A Ranking Based Model for Automatic Image Annotation in a Social Network

Ludovic Denoyer and **Patrick Gallinari**
University Pierre et Marie Curie - LIP6
104 avenue du president Kennedy
75016 PARIS FRANCE
ludovic.denoyer@lip6.fr - patrick.gallinari@lip6.fr

Abstract

We propose a relational ranking model for learning to tag images in social media sharing systems. This model learns to associate a ranked list of tags to unlabeled images, by considering simultaneously content information (visual or textual) and relational information among the images. It is able to handle implicit relations like content similarities, and explicit ones like friendship or authorship. The model itself is based on a transductive algorithm thats learns from both labeled and unlabeled data. Experiments on a real corpus extracted from Flickr show the effectiveness of this model.

Introduction

We consider the problem of image annotation in large social media sharing web sites. These systems allow users to upload and share pictures and videos over the Web. Popular systems like Flickr or Youtube[1] contain billions of images/videos. Retrieving relevant information in such large collaborative systems is usually handled via manual tagging of the corpus resources. Textual tags then allow using simple and fast keyword based search engines.

Tagging large collections is often prohibitive and manual tags are known to be imprecise, ambiguous, inconsistent and subject to many variations (Matusiak 2006). Automatic and collaborative tagging methods, aimed at improving the quality of annotations and at helping users to tag their resources have been the subject of an intensive recent research. However, for most applications, performance is still quite deceiving.

The most common approaches attack the tagging problem as a supervised classification problem, where a visual signal has to be associated to a set of known key words ((Hironobu, Takahashi, and Oka 1999), (Chang et al. 2003), (Li and Wang 2008),...). A recent work proposes to exploit also the visual correlation between images (Wu et al. 2009). Some methods propose to use the dependencies between tags as an additionnal information (Liu et al. 2009). Several unsupervised methods, exploiting the co-occurrence of image cues and keywords, have also been proposed, e.g. (Barnard et al. 2003). All these methods have strong limitations. Classification or latent variable techniques for example can only learn a correspondence between visual information and corresponding keywords and cannot cope with more specific or abstract annotations. A few methods and applications have been developed for exploiting relational information in image collections. For example, Tong et al. in (Tong et al. 2006) propose a semi-supervised method for propagating tags. Cao et al. (Cao, Luo, and Huang 2008) handle metadata and also use label propagation. Both systems only exploit implicit relations computed among visual features and tags. The article by Stone et al. proposes to use social information in facebook for finding people in photos (Stone, Zickler, and Darrell 2008). It is focused on face recognition and not on general tagging.

We introduce here a new relational learning model designed for image annotation in a social media. It is able to exploit simultaneously visual and textual content, metadata, and relational information, either implicit like word or image similarities, or explicit, like friendship links of a social network. In particular the model is able to exploit the rich folksonomy present in social sites. The specific task the model is used for, is the ranking of tags associated to images in a social media collection. Besides, the system being based on a relational ranking algorithm, tags associated to an image will be ranked by relevance and may serve either for fully automatic annotation or for suggesting ordered lists of keywords to a user. Additional material to this paper can be found in (Denoyer and Gallinari 2010)

Ranking Model for Image Annotation

Vectors are denoted in bold; subscripts correspond to components of a vector while superscripts correspond to indexes. For example, $\mathbf{x}_k^j$ corresponds to the k-th component of the j-th vector $\mathbf{x}$. We consider:

- The set of images that are currently in the sharing system : $\mathcal{I} = (i^1, ..., i^n)$ where n is the number of images.
- A feature function $\psi : \mathcal{I} \rightarrow \mathbb{R}^N$. This function will map an image onto a feature space of size N. Different ψ functions will be defined in the experimental section. For simplicity, we will replace $\psi(i^j)$ by the identifier $\mathbf{x}^j$ in the following. $\mathbf{x}_k^j$ thus corresponds to the k-th component of the feature vector of image i^j.

[1] http://www.flickr.com and http://www.youtube.com

- The set of possible tags is $\mathcal{T} = (1, .., T)$ where T is the number of tags.

We define a ranking function f as $f(j,t) = \mathbf{y}_t^j$ where $\mathbf{y}_t^j$ is the score of tag t for image i^j. The higher the score, the more relevant the tag is. The set of images in $\mathcal{I}$ is composed of two subsets:

- A labeled set $\mathcal{I}_l = (i^1, ..., i^\ell)$ composed of ℓ images and their associated annotations $(\mathbf{y}^1, ..., \mathbf{y}^l)$ provided by users where $\mathbf{y}_k^j = 1$ if k is a tag associated to i^j and 0 otherwise
- An unlabeled set $\mathcal{I}_u = (i^{\ell+1}, ..., i^{\ell+u})$ of size u corresponding to images not labeled by any user[2].

The annotation problem amounts at computing from the set of images $\mathcal{I}$ and the set of manual annotations $(\mathbf{y}^1, ..., \mathbf{y}^l)$, a ranking score $f(i^j, t)$ for all images and tags in $\mathcal{I}_u$.

In order to describe the relational ranking model, we proceed in two steps. In the next section, we introduce a ranking model which operates only on the content of images. It will be used in the experiments section as a baseline for performance comparison. It is also a component of the relational ranking model. The latter is presented as a relational extension of the content only ranking model.

Content only Annotation with Pairwise Ranking

Learning problem Several ranking algorithms have been proposed in the machine learning community, and recently, their use has become popular in the IR field leading to a large literature in this domain. The algorithm below is formally similar to RankSVM (Joachims 2002). The only difference is that in the relational learning model, all modules are jointly trained using gradient descent since classical SVM optimization algorithms do not easily extend to this relational setting. We define the pairwise ranking risk for model θ over a labeled image $i^k \in \mathcal{I}_l$ with annotation $\mathbf{y}^k$ as:

$$\Delta_\theta(i^k, \mathbf{y^k}) = \sum_{(t,t'):\mathbf{y}_t^k > \mathbf{y}_{t'}^k} h(f_\theta^{PR}(i,t) - f_\theta^{PR}(i,t')) \quad (1)$$

where $h(.)$ is the classical hinge loss function and $f_\theta^{PR}(i,t)$ is the annotation function. This risk is known to be an upper bound of the error over pairs of examples to rank and its minimization has been shown to be very effective for the minimization of the ranking error, and particularly for the minimization of the average precision (see experimental section).

We introduce a L2-regularization term with hyperparameter λ_{reg}[3] for avoiding overfitting. We then define the ranking objective function as:

$$\mathcal{L}_{PR}(\theta) = \sum_{i^k \in \mathcal{I}_l} \Delta_\theta(i^k, \mathbf{y^k}) + \lambda_{reg}||\theta||^2 \quad (2)$$

Note that only labeled images are considered at this step. The learning problem thus corresponds at finding parameters θ^* that minimize this ranking loss. This model is denoted **PR** for Pairwise Ranking in the following.

[2]Note that $\ell + u = n$

[3]λ_{reg} is usually fixed by cross-validation

Ranking function We will use a linear function $f_\theta(k,t)$ for computing the score of an image i^k for tag t:

$$f_\theta^{PR}(k,t) = \langle \theta, \Phi(\mathbf{x^k}, t) \rangle \quad (3)$$

where $\Phi(\mathbf{x^k}, t)$ is a feature vector for the couple (image i^k ;tag t). $\langle ., . \rangle$ is the classical dot product. The feature function Φ, used here is the multiclass feature function (Har-Peled, Roth, and Zimak 2002) defined as:

$$\Phi(\mathbf{x^k}, t) = \begin{pmatrix} (0 \cdots 0) \cdots \mathbf{x^k} \cdots (0 \cdots 0) \end{pmatrix} \quad (4)$$

$\Phi(\mathbf{x^k}, t)$ is a block vector with sub-vector $\mathbf{x^k}$ at position $N.t + 1$ in $\mathbb{R}^{N*T}$ and 0 anywhere else. This notation allows expressing the ranking function as a simple dot product over an input space: it can be seen that one linear model per tag will be learned, which amounts at T independent linear models in all.

Annotation in a Social Network

In this section, we introduce a new ranking model called GPR - for *Graph Pairwise Ranking* - designed for labeling objects (images) based on the object content itself, on external information sources like metadata, and on different types of relations shared by the objects in the collection.

The GPR model is thus a relational ranking model which considers both the content of images and the structure of a graph defined over the images. This graph can be extracted directly from the content of the images or the surrounding text as in (Tong et al. 2006) (implicit relations) or build over a social network (explicit relations e.g. authors friendship). In the experimental section, we will explore the influence of both types of relations.

Let us denote $\mathcal{R}$ a set of relations between images. Elements of $\mathcal{R}$ are scalars: $\mathcal{R} = \{w_{j,k}, j \times k \in [1..n]^2\}$ where $w_{j,k} > 0$ is the weight of the relation between j and k. The couple $(\mathcal{I}, \mathcal{R})$ thus corresponds to a weighted graph of images.

The GPR model is based on two assumptions. It considers that a good ranking model is both a model that correctly ranks the tags of all images $\mathcal{I}$ (Classical Ranking Assumption) and a model that exploits the graph structure in order to extract regularities over the image collection. In particular, we will exploit here relations which will reinforce the proximity of tag lists, for images that share a strong positive connection. (Regularity Assumption)

Typically, if the structure of the graph is extracted from a friendship network of people sharing common interests, knowing that two authors are friends will probably indicate some important relation about their image collections and tag lists. If the network is inferred from some similarity function between images or surrounding texts, a similar conclusion will held. The first assumption is captured by the objective function of the PR model, the second one is intro-

duced through a relational term:

$$\mathcal{L}_{GPR}(\theta) = \sum_{i^k \in \mathcal{I}_l} \Delta_\theta(x^k, \mathbf{y^k}) + \lambda_{reg}||\theta||^2$$
$$+\lambda_{REL} \sum_{t\in[1..T]} \sum_{(j,k)\in[1..n]^2} w_{j,k}\big(f_\theta^{GPR}(j,t) - f_\theta^{GPR}(k,t)\big)^2$$
$$=\mathcal{L}_{PR}(\theta) + \lambda_{REL}\mathcal{L}_{REL}(\theta) \quad (5)$$

where $f_\theta^{GPR}(k,t)$ is the scoring function of the GPR model with parameters θ.

The term $\mathcal{L}_{REL}(\theta)$ will force increase tag scores similarity for connected pictures according to the connection weight. Such regularity assumptions over a graph structure are commonly used in the context of semi-supervised learning for classification (Abernethy, Chapelle, and Castillo 2008).

GPR Ranking Function In the relational model, we want the score of a node to depend on several available information sources in the social network. In this model, the score of a tag for an image will depend both on the visual information in the image itself and the relational information gathered from neighboring images in the graph.

We define the GPR scoring function as:

$$f_{\theta,\xi}^{GPR}(k,t) = \langle\theta, \Phi(\mathbf{x}^k, t)\rangle + \xi_{k,t} \quad (6)$$

where $\theta \in \mathbb{R}^N$ denotes as before the parameters of a content ranking function, $\xi \in \mathbb{R}^{[1..n]\times[1..T]}$ are additional slack variables. There is one slack variable per image and tag. Globally considering the n images and T tags, they add $n * T$ degrees of freedom w.r.t. the initial PR model. They will allow the model to adjust the tag scores as a function of the neighboring images influence. Note that all the parameters of this new scoring function, i.e. both the θs and ξs, will be learned together according to the global objective function 5. The scoring function will then use all available information present in the social sharing system (content, metadata, relations) in an optimal way, according to the objective function.

GPR Objective Function Inserting this new ranking function in the general objective function 5, we obtain the detailed form of the objective function for our relational model:

$$\mathcal{L}_{GPR}(\theta,\xi) = \sum_{i^k\in\mathcal{I}_l} \sum_{(t,t'):y_t^k > y_{t'}^k} h(f_{\theta,\xi}^{GPR}(k,t) - f_{\theta,\xi}^{GPR}(k,t'))$$
$$+ \lambda_{REL} \sum_t \sum_{(j,k):w_{j,k}>0} w_{j,k}(f_{\theta,\xi}^{GPR}(j,t) - f_{\theta,\xi}^{GPR}(k,t))^2$$
$$+ \lambda_{reg}||\theta||^2 + \lambda_{slack} \sum_{k,t\in[1..\ell]\times[1..T]} \xi_{k,t}^2 \quad (7)$$

where: the first term of the sum is the pairwise hinge loss as defined previously, where $f_{\theta,\xi}^{GPR}$ has been substituted to f_θ^{PR}, the second term is the graph regularity term introduced in the previous section, the third term is the L2 regularization over the content parameters θ, the last term is the L2 regularization term over the slack variables. λ_{slack} is an hyper-parameter fixed by cross validation or by hand that penalizes high slack variables values and encourages final scores to be close to the content-based score $f_\theta^{PR}(k,t)$. Typically, if $\lambda_{slack} = +\infty$, $\forall k,t, \xi_{k,t} = 0$ and the GPR model will rank tags using only the content information.

This objective function is minimized through gradient descent simultaneously for θ and ξ. The experiments presented show that it is stable and reaches good performance. The complexity of the model is $O(R.T.N)$ where R is the number of non-null edges among images in the graph. The model will perform faster using sparse relations (like friendship or authorship) than dense ones like similarities. Moreover, the experiments show that sparse relations work much better.

Corpus	C1	C2	C3
Number of images	519	801	3 183
Number of authors	100	100	1 000
Number of tags	**32**	**326**	**25**
Size of ψ^{text} vectors	990	990	4 460
Number of w^{im} relations	$\approx$ 120 000	$\approx$ 260 000	$\approx$ 2 millions
Number of w^{text} relations	$\approx$ 90 000	$\approx$ 140 000	$\approx$ 1.3 millions
Number of w^{au} relations	$\approx$ 9 000	$\approx$ 20 000	$\approx$ 100 000
Number of w^{fr} relations	$\approx$ 12 000	$\approx$ 30 000	$\approx$ 320 000

Table 1: Statistics about the datasets. The implicit relations have been thresholded so that only the largest are kept.

Features	**Description**
ψ^{im}	Normalized RGB histograms (48 bands)
ψ^{text}	Normalized TF-IDF vectors over the titles and description
Relations	**Weight values (0 if no relation)**
w^{im}	$w_{j,k}^{im} =< \psi^{im}(i^j); \psi^{im}(i^k) >$ (thresholded).
w^{text}	$w_{j,k}^{text} =< \psi^{text}(i^j); \psi^{text}(i^k) >$ (thresholded).
w^{au}	$= \begin{cases} 1 \text{ if image } j \text{ and } k \text{ have the same authors.} \\ 0 \text{ otherwise} \end{cases}$
w^{fr}	$= 1$ if the authors of image j and k are friends

Table 2: Feature functions and relations

Experiments

The experiments have been made on different corpora extracted from the Flickr[4] website. Each corpus is composed by a set of images and a set of users. Each image is described by both its visual content (.jpg files) and by meta-informations (author, date, comments, ...). We have collected 3 corpora that have different characteristics as described in Table 1. For the experiments, the collection was divided into two sets: a labeled set and a so called unlabeled set (i.e. the labels of these images were not considered during training). Training was performed using both labeled and unlabeled sets according to the algorithm. Evaluation was performed on the unlabeled set using as targets the tags associated to these images.

From these collections, we have extracted two different feature vectors that correspond respectively to the visual content of the images (ψ^{im}) and to the textual representation associated with the images (ψ^{text}). The feature functions are detailed in table 2. We have extracted 4 types of

[4]http://www.flickr.com

Corpus:			C1			C2		
Test Size (p) - %:			25	50	75	25	50	75
Feat.	**Edges**	**Model**	Average Precision (.. %)					
ψ^{im}		PR	27	25.6	23.4	8.6	8.1	7.5
-	w^{au}	GPR	**55.7**	**59.3**	**45**	**39.7**	**33.5**	**24.3**
-	w^{fr}	GPR	51.5	49.3	42	25.6	21.3	16.6
-	w^{im}	GPR	28.3	26.9	24.7	8.4	7.9	7.8
-	w^{text}	GPR	29.9	26.6	24.7	8.8	8.2	8.1
ψ^{text}		PR	41.5	38.5	34.4	20.6	18.7	15.3
-	w^{au}	GPR	**59.7**	56.8	51.5	41.4	**39.2**	**32**
-	w^{fr}	GPR	59	**58.8**	**52**	**43.2**	38.9	31.5
-	w^{im}	GPR	32	27.6	27.3	15.9	13.1	12.1
-	w^{text}	GPR	34	35.4	34	15.6	16.8	15.4

Corpus:			C3		
Test Size (p):			25%	50 %	75 %
Features	**Edges**	**Model**	Average Precision (.. %)		
ψ^{text}		PR (Content Only)	33.2	31.7	30.4
	w^{au}	GPR	**40.5**	36.1	33.7
	w^{fr}	GPR	39.1	**37.2**	**35.3**

Table 3: Performances of PR (Content Only) and GPR models on the three datasets.

relations between pictures. Two of them (w^{text} and w^{im}) correspond to implicit relations extracted from the content of the images. Both are proportional to the text and image similarities. w^{au} and w^{fr} correspond to explicit relations extracted from the Flickr social network. For the implicit relations which are very dense, with have used a threshold in order to keep only the strongest weights and reduce the number of edges in the image graph.

We have performed experiments with different hyper-parameters values for both the PR and the GPR models. These hyper-parameters are the number of gradient descent iterations, the gradient descent step, the proportion of unlabeled images denoted p with $p = \frac{u}{l+u}$ and the regularization parameters λ_{reg}, λ_{REL} and λ_{slack}. We have launched three runs for each experiment. We have then computed, for each image, the average precision (APR) obtained with the ordered list of tags returned by the learning machine over the runs. Table 3) clearly show that the combination of textual features with social network based relations (friendship and authorship) outperform the content only model, particularly when the number of possible labels is high (Corpus C2).

Conclusion

We have proposed a model that is able to automatically annotate images. This method handles both the content of images and also the relations between images which can be either implicit (visual similarities for example) or explicitly build upon a social network, while state-of-the-art methods either consider content or structure. We have shown that this model greatly improves a baseline ranking method when using authorship or friendship relations. This model leaves open some questions. It considers only one relation at a time and cannot deal with all the possible relations simultaneously. Moreover, its complexity is still high for now. These two points are currently being investigated.

Acknowledgments

This work was partially supported by the French National Agency of Research (ANR Fragrances and ANR ExDeus/Cedres Projects)

References

Abernethy, J.; Chapelle, O.; and Castillo, C. 2008. Web spam identification through content and hyperlinks. In *AIRWeb '08*. New York, NY, USA: ACM.

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3:1107–1135.

Cao, L.; Luo, J.; and Huang, T. S. 2008. Annotating photo collections by label propagation according to multiple similarity cues. In *MM '08*, 121–130.

Chang, E.; Goh, K.; Sychay, G.; and Wu, G. 2003. Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology* 13:26–38.

Denoyer, L., and Gallinari, P. 2010. A ranking based model for automatic image annotation in a social network. Technical report.

Har-Peled, S.; Roth, D.; and Zimak, D. 2002. Constraint classification: A new approach to multiclass classification. In *ALT*, 365–379.

Hironobu, Y. M.; Takahashi, H.; and Oka, R. 1999. Image-to-word transformation based on dividing and vector quantizing images with words. In *in Boltzmann machines, Neural Networks*, 405409.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *KDD*.

Li, J., and Wang, J. Z. 2008. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(6):985–1002.

Liu, D.; Hua, X.-S.; Yang, L.; Wang, M.; and Zhang, H.-J. 2009. Tag ranking. In *WWW*, 351–360.

Matusiak, K. K. 2006. Towards user-centered indexing in digital image collections. *OCLC Systems & Services* 22(4):283–298.

Stone, Z.; Zickler, T.; and Darrell, T. 2008. Autotagging facebook: Social network context improves photo annotation. In *InterNet08*, 1–8.

Tong, H.; He, J.; Li, M.; Ma, W.-Y.; Zhang, H.-J.; and Zhang, C. 2006. Manifold-ranking-based keyword propagation for image retrieval. *EURASIP J. Appl. Signal Process.*

Wu, L.; Yang, L.; Yu, N.; and Hua, X.-S. 2009. Learning to tag. In *WWW*, 361–370.