Instituto Tecnológico de Monterrey* formed by Octavio Loyola, Miguel Medina and Andrés Gutierrez.
- **MindLab.** The team from *Universidad Nacional de Colombia (Bogotá, Colombia)* formed by Jorge Vanegas and Victor Contreras.
- **Voltaire Project.** Team formed by Mauricio García.
- **Argenis.** The team formed by Argenis Aroche.
- **Other**. Nickname of users[6]: naman, miguelgarcia, barb.

B. Methodologies

Table I provides a comparison of the methods proposed by each team[7]. In general all teams used the features provided by the challenge, some of them introduced additional resources. We can see a wide range of techniques for data preprocessing and dimensionality reduction. For instance, I3GO+ team used its own preprocessed features extracted from the provided textual and visual information. Whilst, other teams relied on the application of PCA and LDA. The considered classifiers were quite diverse as well, although ensemble learning achieved the best results. In the remainder of this section we summarize the methods of the four teams that submitted results for the test phase, followed by their obtained results.

First place (I3GO+): The winner team *pulls out all the stops* by offering a framework that integrates heterogeneous features in order to capture as much information as possible from the raw data. Followed by dimensionality-reduction transformations based on the farthest-first transversal (F-FT) method with Gaussian kernel. And finally, a complete pipeline composed by ensembles of classifiers including approaches from the five main paradigms of machine learning, i.e. SVM, neural networks, genetic programming, Naïve Bayes and k-

[5]http://clopinet.com/CLOP/

[6]Only those that submitted at least one results.

[7]Only those that submitted results to the test phase.

TABLE I
Comparison of the systems developed by each team.

Team	Data processing highlights	Final representation used	Method highlights
T1. I3GO+	Diverse feature representations were extracted and transformed with a method based on farthest-first traversal with Gaussian kernel.	JSON containing 7251 textual features some of them obtained with μTC, 1536 visual features from descriptors, among others.	An ensemble of k-NN/EvoDAG classifiers over heterogeneous schemes of representation.
T2. MIGUE, TAVO & ANDRES	Extraction of emerging patterns by Bhattacharya as function for evaluating candidate splits over ensemble of decision trees	Same representation provided.	Ensemble of decision trees using Bhattacharya distance, followed byPBC4cip for classification.
T3. Voltaire Project	Reduction and feature extraction by using PCA (principal component analysis) and. LDA (linear discriminant analysis)	Representation obtained by PCA where only are considered features that generate less than 90% of variation.	k-NN based on batch learning.
T4. Argenis	A normalization was performed over each feature in the distance function.	Same representation provided.	Lazy learning based in Kmeans and k-NN.

NN. We now describe the techniques used, as well as, some of the findings in this work:

1) Preprocessing. Two techniques were considered: μTC [13] for textual features, and a combination of HoG (histograms of oriented gradients) and LBP (local binary patterns) for visual features (e.g. see a prior work in [14]).).
2) Dimensionality reduction. Using the mentioned features, the team proposed a novel technique for dimensionality reduction called F-FT. It uses a Gaussian kernel and offers a fast and competitive alternative to PCA and other dimensional reduction method. However, it is much faster that kernelized PCA and still supports non linear transformations. Moreover, KPCA (kernelized principal component analysis) was tested as a second dimensional reduction method. Several vector spaces were obtained applying both techniques and at the end they kept the smaller and best performing representation. The complexity of this method is $\Theta(c*n+dim*n)$ where c is the number of coordinates of the reduced representation, dim is the number of nonzero coordinates in the explicit representation, and n the size of the training set.
3) Classification. Here, they selected the best classifiers among a large number of configurations by using a 70%-30% partition of the data set to estimate performance. Classifiers were assembled with a simple voting scheme. The best performances were achieved by k-NN and the EvoDAG genetic programming system, so at the end only these methods were considered. The unexpected superiority of k-NN seems to be linked with the dimensionality reduction technique and the use of cosine similarity. The complexity for k-NN is $\Theta(k*n*log_2 k)$ where the k is for the number of nearest neighbors used, while $log_2 k$ corresponds to the priority of size k. On the other hand, the time to calculate the EvoDAG parameters is large, so they were calculated with the validation data and were used in both phases: validation and test.

Second place (MIGUE, TAVO & ANDRES): This team exploited the provided data through a solid framework divided in two phases: (1) training/filtering, and (2) classification. At the first phase, it is proposed a novel method for filtering useful patterns by using ensemble of decision trees guided by the Bhattacharya method. Candidate splits are evaluated by bagging decision trees for extracting emerging patterns, e.g. see their prior work [15]. Afterwards, in the classification phase, they used a new contrast pattern-based classifier for class imbalance problems (PBC4cip) [16]. At training, the proposed method has a complexity of $\Theta(d*dim*n*log_2(n))$ where dim is the number of features used for building the decision trees, n is the number of objects into the training data set, and d is the number of decision trees to be built. The filtering phase has a complexity of $\Theta(p^2)$, being p the number of patterns to be filtered. Finally, the classification phase has a complexity of $\Theta(p)$. The runtime for the methods of this team ranged between 19 and 31 hours.

Third place (Voltaire Project): This team presented a strategy based on exploiting contextual information. The proposed method tries to infer whether a word is relevant for an image by using information shared among images (images labeled with the same word). The strategy uses this contextual information in order to increase reliability, by inferring whether a word is relevant for a given image comparing with the closest image set. Despite its simplicity, the adopted strategy take advantage of traditional methods, i.e. LDA and PCA, offering a competitive method with a trade-off between efficiency/effectiveness. The time estimated runtime of this method was of 2.5 hours.

Fourth place (Argenis): Unlike the other teams, the last team opted for a simple but more efficient solution. Its solution was based on joint application of two well-known algorithms, Kmeans and k-NN. The strategy was to use a weighted distance in order to give the same importance to the visual and textual features. To accomplish such strategy, clustering techniques were applied off-line, reducing time at the classification phase. The complexity at the training phase for this method was $\Theta(m*t*n)$ where m is the number of clusters, t is the number of iterations until clusters stop changing, and n is the number of training instances to classify. While, for the classification phase its complexity is reduced to $\Theta(m*k)$ being m the number of clusters considered. The runtime of this method was around 15-20 minutes.

It was encouraging to have methods completely different, covering both traditional and advanced techniques from machine learning. It is also quite interesting that the obtained solutions had quite similar performances, in fact, the differences in performance can only be appreciated, in some cases, at the third decimal of the accuracy results (see Table II). As we can see, the participants that exploit jointly visual and textual features obtained the best performance.

C. Discussion

The feasibility of the challenge has been successfully demonstrated, achieving results on accuracy greater 0.8. Both baselines were improved, the first one by a large margin, whereas the second proved to be a quite strong baseline. Please note that the second baseline is an improved implementation that has won several challenges in the past. The solutions contributed by participants were both, diverse and quite competitive. Ensembles obtained the best performance and feature preprocessing proved to be very helpful. Learned representations benefited from an additional feature extraction process. This is an interesting finding worth of further study.

Regarding the proposed task, results gave evidence that it is promising to approach the AIA problem in this way. A further analysis of results is ongoing work to explore the limitations and benefits of this formulation, and the existing solutions so far. To the best of our knowledge, this is the first work approaching the AIA problem as one of textual-visual representation matching, explicitly taking advantage of representation learning. In addition to image labeling, we foresee this approach is promising for multimodal data

TABLE II
PERFORMANCE MEASURES (IN %) FOR THE ALL TECHNIQUES.

System (TEAM)	Accuracy		F1	
	dev.	test	dev.	test
Organizing team				
baseline 1	0.644	0.639	0.69	0.64
baseline 2	0.760	0.818	0.76	0.79
I3GO+ team				
job80 (T1)	**0.831**	0.838	**0.85**	**0.84**
ClaudiaSanchez (T1)	0.816	0.823	0.84	0.82
mgraffg (T1)	0.813	0.824	0.84	0.82
dmocteo (T1)	0.684	0.833	0.71	0.83
sadit (T1)	0.797	**0.844**	0.82	**0.84**
MIGUE, TAVO & ANDRES team				
migue (T2)	0.811	0.830	0.83	0.82
octavioloyola (T2)	0.807	0.834	0.83	0.83
andres (T2)	0.801	0.824	0.82	0.82
Voltaire Project team				
Phenix (T3)	0.823	0.828	0.85	0.83
Argenis team				
argenis (T4)	0.758	0.780	0.78	0.78
Individual participants				
naman	0.656	—	0.72	—
miguelgarcia	0.600	—	0.75	—
barb	0.479	—	0.48	—
victor	0.499	—	0.55	—

embedding as well. We consider the aforementioned data set to be a valuable resource, that will be made publicly available.

On the other hand, our academic challenge has attracted the participation from both local and foreign participants. It has boosted and promoted collaborations among RedICA members. With these positive results, we plan to organize future editions of this challenge in the near future.

IV. Conclusions and Future Work

We formulated the AIA problem as one of text-image matching, in turn casted as a binary classification task, and organized an academic challenge around this problem. Pretrained representation learning methods were used to characterize instances in our problem, and three strategies for determining the text-image relevance were considered. A mid size data set for the task was made available together with an evaluation protocol.

Participation during the challenge was numerous (43 participants), with several teams participating until the very last stage of the challenge. Participants provided diverse solutions showing very competitive performances, encouraging the development of novel methodologies with aims to exploit heterogeneous representations.

Future work includes taking advantage of the models developed by the participants. with the aim to approach them in related task with images, e.g. annotation, description and text-illustration. Also, we plan to extend our data set by considering new approaches in order to produce new training instances, for instance, generating synthetic instances from a word embedding perspective rather than image perspective (presented here).

Acknowledgment

This work was supported by CONACYT under project grant CB-2014-241306 (Clasificación y recuperación de imágenes mediante técnicas de minería de textos). The first author was supported by the CONACyT with scholarship No. 214764. The authors would like to thank CodaLab (running on MS Azure) and ChaLearn, and also to thank sponsors *Red temática en Inteligencia Computacional Aplicada* (RedICA), CONACyT and INAOE.

References

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[2] L. Pellegrin, H. J. Escalante, M. Montes-y Gómez, and F. A. González, "Local and global approaches for unsupervised image annotation," *Multimedia Tools and Applications*, pp. 1–26, 2016.

[3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 5 2015.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Mar. 2003.

[5] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27 – 48, 2016, recent Developments on Deep Big Vision.

[6] E. Bruni, N. K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Int. Res.*, vol. 49, no. 1, pp. 1–47, Jan. 2014.

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS'14, 2014, pp. 3320–3328.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, ser. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119.

[9] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *The Proceedings of the International Workshop OntoImage'06 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06,*, 2006, pp. 13–23.

[10] H. J. Escalante, C. Hernández, J. Gonzalez, A. López-López, M. Montes, E. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger, "The segmented and annotated {IAPR} tc-12 benchmark," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 419 – 428, 2010, special issue on Image and Video Retrieval Evaluation.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[13] E. S. Tellez, D. Moctezuma, S. Miranda-Jiménez, and M. Graff, "An automated text categorization framework based on hyperparameter optimization," *CoRR*, vol. abs/1704.01975, 2017.

[14] C. Conde, D. Moctezuma, I. M. D. Diego, and E. Cabello, "Hogg: Gabor and hog-based human detection for surveillance in non-controlled environments," *Neurocomputing*, vol. 100, pp. 19 – 30, 2013, special issue: Behaviours in video.

[15] M. Garca-Borroto, J. F. Martnez-Trinidad, and J. A. Carrasco-Ochoa, "Finding the best diversity generation procedures for mining contrast patterns," *Expert Systems with Applications*, vol. 42, no. 11, pp. 4859 – 4866, 2015.

[16] O. Loyola-Gonzlez, M. A. Medina-Prez, J. F. Martnez-Trinidad, J. A. Carrasco-Ochoa, R. Monroy, and M. Garca-Borroto, "Pbc4cip: A new contrast pattern-based classifier for class imbalance problems," *Knowledge-Based Systems*, vol. 115, pp. 100 – 109, 2017.