CEA LIST's participation to the Concept Annotation Task of ImageCLEF 2012

Amel Znaidia, Aymen Shabou, Adrian Popescu, and Hervé Le Borgne

CEA, LIST, Laboratory of Vision and Content Engineering, France

amel.znaidia@cea.fr , aymen.shabou@cea.fr, adrian.popescu@cea.fr , herve.le-borgne@cea.fr

Abstract. This paper describes our participation to the ImageCLEF2012 Photo Annotation Task. We focus on how to use the tags associated to the images to improve the annotation performance. We submitted one textual-only and three multimodal runs. Our first textual model [14] is based on the local soft coding of images tags over a dictionary of most frequent tags. A second model of tag is an adaptation of the TF-IDF model to the social space in order to compute the social relatedness of two tags[9]. For the fusion we used a trainable combiner, called stacked generalization [12] which uses predictions from base classifiers to learn a new model. Results have shown that combination of textual and visual features can improve the annotation performance significantly. Our best run achieves 41.59 % in terms of MAP, allowing us to rank 3^{rd} team.

Keywords: Multimedia fusion, Bag-of-Visual-Words, Bag-of-Multimedia-Words, image annotation, classification.

1 Introduction

The ImageCLEF 2012 Photo Annotation Task [8] is a multi-label classification problem, with 15.000 image for training, 10.000 for testing and 94 concepts to detect. Images are extracted from the MIR Flickr dataset [4] and the Flickr user tags and/or EXIF information are available for most photos.

In our participation to the ImageCLEF Photo Annotation Task, we focus on how to use the tags associated to the images to enhance the annotation performance. We propose three different models: textual only and two multimodal models.

This paper is organized as follows. In Section 2 we describe our visual features. In Section 3 we give an overview of our textual model which uses user tags. Then in Section 4 we present in more detail the experiments we did, the submitted runs and the obtained results.

2 Visual Features

2.1 Bag-of-Visual-Words model

In recent works addressing object recognition and scene classification tasks, the Bag-of-Visual-Words (BoVW) is one of the most popular model for feature design. Given an

image, its visual features are built in three steps (i) codebook learning, (ii) local features coding and (iii) pooling.

1. **Codebook Learning**
   The codebook, which entries are termed codewords, is a collection of basic patterns used to reconstruct the input local features. A simple way to generate the codebook is to use clustering based methods such as K-means [5]. In the rest of this paper, we denote by $\mathbf{B} = \left\{\mathbf{b}_k\,;\,\mathbf{b}_k \in \Re^d\,;\,k = 1, ..., K\right\}$ a codebook of K codeword vectors, which is learned on a training subset of local features $\left\{\mathbf{x}_i\,;\,\mathbf{x}_i \in \Re^d\,;\,i = 1, ..., N\right\}$ extracted from the learning dataset.
2. **Coding**
   For each image dense local descriptors (such as SIFTS [7]) are extracted and mapped to codes. Following recent observations in scene classification, we chose to implement the locality-constraint coding based on local soft coding [6], because of its effectiveness and robustness toward quantization errors. In [6], authors propose another efficient implementation of the locality-constrained coding by restricting the probabilistic soft coding approach [3] to only the M-nearest-codewords to a local feature, i.e.,

$$z_{i,j} = \begin{cases} \frac{\exp\left(-\beta||\mathbf{x}_i - \mathbf{b}_j||_2^2\right)}{\sum_{k=1}^{M}\exp\left(-\beta||\mathbf{x}_i - \mathbf{b}_k||_2^2\right)} & \text{if } \mathbf{b}_j \in \mathcal{N}_M(\mathbf{x}_i)\,, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

   where $\mathcal{N}_M(\mathbf{x}_i)$ denotes the M-nearest neighborhood of $\mathbf{x}_i$, under the Euclidean distance for instance.
3. **Pooling**
   Given the coding coefficients of all local features within one image, a pooling operation has to be performed to obtain a compact signature $\mathbf{h}$, while preserving important information and discarding irrelevant details. This operation can be formulated as the following:

$$h_j = g\Big(\left\{z_{i,j}\,;i \in \{1, ..., N\}\right\}\Big)\,;\forall j \in \{1, ..., K\}\,, \tag{2}$$

   with g a pooling function such as the average, the sum or the maximum functions. The sum-pooling is the sum of the coding coefficients obtained on local features while the average-pooling is its normalized form. Both have been usually considered in the original BoW model. Recent works [2, 6] show, both theoretically and empirically, that max-pooling is best suited to the recognition task. Max-pooling is obtained by selecting the maximum coding coefficient (or codeword response) over local features for each codeword.
   Furthermore, since the classic BoVW is an orderless signature that disregards the location of the visual words in the image, the spatial pyramid matching (SPM) [5] is an interesting way to incorporate some global spatial contextual information into the signature. The image is divided into P different regions and a pooling is conducted in each of them. The final signature is then obtained by a concatenation of all the region-relative R_i signatures, i.e.,

$$\mathbf{h} = [\mathbf{h}_{R_1}^T, \mathbf{h}_{R_2}^T, ..., \mathbf{h}_{R_P}^T]^T\,. \tag{3}$$

2.2 Bag-of-Multimedia-Words model

The Bag-of-Multimedia-Words is a method of early fusion that combines textual and visual features. Since the late fusion method presented in section 2.1 gives better result, we do not present this method in this paper and refer to [15] for further details.

3 Textual Features

It is commonly accepted [10] that visual features alone do not convey a high level semantic description of image content. In order to build robust BoW based tag-signatures toward quantization errors, we rely on the locality-constrained coding method that has proved to be effective for visual features when paired with max-pooling. This model is detailed in [14]. The coding step of a given tag over a codebook requires a tag-similarity measure.

The two similarity measures that we detail below, capture complementary facets of tags and their combination improves the quality of predicted tags.

- ***Hierarchical similarity:***
  WordNet concepts are structured as synsets (sets of synonyms) that are arranged as a hierarchy whose main structural axis is defined by conceptual inheritance. Wu-Palmer measure [13] gives a similarity between two tags as their distance in the WordNet hierarchy.
  Since a tag can belong to more than one synset in WordNet (i.e., can have more than one meaning), we opt to determine the semantic relationship between two tags $\mathbf{t}_1$ and $\mathbf{t}_2$ as the maximum Wu-Palmer similarity between the synset or the synsets that include syns($\mathbf{t}_1$) and syns($\mathbf{t}_2$):

$$\text{sim}_{\text{hierarchical}}(\mathbf{t}_1, \mathbf{t}_2) = \max \left\{ \text{sim}_{\text{wup}}(\mathbf{s}_1, \mathbf{s}_2) \, ; \; (\mathbf{s}_1, \mathbf{s}_2) \in \text{syns}(\mathbf{t}_1) \times \text{syns}(\mathbf{t}_2) \right\}, \quad (4)$$

  where sim_{wup} is the Wu-Palmer similarity.
- ***Contextual similarity***
  In [9], an adaptation of the TF-IDF model to the social space is proposed in order to compute the social relatedness of two tags.
  Let $\mathbf{S}$ be the matrix of size $N \times K$ defined by:

$$\mathbf{S}(i,j) = \text{users}(\mathbf{t}_i, \mathbf{t}_j) \times \log\left(\frac{\text{users}_{\text{collection}}}{\text{users}_{\text{collection}(\mathbf{t}_j)}}\right), \quad (5)$$

  where $\mathbf{t}_i$ is the target tag, $\mathbf{t}_j$ is an element of the codebook, users($\mathbf{t}_i, \mathbf{t}_j$) is the number of distinct users who associate the tag $\mathbf{t}_i$ to the tag $\mathbf{t}_j$ among the top results returned by the Flickr API for $\mathbf{t}_i$; $\text{users}_{\text{collection}(\mathbf{t}_j)}$ is the number of distinct users from a pre-fetched subset of Flickr users that have tagged photos with tag $\mathbf{t}_j$, and N is the number of unique tags associated to photos of the dataset and K is the size of the codebook.

Relying on this matrix, a Flickr model for a given tag $\mathbf{t}_i$ is proposed in [9] as the following vector of weights:

$$\mathbf{w}_i = [w_{i,1}, w_{i,2}, ..., w_{i,K}]^T , \tag{6}$$

with $w_{i,j}$ the normalized social weight defined by:

$$w_{i,j} = \frac{\mathbf{S}(i,j)}{max\{\mathbf{S}(i,k)\,, k = 1, ..., K\}} . \tag{7}$$

Thereby, given two tag-Flickr models $\mathbf{w}_i$ and $\mathbf{w}_j$, we compute the contextual similarities between their related tags $\mathbf{t}_i$ and $\mathbf{t}_j$ using the cosine similarity:

$$\text{sim}_{\text{contextual}}(\mathbf{t}_i, \mathbf{t}_j) = \frac{\mathbf{w}_i^T \mathbf{w}_j}{||\mathbf{w}_i|| ||\mathbf{w}_j||} . \tag{8}$$

Coding/pooling Once the similarity measures are calculated, we perform local soft coding for each $\mathbf{t}_i$ in order to achieve the assignment step. Consequently, a tag is mapped to only its M-nearest tags under a similarity measure.

$$z_{i,j} = \begin{cases} \text{sim}(\mathbf{t}_i, \mathbf{b}_j) & \text{if } \mathbf{b}_j \in \mathcal{N}_M(\mathbf{t}_i)\,, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathcal{N}_M(\mathbf{t}_i)$ denotes the M-nearest neighbors of $\mathbf{t}_i$, under the hierarchical or the contextual similarity denoted by $\text{sim}(\mathbf{t}_i, \mathbf{b}_j)$. The locality assumption in the tag-space induces sparse codes while reducing the reconstruction errors, mainly in terms of semantic reconstruction.

Given the tag-related codes within one image, a max-pooling is performed in order to obtain the final tag-signature vector. In our case, separate signatures are generated considering each similarity measure.

3.1 Classifier Fusion

A linear SVM classifier is used for the features obtained from each modality. To combine classifiers learned on different modalities and/or features, we use a trainable combiner, called **stacked generalization**, originally introduced in [12]. It is an ensemble learning technique, which aims to increase the performance of individual classifiers by combining them in a hierarchical architecture. The key idea is to learn a meta-level (level-1) classifier based on the outputs of base-level (level-0) classifiers, estimated via cross-validation. An example of combination of one visual and two textual classifiers is presented in Figure 1.

Given a training dataset $\mathbf{D} = \{(\mathbf{x}_i^F, \mathbf{y}_i), i = 1, ..., n\}$ where $\mathbf{x}_i^F$ is the F-feature vector among the visual (V-feature), the contextual tag (C-feature) and the hierarchical tag (S-feature) and $\mathbf{y}_i$ is the associated vector of labels, the algorithm operates as follows:

1. A K-fold cross-validation process randomly splits $\mathbf{D}$ into disjoint parts of almost equal size $\mathbf{D}_1, ..., \mathbf{D}_K$;

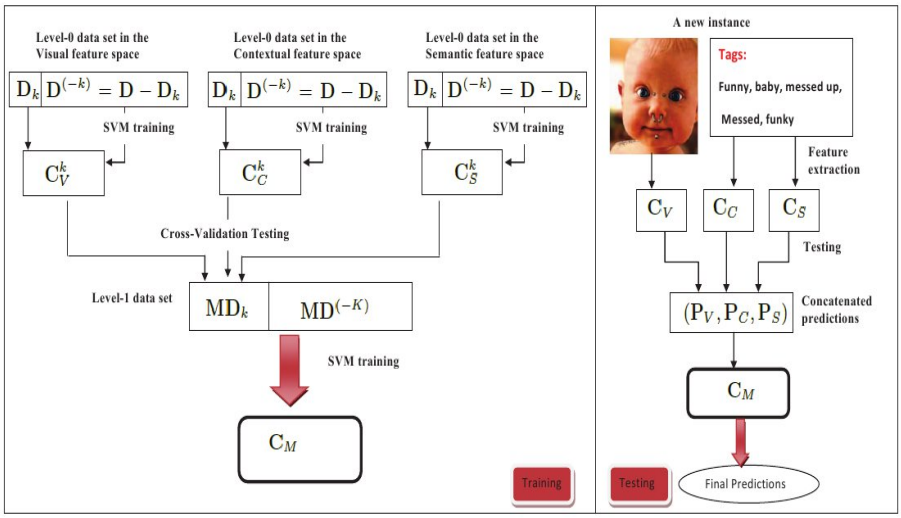

Fig. 1. Left Part is the training scheme of the meta-level classifier and right part is the classification using stacking framework.

2. At each k^{th} fold, $\mathbf{D}_k$ and $\mathbf{D}^{(-k)} = \mathbf{D} - \mathbf{D}_k$ are used respectively as test and train parts. Here 3 linear SVMs are applied to $\mathbf{D}^{(-K)}$ giving 3 level-0 confidence matrix for the visual, contextual and semantic (hierarchical) modalities, denoted by $\mathbf{C}_V^k, \mathbf{C}_C^k$ and $\mathbf{C}_S^k$ respectively;
3. Given the concatenated predictions of these outputs on each sample of $\mathbf{D}_k$ and their class labels, a new set $\mathbf{MD}_k$ is then formed. At the end of the cross-validation process, the union $\mathbf{MD} = \cup_{k=1}^{K} \mathbf{MD}_k$ constitutes the meta-level data set that is used to train the meta-level classifier $\mathbf{C}_M$;
4. The three modality based linear SVMs are now trained on the entire dataset to induce the final base-classifiers $\mathbf{C}_V, \mathbf{C}_C$ and $\mathbf{C}_S$ required by the classification task;
5. Finally, given a new instance, the concatenated predictions of all level-0 classifiers $\mathbf{C}_V, \mathbf{C}_C$ and $\mathbf{C}_S$ are used as input for the level-1 classifier $\mathbf{C}_M$ to compute the final prediction scores.

4 Experiments

4.1 Submitted runs

We submitted four runs to the campaign, allowing relevant comparison between methods:

– **textual_tagflickr_tagwordnet** uses only the textual feature described in section 3. The codebook size is fixed to 2500 (resp. 5134) for the hierarchical (resp. contextual) similarity. the codewords for the soft assignment. The optimal size of the

neighborhood has been estimated by cross-validation on the training dataset leading to a number of 5 (resp. 50) neighboring codewords for the hierarchical (resp. contextual) based tag-distance measures. A one-versus-all linear kernel based Support Vector Machine (SVM) classier is used, for each measure. They are combined using the stack generalization using a 10-fold cross validation.

– **multimedia_visualrootsift_tagflickr_tagwordnet** This run is a combination of the previous textual run and the Bag-Of-Visual-Words model detailed in section 2.1. The pipeline is as follows :
  - **Local visual descriptors:** dense SIFTs of size 128 are extracted within a regular spatial grid and only one scale. The patch-size is fixed to 16×16 pixels and the step-size for dense sampling to 6 pixels;
  - **Codebook:** a visual codebook of size $4,000$ is created using the K-means clustering method on a randomly selected subset of SIFTs from the training dataset ($\sim 10^5$ SIFTs).
  - **Coding/pooling:** for coding the local visual descriptors SIFTS, we also fix the patch-size to 16×16 pixels and the step-size for dense sampling to 6 pixels. Then for the extracted visual descriptors associated to one image, we consider a neighborhood in the visual feature space of size 5 for local soft coding and the softness parameter β is set to 10. The max-pooling operation is performed to aggregate the obtained codes and a spatial pyramid decomposition into 3 levels ($1 \times 1, 2 \times 2, 3 \times 3$) is adopted for the visual-signature.

    A one-versus-all linear kernel based Support Vector Machine (SVM) classifier is used, since it has shown good performances in scene categorization task when paired with the max-pooling operation on local features [11, 6].
  - **Classifier fusion:** base classifiers are trained on the considered modalities (visual, contextual and hierarchical) and combined by the stack generalization approach using 10-cross-validations on the training set as shown in figure 1.
– **multimedia_visualcsift_tagflickr_tagwordnet** Is the same run as the previous one except the SIFT version. In this run, we use a colored SIFT (CSIFT) [1].

4.2 Results and Discussion of Submission Outcomes

The official results of our runs are illustrated in Table 1. Among the multimodal runs (2 and 3), we notice that using a Colored SIFT works better than the conventional SIFT. The multimodal run (run 4) scores shows the competitive performances of the Bag-of-Multimedia-Words, ensuring a trade-off between classification accuracy and computation cost. This model is easier to scale for large-scale datasets since it achieves comparable performances compared to the other multimodal runs while using only a feature vector of size 512.

The first purely textual submission is the combination of the semantic and the contextual classifiers detailed in section 3. Its performance was almost identical to the best textual submission of *LIRIS ECL* Group (the best MAP in the textual modality) as shown in Table 2.

Run	Modality	MAP	GMAP	F-ex
1: textual_tagflickr_tagwordnet	Textual	0.3314	0.2698	0.4452
2: multimedia_visualrootsift_tagflickr_tagwordnet	Multimodal	0.4086	0.3472	0.5374
3: multimedia_visualcsift_tagflickr_tagwordnet	Multimodal	0.4159	0.3615	0.5404
4: multimedia_bomw	Multimodal	0.4084	0.3487	0.5295

Table 1. Overview of the different submissions.

Run	MAP	GMAP	F-ex
Our textual run	0.3314	0.2698	0.4452
LIRIS ECL textual run	0.3338	0.2759	0.4691

Table 2. Comparison of our textual submission and the best textual one.

5 Acknowledgment

This work is supported by grants from DIGITEO and Région Ile-de-France, and has been partially funded by I2S in the context of the project Polinum. We acknowledge support from the French ANR (Agence Nationale de la Recherche) via the YOJI (ANR-09-CORD-104) and PERIPLUS (ANR-10-CORD-026) projects.

References

1. Abdel-Hakim, A.E., Farag, A.A.: Csift: A sift descriptor with color invariant characteristics. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2. pp. 1978–1983. CVPR '06, IEEE Computer Society, Washington, DC, USA (2006)
2. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR. pp. 2559–2566 (2010)
3. van Gemert, J., Veenman, C., Smeulders, A., Geusebroek, J.: Visual word ambiguity. PAMI pp. 1271–1283 (2009)
4. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval. ACM, New York, NY, USA (2008)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2169–2178 (2006)
6. Liu, L., Wang, L., Liu, X.: In Defense of Soft-assignment Coding. In: ICCV (2011)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision (IJCV) 60(2), 91–110 (2004)
8. Nowak, S., Nagel, K., Liebetrau, J.: The clef 2011 photo annotation and concept-based retrieval tasks. In: CLEF 2011 working notes (2011)
9. Popescu, A., Grefenstette, G.: Social media driven image retrieval. In: ACM International Conference on Multimedia Retrieval (ICMR). pp. 33:1–33:8 (2011)
10. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 22, 1349–1380 (2000)

11. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)
12. Wolpert, D.H.: Stacked generalization. Neural Networks 5, 241–259 (1992)
13. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Annual Meeting of the Association for Computational Linguistics. pp. 133 –138 (1994)
14. Znaidia, A., Shabou, A., Borgne, A.P.H.L., Hudelot, C.: Multimodal Feature Generation Framework for Semantic Image Classification. In: ACM International Conference on Multimedia Retrieval (ICMR), Hong Kong (Jun 2012)
15. Znaidia, A., Shabou, A., Le Borgne, H., Hudelot, C., Paragios, N.: Bag-of-Multimedia-Words for Image Classification. In: International Conference on Pattern Recognition ICPR'12, Tsukuba, JAPAN (Nov 2012)