XRCE's Participation at Medical Image Modality Classification and Ad-hoc Retrieval Tasks of ImageCLEF 2011

Gabriela Csurka[1], Stéphane Clinchant[1,2] and Guillaume Jacquet[1]

[1] Xerox Research Centre Europe, 6 chemin de Maupertuis 38240, Meylan France
firstname.lastname@xrce.xerox.com
[2] LIG, Univ. Grenoble I, BP 53 - 38041 Grenoble cedex 9, Grenoble France

Abstract.
The aim of this document is to describe our methods used in the Medical Image Modality Classification and Ad-hoc Image Retrieval Tasks of ImageClef 2011.

The main novelty in medical image modality classification this year was, that there were more classes (18 modalities) organized in a hierarchy and for some categories only few annotated examples were available. Therefore, our strategy in image categorization was to use a semi-supervised approach. In our experiments, we investigated mono-modal (text and image) and mixed modality based classification. The image classification was based on Fisher Vectors built on SIFT-like local orientation histograms and local color statistics. For text representation we used a binarized bag-of-words representation where each element indicated whether the term appeared in the image caption or not. In the case of multi-modal classification, we simply averaged the text and image classification scores.

For the ad-hoc retrieval task, we used the image captions for text retrieval and Fisher Vectors for visual similarity and modality detection. Our text runs were based on a late fusion of different state of the art text experts and the Lexical Entailment model. This Lexical Entailement model used the last year articles to compute similarities between terms and rank first at the previous challenge.

Concerning the submitted runs, we realized that we forgot by inadvertance[3], to submit our best run from last year [3]. We did not submit either improvement over this run, which was proposed in [6]. Overall, this explain the medium performance of our submitted runs. In this document, we show that our system from last year and its improvements would have achieve top performance. We have not tuned the parameter of this system for this year task, we have just evaluated the runs we did not submit !.

Finally, we experimented with different fusion strategies of our textual expert, visual expert and image modality classification scores, which gives consistent results to last year results and to our analysis presented in [6].

Keywords

Multi-modal Information Retrieval, Medical Image Modality Classification, Ad-hoc Retrieval, Semi-supervised learning, Fisher Vector

[3] and because we participated in parallel at several ImageCLEF Task

1 Introduction

This year the medical retrieval task of ImageCLEF 2011 uses a subset of PubMed Central containing 231,000 images [9]. As it was indicated by the clinicians that modality is one of the most important filters to limit their search, a first subtask of the Medical Challenge was the Medical Image Modality Classification. Participants were therefore provided a training set of 1K images that have been classified into one out of 18 modalities organized in a hierarchy (see Figure 1).

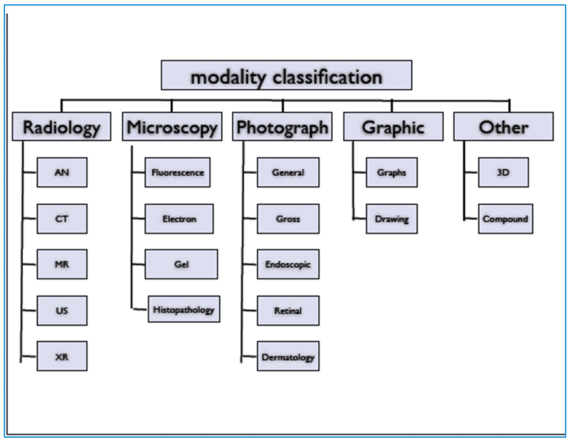

Fig. 1. Medical image modalities.

The main novelty in medical image modality classification this year was, that we had more classes and less annotated data. Furthermore, for some of the categories only few annotated examples were available. Therefore, our main stategy in image categorization was to first automatically augment the training set. We basically used two main approaches, a semi-supervised learning approach and an approach based on an CBIR retrieval scenario (see section 2).

In both cases, we experimented with mono-modal and multi-modal strategies. In the case of visual modality, we used as image representation Fisher Vectors built on SIFT-like local orientation histograms and local color statistics (see for details [10, 7], while text (in our case image captions) were represented by a binarized bag-of-words representation, where each element indicated whether the term appeared in the image caption or not. In the case of multi-modal classification, we simply averaged the text and image classification scores.

Concerning the ad-hoc medical image retrieval task, the only information we used this year was image captions and visual representation. Our text expert was based on a late fusion four textual model built on image captions: a Dirichlet Smoothed language model (DIR), two Power Law Information-Based Model [4] (SPL and LGD) and the Lexical Entailment IR Model [5] (AX). The only model that used other information than the provided image captions was the Lexical Entailment, as it used the last years articles to compute similarities between terms. This text expert was combined with our Fisher Vector based visual model, with modality class predictions or both using different fusion strategies (see section 3).

2 Medical Image Modality Classification Task

In our experiments we investigated both mono-modal and mixed modality based classification. Concerning the pure visual-based classifiers, we used Fisher Vector [10] representation of the images as decribed also in [7]. Note that in the case of medical images we used only a single FV per image and per feature without using the spatial pyramid. The low level features we used were similar to the features used in our Wikipedia runs [7], *i.e.* SIFT-like local orientation histograms (ORH) and local RGB statistics.

Note, that in the medical corpus a large amount of images were 1 channel gray-scale image.To be able to compute the COL features for this images, we first transformed them into 3 channel images where the R, G, B channels were simply made equal each with the luminance channel of the gray scale image. This allowed us to obtain low level features of the same size for grayscale and color RGB images, and hence, to build a common COL visual vocabulary.

Concerning our text representation (TXT), we used a binarized bag-of-words representation, where each element indicated whether the term appeared in this document or not (in our case image caption). Similarly to the Fisher Vectors, we further normalized this vector with Power Norm ($\alpha = 0.5$) and L2 normalization.

To train the classifiers, we used our own implementation of the Sparse Logistic Regression (SLR) [8], (i.e. logistic regression with a Laplacian prior). We trained a classifier per class (one-versus-all) and per feature type (ORH, COL or TXT). Finally, we used the late fusion to combine them where the scores were simply averaged.

However, when we tested this system on the training data, using a 5 fold cross-validation scheme, the results we obtained were rather poor (see Table 1.

Table 1. Modality classification experiment using the provided training data with a 5 fold cross-validation scheme. The accuracy results (average of the diagonal of the confusion matrix) are in %.

Features	ORH	COL	TXT	ORH+COL	ORH+TXT	ORH+COL+TXT
ACC	57.02	55.73	49.23	62.2	61.5	64.66

Analyzing our results per class, we realized that the low performance might be due to the fact that for some of the categories only few annotated examples were available. Therefore, we decided to automatically increase the training set. We basically used two main approach for this a semi-supervised learning approaches and a visual retrieval based approach and the image collection from the medical retrieval task.

The main idea of the semi-supervised learning approaches was to use the modality classifiers trained on the provided training data to rank the 231K images of the collection based on the classification score. Hence for each modality, the corresponding top K documents were considered as most probably correctly classified, labeled with the given modality and added to the training set.

In the case of the second scenario, we first built an image query with a random set of images labeled with the given modality. The images in the collection were ranked based on their average similarity (dot product between Fisher Vectors) to the query image set. The top K documents ranked as most similar to the query set were added to the training set labeled with the given modality.

Finally, the modality classifiers were retrained with the increased training using different feature types and combined as described above. We submitted different runs (detailed below) using either only visual information or both visual and textual.

2.1 *Visual only based runs:*

Table 2 describes the results of the following "visual only" based runs:

- **V1**: For this run we used the semi-supervised approach with the visual classifier using COL+ORH features in both steps (training with the original set and training with the increased training set). After the first step we added the top 25 images for each modality classifier.
- **V2** and **V3**: For this run we used the visual retrieval to increase the training set. To build our queries, we used the labeled images from two of the 5 folds used in our first experiments

Table 2. Overview of our visual only runs at the Medical Image Modality Classification task.

	RUN	Modality	ACC
V1	XRCE_all_VIS_semiL25	Visual	0.8359
V2	XRCE_Testset_VIS_semi20_CBIR	Visual	0.8349
V3	XRCE_all_VIS_semi20_CBIR	Visual	0.8339

(one for each run). The top 20 images for each query were added to the training set and a COL+ORH visual categorization system trained.

Note that both stategies leaded to similar performances. The choice of k might be non-optimal and a better strategy would be to learn a different k for each modality based on some confidence of the classification scores.

2.2 *Mixed modality based runs:*

Table 3. Overview of our mixed modality runs at the Medical Image Modality Classification task.

	RUN	Modality	ACC
M1	XRCE_all_MIX_semiLM	Mixed	0.8691
M2	XRCE_Testset_MIX_semiL50	Mixed	0.8642
M3	XRCE_Testset_MIX_semiL25	Mixed	0.8593

Table 3 describes the results of the following "mixed modality" runs:

- **M1**: For this run we used the semi-supervised approach with the multi-modal classifier (COL+ORH+TXT) in both steps (training with the original set and training with the increased training set). After the first step we added the top 50 images for each modality classifier.
- **M2** and **M3**: For this run we used the semi-supervised approach with the visual classifier (COL+ORH) in the first step (same as for V1) and the multi-modal classifier (COL+ORH+TXT) in the final step. As their names show, for M2 we added the top 50 and in the case of M3 we added the top 25 images for each classifier.

As the table 3 shows, all the "mixed modality" runs outperformed the "visual only" runs with about 2-3% in accuracy (see for example V1 and M3, where the enriched training set was the same). Finally, comparing with the results in table 1, we can see that both strategy to increase the training set was extremely useful leading to an absolute 25% over the baseline using only the annotated training set.

3 Medical Ad-hoc Image Retrieval Task

Concerning the ad-hoc medical image retrieval task, the only information we used this year was image captions (not the full articles) and visual representation. Using these information, we built several mono-modal and multi-modal systems. In addition, we also integrated the modality classification scores with these systems. In which follows, we give further details on these methods and the corresponding runs.

3.1 *Visual only retrieval system*

As visual representation, we used exactly the same representation as in the modality classification, namely Fisher Vectors with ORH and COL features. As we use the dot product as similarity measure, the sum of similarities (used in our case) between the FVs of the corresponding features is equivalent to the dot product between image signatures built as a concatenation of the FV ORH with FV COL. As the Table 4 shows, adding the color information slightly improve the retrieval results, however as expected the visual similarity alone is not sufficient to handle the semantic queries.

Table 4. Medical ad-hoc retrieval: using visual only information (not submitted). MAP and P10 are shows using percentages.

	Model	MAP	P@10
V1	ORH	1.18	7.00
V2	ORH+COL	1.81	11.00

3.2 *Text based retrieval systems*

Concerning the text representation, we used only the image captions that were first pre-processed including tokenization, lemmatization, and standard stopword removal. As in some cases lemmatization might lead to a loss of information, when we constructed the dictionary, we kept for each term its lemmatized and non lemmatized version. Then, each caption was transformed first into a bag-of-words representation on which we built basically four textual models. These models (summerized in Table 5) were Diriclet Smoothed standard language model (similar to the techniques used in our past participation in other tasks of ImageCLEF [1]), two Power Law Information-Based Model [4] (LGD and SPL) and finally the Lexical Entailment IR Model [5].

Table 5. Notations for the Runs Name

Notations	Descriptions
DIR	Dirichlet Smoothed Language Model [11]
LGD	Log-logistic Information-Based Model [4]
SPL	Smoothed Power Law Information-Based Model [4]
AX	Lexical Entailment based IR Model [5]

In the first model (DIR), we used with standard Language Model representation the Dirichlet smoothing that gives the following retrieval model [11]:

$$RSV(q,d) = \sum_{w \in Q, x_{dw} > 0} x_w^q \log(1 + \frac{x_w^d}{\mu p(w|C)}) + l_q \log \frac{\mu}{l_d + \mu} \tag{1}$$

where x_w^d is the number of occurrences of word w in document d, l_d is the length of d in tokens after lemmatization, μ the smoothing parameter and $p(w|C)$ is the corpus language model.

The Log-logistic model (LGD) has the same steps as the Smoothed Power Law model (SPL) (described in [7]). The only difference is that Relevance Score Vector in the Ranking Model becomes (see details in [4]):

$$RSV(q,d) = \sum_{w \in q \cap d} x_w^q \left[-\log P(Tf_w > t_w^d) \right] \tag{2}$$

Our last model, the Lexical Entailment based IR Models (AX) is also described in [7]. For this model, we need to compute the probabilistic term similarities between terms. However, using only the image captions, gives a context rather poor for the words. As we didn't processed the full articles from this year, we used the processed articles from last year's medical corpus.

Finally, expecting to bring complementary information, we averaged (with equal weighting) several models to get a single text expert. However, as the Table 6, our expectation was rather wrong, and instead we bringed more noise than useful information. Comparing the four models shows that without any external information, using only the image captions is unsufficient. However, using the AX model alone (not submitted [4]) gives better performance than all other runs, including the ones submitted to the challenge by other teams. Note that only T6 and T7 (corresponding to the runs XRCE_RUN_TXTax_dir_spl respectively XRCE_RUN_TXT_noMOD) were submitted (results in red).

Table 6. Medical ad-hoc retrieval: overview of the performances of using different text models. Finally for comparison, we added the best textual run in the challenge.

	Model	MAP	P@10
T1	DIR	16.62	27.00
T2	LGD	16.67	30.33
T3	SPL	17.25	29.00
T4	AX	**22.95**	38.67
T5	AX+DIR	21.86	36.67
T6	AX+DIR+SPL	**18.70**	**32.33**
T7	AX+DIR+SPL+LGD	**18.02**	**31.00**
	best TXT run	21.72	34.67

3.3 *Multi-modal retrieval systems*

As multi-modal retrieval system, we used the *Late Semantic Combination (LSC)* proposed in [2], but also described in [7]. The main idea of this late fusion is that first we use the text expert to select the top $N = 1000$ semantically relevant documents, and then we average their textual and visual scores. Results for different text experts using equal weighting ($w_T = w_V = 0.5$) and unbalanced weighting between scores images and score texts ($w_T = 0.9$ and $w_V = 0.1$) are shown in Table 7. Note that we submitted only M6 (corresponding to XRCE_RUN_MIX_SFL_noMOD_ax_dir_spl) and M7 (corresponding to XRCE_RUN_MIX_SFL_noMOD_ax_dir_spl_lgd) with equal weigthing (results in red in the table).

We can see first that using LSC with equal weighting leaded to a decrease of the MAP but increased in most cases the P10 value. On the contrary, giving a much important weighting for the text scores compared to the image similarities ($w_T = 0.9$ and $w_V = 0.1$), we are able in some cases to improve over our text results such as the model AX. As we obtain similar performances or below using the late fusion, we can say that for this task the image similarities did not really help to improve the retrieval.

3.4 *Multi-modal retrieval systems with image modality prediction*

In this section we show the performance of our retrieval system when we combined them with the modality prediction. Therefore, for each topic each image and the query text was individually clssified by our mono-modal modality classifier, and we retained the modality (or modalities) we

[4] As explained in our abstract, we realized that we forgot by inadvertance (and because we participated in several ImageCLEF Task), to submit our best run from last year

Table 7. Medical ad-hoc retrieval: overview of multi-modal runs using the Late Semantic Combination method without modality predictions.

LCS Model /(w_T, w_V)		(0.5, 0.5)		(0.9, 0.1)	
MIX	TXT expert	MAP	P@10	MAP	P@10
M1	DIR	15.22	33.67	16.44	27.33
M2	LGD	13.32	34.33	16.99	29.33
M3	SPL	13.44	32.67	16.48	29.33
M4	AX	16.98	37.67	**23.28**	37.67
M5	AX+DIR	**17.30**	36.33	21.87	37.67
M6	AX+DIR+SPL	**14.72**	**34.33**	18.82	32.00
M7	AX+DIR+SPL+LGD	**14.29**	**33.67**	18.10	30.33
	best MIXED run	**23.72**	**39.33**	**23.72**	**39.33**

obtained. Note that in the case of topics where any type of images were allowed, we also considered only modalities obtained by this automatic model. Hence our model was sub-optimal.

Then, we experimented with two strategies. In the first case (FILT), for each document in the dataset that corresponded to the selected modality of the topic we boosted its retrieval score (multiplied by 2), while all other scores were retained unchanged. Hence if a score was high and with the desired modality, it was significantly increased, while if the retrieval score was low, the modality classifier had smaller effect on it.

In the second case (Mscore), for each document we added to the retrieval score the classification score of the query modality. When several modalities were retrained, we used the maximum of all thoses scores. Results for both strategies applied to different text runs are shown in Table 8.

Table 8. Medical ad-hoc retrieval: overview of multi-modal runs using text based expert and modality predictions using different combination strategies.

Model /strategy		Filt		Mscore	
Ti+Mod	TXT expert	MAP	P@10	MAP	P@10
TM1	DIR	20.27	38.67	18.48	36.67
TM2	LGD	19.29	35.67	18.92	34.33
TM3	SPL	19.77	36.67	19.27	34.00
TM4	AX	24.15	40.33	21.31	37.67
TM5	AX+DIR	**24.54**	**41.00**	21.55	38.67
TM6	AX+DIR+SPL	21.10	38.33	20.78	36.00
TM7	AX+DIR+SPL+LGD	20.54	37.33	20.12	40.00
	best MIXED run	**23.72**	**39.33**	**23.72**	**39.33**

Unfortunately, we submitted only our equal weighted mixed runs M6 and M7 combined with the modality classifier instead of our text runs, shown in table 9. While these results confirm that the modality classification helps, their performance is significantly worse that the performances obtained with TM4 and TM5.

4 Conclusion

In this document, we describe the methods we used in Medical Image Modality and Medical Ad-hoc Retrieval Tasks at ImageClef 2011. We have shown that while for some classes only few examples were available, using a semi-supervised approach to increase the training data can lead to very significant improvement of the results. With this method our method came again as best performing in the challenge. Concerning the ad-hoc retrieval task, our strategy to average

Table 9. Medical ad-hoc retrieval: overview of multi-modal runs using the Late Semantic Combination method with modality predictions.

Model /strategy		NoMod		Filt		Mscore	
Mi+Mod	TXT expert	MAP	P@10	MAP	P@10	MAP	P@10
MM6	AX+DIR+SPL	14.72	34.33	15.20	36.33	16.43	38.00
MM7	AX+DIR+SPL+LGD	14.29	33.67	15.12	36.67	15.45	38.00

several text models instead of using only the Lexical Entailment IR Model [5] (AX) leaded to a text expert that had a medium performance. Further combining this with a equally weighted Late semantic combination leaded to a retrieval system that performed even poorer than our text expert. While the combination of this expert allowed to increase slightly the performance of the system, it remained far behind the best performing systems in the challenge. However, after testing our not submitted runs we realized that our AX text model alone performed better that the best performing text expert in the challenge, and when we further combine it with the modality classifier, the system out-performs the best submitted run.

Acknowledgments

We would like also to acknowledge Florent Perronnin and Jorge Sánchez for the efficient implementation of the Fisher Vector computation and the of the Sparse Logistic Regression (SLR) we used in our experiments.

References

1. J. Ah-Pine, S. Clinchant, G. Csurka, F. Perronnin, and J-M. Renders. *Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval*, volume The Information Retrieval Series, chapter 3.4. Springer, 2010. ISBN 978-3-642-15180-4.
2. Stéphane Clinchant, Julien Ah-Pine, and Gabriela Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2011.
3. Stéphane Clinchant, Gabriela Csurka, Julien Ah-Pine, Guillaume Jacquet, Florent Perronnin, Jorge Sanchez, and Keyvan Minoukadeh. Xrce's participation in wikipedia retrieval, medical image modality classification and ad-hoc retrieval tasks of imageclef 2010. In *Working Notes of CLEF 2010, Padova, Italy*, 2010.
4. Stéphane Clinchant and Eric Gaussier. Information-based models for ad hoc ir. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241, New York, NY, USA, 2010. ACM.
5. Stéphane Clinchant, Cyril Goutte, and Éric Gaussier. Lexical entailment for information retrieval. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12*, pages 217–228, 2006.
6. Gabriela Csurka, Stéphane Clinchant, and Guillaume Jacquet. Medical image modality classification and retrieval. In *International Workshop on Content-based Multimedia Indexing*, 2011.
7. Gabriela Csurka, Stéphane Clinchant, and Guillaume Jacquet. Xrces participation at wikipedia retrieval of imageclef 2011. In *Working Notes of CLEF 2011, Amsterdam, The Netherlands*, 2011.
8. B. Krishnapuram and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *PAMI*, 27(6), 2005.
9. Henning Müller, Jayashree Kalpathy-Cramer, and Steven Bedrick. Overview of the clef 2011 medical image retrieval track. In *Working Notes of CLEF 2011, Amsterdam, The Netherlands*, 2011.
10. F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.
11. C. Zhai and J lafferty. A study of smoothing methods for language models applied to ad hoc to information retrieval. In *Proceedings of SIGIR'01*, pages 334–342. ACM, 2001.