INAOE-UNAL at ImageCLEF 2015: Scalable Concept Image Annotation

Luis Pellegrin[1], Jorge A. Vanegas[2], John Arevalo[2], Viviana Beltrán[2], Hugo Jair Escalante[1], Manuel Montes-y-Gómez[1], and Fabio A. González[2]

[1]Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico
[2]MindLab research group, Universidad Nacional de Colombia (UNAL), Colombia.
[1]{pellegrin,hugojair,mmontesg}@ccc.inaoep.mx
[2]{javanegasr,jearevaloo,lvbeltranb,fagonzalezo}@unal.edu.co

Abstract. This paper describes the joint participation of the TIA-LabTL (INAOE) and the MindLab research group (UNAL) at the ImageCLEF 2015 Scalable Concept Image Annotation challenge subtask 2: *generation of textual descriptions of images - noisy track*. Our strategy relies on a multimodal representation that is built in an unsupervised way by using the associated text to images and the visual features that represent them. In the multimodal representation for every word extracted from the indexed web pages a visual prototype is formed, each prototype being a distribution over visual descriptors. Then, the process of generation of a textual description is formulated as a two-step IR problem. First, the image to be described is used as visual query and compared with all the visual prototypes in the multimodal representation; next the $k-$nearest prototypes are used as a textual query to search for a phrase in a collection of textual descriptions, the retrieved phrase is then used to describe the image.

Keywords: multimodal representation, textual description of images, visual prototype.

1 Introduction

In the 4th edition of the Scalable Concept Image Annotation challenge [2], there are two main subtasks: 1) image concept detection and localization, and 2) generation of textual description of images. There are two separate tracks for SubTask 2: *clean* and *noisy tracks*. The goal of the *noisy track* is to develop a system that generates a textual description for an input image, by using only the information that provides the visual representation of the image. Additionally, we have a collection of images downloaded from Internet, for which we have the web page where images appear and the keywords that were used for searching the images.

This paper describes the joint participation of the TIA-LabTL[1] (INAOE) and the MindLab[2] research group (UNAL) in the 2015 Scalable Concept Image

[1] http://ccc.inaoep.mx/labtl/
[2] https://sites.google.com/a/unal.edu.co/mindlab/

Annotation challenge subtask 2: *generation of textual descriptions of images - noisy track*.

Our approach to generate textual descriptions relies on the use of the textual information of web pages to describe the visual content. First, we associate visual descriptors to text and build a multimodal representation where we have a visual prototype for every word extracted from the reference text-collection. Second, the multimodal representation is used to retrieve words related to an input image to be described, then the $k-$nearest words are used as query to retrieve captions in a reference textual description set. Therefore, our method of image captioning can be seen as a two-step information retrieval (IR) task (first retrieve words, then retrieve descriptions). An important remark, is that we build the multimodal representation following an unsupervised approach where we do not require labeled data at all, and can describe images with any term in our vocabulary. The official results are encouraging and show us that there are ways to improve, including to explore different visual features and different reference captioning collections.

The remainder of the paper is organized as follows: Section 2 describes our method; Section 3 shows the experimental settings that we used; Section 4 shows the experimental results obtained; finally, in Section 5 some conclusion of this work are presented.

2 Two-step IR process for Automatic Image Captioning

As mentioned above, our proposed method is divided in two IR stages: word-retrieval and caption-retrieval. Intuitively, the second stage requires a textual query to search for a caption, and the first stage produces as output a textual query. Where the first stage matches a query image with words. Hence, our method requires a way to associate images with words, more specifically, we require a representation for words in the visual feature space. In addition to the two IR steps, we have a preliminary multimodal-indexing (MI) step, but first we preprocess text and images in order to build the multimodal representation. Fig. 1 shows a diagram of the overall process (including the MI step). In the following we describe in detail each of these steps.

2.1 Preprocessing

As we can see in the Fig. 1 the preprocessing is carried in two datasets, a reference image collection and reference textual description set. In the reference web-image collection $\mathcal{C}$ we have represented every image $\mathcal{I}_i$ by means of visual $\mathcal{V}_i$ and textual $\mathcal{T}_i$ features defined by *bags-of-words* (textual and visual[3]):

[3] To obtain a bag of visual words representation first, a corpus of images are used to extract visual features, i.e. points of interest in images; second, the sampled features are clustered and the centroids quantize in a discrete number of visual words; third, the nearest visual words are identified in images using their visual features; fourth, a bag-of-visual-words histogram is used to represents images.

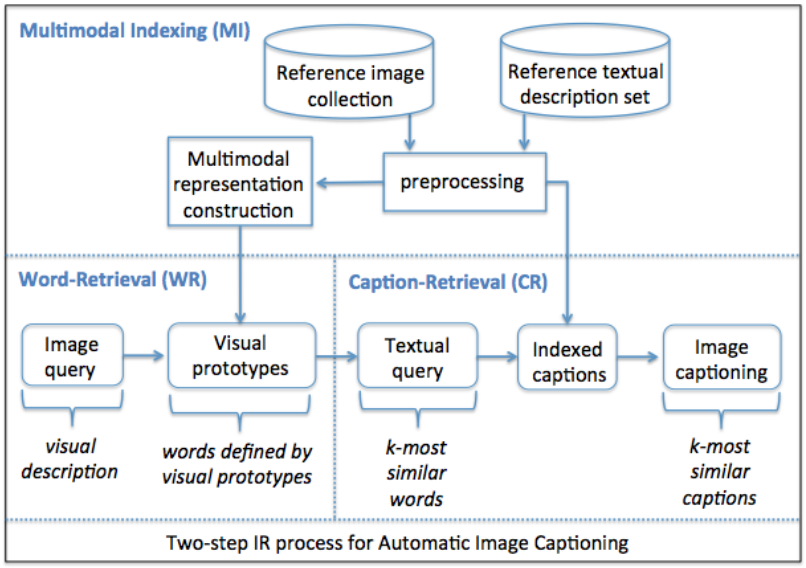

Fig. 1: Diagram of the proposed approach to describe images.

$$\mathcal{C} = \mathcal{I}_i : \{\mathcal{V}_i, \mathcal{T}_i\}_{i=1,\ldots,n}$$

The visual and textual representations of images $\mathcal{I}_i$ can be denoted by:

$$\mathcal{T}_i = \langle t_{i,1}, \ldots, t_{i,|X|} \rangle$$

$$\mathcal{V}_i = \langle v_{i,1}, \ldots, v_{i,|Y|} \rangle$$

where $t_{i,j} \in \mathbb{R}^{|X|}$ and $v_{i,j} \in \mathbb{R}^{|Y|}$ respectively, where $|X|$ and $|Y|$ denote the sizes of the textual and visual vocabularies. Where $\mathcal{T}$ and $\mathcal{V}$ are matrices whose rows correspond to the text and visual descriptions of each image with size $n \times |X|$ and $n \times |Y|$ respectively. The corresponding preprocessing is detailed as follows:

Textual preprocessing Textual information (i.e., the web pages associated to images) is represented with a standard term-frequency bag of words representation, then classic scheme of weighting TF-IDF is applied. We consider that to express the visual content of the images is more important the use of nouns than the use of pronouns or verbs, so we eliminate these latter.

Visual preprocessing Visual information (i.e., the visual descriptors that represented to images) is normalized and treated as a standard frequency bag-of-visual-words representation, then classic scheme of weighting TF-IDF is applied,

the purpose of this weighting here is to prioritize to those visual words that occur not in the most of the images. Here we did not consider the elimination of any visual words.

Furthermore, our method also uses as input a reference textual description set $\mathcal{D}$ (see Fig. 1) that contains a set of captions that can be used to describe to images. We use three different set of sentences as reference in our tests for the generation of the textual descriptions (see Table 1 for examples):

- **Set A**. The set of sentences given in the evaluation ground truth for the subtask 2 by the organizers at the ImageCLEF, with ≈19,000 sentences.
- **Set B**. A set of sentences used in the evaluation of MS COCO 2014 dataset [6], with ≈200,000 sentences.
- **Set C**. A subset of PASCAL dataset [3], with ≈5,000 sentences.

Table 1: Examples of sentences extracted from three reference caption sets.

dataset	**sentences**
set A	*'A person shaking another persons hand.'* *'Bunk beds in a room by a window with vibrant flowered bed covers.'*
set B	*'A Honda motorcycle parked in a grass driveway.'* *'Well kept kitchen with marble counter tops and stainless steel fridge.'*
set C	*'One jet lands at an airport while another takes off next to it.'* *'Duck grooming itself while standing in the water.'*

Textual description preprocessing The sentences of the reference textual description set are indexed and represented with a standard term-frequency of *bag-of-words* representation removing only *stop words*. For the case of this indexing we did not consider the elimination of pronouns or verbs, the reason is that by including the phrases can be weighted by frequency considering the amount of words.

2.2 Multimodal Indexing

Once it performed the preprocessing, the preprocessed textual and visual features are used in the construction of the multimodal representation (as we can see in the diagram of the Fig. 1). As we mentioned before, we require a way to associate images with words in order to perform the first stage of our method. We hypothesized a space where each word can be associated with a visual prototype, a multimodal representation. In this way, any query image can be compared with prototypes and we can determine what words are more related to the query. For the construction of our multimodal representation we rely on term-occurrence statistics, where the terms are both textual and visual features. The main idea behind it is that both the textual view as the visual view of an image have a salience in the same objects represented by two different features (textual and visual), that is, if in a web page the main topic in text is about 'dogs' there exist

a likelihood that the visual descriptors or images in the web page are associated to a 'dog'. This assumption can be confirmed when a majority of images share the same characteristics forming a distinctive visual prototype. The hypothesis is that one word has predominant visual descriptors associated to it, so every word can be seen as a prototypical of image that represents to the word.

The multimodal representation associates each word with a distribution of weights over the visual vocabulary forming a visual prototype for every word. Our multimodal representation is obtained as follows:

$$\mathcal{M} = \mathcal{T}^T \cdot \mathcal{V}$$

Therefore, we can see that $\mathcal{M}_{i,j} = \sum_{k=1}^{n} \mathcal{T}_{i,k} \cdot \mathcal{V}_{k,j}$ is a scalar value that express the degree of association between word i and visual-word j, across the whole collection of documents. In this way, each row of the matrix $\mathcal{M}$ (each associated to a word), can be seen as a visual prototype. Finally, our $\mathcal{M}$ is a matrix of size $|X| \times |Y|$, that is, the dimension is determined by the sizes of the vocabularies that represent both textual and visual features.

2.3 Step 1: Word-Retrieval (WR)

The multimodal representation is used in this step in a similar way to a IR task, first we use the visual representation of an image to be described as a query, next using a cosine similarity we score the distances with all the visual prototypes from the multimodal representation (see WR step in Fig 1). The cosine similarity is defined by:

$$sim(\mathcal{I}_q, \mathcal{M}_w) = \frac{\mathcal{I} * \mathcal{M}_w}{||\mathcal{I}||||\mathcal{M}_w||}$$

where $\mathcal{I}_q$ is the input image represented by a vector of visual words, and $\mathcal{M}_w$ is one visual prototype from the multimodal representation, that is, the row w of the matrix $\mathcal{M}$ above, remember that this formula is applied to every word, that is, $w = \{1, \ldots, |X|\}$. Finally, the k words associated to the most similar prototypes are used in the next stage to describe the visual content of the input image.

2.4 Step 2: Caption-Retrieval (CR)

Using the output of the WR step, with the k words associated to the k most similar visual prototypes, we formulate a second query, this time, it is a textual query (represented by a TF weighting scheme as the reference captions). We measure a cosine similarity between the textual query and the indexed captions, where Q_T is a vector of size $|X|$ that contains the values of k words with the highest scores obtained in the WR step, and the $\mathcal{D}_c$ is one sentence in the set of indexed captions. The cosine similarity is applied to every caption in $\mathcal{D}$, that is, $c = \{1, \ldots, d\}$. The quantity of words used in the textual query determines the

specificity of the textual description to retrieve, that is, we can obtain a general textual description using only a few words. Finally, the m indexed captions that maximize the cosine similarity can be used to describe the input image. We used the caption with highest similarity as textual description for the image $\mathcal{I}_q$. The CR step is depicted the Fig. 1.

3 Experimental Settings

The dataset for the approached task is composed by 500,000 images extracted from a database of millions of images downloaded from the Internet. Every image in the dataset has an associated text, i.e. the web pages where the images were found it or the keywords used for the searching at the Internet.

Several pre-processed visual features are provided by the organizers of the task, we experimented with three different: OPPONENT-SIFT, relu7 layer of activations in a CNN model of 16-layer, and fc8 layer of activations in a CNN model of 16-layer. At the end, we chose relu7 layer as visual features for the representation of the images in the training set, we refer to relu7 layer as visual words[4] but strictly this representation is an activation of a CNN model.

Table 1 describes the 10 runs that we submitted for evaluation. The first row of the Table 1 shows the set of reference captions used: denoted as A the set of sentences given by organizers in the ImageCLEF; denoted as B the set of sentences of MS COCO dataset; and denoted as C a subset of sentences of PASCAL.

From the output of the WR step, we formed a textual query as we mentioned before by the k most similar words defined by visual prototypes from our multimodal representation. We experimented with two ways to select words, in the second row of the Table 2 with the letter w corresponds to use the k most similar words; and with the letter c we filtered visual prototypes corresponding to words that matches concepts taken from a list of 251 (the concept list provided by the organizers, for more info. see [2]), so in this latter case the k most similar concepts are used.

Selection of k value was established empirically, that is, the number of words that we used in textual query. We established a threshold for the k value using information from the visual descriptor of the images. We noted that if image to be described was represented by few visual words, that is, if it had great number of zeros values in visual words (see Fig. 2 for some examples of images) then for these kind of images we can use a minimum k value to form the textual query. On the other hand, the images with a low sparsity (lower number of zeros) in their visual representation we need to use more data. The ranges of threshold for k values are summarized follows:

[4] Although, relu7 layer is not based on visual words, for our purpose we can see this representation like one based in visual words. We normalize every vector of activation and use the normalized activation of neurons like frequencies of a group of visual words. Using the activation of neurons at the end is a kind of codebook, because all the images share a common space of representation.

– Concepts: 3 for >40% sparsity; and 5 for <40% sparsity.
– Words: 10 words for >50% sparsity; 30 words between 40% and 50% sparsity, and 50 words for <40% sparsity.

Another aspect considered in our settings was to use different representation for the textual query: in the third row in the Table 2 with weights expressing real values with the letter r, that is, the score obtained by cosine similarity for the textual query and the normalized frequency for the sentences in the reference textual description set; and denoted by the letter b a binary value for both textual query and indexed captions; the reason is because when we use a binary representation we are giving importance to sentences with a greater number of words or concepts, on the other hand, if we use the weights then the privileged sentences are those with a better matching.

Table 2: Settings of submitted runs.

	runs →									
↓ settings	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
sentences set used	A	A	A	A	B	B	B	B	*	A,B,C
data used	c	w	c	w	c	c	w	w	*	w
query representation	r	r	b	b	r	b	r	b	*	b

∗ Only using top 5 concepts from defined list.

4 Experimental Results

Table 3 shows the results obtained by each of the submitted runs. For comparison, we also show the result of the best run submitted to the task (RUC_{run3}). The reported values corresponds to the Meteor Score [1] when using a minimum of five human-authored textual descriptions as the gold standard reference.

Columns of Table 3 express across all test images: the average, median, min and max Meteor scores. According to the mean and median performance when we used only concepts in the textual query (runs 1, 3, 5 and 6) we obtained a better score than using words (runs 2, 4, 7, 8 and 10), we believe it is because of the confidence of the detected concept and to the existence of sentences that describe the concept.

In ours runs the best mean score was using the set A (run 3), we believe that this set A (sentences of evaluation of ImageCLEF dataset) is more controlled in the quality of the description due to the amount of sentences in comparison to set B (MS COCO dataset) where there are 10 times more of sentences. Another characteristic of the run 3 is that uses a binary representation and concepts as data, we believe that exist a relationship between the amount of data used in the textual query, when the textual query is short is beneficial a binary representation to retrieve sentences with the more quantity of concepts.

The results in comparison with the best score of RUC team show us that we are not so far, we would like to evaluate our method with a list of concepts bigger and others sets of sentences that can complement to those used already.

Table 3: METEOR score of the our submitted runs.

RUN	MEAN +− STDDEV	MEDIAN	MIN	MAX
run1	0.1255+−0.0650	0.1140	0.0194	0.5679
run2	0.1143+−0.0552	0.1029	0.0175	0.4231
run3	**0.1403+−0.0564**	**0.1342**	**0.0256**	**0.3745**
run4	0.1230+−0.0531	0.1147	0.0220	0.5256
run5	0.1192+−0.0521	0.1105	0.0000	0.4206
run6	0.1260+−0.0580	0.1172	0.0000	0.4063
run7	0.1098+−0.0527	0.1005	0.0000	0.4185
run8	0.1079+−0.0498	0.1004	0.0000	0.3840
run9	0.0732+−0.0424	0.0700	0.0135	0.2569
run10	0.1202+−0.0528	0.1123	0.0000	0.5256
RUC_run3	*0.1806+−0.0817*	*0.1683*	*0.0192*	*0.5696*

In Fig. 2, we can see two examples of images annotated with textual description using our method. For both images, we can see the output of the WR step, the textual query, and the final output using two sets of sentences (CR step). In image a, we can see that the top two concepts can describe the image completely, and that the difference between the two generated textual descriptions seems to be that in set A are not enough phrases that describing 'watches' or 'clocks'.

The generated textual description for image b fails because the image shows 'two spatulas' but the detected concept is 'knife' (there is not 'spatula' concept in the list), using words the 'spatula' could be detected, however, there is not a textual description with this particular word in both A or B sets. One natural way to correct the description of the image b is to add new set of sentences that contain the words that have been detected.

In Fig. 3, we show two images with scenes that are ambiguous and can be described in several ways. For both images, we can see the final output using two sets of sentences (corresponding to the 4 and the 8 runs respectively). We use words as textual query to find the textual description with binary values. For image c, the two generated captions could be used as description, the sentences describe different aspect in the image. On the other hand, the image d can be described by human annotator as a 'videogame' or 'animation', here our method describe the image using the detected objects like 'car' or 'alleyway' but not as 'animation'.

5 Conclusions

We described the joint participation of the TIA-LabTL (INAOE) and the MindLab research group (UNAL) at the ImageCLEF 2015 Scalable Concept Image Annotation challenge subtask 2: *generation of textual descriptions of images* -

image a

Textual query (WR step): 'watch', 'clock', 'flashlight', 'lotion', 'radio'.

Textual description generated (CR step):
Set A) *'There are two people watching through and structure.'*
Set B) *'A clock shows it is just after 10:00 o'clock.'*

image b

Textual query (WR step): 'knife', 'flashlight',, 'pen', 'sword', 'bathtub'.

Textual description generated (CR step):
Set A) *'Two Ostriches in a pen.'*

Set B) *'A knife being pushed into a knife block.'*

Fig. 2: Two images with their text description generated under two different sets of sentences, and the top 5 concepts, where the firsts three were used for the query. These examples correspond to the 1 and the 5 runs respectively.

image c

Textual query (WR step): 'scoreboard', 'alleyway', 'limo', 'tram', 'nightclub', 'avenida', 'ballpark', 'billboard', ...

Textual description generated (CR step):
Set A) *'A picture of an empty movie theatre or stage and or stadium.'*
Set B) *'A downtown area of a city with buildings and billboards.'*

image d

Textual query (WR step): 'ambulance', 'alleyway', 'tram', 'saleen', 'streetcar', 'hatchback', 'saloon', 'limo', 'maserati', ...

Textual description generated (CR step):
Set A) *'A blue Dodge Viper with white stripes parked in an alleyway.'*
Set B) *'A Volkswagen sedan lies still in a parking space.'*

Fig. 3: Two images with their text description generated under two different sets of sentences. These examples correspond to the 4 and the 8 runs respectively.

noisy track. The proposed method works in an unsupervised way using the information of textual and visual features to build a multimodal representation. The overall process of generation of textual description is formulated as two-step IR task. Our method is flexible and can be applied with different visual features encouraging us to explore visual features learned by using different approaches.

The experimental results showed the competitiveness of our technique and that it can be improved, including refined reference sentences for the textual description or filtering of words in the vector representation by means of measures of relatedness.

Acknowledgments. This work was supported by CONACYT under project grant CB-2014-241306 (Clasificación y recuperación de imágenes mediante técnicas de minería de textos). Also this work was partially supported by the LACCIR programme under project ID R1212LAC006.

References

1. Denkowski M., and Lavie A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, (2014)
2. Gilbert A., Piras L., Wang J., Yan F., Dellandrea E., Gaizauskas R., Villegas M., and Mikolajczyk K.: Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: CLEF2015 Working Notes Workshop Proceedings, Toulouse, France, (2015)
3. Rashtchian C., Young P., Hodosh M., and Hockenmaier J.: Collecting Image Annotations Using Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, (2010)
4. Villegas M., Müller H., Gilbert A., Piras L., Wang J., Mikolajczyk K., García Seco de Herrera A., Bromuri S., Amin M. A., Mohammed M. K., Acar B., Uskudarli S., Marvasti N. B., Aldana J. F., and Roldán García M. M.: General Overview of ImageCLEF at the CLEF 2015 Labs. In: CLEF2015 Working Notes Workshop Proceedings. LNCS, Springer International Publishing (2015).
5. Caffe: An open source convolutional architecture for fast feature embedding. `htt://caffe.berkeleyvision.org/`
6. Microsoft COCO (Common Objects in Context), `http://mscoco.org/dataset/`