Comparing and Combining Unimodal Methods for Multimodal Recognition

Satoru Ishikawa and Jorma Laaksonen
Department of Computer Science
Aalto University School of Science
P.O.Box 15400, FI-00076 Aalto, Finland
Email: firstname.lastname@aalto.fi

***Abstract*—Multimodal recognition has recently become more attractive and common method in multimedia information retrieval. In many cases it shows better recognition results than using only unimodal methods. Most of current multimodal recognition methods still depend on unimodal recognition results. Therefore, in order to get better recognition performance, it is important to choose suitable features and classification models for each unimodal recognition task. In this paper, we research several unimodal recognition methods, features for them and their combination techniques, in the application setup of concept detection in image-text data. For image features, we use GoogLeNet deep convolutional neural network (DCNN) activation features and semantic concept vectors. For text features, we use simple binary vectors for tags and word2vec vectors. As the concept detection model, we apply the Multimodal Deep Boltzmann Machine (DBM) model and the Support Vector Machine (SVM) with the linear homogeneous kernel map and the non-linear radial basis function (RBF) kernel. The experimental results with the MIRFLICKR-1M data set show that the Multimodal DBM or the non-linear SVM approaches produce equally good results within the margins of statitistical variation.**

I. Introduction

Recently, multimodal recognition has become more attractive and common method in multimedia information retrieval research. For example, object or scene detection from multimedia resources, such as textually described images, can be done by combining search results of related words from text description and visual object recognition. In many cases, multimodal models show better recognition results than using only unimodal recognition [1], [2]. However, current multimodal recognition methods depend on individual unimodal recognition results and their efficient combination.

In this paper, we research several combinations of unimodal methods for the concept detection task in image–text data. We apply the Multimodal Deep Boltzmann Machine (DBM) model, the Support Vector Machine (SVM) with the linear homogeneous kernel map and the non-linear RBF kernel, semantic concept detectors, and the word2vec approach [3]. We perform concept detection experiments with the MIRFLICKR-1M dataset, where one million images are combined with zero or more content-describing tags, and 25,000 of the images have additionally been annotated with visual concept labels of 38 and 94 content classes.

The rest of this paper is organized as follows: In Section II we give an overview of previous related works, then we present the models and features used in our study in Section III. The data used in our experiments is described in Section IV and the experiments and results thereof in Section V. In Section VI, we discuss some of our findings in more detail and, finally, our conclusions are presented in Section VII.

II. Related Works

Multimodal content recognition can consist of many unimodal recognition tasks. In this paper, we especially focus on combining visual and textual information retrieval. Various statistical methods have been applied to extract the semantic information from each data modality separately. For instance, the structured Vector Space Model [4], ontology based semantic indexing model [5], syntactic topic model [6] and word2vec [3] have shown great success when used for textual search in general purpose search engines. Recently, the deep neural network approach has gained popularity in many application areas, especially in the visual data domain. For example, according to [1], the Deep Boltzmann Machine (DBM) outperformed the linear and non-linear SVMs in both unimodal and multimodal recognition tasks.

It is also common to use several different models together for improving the recognition results. For example in [7], the TagProp model, a weighted nearest neighbor model that predicts the term relevance of images with a weighted sum of the annotations of most similar images in the training set, was combined with an SVM classifier, showing significant improvements in the recognition results.

In any recognition task, it is important to choose the best-performing features for improving the recognition. In image classification, it has been common to use a combination of SIFT-based [8] and other hand-crafted features, but this approach has now largely been replaced by the use of deep convolutional neural network (DCNN) activation features. In the experiments of [1], concatenated Pyramid Histogram of Words (PHOW) features [9], Gist [10] and MPEG-7 descriptors (EHD, HTD, CSD, CLD, SCD) [11] were used as the input image features. [12] and [7] combined Gist, local SIFT features [8], RGB, LAB, HSV histograms, and hue descriptors [13]. A drawback of such pre-classification

978-1-4673-8695-1/16/$31.00 ©2016 IEEE

combination of different features is that the dimensionality usually becomes quite large and the computational cost will be rising accordingly. In the experiments of this paper, we use DCNN activation features calculated using a pre-trained GoogLeNet [14] network.

We also use semantic concept vectors as image features. The approach is based on our earlier work [15] where we were using the same MIRFLICKR-1M database that is used also in the current experiments, but SIFT features instead of the DCNN features utilized here. The basic idea of using the outputs of a bank of visual detectors as feature inputs to other detectors has been used by many researchers for both image and video content analysis. Examples of successful uses of the method include e.g. [16]–[18].

III. Methods

In this section, we describe the models and features used in our experiments.

A. Models

For modeling the unimodal data distributions, we have used two approaches: the Deep Boltzmann Machine (DBM) and the Support Vector Machine (SVM). For the multimodal case, we have used the multimodal extension of DBM and post fusion of the unimodal SVM outputs.

1) DBM: In [1], Gaussian-Bernoulli Restricted Boltzmann Machine (RBM) [19] is used for modeling the image classification layer, and the replicated softmax model [20] for modeling the text classification layer. Then, a Multimodal DBM is used over the joint distribution of those two layers. Because the Gaussian RBM [19] was designed for modeling real-valued vectors, it is suitable for modeling feature vector values as the input image representation. Let $\mathbf{v} \in \mathcal{R}^D$ be the real-valued input features, and $\mathbf{h} \in \{0,1\}^F$ be binary stochastic hidden units. Then, the energy of the state $\{\mathbf{v},\mathbf{h}\}$ for Gaussian-Bernoulli RBM is:

$$E(\mathbf{v},\mathbf{h};\theta) = \sum_{i=1}^{D}\frac{(v_i-b_i)^2}{2\delta_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F}\frac{v_i}{\delta_i}W_{ij}h_j - \sum_{j=1}^{F}a_jh_j \quad (1)$$

where $\theta = \{\mathbf{a},\mathbf{b},\mathbf{W},\delta\}$ are the model parameters. The probability density that the model assigns to $\mathbf{v}$ is given by

$$\begin{aligned} P(\mathbf{v};\theta) &= \frac{1}{Z(\theta)}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h};\theta)), \\ Z(\theta) &= \int_{\mathbf{v}}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h};\theta))dv. \end{aligned} \quad (2)$$

The replicated softmax model [20] is suitable for modeling with sparse count data because it is automatically extracting low-dimensional latent semantics from a large unstructured collection of documents. Assume we can ignore the order of words and consider a document that contains D words. Let $v \in \mathcal{N}^K$ be a vector of visible units where v_k is the number of times the word k appears in the document with vocabulary of size K, and $\mathbf{h} \in \{0,1\}^F$ be binary stochastic hidden units. Then, the energy of the state $\{\mathbf{v},\mathbf{h}\}$ is:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{k=1}^{K}\sum_{j=1}^{F}v_kW_{kj}h_j - \sum_{k=1}^{K}b_kv_k - M\sum_{j=1}^{F}a_jh_j \quad (3)$$

where $\theta = \{\mathbf{a},\mathbf{b},\mathbf{W}\}$ are the model parameters and M is the total number of words in document. The probability density that the model assigns to $\mathbf{v}$ is:

$$\begin{aligned} P(\mathbf{v},\mathbf{h};\theta) &= \frac{1}{Z(\theta)}\exp(-E(\mathbf{v},\mathbf{h};\theta)), \\ Z(\theta) &= \sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp(-E(\mathbf{v},\mathbf{h};\theta)). \end{aligned} \quad (4)$$

In order to reduce the computational cost, the above unimodal DBM models were trained with the Contrastive Divergence (CD) algorithm [21].

Figure 1 shows a two-layer DBM for the text and image modalities and an additional layer of binary hidden units above them to join the two modalities together. The DBM contains a set of visible units $\mathbf{v} \in \{0,1\}^D$, and sequence of hidden units layers $\mathbf{h}^{(1)} \in \{0,1\}^{F_1}, \mathbf{h}^{(2)} \in \{0,1\}^{F_2}, \ldots, \mathbf{h}^{(L)} \in \{0,1\}^{F_L}$. The energy of the joint configuration $\{\mathbf{v},\mathbf{h}\}$ is defined as:

$$E(\mathbf{v},\mathbf{h};\theta) = -\mathbf{v}^\top\mathbf{W}^{(1)}\mathbf{h}^{(1)} - \mathbf{h}^{(1)\top}\mathbf{W}^{(2)}\mathbf{h}^{(2)} \quad (5)$$

The joint probability distribution of the text–image input is then modeled as:

$$\begin{aligned} P(\mathbf{u}_a,\mathbf{v}_b|\theta) = &\sum_{\mathbf{h}_a^{(2)},\mathbf{h}_b^{(2)},\mathbf{h}^{(3)}} P(\mathbf{h}_a^{(2)},\mathbf{h}_b^{(2)},\mathbf{h}^{(3)}) \cdot \quad (6) \\ &(\sum_{\mathbf{h}_a^{(1)}} P(\mathbf{u}_a,\mathbf{h}_a^{(1)},\mathbf{h}_a^{(2)})) \cdot (\sum_{\mathbf{h}_b^{(1)}} P(\mathbf{v}_b,\mathbf{h}_b^{(1)},\mathbf{h}_b^{(2)})) \end{aligned}$$

where $\mathbf{u}_a \in \mathbb{R}^L$ denotes the image input I_a represented in an L-dimensional feature space and $\mathbf{v}_b \in \mathbb{N}^K$ denotes the representation of the text query Q_b consisting of keywords in a K-dimensional vector space.

2) SVM: In the experiments, we used both linear and non-linear SVMs. For the linear case, we apply the homogeneous kernel map approximation of the intersection kernel [22], and use the LIBLINEAR [23] library with the L_2-regularized logistic regression solver. The implementation of the homogeneous kernel maps for the intersection kernels is available in the VLFeat library [24].

For the non-linear case, we used the non-linear radial basis function (RBF) kernel. It is a popular SVM kernel in many computer vision tasks and has often been reported to achieve a good performance. The RBF kernel can be represented as:

$$K_{\mathrm{RBF}}(\mathbf{x},\mathbf{z}) = \exp\left(-\gamma\left\|\mathbf{x}-\mathbf{z}\right\|_2^2\right) \ , \quad (7)$$

where γ is the kernel width. To train the non-linear RBF kernel SVM, we used the C-SVC classifier of the LIBSVM software library [25].

Some form of post fusion of the unimodal SVM classifier outputs is needed to obtain classification results for multimodal objects, which in our case consist of pairs of one image and a

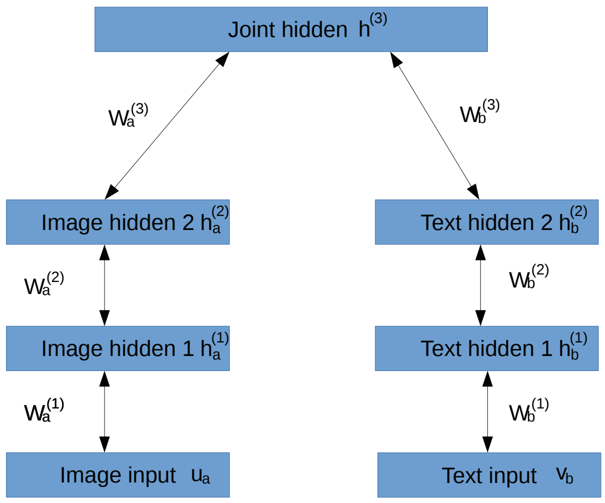

Fig. 1: Multimodal DBM [1]. The left side layers are image-specific DBM and the right side layers are text-specific DBM.

(possibly empty) set of textual tags. In our current experiments we have used the simple *weighted arithmetic mean* fusion rule where a weighted sum of the SVM outputs is assigned as the fusion output to the multimodal object in question.

B. Features

We have used three types of features in our experiments: 1) purely visual, 2) semantic concept vectors that combine visual information and image tags, and 3) purely textual features.

1) Visual features: Motivated by the good results obtained by using deep convolutional neural network (DCNN) activation value features for object and scene recognition tasks, we have replicated the experiments of [1] and performed our new experiments with state-of-the-art deep net features. In particular, we use reverse spatial pyramid pooled [26] activations with two scale levels from the *5th Inception module* in the GoogLeNet [14] network, implemented with the Caffe library [27]. The resulting features are then 2048 dimensional.

2) Semantic concept vectors: Semantic concept vectors incorporate semantic background information from auxiliary training data where either accurate class information is available for images, or less accurate, but numerous textually tagged images exist. Our case is the latter as we are using the image tags of the 975,000 unannotated images in the MIRFLICKR-1M dataset.

The semantic concept vectors are produced in three steps: 1) a large number of background concept classifiers are trained, 2) the concept classifiers are applied to the training and testing images and the image-wise concatenated classifier outputs are treated as novel visual features, and 3) the training data part of these features is used to create new classifiers which are then applied to the testing data in the original classification problem.

The background concept detectors are traditional visual concept detectors, which in our case were trained with the RBF kernel SVM from low-level DCNN visual features and $K_s = 500$ most common tags in the MIRFLICKR-1M dataset. Then, semantic concept vectors were produced by using those prediction outcomes. Let $C_1, ..., C_{K_s}$ be the background concept vocabulary, the semantic concept vector $\mathbf{c}_i$ for each image $x_i, i = 1, \ldots, N$, is then constructed as follows:

$$\mathbf{c}_i = \begin{pmatrix} p_{i,1} \\ \vdots \\ p_{i,K_s} \end{pmatrix} \tag{8}$$

where $p_{i,j} \in [0, 1]$ is the concept membership score of image x_i in concept C_j, generated as the prediction output score of the corresponding semantic concept classifier.

3) Textual features: We used two types of textual features. First, the 2000-dimensional term-frequency-type *text* feature consisting of the 2000 most frequent tags in the dataset was used, similarly to [1]. Each component of the vectors was 1 or 0, indicating whether the corresponding tag had been given to that image or not, respectively. If no tag had been given, then values were set by Gibbs sampling in the feature vector as described in [1].

Second, we used word2vec [3] which can produce high-dimensional semantically meaningful vectorial representations for words. The word embedding reflects the semantic similarities of the words and can be trained by using different natural language processing models on large text corpora. In our experiments, we used a pre-trained 200-dimensional word2vec model created from 17 million words of the ”text8” Wikipedia corpus. We again used the 2000 most frequent tags and summed the word2vec vectors of the tags to represent the textual information associated with each image.

IV. Data

MIRFLICKR-1M dataset: The MIRFLICKR-1M dataset consists of 1,000,000 images with user-given tags and EXIF meta data. 25,000 of them have two annotations from sets of 38 and 94 concepts. The rest 975,000 images have not been annotated with these concepts, but most of them have textual tags. The images were originally downloaded from the social photography site flickr.com [28]. The 38 concept categories include scene categories such as ”sky,” ”river,” ”lake” and object categories such as ”portrait,” ”people,” ”car.” The 94 concept categories have 19 super categories such as ”timeofday,” ”weather,” ”age,” ”gender,” and 94 child categories such as ”day,” ”sun,” ”baby,” ”male,” under the corresponding super categories. In this paper, we perform experiments on the 38 concepts to get results commensurable with those in [1], [2], and also with the 94 concepts set for completeness and future reference.

V. Experiments and Result

For the empirical evaluation, we implemented a similar setting as Srivastava *et al* did in [1], [2]. For the text feature inputs v_b, we used the same $K = 2000$ vocabulary words as used in their work and additionally 200-dimensional word2vec features. In order to compare the image classification results, we used the PHOW, Gist and MPEG-7 based features ($L = 3857$) provided in [1] and our DCNN GoogLeNet activation features ($L = 2048$) as the image input features.

The number of hidden units in each DBM layer were the same as in [1]. Following their procedure, we used the DBM model with and without pre-training with the 975,000 images with tags only. We performed each experiment five times, always using 10,000 objects for training, 5000 objects for validation and the remaining 10,000 objects for testing.

The results of the experiments are shown in Table I, measured as the mean average precision (MAP) and the precision at rank 50 (Prec@50). The rows 1–3 show the results with the image only unimodal models trained without using the 975,000 unannotated images in any way. The rows 4–8 show the same unimodal image models, but now making use of also the extra 975,000 images. The rows 9–13 are the results with the text-only unimodal models by using the tag information from the 25,000 annotated images only. The rows 14 and 15 show the unimodal text-only concept detections where information from the 975,000 tag-only annotated images has been used additionally. Similarly, the rows 16–22 are the results of multimodal concept detections with joined image–text models with and without the 975,000 unannotated images.

In all cases where SVM detectors of multiple modalities or features have been combined (i.e. the rows 8, 15, 17, and 22), we used the weighted arithmetic mean fusion rule with the weight percentages shown in the parentheses. The multimodal combination of the DBM results was always performed by using the Multimodal DBM model of eq. (6).

The row 19 shows the best multimodal model in [2]. In this case, the text input was not clamped and the model was allowed to update the text input layer when performing the mean-field update. Similarly, the row 20 is the best multimodal result of [1] where various additional techniques have been used to improve the MAP result.

Comparing the performances of the different features, on the rows 5 vs. 6 and 20 vs. 21, it is clear but unsurprising that the GoogLeNet features outperformed the PHOW-based and other hand-crafted features. In the text modality, rows 10 vs. 12 and 11 vs. 13, the 200-dimensional word2vec features gave disappointing result compared to the full 2000-dimensional text features. Also, the semantic concept vector features, when combined with either the visual or textual unimodal model, give significant improvement in the MAP results of the rows 3 vs. 8, 11 vs. 15, and 17 vs. 22.

Comparing the classification models, the non-linear RBF kernel SVM outperformed the linear homogeneous kernel map SVM, on rows 2 vs. 3, 10 vs. 11 and 12 vs. 13. Comparing the RBF SVM and the DBM models is not as straightforward. The RBF SVM tends to show slightly better performance in the mean average precision measure, especially in the case of 94 concepts. On the other hand, DBM seems to be better in the rank 50 precision measure. We can also observe that the DBM approach performs slightly better than RBF SVM in the text modality on rows 9 vs. 11.

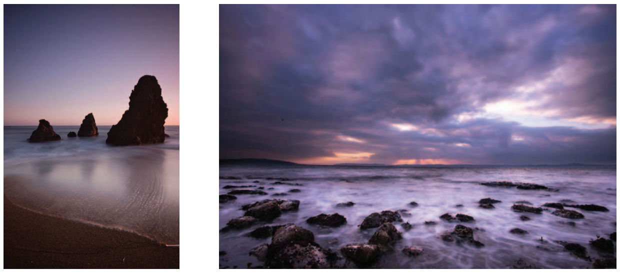

Fig. 2: Two positive example images for concept ”sea_r1”. **Left:** Ranking improved with multimodal approach. **Right:** Ranking worsened with multimodal approach. See text for details.

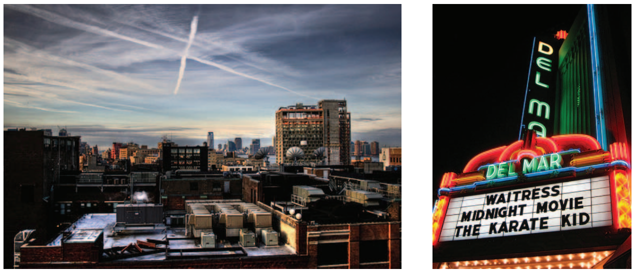

Fig. 3: Two false positive example images for concept ”sea_r1.” **Left:** False recognition became less probable with multimodal approach. **Right:** False recognition became more probable with multimodal approach. See text for details.

Finally, based on the results on the rows 16–22 it is evident that the multimodal results are better than either visual or textual unimodal result alone, with respect to the 38 concepts MAP and both 94 concepts measures. However, it seems on the rows 16 vs. 21 that the DBM pretrainig with the extra 975,000 images is not necessarily beneficial. Overall we can conclude that the DBM and RBF SVM methods are performing equally well within the margins of statitistical variation.

VI. Discussion

Table II shows some examples of concept-wise differences between the unimodal and multimodal results. The columns ”row 3” to ”row 21” show the MAP values of the corresponding row in Table I. The ”diff” columns show the differences between those two values for the corresponding concepts.

On the ”baby” and ”sea_r1” rows, the differences are positive, which means that the multimodal mean average precision is higher than the visual unimodal. On the the other hand, the multimodal approach effects slightly negatively for the ”clouds” and ”tree” concepts. This observations holds for both the RBF SVM method (the rows 3 vs. 22) and the Multimodal DBM (the rows 6 vs. 21). Actually, we picked in Table II those concepts among the set 38 which displayed the largest absolute positive or negative change between the results of the rows 3 vs. 22. So we can see that even some concepts suffer in MAP from the multimodal fusion, this effect is negligible compared to the benefit that some other concepts obtain. Nevertheless, the multimodal approach seems not to be beneficial for all types of image–text contents.

TABLE I: MIRFLICKR-1M 38 and 94 concept classification results with different models. RBF = non-linear RBF kernel SVM, linear = linear homogeneous kernel map SVM for intersection kernel, text = 2000-dimensional 0/1 tag features, word2vec = 200-dimensional word2vec features. DBM p.t. = DBM pre-training performed with 975,000 unannotated images and/or tags. sem. = semantic concept vectors.

	model	image features	text features	975,000	38 MAP	38 Prec@50	94 MAP	94 Prec@50
1	DBM	GoogLeNet	—	—	**0.723** ± 0.004	**0.915** ± 0.003	0.405 ± 0.004	0.550 ± 0.006
2	linear	GoogLeNet	—	—	0.702 ± 0.007	0.903 ± 0.005		
3	RBF	GoogLeNet	—	—	**0.721** ± 0.004	0.905 ± 0.004	**0.439** ± 0.006	**0.570** ± 0.003
4	DBM [1]	PHOW, Gist, MPEG-7	—	DBM p.t.	0.469 ± 0.005	0.803 ± 0.005		
5	DBM	PHOW, Gist, MPEG-7	—	DBM p.t.	0.475 ± 0.002	0.753 ± 0.002		
6	DBM	GoogLeNet	—	DBM p.t.	0.727 ± 0.003	**0.918** ± 0.004	0.437 ± 0.004	**0.573** ± 0.005
7	RBF	sem.	—	500 tags	0.720 ± 0.003	0.901 ± 0.005	0.429 ± 0.005	0.559 ± 0.001
8	RBF	GoogLeNet (50%) + sem. (50%)	—	500 tags	**0.735** ± 0.003	0.909 ± 0.004	**0.449** ± 0.005	**0.577** ± 0.002
9	DBM	—	text	—	**0.488** ± 0.004	**0.829** ± 0.008	**0.270** ± 0.003	**0.456** ± 0.007
10	linear	—	text	—	0.421 ± 0.010	0.709 ± 0.016		
11	RBF	—	text	—	**0.490** ± 0.006	0.805 ± 0.014	0.262 ± 0.007	0.430 ± 0.007
12	linear	—	word2vec	—	0.267 ± 0.004	0.420 ± 0.008		
13	RBF	—	word2vec	—	0.466 ± 0.003	0.798 ± 0.008		
14	DBM	—	text	DBM p.t.	0.511 ± 0.004	0.834 ± 0.005	0.287 ± 0.002	0.463 ± 0.007
15	RBF	—	text (25%) + sem. (75%)	500 tags	**0.740** ± 0.002	**0.909** ± 0.006	**0.449** ± 0.004	**0.579** ± 0.005
16	DBM	GoogLeNet	text	—	**0.745** ± 0.003	**0.923** ± 0.003	0.458 ± 0.004	**0.594** ± 0.008
17	RBF	GoogLeNet (70%)	text (30%)	—	0.741 ± 0.003	0.911 ± 0.005	0.458 ± 0.003	0.582 ± 0.002
18	DBM [2]	PHOW, Gist, MPEG-7	Generated text	DBM p.t.	0.531 ± 0.005	0.832 ± 0.004		
19	DBM [2]	PHOW, Gist, MPEG-7	text	DBM p.t.	0.609 ± 0.004	0.873 ± 0.004		
20	DBM [1]	PHOW, Gist, MPEG-7	text	DBM p.t.	0.641 ± 0.004	0.888 ± 0.004		
21	DBM	GoogLeNet	text	DBM p.t.	**0.748** ± 0.003	**0.919** ± 0.005	0.459 ± 0.003	**0.599** ± 0.007
22	RBF	GoogLeNet (37.5%) + sem.(37.5%)	text (25%)	500 tags	**0.752** ± 0.002	**0.915** ± 0.006	**0.467** ± 0.003	0.591 ± 0.003

TABLE II: Examples of concept-wise MAP measure differences between unimodal and multimodal results. The row numbers refer to the corresponding results in Table I.

concept	row 3	row 22	diff	row 6	row 21	diff
”baby”	0.451	0.523	0.072	0.449	0.521	0.072
”clouds”	0.807	0.801	–0.006	0.798	0.796	–0.002
”sea_r1”	0.451	0.589	0.138	0.488	0.572	0.084
”tree”	0.773	0.773	–0.000	0.760	0.751	–0.009

Figure 2 shows two example images of the concept ”sea_r1” where the ranking of the image improved (left) and worsened (right) when moving from the visual unimodal method (the row 3) to the multimodal fusion (the row 22). The user-given tags for the left image are *beach, coast, ocean, pacific, shore*, etc. Most of them really are related to the sea, hence the tags affect positively and lead to better ranking of the image. The tags of the right image are *shutter, slow*, and *speed*, and they are not related to sea at all. Therefore, the tag information can be regarded as noise and it affects the image's ranking negatively in this case.

Figure 3 shows two example images where false recognition to concept ”sea_r1” is becoming either less or more probable due to the multimodal approach. For the left image, the user-given tags include *buildings, city, newyork, streets, urban*, which are clearly not related to sea and make it less probable to classify the image as a sea view On the other hand, for the right image, the tags include *beach, cinema, coast, ocean* and *pacific*, which are related to the sea. The tag information thus misleads the multimodal classification and increases the false recognition rate from the visual unimodal case.

Our examples on the concept and individual image levels show that, inevitably, some concepts and some images benefit and some suffer from the tag-based textual input to the multimodal recognition system. On the average, however, the gains are larger in magnitude than the losses.

VII. Conclusions

In this paper, we compared between multimodal DBM models and linear and non-linear SVM classifiers in a multimodal recognition task with image–text data of the MIRFLICKR-1M database. We also studied the performance of different visual and textual features.

For the visual features, we found out that the GoogLeNet-based DCNN features outperform the pre-classifier fusion of PHOW-based and other traditional hand-crafted features. The semantic concept vectors, trained by using auxiliary image–tag data, also brought improvement in the results. For the textual features, the 2000-dimensional binary tag vector was better than the lower-dimensional word2vec representation. The combination of the semantic concept vectors and the binary vectors of tags clearly outperformed the use of binary tag vectors only.

According to the mean average precision results, in the post fusion of the visual and textual classification, the combination of DCNN, semantic features and the binary tag features with the RBF SVM classifier achieved the same performance level as the corresponding Multimodal DBM model. When using the precision at rank 50 as the performance criterion, the Multimodal DBM model showed slightly better results.

Overall, the multimodal approaches always gave better results than any unimodal approach alone. In the particular case of the MIRFLICKR-1M database, where the user-given image tags are quite unreliable, the visual domain proved to be the more reliable one in the multimodal recognition task. In the future, we will study the visual semantics of each tag, taking into account only those tags which are relevant for visual classification. In that way, we could concentrate on using only visually meaningful and thus more reliable tags in multimodal recognition.

ACKNOWLEDGMENT

This work has been funded by the grant 251170 of the Academy of Finland and the Data to Intelligence D2I DIGILE SHOK program. The calculations were performed using computer resources within the Aalto University School of Science ”Science-IT” project.

REFERENCES

[1] N. Srivastava and R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” *Journal of Machine Learning Research*, vol. 15, pp. 2949–2980, 2014.

[2] ——, “Multimodal learning with deep boltzmann machines,” in *Advances in neural information processing systems. 2012.*, 2012, pp. 2222–2230.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean., “Efficient estimation of word representations in vector space.” *CoRR*, vol. abs/1301.3781, 2013.

[4] K. Erk and S. Padó, “A structured vector space model for word meaning in context,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 897–906.

[5] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan, “An ontology-based retrieval system using semantic indexing,” *Information Systems*, vol. 37, no. 4, pp. 294–305, 2012.

[6] J. L. Boyd-Graber and D. M. Blei, “Syntactic topic models,” in *Advances in neural information processing systems*, 2009, pp. 185–192.

[7] J. J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, “Image annotation with tagprop on the mirflickr set,” in *MIR ’10 Proceedings of the international conference on Multimedia information retrieval*. New York, NY, USA: ACM, 2010, pp. 537–546.

[8] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[9] A. Bosch, A. Zisserman, and X. Munoz, “Representing shape with a spatial pyramid kernel,” in *Proceedings of ACM ICVR 2007*, 2007, pp. 401–408.

[10] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1011139631724

[11] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703 – 715, June 2001.

[12] M. Guillaumin, J. Verbeek, and C. Schmid, “Multimodal semi-supervised learning for image classification,” in *CVPR 2010 - 23rd IEEE Conference on Computer Vision & Pattern Recognition*. San Francisco, United States: IEEE Computer Society, June 2010, pp. 902–909.

[13] J. van de Weijer and C. Schmid, “Coloring local feature extraction,” in *Proc. ECCV 2006*, May 2006.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *arXiv*, vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[15] M. Sjöberg and J. Laaksonen, “Using semantic features to improve large-scale visual concept detection,” in *Proceedings of the 12th International Workshop on Content Based Multimedia Indexing (CBMI 2014)*. Klagenfurt, Austria: IEEE, June 2014, pp. 1–6.

[16] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, “Object bank: A high-level image representation for scene classification & semantic feature sparsification,” in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[17] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, “Semantic model vectors for complex video event recognition,” *Trans. Multi.*, vol. 14, no. 1, pp. 88–101, Feb. 2012. [Online]. Available: http://dx.doi.org/10.1109/TMM.2011.2168948

[18] A. Habibian, K. E. van de Sande, and C. G. Snoek, “Recommendations for video event recognition using concept vocabularies,” in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ser. ICMR ’13. New York, NY, USA: ACM, 2013, pp. 89–96. [Online]. Available: http://doi.acm.org/10.1145/2461466.2461482

[19] G. E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation*, vol. 18, pp. 1527 – 1554, July 2006.

[20] G. E. Hinton and R. R. Salakhutdinov, “Replicated softmax: an undirected topic model,” in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 1607–1614. [Online]. Available: http://papers.nips.cc/paper/3856-replicated-softmax-an-undirected-topic-model.pdf

[21] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural Computation*, vol. 14, pp. 1771 – 1800, August 2002.

[22] A. Vedaldi and A. Zisserman, “Efficient additive kernels via explicit feature maps,” in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010.

[23] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, “LIBLINEAR: A library for large linear classification,” *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[24] A. Vedaldi and B. Fulkerson, “VLFeat: A library of computer vision algorithms,” *http://www.vlfeat.org/*.

[25] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[26] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, “Multi-scale orderless pooling of deep convolutional activation features,” March 2014, arXiv.org:1403.1840.

[27] Y. Jia, “Caffe: An open source convolutional architecture for fast feature embedding,” http://caffe.berkeleyvision.org/, 2013.

[28] M. J. Huiskes and M. S. Lew, “The MIR Flickr retrieval evaluation,” in *MIR ’08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*. New York, NY, USA: ACM, 2008.