A two-step retrieval method for Image Captioning

Luis Pellegrin*, Jorge Vanegas, John Arevalo, Viviana Beltrán, Hugo Escalante, Manuel Montes-Y-Gómez and Fabio González

Computer Science Department
National Institute of Astrophysics, Optics and Electronics
Tonantzintla, Puebla, 72840, Mexico
*pellegrin@inaoep.mx

Clef2016 5-8 September, Évora, Portugal

Content

1. Introduction
2. Our approach
3. Experimental Results
4. Conclusions
5. References

Automatic Description Generation

The goal is to develop systems that can automatically generate sentences that verbalize information about images.

'A man is standing on a cliff high above a lake.'

'Boats in the water with a city in the background.'

- Descriptions: things that what can be seen in the image.
- Captions: information that cannot be seen in the image.

Automatic Description Generation

The goal is to develop systems that can automatically generate sentences that verbalize information about images.

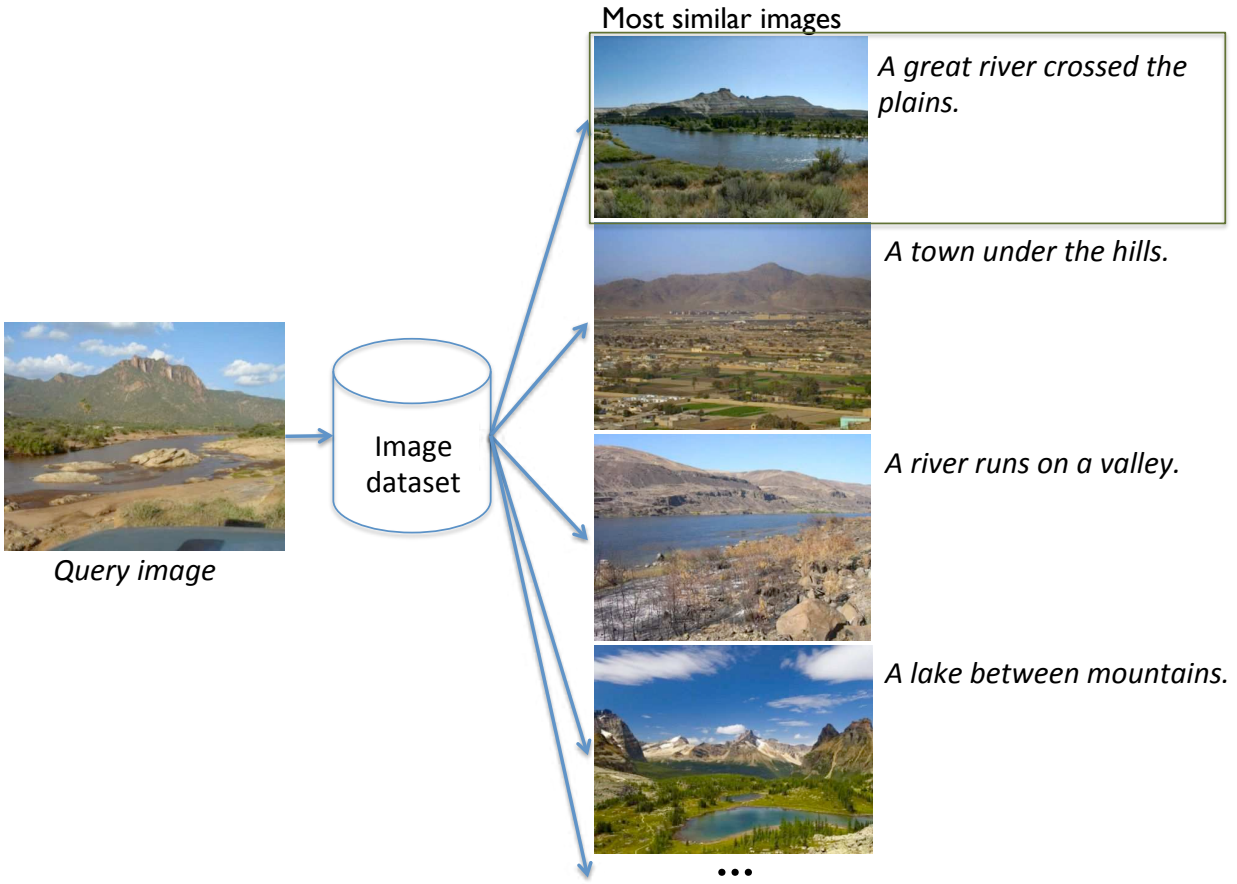
'A man is standing on a cliff high above a lake.'

'Boats in the water with a city in the background.'

- Descriptions: things that what can be seen in the image.
- Captions: information that cannot be seen in the image.

Related Work: main approaches (1/2)

- Traditional approaches assign (transfer) or synthesize the sentences from the most similar images to the query image.



Related Work: main approaches (2/2)

- Recent methods rely on sentence generation systems that learn a joint distribution over training pairs of images and their descriptions/captions.

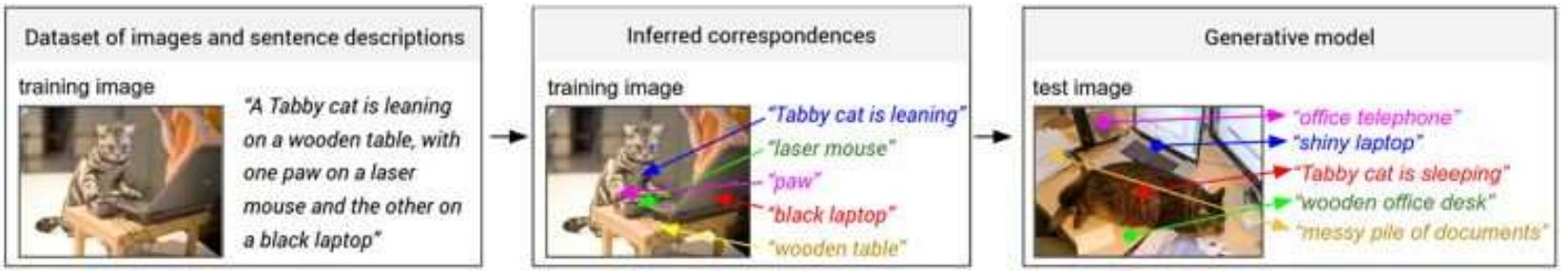


Figure taken from [Karpathy & Fei-Fei, 2015]

Disadvantages of main approaches

- A drawback of the traditional approach is that a **great variety** of images are **necessary** to have enough coverage in terms of the sentence assignation.
- In general sentence generation systems rely on great quantity of **manually labeled data** for learning models, an **expensive** and **subjective** labor due to the great variety of images.

Disadvantages of main approaches

- A drawback of the traditional approach is that a **great variety** of images are **necessary** to have enough coverage in terms of the sentence assignation.
- In general sentence generation systems rely on great quantity of **manually labeled data** for learning models, an **expensive** and **subjective** labor due to the great variety of images.

Overview of the proposed approach

- The proposed method to generate textual descriptions **does not require labeled** images.
- Motivated by the large number of images that can be found and gathered from the Internet. **It uses textual-visual information** derived from webpages containing images.
- Our strategy relies on a multimodal indexing of words, where for **each word in the vocabulary** extracted from webpages, it is built **a visual representation**.

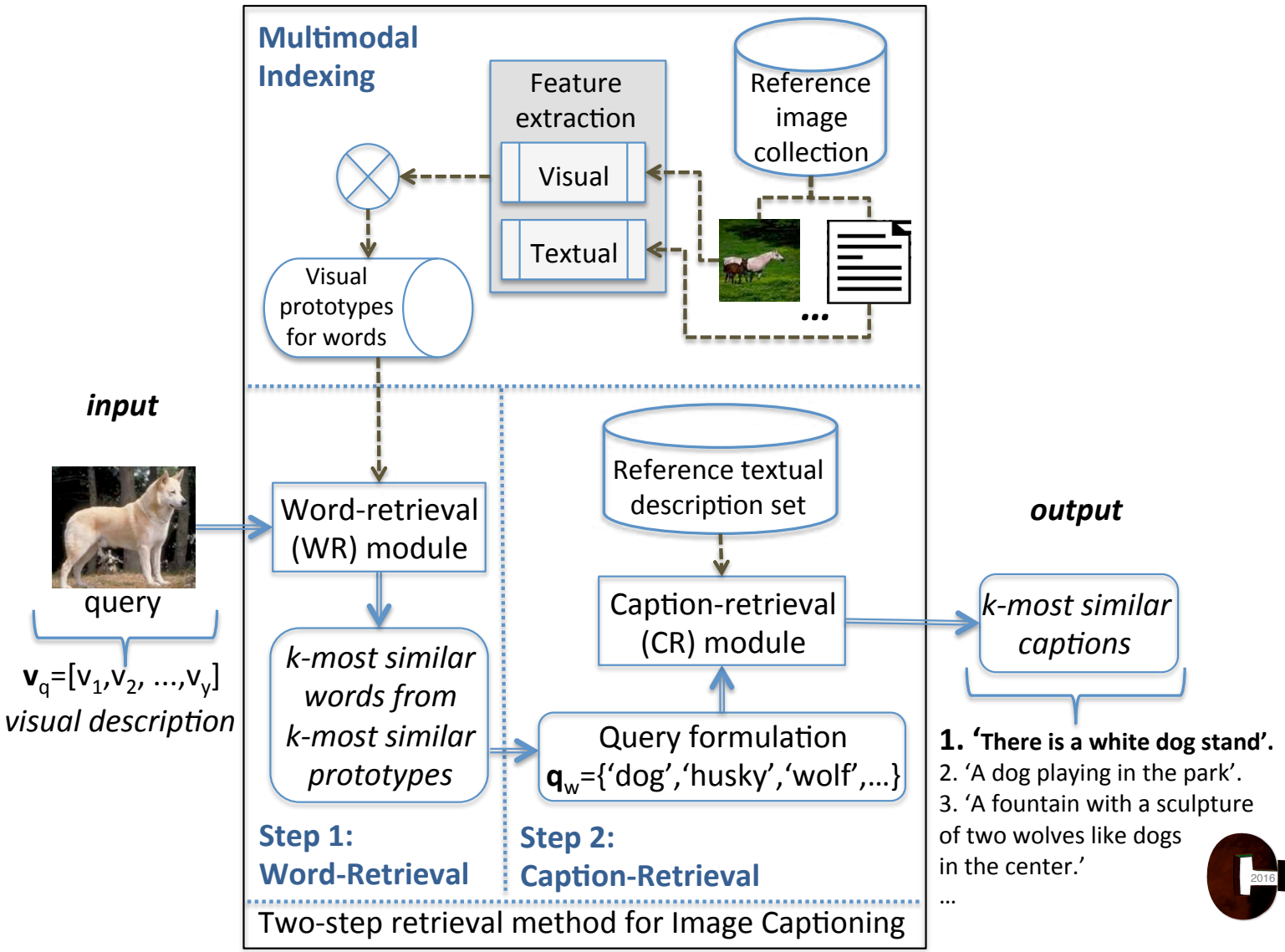
Overview of the proposed approach

- The proposed method to generate textual descriptions **does not require labeled** images.
- Motivated by the large number of images that can be found and gathered from the Internet. **It uses textual-visual information** derived from webpages containing images.
- Our strategy relies on a multimodal indexing of words, where for **each word in the vocabulary** extracted from webpages, it is built **a visual representation**.

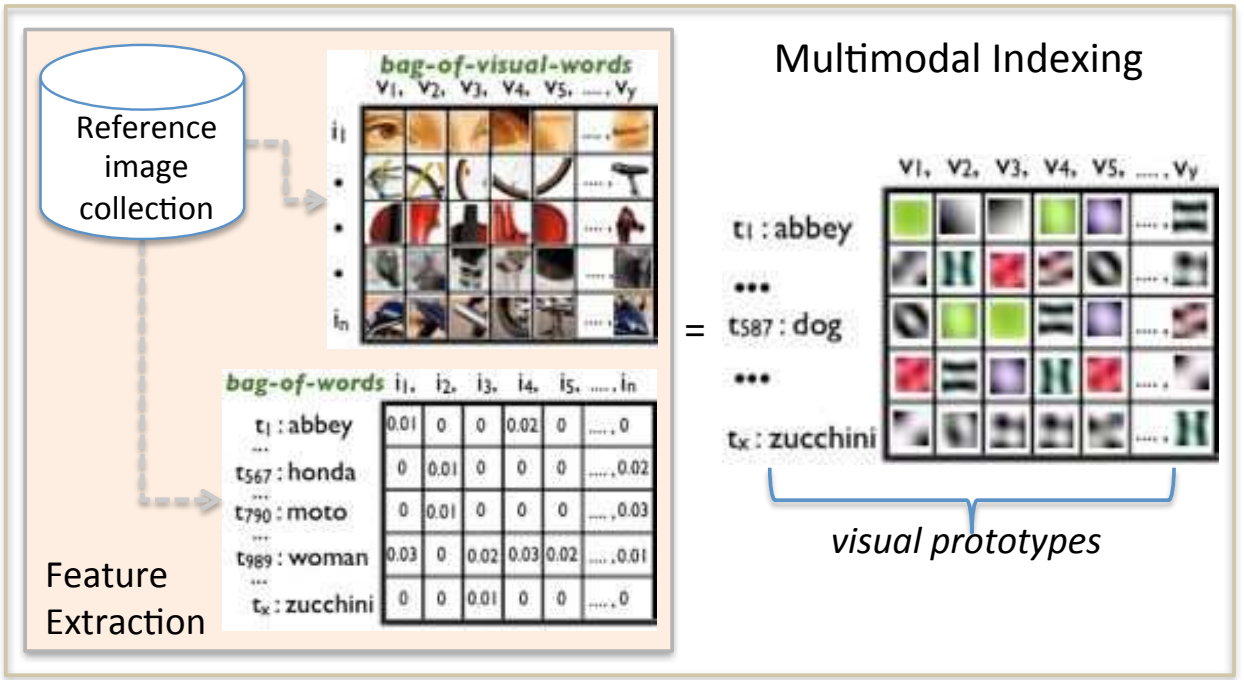
Overview of the proposed approach

- The proposed method to generate textual descriptions **does not require labeled** images.
- Motivated by the large number of images that can be found and gathered from the Internet. **It uses textual-visual information** derived from webpages containing images.
- Our strategy relies on a multimodal indexing of words, where for **each word in the vocabulary** extracted from webpages, it is built **a visual representation**.

Sentence generation for images (architecture)



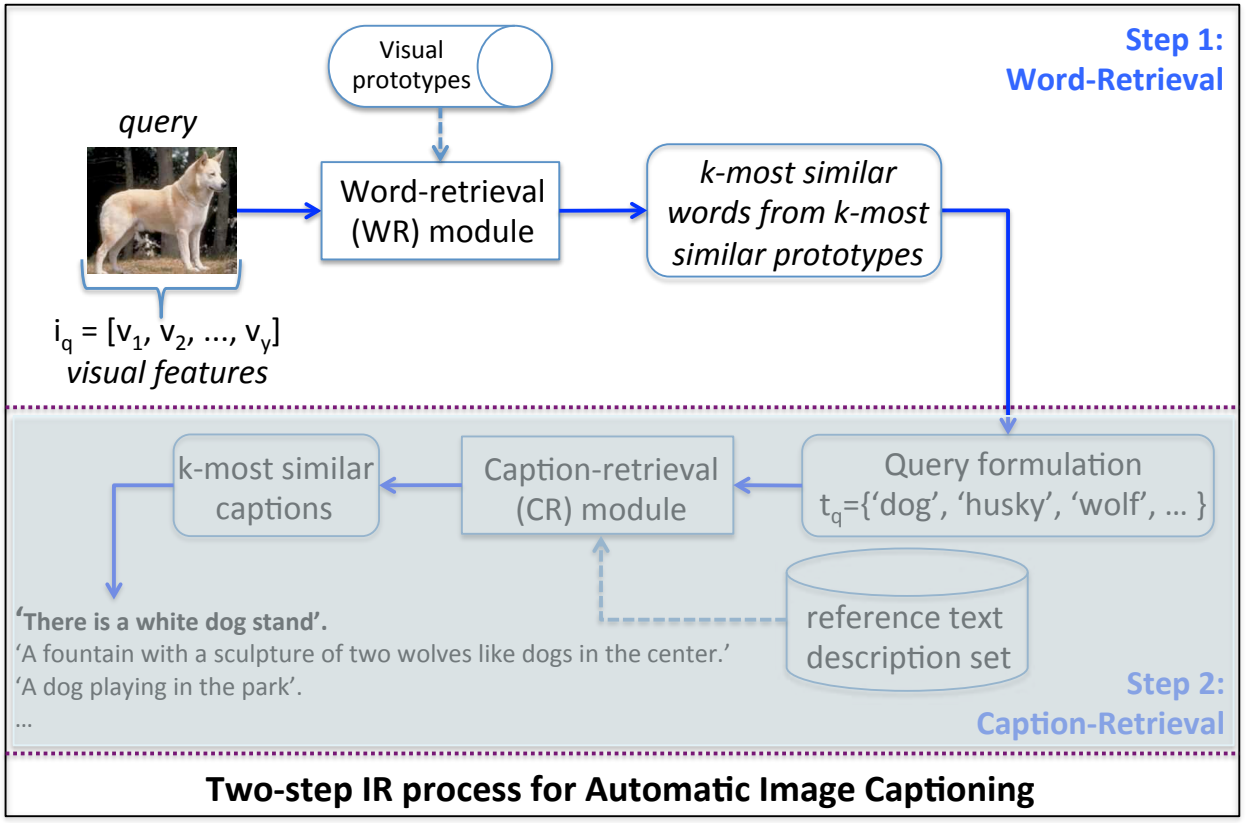
Multimodal Indexing: feature extraction



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

$$\mathbf{M}_{i,j} = \sum_{k=1}^{n} \mathbf{T}_{i,k} \cdot \mathbf{V}_{k,j}$$

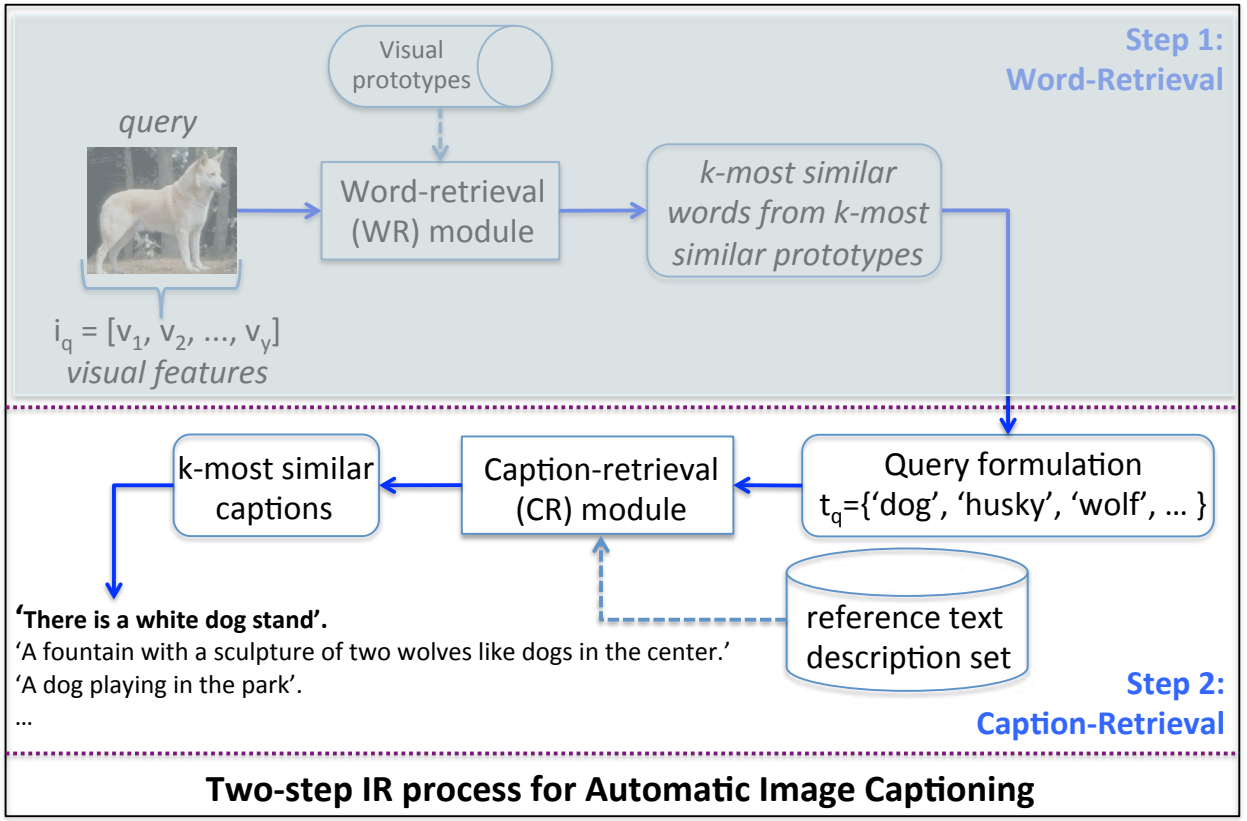
Step 1: Word-Retrieval (WR)



Two-step IR process for Automatic Image Captioning

$$WR : score(\mathcal{W}_i) = cosine(\mathbf{v}_q, \mathbf{M}_i)$$

Step 2: Caption-Retrieval (CR)



$$CR : score(\mathcal{C}_i) = cosine(\mathbf{t}_q, \mathbf{C}_i)$$

Remarks about our MI

- It can match query images with words by simply measuring visual similarity.
- In principle, it can describe images using any word from the extracted vocabulary.
- It is possible to change the direction of the retrieval process, that is, it can be used to illustrate a sentence with images.
- Although the relatedness of an image with the text in the web page varies greatly, the MI is able to take advantage of multimodal redundancy.

Remarks about our MI

- It can match query images with words by simply measuring visual similarity.
- In principle, it can describe images using any word from the extracted vocabulary.
- It is possible to change the direction of the retrieval process, that is, it can be used to illustrate a sentence with images.
- Although the relatedness of an image with the text in the web page varies greatly, the MI is able to take advantage of multimodal redundancy.

Remarks about our MI

- It can match query images with words by simply measuring visual similarity.
- In principle, it can describe images using any word from the extracted vocabulary.
- It is possible to change the direction of the retrieval process, that is, it can be used to illustrate a sentence with images.
- Although the relatedness of an image with the text in the web page varies greatly, the MI is able to take advantage of multimodal redundancy.

Remarks about our MI

- It can match query images with words by simply measuring visual similarity.
- In principle, it can describe images using any word from the extracted vocabulary.
- It is possible to change the direction of the retrieval process, that is, it can be used to illustrate a sentence with images.
- Although the relatedness of an image with the text in the web page varies greatly, the MI is able to take advantage of multimodal redundancy.

Datasets

ImageCLEF 2015: *Scalable Concept Image Annotation* benchmark

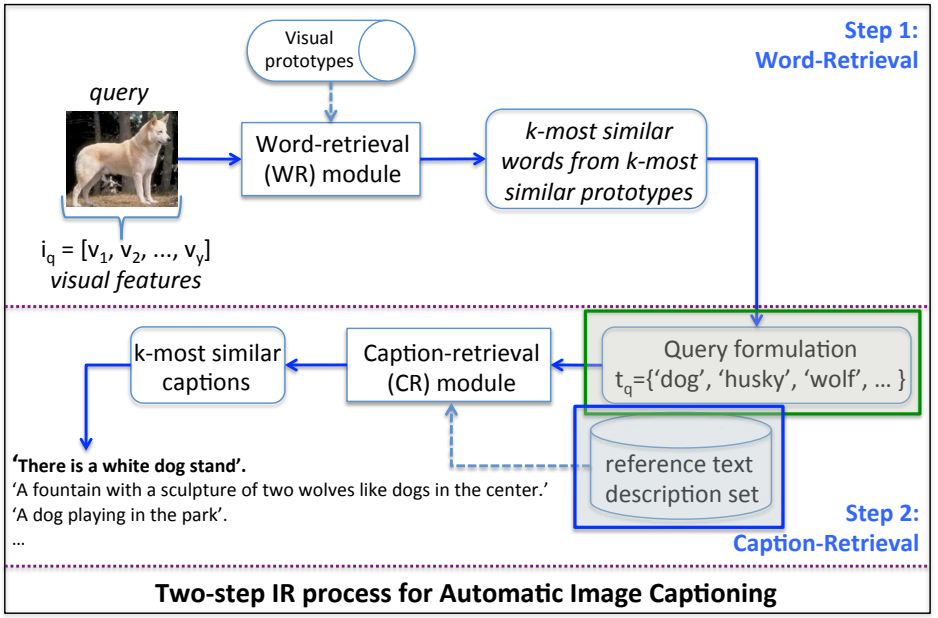
500,000 documents:

- The complete web page (**textual** information).
- Images (**visual** information) represented by visual descriptors: GETLF, GIST, color histogram, a variety of SIFT descriptors, activations of a CNN model of 16 layers. ReLU7 layer were chosen.

Reference description sets:

- **Set A**. The set of sentences from the development data of ImageCLEF 2015, with ≈19,000 sentences.
- **Set B**. Set of sentences used in the evaluation of MS-COCO 2014 dataset [8], with ≈200,000 sentences.

Settings



Two-step IR process for Automatic Image Captioning

Output of the WR step to CR step:

- Number of terms: words, concepts.
- Values: binary, real.

Reference sentence sets:

- Set A - ImageCLEF15.
- Set B - MS-COCO 2014.

Quantitative results (1)

Table: METEOR[1] scores of our method and other approaches.

values	terms	set	RUN	MEAN (STDDEV)	MIN	MAX
real	cpts	A	run1	0.125 (0.065)	0.019	0.568
real	words	A	run2	0.114 (0.055)	0.017	0.423
binary	cpts	A	**run3**	**0.140 (0.056)**	**0.026**	**0.374**
binary	words	A	run4	0.123 (0.053)	0.022	0.526
real	cpts	B	run5	0.119 (0.052)	0.000	0.421
binary	cpts	B	run6	0.126 (0.058)	0.000	0.406
		A,B	RUC-Tencent* [8]	**0.180 (0.088)**	**0.019**	**0.570**
			UAIC$^+$ [1]	0.081 (0.051)	0.014	0.323
			Human [12]	0.338 (0.156)	0.000	0.000

*Long-Short Term Memory based Recurrent Neural Network (LSTM-RNN) trained using MS-COCO14, then fine-tune the model on ImageCLEF development set.
$^+$Template-based approach.

[1]F-measure of word overlaps with a fragmentation penalty on gaps and order.

Qualitative results (1): outputs from the two steps

(1) Query image and its generated description using set A under different settings.

- **WR step:**
  [c]: *helicopter, airplane, tractor, truck, tank, ...*
  [w]: *airbus, lockhe, helicopter, airforce, aircraft, warship, biplane, refuel, seaplane, amphibian, ...*
- **CR step:**
  [c_b]: *A helicopter hovers above some trees.*
  [c_r]: *A helicopter that is in flight.*
  [w_b]: *A large vessel like an aircraft carrier is sat stationary on a large body of water.*
  [w_r]: *A helicopter that is in flight.*

Qualitative results (2): outputs from the two steps

(2) Query image and its generated description using set A under different settings.

- **WR step:**
  $[c]$: *drum, piano, tractor, telescope, guitar, ...*
  $[w]$: *sicken, drummer, cymbal, decapitate, remorse, conga, snare, bassist, orquesta, vocalist, ...*
- **CR step:**
  $[c_b]$: *A band is playing on stage, they are playing the drums and guitar and singing, a crowd is watching the performance.*
  $[c_r]$: *Two men playing the drums.*
  $[w_b]$: *A picture of a drummer drumming and a guitarist playing his guitar.*
  $[w_r]$: *A picture of a drummer drumming and a guitarist playing his guitar.*

Text illustration: reverse problem

Using MI, it is possible to change the direction of the retrieval process, that is, it can be used to illustrate a sentence with images.

The goal is to find an image that best illustrates a given document.

Given a document

I want to travel. A lot. My whole life if I have to.
I want to watch beautiful sunsets and see beautiful flowers and feel the sun on my skin and the salt water in my eyes. I want to meet people.
Lots of them. Kind people. Interesting people.
I want to see the world and get lost in it. I want to look at the stars until I forget everything. I want to know what it's like elsewhere. I want to experience things for myself and decide where I want to live one day. I don't want to feel like I'm trapped in a lonely, gray town.

- m.s.

Extracting topics

Illustrating

Text illustration: qualitative results (1)

1. A sentence is taken as query and used to retrieve images from a reference image collection.
   'Some people are standing on a crowd sidewalk'.
2. Keywords are extracted: *'crowd'*, *'people'*, *'sidewalk'* and *'stand'*.
3. An average visual prototype is formed that is used to retrieve related images.

Text illustration: qualitative results (1)

1. A sentence is taken as query and used to retrieve images from a reference image collection.
   'Some people are standing on a crowd sidewalk'.
2. Keywords are extracted: *'crowd'*, *'people'*, *'sidewalk'* and *'stand'*.
3. An average visual prototype is formed that is used to retrieve related images:

Some of the top images retrieved.

Text illustration: qualitative results (2)

Given the phrase *'A grilled ham and cheese sandwich with egg on a plate'*

The average visual prototype was formed by *'cheese'*, *'egg'*, *'grill'*, *'ham'*, *'plate'* and *'sandwich'*.

Some of the top images retrieved.

Conclusions

- Our method works in an **unsupervised** way using the information of textual and visual features in a multimodal indexing.
- The experimental results show the competitiveness of the proposed method in comparison with state of the art methods that are **more complex** and require **more resources**.
- The multimodal indexing is flexible and can be used for sentence generation for images and text illustration .
- As future work, we will focus on improving our method of multimodal indexing, and also including refined reference sentence sets.

Conclusions

- Our method works in an **unsupervised** way using the information of textual and visual features in a multimodal indexing.
- The experimental results show the competitiveness of the proposed method in comparison with state of the art methods that are **more complex** and require **more resources**.
- The multimodal indexing is flexible and can be used for sentence generation for images and text illustration .
- As future work, we will focus on improving our method of multimodal indexing, and also including refined reference sentence sets.

Conclusions

- Our method works in an **unsupervised** way using the information of textual and visual features in a multimodal indexing.
- The experimental results show the competitiveness of the proposed method in comparison with state of the art methods that are **more complex** and require **more resources**.
- The multimodal indexing is flexible and can be used for sentence generation for images and text illustration .
- As future work, we will focus on improving our method of multimodal indexing, and also including refined reference sentence sets.

Conclusions

- Our method works in an **unsupervised** way using the information of textual and visual features in a multimodal indexing.
- The experimental results show the competitiveness of the proposed method in comparison with state of the art methods that are **more complex** and require **more resources**.
- The multimodal indexing is flexible and can be used for sentence generation for images and text illustration .
- As future work, we will focus on improving our method of multimodal indexing, and also including refined reference sentence sets.

Calfa A. and Iftene A. (2015)

Using textual and visual processing in scalable concept image annotation challenge.
In: *CLEF 2015 Evaluation Labs and Workshop*, Online Working Notes.

Denkowski M., and Lavie A. (2014)

Meteor universal: Language specific translation evaluation for any target language.
In: *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation.*

Farhadi A., Hejrati M., Sadeghi M.A., Young P., Rashtchian C., Hockenmaier J., and Forsyth D. (2010)

Every picture tells a story: Generating sentences from images.
In: *Proceedings of the 11th European conference on Computer Vision*, Part IV, 15-29.

Hodosh M., Young P., and Hockenmaier J.(2013)

Framing image description as a ranking task: Data, models and evaluation metrics.
In: *J. Artif. Int. Res.*, 47, 853-899.

Karpathy A., and Fei-Fei L. (2015)

Deep visual-semantic alignments for generating image descriptions.
In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, USA, 3128-3137.

Krizhevsky A., Sutskever I., and Hinton G.E. (2012)

Imagenet classification with deep convolutional neural networks.
In: *Advances in Neural Information Processing Systems*, 25, Curran Associates, Inc., 1097-1105.

Kulkarni G., Preraj V., Dhar S., Li S., Choi Y., Berg A.C., and Berg T.L. (2011)

Baby talk: Understanding and generating image descriptions.
In: *Proceedings of the 24th CVPR.*

Li X., Jin Q., Liao S., Liang J., He X., Huo Y., Lan W., Xiao B., Lu Y., and Xu J. (2015)

Ruc-tencent at imageclef 2015: Concept detection, localization and sentence generation.
In: *CLEF 2015 Evaluation Labs and Workshop*, Online Workings Notes.

Lin T., Maire M., Belongie S.J., Bourdev L.D., Girshick R.B., Hays J., Perona P., Ramanan D., Dollár P., and Zitnick C.L. (2014)

Microsoft COCO: common objects in context.
In: *CoRR abs/1405.0312*.

Ordonez V., Kulkarni G., and Berg T.L. (2011)
Im2text: Describing images using 1 million captioned photographs.
In: *NIPS*, 1143-1151.

Srivastava N., Salakhutdinov R. (2014)
Multimodal learning with deep boltzmann machines.
In: *Journal of Machine Learning Research*, 15, 2949-2980.

Villegas M., Müller H., Gilbert A., Piras L., Wang J., Mikolajczyk K., de Herrera A.G.S., Bromuri S., Amin M.A., Mohammed M.K., Acar B., Uskudarli S., Marvasti N.B., Aldana J.F., del Mar Roldán García M. (2015)
General Overview of ImageCLEF at the CLEF 2015 Labs.
In: *LNCS*, Springer.

Thank you for your attention, questions?

pellegrin@inaoep.mx