LEARNING SPARSE LATENT REPRESENTATION AND DISTANCE METRIC FOR IMAGE RETRIEVAL

Tu Dinh Nguyen†, Truyen Tran†‡, Dinh Phung† and Svetha Venkatesh†

†Center for Pattern Recognition and Data Analytics
School of Information Technology, Deakin University, Geelong, Australia
‡Institute for Multi-Sensor Processing and Content Analysis
Curtin University, Australia
{ngtu,truyen.tran,dinh.phung,svetha.venkatesh}@deakin.edu.au

ABSTRACT

The performance of image retrieval depends critically on the semantic representation and the distance function used to estimate the similarity of two images. A good representation should integrate multiple visual and textual (e.g., tag) features and offer a step closer to the true semantics of interest (e.g., concepts). As the distance function operates on the representation, they are interdependent, and thus should be addressed at the same time. We propose a probabilistic solution to *learn* both the representation from multiple feature types and modalities and the distance metric from data. The learning is regularised so that the learned representation and information-theoretic metric will (i) preserve the regularities of the visual/textual spaces, (ii) enhance structured sparsity, (iii) encourage small intra-concept distances, and (iv) keep inter-concept images separated. We demonstrate the capacity of our method on the NUS-WIDE data. For the well-studied 13 animal subset, our method outperforms state-of-the-art rivals. On the subset of single-concept images, we gain 79.5% improvement over the standard nearest neighbours approach on the MAP score, and 45.7% on the NDCG.

***Index Terms*—** Image retrieval; Mixed-Variate; Restricted Boltzmann Machines; metric learning; sparsity; NUS-WIDE.

1. INTRODUCTION

Images are typically retrieved based on the distance from the query image. Thus the retrieval quality depends critically on how the images are represented and the distance metric operating on the representation space. Standard vector-based representations may involve colour histograms or visual descriptors; and distance metrics can be those working in the vector space. However, they suffer from important drawbacks. First, there are no simple ways to integrate multiple representations (e.g., histograms and bag-of-words) and multiple modalities (e.g., visual words and textual tags). Furthermore, designing a representation separately from distance metric is sup-optimal – it takes time to search for the best distance metric for a given representation. And third, using low-level features may not capture well the high-level semantics of interest, leading to poor retrieval quality if the visual features are similar and the objects are different. For example, it can be easy to confuse between a lion and a wolf if we rely on the textures and colours alone.

Our solution is to *learn* both the higher representation and the distance metric specifically for the retrieval task. The higher representation would capture the regularities and factors of variation in the data space from multiple lower feature types and modalities. At the same time, the representation would lead to small distances between conceptually related objects and large distances between those unrelated. To that end, we extend our recent versatile machinery known as Mixed-Variate Restricted Boltzmann Machine (MV.RBM) [1]. The MV.RBM is a probabilistic architecture capable of integrating several data types into a homogeneous "latent" representation in an unsupervised fashion. Our extensions include the introduction of the counting of visual/textual words [2] and the group-wise sparsity [3] into the model. During the training phase, the model learning is regularised in the way that the information theoretic distances on the latent representation between intra-concept images are minimised and those between inter-concept images are maximised. During the testing phase, the learned distance metrics are then used for retrieving similar images.

Our method differs substantially from the recent metric learning methods such as those in [4, 5, 6, 7, 8]. The main difference is the focus on the probabilistic representation itself which has not been presented in previous work. Our goal is not only to learn a good distance function but also to capture as much information from the heterogeneous data as possible. The regularisation using intra-concept distances under the RBM framework has been studied in our previous work [9], but that work is limited to face recognition with single type input.

We demonstrate the effectiveness of the proposed method on the NUS-WIDE data. This data is particularly rich: Each

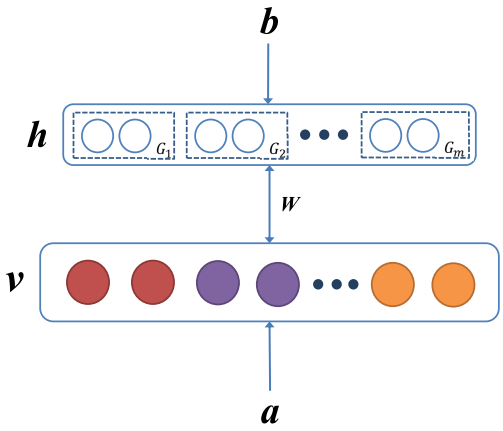

Fig. 1: Image latent representation using Mixed-Variate RBM with sparse group. The bottom layer contains visible units which represent image features. Different colours denote varied types. The top layer represents stochastic hidden binary units. The hidden units within dotted squares mean that they are arranged into their individual groups $\{G_1, G_2, ..., G_m\}$.

image has multiple visual representations, sometimes social tags, and one or more manually annotated high-level concepts. We run experiments on two subsets. The first is the well-studied 13 animal subset in which we show that our method is competitive against recent state-of-the-arts. The second subset contains $20,000$ single-concept images. We obtain 79.5% improvement on mean average precision (MAP) over the standard nearest neighbours approach and 45.7% increase on the normalised discounted cumulative gain (NDCG).

In short, our main contributions are: (i) a novel extension of the recently introduced representation learning machinery, Mixed-Variate RBM, to construct a robust latent image representation with sparsity structures; (ii) enhancing the distance metric effectiveness by imposing intra-concept and inter-concept constraints; (iii) demonstrating that the proposed method outperforms recent state-of-the-art methods on the well-known NUS-WIDE data.

The remaining paper is structured as follows. Sec. 2 presents the Mixed-Variate RBM with group-wise sparsity and metric learning for image retrieval. We show empirical evaluation on the two NUS-WIDE subsets in Sec. 3. Finally, Sec. 4 concludes the paper.

2. LEARNING SPARSE LATENT REPRESENTATION AND DISTANCE METRIC

In this section we present our framework of simultaneous learning of sparse data representation and distance metric for image retrieval tasks. The framework has three components: (i) a mixed-variate machine that maps multiple feature types and modalities into a homogeneous higher representation, (ii) a regulariser that promotes structured sparsity on the learned representation, and (iii) an information-theoretic distance operating on the learned representation.

2.1. Learning Homogeneous Representation

Our representation is based on our recently introduced Mixed-Variate Restricted Boltzmann Machine (MV.RBM) [1]. A standard RBM is a bipartite network with two layers where the first layer consists of binary visible units and the second layer binary hidden units [10]. A MV.RBM has the same hidden layer as the RBM but integrates a variety of visible types, and thus it is particularly suitable for representing multimedia objects. See Fig 1 for an illustration of the MV.RBM.

More formally, let $\boldsymbol{v}$ denote the joint set of mixed-type features: $\boldsymbol{v} = (v_1, v_2, ..., v_N)$, $\boldsymbol{h}$ the set of binary hidden variables: $\boldsymbol{h} = (h_1, h_2, ..., h_K) \in \{0,1\}^K$. The MV.RBM admits the Boltzmann distribution over all variables, i.e., $P(\boldsymbol{v}, \boldsymbol{h}) \propto \exp\{-E(\boldsymbol{v}, \boldsymbol{h})\}$, where $E(\boldsymbol{v}, \boldsymbol{h})$ is the model energy defined as

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\left(\sum_i F(v_i) + \boldsymbol{a}^\top \boldsymbol{v} + \boldsymbol{b}^\top \boldsymbol{h} + \boldsymbol{h}^\top \boldsymbol{W} \boldsymbol{v}\right)$$

where $F(v_i)$ is the type-specific function, $\boldsymbol{a} = (a_1, a_2, ..., a_N)$ and $\boldsymbol{b} = (b_1, b_2, ..., b_K)$ are biases of visible and hidden units respectively, and $\boldsymbol{W} = [w_{ij}]$ are the weights connecting hidden and visible units.

Due to the bipartite structure of the RBM, the conditional distributions over hidden and visible units can be factorised as

$$P(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_{j=1}^{K} P(h_j \mid \boldsymbol{v}) \quad (1)$$

$$P(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_{i=1}^{N} P(v_i \mid \boldsymbol{h}) \quad (2)$$

The posterior representing the data in the latent space is *homogeneous*, i.e., $P(h_j = 1 \mid \boldsymbol{v}) = \sigma\left(b_j + \sum_{i=1}^{N} v_i w_{ij}\right)$ where $\sigma(x) = (1 + \exp(-x))^{-1}$. The data generating distribution $P(v_i \mid \boldsymbol{h})$ is, on the other hand, *type-specific*. For example, let $\mu_i = a_i + \sum_{j=1}^{K} h_j w_{ij}$, *binary* units (e.g., textual tags) would be specified as: $P(v_i = 1 \mid \boldsymbol{h}) = \sigma(\mu_i)$ and *Gaussian* units (e.g., histograms) as: $v_i \mid \boldsymbol{h} \sim \mathcal{N}(\mu_i; 1)$. To represent *counts* (e.g., bag-of-visual words), we adopt the constrained Poisson model by [2] in that $P(v_i = n \mid \boldsymbol{h}) = \text{Poisson}\left(n, \frac{\exp\{\mu_i\}}{\sum_k \exp\{\mu_k\}} L\right)$, where L is the "document" length. Note that the integration of counts is not readily available in the original MV.RBM.

Once the model is fully specified, the *latent representation* can be achieved by transforming feature space of input image into hidden posterior vector $\hat{\boldsymbol{h}} = \left(\hat{h}_1, \hat{h}_2, ..., \hat{h}_K\right)$, where $\hat{h}_j = P(h_j = 1 \mid \boldsymbol{v})$. This representation is highly interpretable: Each $\hat{h}_j$ is the probability that a particular latent feature is activated. It also has nice numerical properties: The posteriors are homogenous and bounded within $(0, 1)$. Thus the distance measures computed on latent representations do not suffer from the heterogeneity and different scales of the features.

2.2. Enhancing the Representation by Structured Sparsity

The latent representation learned from the MV.RBM captures the regularities in the data space. However, it is largely unstructured and may not readily disentangle all the factors of variation (e.g., those due to different object types in the image). One way to improve the representation is to impose some structured sparsity, which may lead to better separation of object groups and easier interpretation. Following the previous work in [3], we impose a mixed-norm regulariser on the latent representation. In particular, hidden units are equally arranged into M non-overlapped groups. Let $\mathcal{G}_m$ denote the indices of hidden units in the m^{th} group, the regulariser reads

$$\mathcal{R}(\boldsymbol{v}) = \sum_{m=1}^{M} \sqrt{\sum_{j \in G_m} P(h_j = 1 \mid \boldsymbol{v})^2} \tag{3}$$

During learning, this regulariser is minimised and this leads to group-wise sparsity, i.e., only few groups of hidden units will be activated (see the last column of Fig. 2).

2.3. Learning Distance Metric

Latent representation may not fully capture *intra/inter-concept variation*s, and thus it may not result in a good distance metric for retrieval tasks. It is better to directly learn a distance metric that suppresses intra-concept variation and enlarges inter-concept variation. Given the probabilistic nature of our representation, a suitable distance is the symmetric Kullback-Leibler divergence, also known as Jensen-Shannon divergence:

$$\mathcal{D}(g, f) = \frac{1}{2}\left(\mathrm{KL}(g \| f) + \mathrm{KL}(f \| g)\right) \tag{4}$$

where $\mathrm{KL}(g\|f) = \sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid g) \log \frac{P(\boldsymbol{h}|g)}{P(\boldsymbol{h}|f)}$. Let $\mathsf{N}(f)$ denote the set of other images that share the same concept with the image f, and $\bar{\mathsf{N}}(f)$ denotes those do not. The mean distance to all other images in $\mathsf{N}(f)$ should be minimised:

$$\mathcal{D}_{\mathsf{N}(f)} = \frac{1}{|\mathsf{N}(f)|} \sum_{g \in \mathsf{N}(f)} \mathcal{D}\left(P\left(\boldsymbol{h} \mid \boldsymbol{v}^{(g)}\right), P\left(\boldsymbol{h} \mid \boldsymbol{v}^{(f)}\right)\right) \tag{5}$$

On the other hand, the mean distance to all images in $\bar{\mathsf{N}}(f)$ should be enlarged:

$$\mathcal{D}_{\bar{\mathsf{N}}(f)} = \frac{1}{|\bar{\mathsf{N}}(f)|} \sum_{g \in \bar{\mathsf{N}}(f)} \mathcal{D}\left(P\left(\boldsymbol{h} \mid \boldsymbol{v}^{(g)}\right), P\left(\boldsymbol{h} \mid \boldsymbol{v}^{(f)}\right)\right) \tag{6}$$

We note that the idea of intra-concept distance has been studied in our recent work [9], but the inter-concept distance is new.

2.4. Putting Things Together

Our learning has three goals: Capturing the joint representation of visual and textual features by maximising the data likelihood $\mathcal{L}(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P(\boldsymbol{h}, \boldsymbol{v})$, enhancing structural sparsity through minimising $\mathcal{R}(\boldsymbol{v})$, and regularising intra-concept and inter-concept distance metrics $\mathcal{D}_{\mathsf{N}(f)}$ and $\mathcal{D}_{\bar{\mathsf{N}}(f)}$. The objective function is now the following regularised likelihood

$$\mathcal{L}_{reg} = \sum_{f} \log \mathcal{L}\left(\boldsymbol{v}^{(f)}\right) - \alpha \mathcal{R}(\boldsymbol{v}) - \beta\left(\sum_{f} \mathcal{D}_{\mathsf{N}(f)} - \sum_{f} \mathcal{D}_{\bar{\mathsf{N}}(f)}\right)$$

where $\alpha \geq 0$ is the regularising constant for sparsity, $\beta \geq 0$ is the coefficient to control the effect of distance metrics. Maximising this regularised likelihood is equivalent to simultaneously maximising the data likelihood $\mathcal{L}(\boldsymbol{v})$, minimising the regularisation function $\mathcal{R}(\boldsymbol{v})$, minimising the neighbourhood distance $\mathcal{D}_{\mathsf{N}(f)}$ and maximising the non-neighbourhood distance $\mathcal{D}_{\bar{\mathsf{N}}(f)}$.

Let $\psi = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{W}\}$ be the set of parameters. The gradient of the log-likelihood function is

$$\frac{\partial \log \mathcal{L}(\boldsymbol{v})}{\partial \psi} = \mathbb{E}_{\boldsymbol{v},\boldsymbol{h}}\left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \psi}\right] - \mathbb{E}_{\boldsymbol{h}|\boldsymbol{v}}\left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \psi}\right]$$

where $\mathbb{E}_{\boldsymbol{v},\boldsymbol{h}}$ is the model expectation and $\mathbb{E}_{\boldsymbol{h}|\boldsymbol{v}}$ is the data expectation. Whilst it is simple to compute the data expectation using Eq. (1), it is intractable to exactly estimate the model expectation. We hereby choose an stochastic method known as Contrastive Divergence (CD) [11] which runs short Markov chains started from the data to approximate the model expectation.

The gradient of sparsity term in Eq. (3) reads

$$\frac{\partial \mathcal{R}(\boldsymbol{v})}{\partial \psi_{\bullet j}} = \frac{P(h_j = 1 \mid \boldsymbol{v})}{\sqrt{\sum_{t \in G_m} P(h_t = 1 \mid \boldsymbol{v})^2}} \frac{\partial P(h_j = 1 \mid \boldsymbol{v})}{\partial \psi_{\bullet j}}$$

where $\psi_{\bullet j}$ is the parameter associated with hidden units h_j and the j^{th} hidden unit belongs to the m^{th} group.

To compute the gradient of the mean distances $\mathcal{D}_{\mathsf{N}(f)}$ and $\mathcal{D}_{\bar{\mathsf{N}}(f)}$ defined in Eqs. (5,6), we need the gradient for each pairwise distance $\mathcal{D}(g, f) = \mathcal{D}\left(P\left(\boldsymbol{h} \mid \boldsymbol{v}^{(g)}\right), P\left(\boldsymbol{h} \mid \boldsymbol{v}^{(f)}\right)\right)$. Taking derivative of the metric distance function with respect to parameter $\psi_{\bullet j}$, we have

$$\frac{\partial \mathcal{D}(g, f)}{\partial \psi_{\bullet j}} = \frac{\partial \mathcal{D}(g, f)}{\partial P\left(h_j^1 \mid f\right)} \frac{\partial P\left(h_j^1 \mid f\right)}{\partial \psi_{\bullet j}} + \frac{\partial \mathcal{D}(g, f)}{\partial P\left(h_j^1 \mid g\right)} \frac{\partial P\left(h_j^1 \mid g\right)}{\partial \psi_{\bullet j}}$$

in which $P\left(h_j^1 \mid f\right)$ is the shorthand for $P\left(h_j = 1 \mid \boldsymbol{v}^{(f)}\right)$. Recall from Sec. 2.1 that this probability is a sigmoid function. Thus the partial derivative with respect to the mapping column $\boldsymbol{W}_{\bullet j}$ is then

$$\frac{\partial P\left(h_j^1 \mid f\right)}{\partial \boldsymbol{W}_{\bullet j}} = P\left(h_j^1 \mid f\right)\left(1 - P\left(h_j^1 \mid f\right)\right)\boldsymbol{v}^{(f)} \tag{7}$$

As defined in Eq. (4), the derivative $\frac{\partial \mathcal{D}(g,f)}{\partial P\left(h_j^1|f\right)}$ depends on the derivative of the KL-divergence, which reads

$$\frac{\partial \mathrm{KL}(g\|f)}{\partial P\left(h_j^1 \mid g\right)} = \sum_{j}\left(\log \frac{P\left(h_j^1 \mid g\right)}{P\left(h_j^1 \mid f\right)} - \log \frac{1 - P\left(h_j^1 \mid g\right)}{1 - P\left(h_j^1 \mid f\right)}\right) \tag{8}$$

Finally, parameters are updated using stochastic gradient ascent as follows

$$\psi \leftarrow \psi + \lambda\left(\frac{\partial}{\partial \psi}\mathcal{L}_{reg}\right)$$

for some learning rate $\lambda > 0$.

3. EXPERIMENTS

In this section, we quantitatively evaluate our method on two real datasets. Both datasets are subsets selected from the NUS-WIDE dataset [12], which was collected from Flickr. The NUS-WIDE dataset includes 269,648 images which are associated with 5,018 unique tags. There are 81 concepts in total. For each image, six types of low-level features [12] are extracted, including 64-D color histogram in LAB color space, 144-D color correlogram in HSV color space, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise LAB-based color moments extracted over 5×5 fixed grid partitions and 500-D bag-of-word (BOW) based on SIFT descriptions.

For training our model, mapping parameters $\boldsymbol{W}$ are randomly initialised from small normally distributed numbers, i.e. Gaussian $\mathcal{N}(0; 0.01)$, and biases $(\boldsymbol{a}, \boldsymbol{b})$ are set to zeros. To enhance the speed of training, we divide training images into small "mini-batches" of $B = 100$ images. Hidden and visible learning rates are fixed to 0.02 and 0.3, respectively[1]. Parameters are updated after every mini-batch and the learning finishes after 100 scans through the whole data. Once parameters have been learned, images are projected onto the latent space using Eq. (1). We set the number of hidden groups M to the expected number of groups (e.g., concepts), and the number of hidden units K is multiple of M. The retrieved images are ranked based on the negative KL-div on these latent representations. We repeat 10 times and report the mean and standard deviation of the performance measures.

3.1. Retrieving Animals

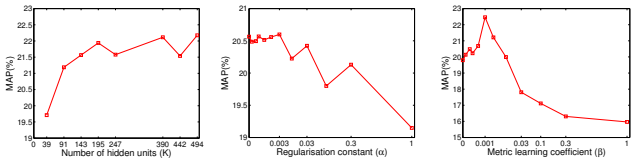

Fig. 3: The MAP performance (%) as functions of the hyper-parameters: (Left) The number of hidden units K (with $\alpha = 0.003$ and $\beta = 0.001$); (Middle) The regularisation constant α (with $K = 195$); (Right) The metric learning coefficient β (with $K = 195$ and $\alpha = 0.003$).

The first subset is the NUS-WIDE animal dataset which contains 3,411 images of 13 animals - *squirrel, cow, cat, zebra, tiger, lion, elephant, whale, rabbit, snake, antler, wolf and hawk*. Fig. 2 shows example images of each category. Out of 3,411 images, 2,054 images are used for training and the remaining for testing. In the testing phase, each test image is used to query images in training set to receive a list of images ranked basing on similarities. These settings are identical to those used in previous work [14, 15, 13]. For our methods, the similarity measure is negative symmetric Kullback-Leibler divergence (KL-div) learned from data (Sec. 2.4). The retrieval performance is evaluated using Mean Average Precision (MAP) over all received images in training set. Two images are considered similar if they depict the same type of animal.

Table 1: Image retrieval results to compare with recent state-of-the-art multiview learning and hierarchical modelling techniques on NUS-WIDE animal dataset. RBM+SG+ML is RBM with latent sparse groups and metric learning.

Method	*MAP*
DWH [13]	0.153
TWH [13]	0.158
MMH [13]	0.163
NHFA-GGM (approx.) [14]	0.179±0.013
Proposed NHFA-GGM [15]	0.195±0.013
RBM	*0.199±0.001*
RBM+SG	*0.206*±0.002
RBM+SG+ML	***0.252±0.002***

In this experiment, we concatenate first five histogram features of each animal image into a long vector and ignore BOW features to fairly compare with recent work. Thus we treat elements of the vector as Gaussian units and normalise them across all training images to obtain zeros mean and unit standard variance. Note that the MV.RBM here reduces to the plain RBM with single Gaussian type.

To find the best setting of the hyper-parameters α, β and K, we perform initial experiments with varying values. Fig. 3 reports the MAP performance (%) with respect to these values. Here $\alpha = 0$ means no sparsity constraint and $\beta = 0$ means no metric learning. As can be seen from the left figure, the performance stops increasing after some certain hidden size. Adding certain amount of sparsity control slightly improves the result (see the middle figure). A much stronger effect is due to metric learning, as shown in the right figure. From these observations, we choose $K = 195$ (15 units per group), $\alpha = 0.003$ and $\beta = 0.001$.

Fig. 4 shows how structured sparsity and metric learning contributes to the higher retrieval quality. The naive nearest neighbour on concatenated normalised features[2] would confuse a wolf with the query of lion, possibly due to the similar colour profiles. The standard RBM admits the same error suggesting that learning just regularities is not enough. Adding structured sparsity (RBM+SG) corrects one error and using learned metric (RBM+SG+ML) would correct all the errors.

Finally, Table 1 presents the MAP results of our methods (RBM, RBM with sparse group (SG) and with metric learning (ML)) in comparison with recent work [13, 14, 15] on the NUS-WIDE animal dataset. It is clear that RBM and RBM

[1] Learning rate settings are different since the hidden are binary whilst the visible are unbounded.

[2] Each feature is normalised to zero mean and unit standard variance over images.

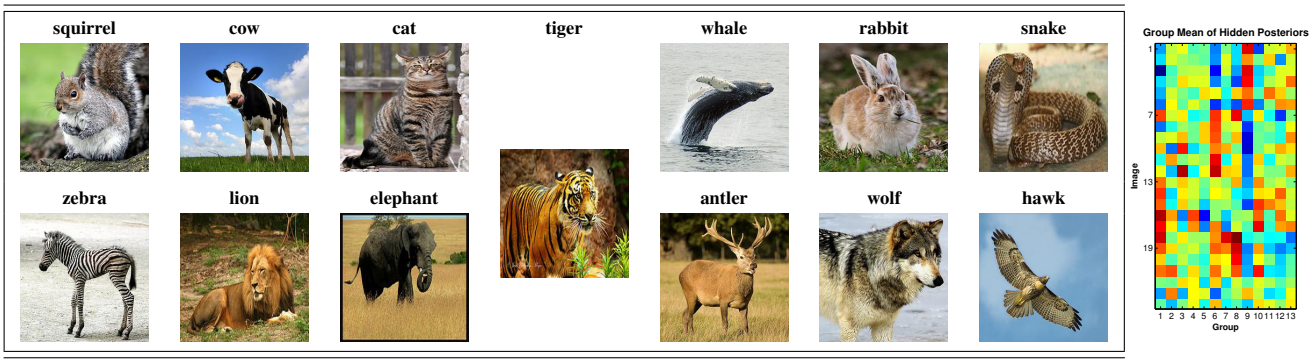

Fig. 2: Example images of each species in NUS-WIDE animal dataset. The last column shows the group mean of hidden posteriors by colours, one line per image. The red cells illustrate higher values whilst the blue denote the lower. It is clear that 4 groups of 6 consecutive images form 4 strips in different groups (9,6,1,8).

Table 2: Comparison of image retrieval results with 4 baselines on NUS-WIDE single label subset. (*model*)+SG+ML means *(model)* is integrated with sparse groups and metric learning. *MAP@100* is evaluated at top 100 similar images. *N@10* = NDCG estimated at top 10 results. (↑%) denotes improved percentage.

Method	*MAP@100* (↑%)	*N@10* (↑%)
kNN	0.283	0.466
RBM	0.381±0.001(+34.6)	0.565±0.001(+21.2)
RBM+SG	0.402±0.035(+42.1)	0.584±0.001(+25.3)
MV.RBM	0.455±0.002 (+60.8)	0.631±0.002 (+35.4)
MV.RBM+SG	0.483±0.002 (+70.7)	0.668±0.002 (+43.4)
MV.RBM+SG+ML	**0.508±0.002(+79.5)**	**0.679±0.001(+45.7)**

with SG are competitive against all previous methods; and RBM integrated with SG and ML significantly outperforms state-of-the-art approaches.

3.2. Retrieving Individual Concepts

In the second experiment, we aim to demonstrate the capability of our method to handle heterogeneous types of features and larger data. We randomly pick 10,000 images for training and 10,000 for testing. Each image in this subset has exactly one concept and altogether, they cover the entire 81 concepts of the NUS-WIDE dataset. Six visual features (1 bag-of-word and 5 histogram-like) and associated social tags, limited to 1,000, of each image are taken. The MV.RBM encodes 5 histogram features as Gaussian, social tags as binary and BOW as Poisson units. We further transform counts into log space using $[log\,(1 + \text{count})]$.

Besides MAP score, we also compute the Normalized Discounted Cumulative Gain (NDCG) [16] for evaluation. Here we only use the top 100 similar images for calculating MAP and top 10 images for computing NDCG. We create 2 baselines and 4 versions of our approach to show the improvement of the MV.RBM when adding sparse groups and metric learning (MV.RBM+SG+ML). The first baseline is to employ k-NN method on concatenated feature vectors. First features are normalised to zeros mean and unit vector over images to eliminate the differences in dimensionality and scale. The second baseline is fusion of multiple plain RBMs, each of which is type specific, i.e., BOW as Poisson, visual histograms as Gaussian and textual tags as binary. For each type of RBM, visible input data is mapped into binary latent representation. Then these latent representations are concatenated into a single latent representation.

The first version (RBM+SG) is the second baseline with group-wise sparsity (Sec. 2.2). The second version (MV.RBM) jointly models all 7 types of features. The third version (MV.RBM+SG) is MV.RBM with 81 sparse groups (Sec. 2.2). And finally, the proposed solution (MV.RBM+SG+ML) integrates both the sparsity and the metric learning into the MV.RBM.

Different from the first experiment, we query within testing set for each testing image[3]. Table 2 reports the retrieval results of all RBM models. Again, it demonstrates that (i) representation learning, especially when it comes to fusing multiple feature types and modalities, is highly important in image retrieval, (ii) adding structured sparsity can improve the performance, and (iii) distance metric, when jointly learned with representation, has significant effect on the retrieval quality. In particular, the improvement over the k-NN when using the proposed method is significant: MAP score increases by 79.5% and NDCG score by 45.7%.

4. CONCLUSION

We have presented a novel probabilistic image retrieval framework that simultaneously learns the image representation and the distance metric. The framework is based on

[3]This way of testing is more realistic since we do not always have all images for training.

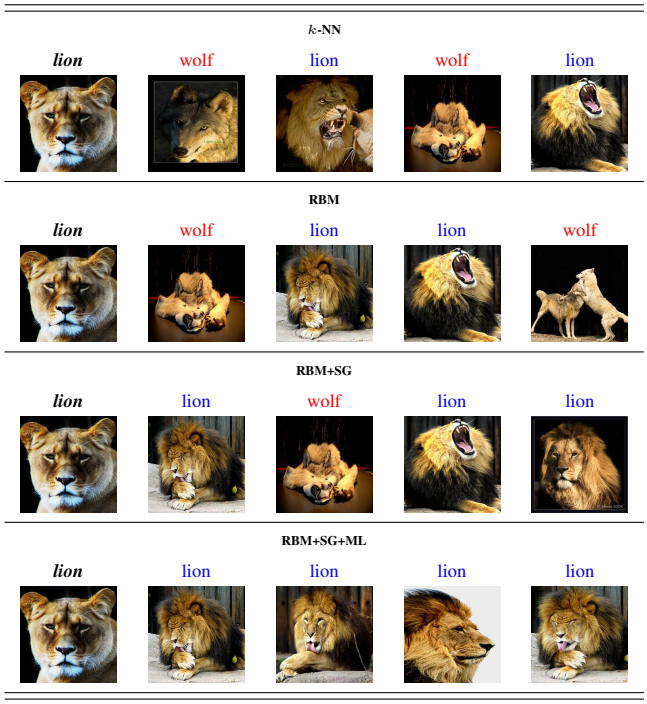

Fig. 4: Retrieved images for query image of a lion in testing set. k-NN: $k-$nearest neighbours, RBM+SG+ML: RBM with sparse group, metric learning. First column is the queried images. Blue titles are right retrieval whilst the red are wrong. Four retrieved images are sorted in descent order of similarities from left to right.

our recent architecture known as Mixed-Variate Restricted Boltzmann Machine which can seamlessly integrate multiple feature types and modalities. Our main extensions are the handling of visual/textual word counts, and a regularisation scheme that promotes structured sparsity in the learned representation, suppresses intra-concept information-theoretic distances, and enlarges inter-concept distances. Our experiments on the NUS-WIDE data confirm that (i) effective representations for image retrieval can be learned from multiple raw feature sets and modalities, (ii) performance can be further improved by appropriate structured sparsity in the learned representation, and (iii) distance metric should be learned jointly with the representation.

5. REFERENCES

[1] Truyen Tran, Dinh Phung, and Svetha Venkatesh, "Mixed-Variate Restricted Boltzmann Machines," *Proc. of ACML*, 2011.

[2] R. Salakhutdinov and G. Hinton, "Semantic hashing," in *SIGIR Workshop on Information Retrieval and Applications of Graphical Models*, 2007, vol. 500.

[3] Heng Luo, Ruimin Shen, Changyong Niu, and Carsten Ullrich, "Sparse Group Restricted Boltzmann Machines," in *Proc. of AAAI*, 2011.

[4] T. Hertz, A. Bar-Hillel, and D. Weinshall, "Learning distance functions for image retrieval," in *Proc. of CVPR*, 2004, vol. 2, pp. 570–577.

[5] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov, "Neighbourhood components analysis," in *Proc. of NIPS*, 2004, pp. 513–520.

[6] J. Yu, J. Amores, N. Sebe, and Q. Tian, "A new study on distance metrics as similarity measurement," in *Proc. of ICME*, 2006, pp. 533–536.

[7] Dacheng Tao, Xiaoou Tang, Xuelong Li, and Yong Rui, "Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 716–727, 2006.

[8] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. of NIPS*, 2006.

[9] T. Tran, D.Q Phung, and S. Venkatesh, "Learning boltzmann distance metric for face recognition," in *Proc. of ICME*, 2012.

[10] Yoav Freund and David Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," Tech. Rep., 1994.

[11] Geoffrey E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[12] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng, "NUS-WIDE: A Real-World Web Image Database from National University of Singapore," in *Proc. of CIVR*, 2009.

[13] N. Chen, J. Zhu, and E.P. Xing, "Predictive subspace learning for multi-view data: a large margin approach," *Proc. of NIPS*, vol. 24, 2010.

[14] S. Gupta, D. Phung, B. Adams, and S. Venkatesh, "A bayesian framework for learning shared and individual subspaces from multiple data sources," *Proc. of KDD*, pp. 136–147, 2011.

[15] S.K. Gupta, D. Phung, and S. Venkatesh, "A Slice Sampler for Restricted Hierarchical Beta Process with Applications to Shared Subspace Learning," in *Proc. of UAI*, 2012, pp. 316–325.

[16] Kalervo Järvelin and Jaana Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.