Probabilistic graphical models **2014/2015**

Final project : 3 – Final report — January 11, 2014

Student: Shuyu Dong

Joint project with Object recognition and computer vision (RecVis Topic 1.B)

Abstract — We carry out a study on canonical correlation analysis for image-to-image, text-to-image, and image-to-text search. Canonical correlation analysis is a natural framework for multimodal matching as it considers a common latent space in correlation with features of different modals/views of an image. These views are the visual features, the textual features related to the image or a possible underlying semantic of the image content. In this project, We extracted neural network features for the visual and textual representations with Matconvnet and word2vec respectively. We implement the image-text retrieval system based on the visual and textual views within the INRIA-webqueries dataset and evaluate the performance of the system.

1.1 Canonical correlation analysis

We first review the classic technique of canonical correlation analysis (*Hardoon et al.*, 2007).

Consider two views X(visual) and T(text): $X \in \mathbf{R}^{p\times n}$ and $T \in \mathbf{R}^{q\times n}$ are one realization of n observations of the two features.

We want to maximize correlation between two projections along w_x and w_t of the observations: $\max_{w_x,w_t} \mathrm{Cor}(w_x^\top X, w_t^\top T) \Leftrightarrow \max_{w_x,w_t} = \frac{w_x^\top \mathbb{E}(XT^\top) w_t}{\|(w_x^\top X)^\top\| \|(w_t^\top T)^\top\|}$, where $\mathbb{E}(XT^\top)$ is estimated by $\frac{1}{n}\sum_{i=1}^n xt^\top = \frac{1}{n} XT^\top$.

- In fact, $\mathbb{E}(XT^\top) = C_{xt}$, where C_{xt} is one of the non-diagonal bloc of the covariance $C(X,T) = \mathbb{E}[(x^\top t^\top)^\top (x^\top, t^\top)]$
- Kernel CCA : a generalization of CCA by performing the analysis in the kernel space to get richer descriptors.
- Resolution: we notice that the objective function is invariant by dilatation along the directions w_x and/or w_y(either one or both), so we have a maximization of a bilinear form within the unit-ellipsoid couple $\boldsymbol{S}_{C_{xx}}^{p-1} \times \boldsymbol{S}_{C_{tt}}^{q-1}$, which leads to a generalized form of eigen-problem :

$$C_{xt} C_{tt}^{-1} C_{tx} w_x = \lambda^2 C_{xx} w_x, \tag{1.1}$$

we get a sequence of eigen-vectors (w_x^k) with λ^k and afterwards (w_t^k) by the identity

$$w_t = \frac{C_{tt}^{-1} C_{tx} w_x}{\lambda}.$$

The maximal number of independent canonical directions is $\min(p, q)$.

The canonical correlation analysis has a probabilistic interpretation (*Bach and Jordan*, 2005), moreover, it provides a more intuitive insight to the solving of the problem in practical. First, the latent common space is the space of a latent variable z that leads to the model of a binary tree with two leaves x and t such that, for modals $m =' x'$ and $'t'$:

- $z \sim \mathcal{N}(0, I_d)$, where d is the dimension of a subspace of the canonical directions $\text{Vect}(w_m^k)_{k=1,..,\min(p,q)}$,
- $x_m|z \sim \mathcal{N}(W_m z + \mu_m, \Phi_m)$, $W_m \in \mathbf{R}^{p_m \times d}, \Phi_m \succeq 0$. And Theorem 2 in (*Bach and Jordan*, 2005) gives the maximum likelihood estimates and interpretations. Particularly, we have an equivalent formulation for the generalized eigen-problem of (Eq.1.1):

$$\begin{pmatrix} C_{xx} & 0 \\ 0 & C_{tt} \end{pmatrix}^{-1} \begin{pmatrix} 0 & C_{xt} \\ C_{tx} & 0 \end{pmatrix} \hat{w} = \lambda \hat{w}. \tag{1.2}$$

  where

$$\hat{w} = \begin{pmatrix} \mu_x & 0 \\ 0 & \mu_t \end{pmatrix} \begin{pmatrix} w_x \\ w_t \end{pmatrix} = \begin{pmatrix} \mu_x w_x \\ \mu_t w_t \end{pmatrix}$$

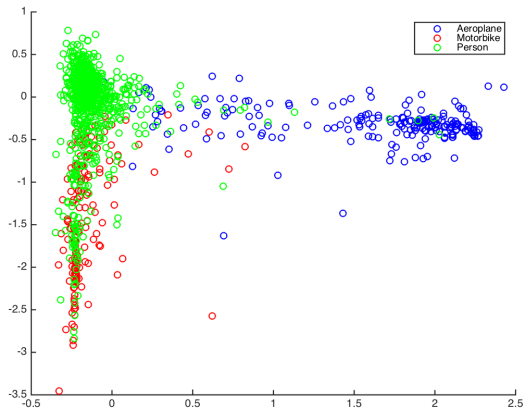
  is the weighted concatenation of the canonical directions. We find clearly by multiplying the two (co)variance bloc matrices in the lefthand side and operate two times the resulting matrix on $\hat{w}$, that μ_x and μ_t dispear and the equation becomes identical to (Eq.1.1).

  As a result, we are put to resolve the eigen-problem (Eq.1.2) in the implementation, note that the canonical directions $(w_m^k, m = x, t)$ obtained in this implementation are implicitely weighted by their respective $\mu_m^k, \quad m = x, t$.

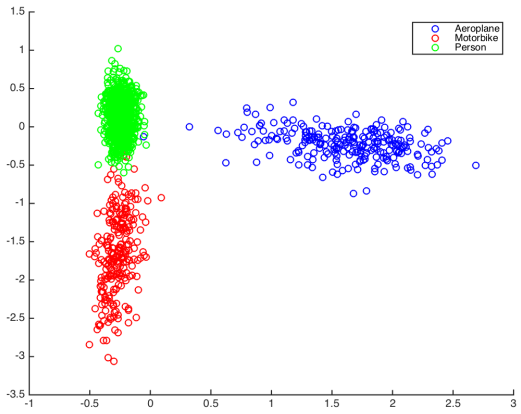
1.1.1 Preliminary tests on half-synthetic data

We first visualize the most important low-dimensional projections of the two views. The solutions $(w_x^k), (w_t^k)$ are ordered by decreasing eigenvalues. So we visualize two or three of the first several dimensions of the canonical variates, for example, point clouds representing the second and third dimensions: $(X^\top w_x^2, X^\top w_x^3)$. We note that the half-synthetic data is composed of real CNN features of images extracted with **Matconvnet** and the textual feature are synthetic:

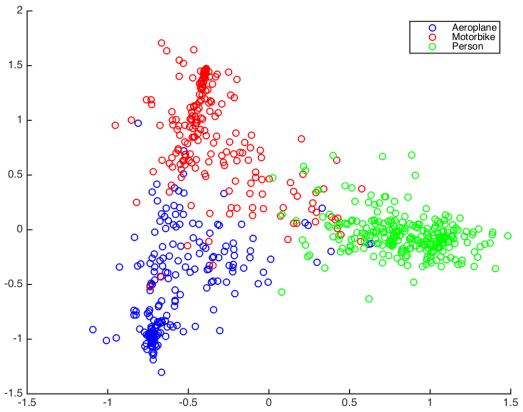
Visual features: CNN features of three classes(*aeroplane, motorbike, person*), there are 200 ~ 300 features for classes *aeroplane* and *motorbike* and around 1000 for *person*; Text features: T is composed of toy synthetic(random within the non-zero support) word vectors of dimension 200(like for `word2vec` vectors), the three classes have different non-zero supports within the 200 dimensions.



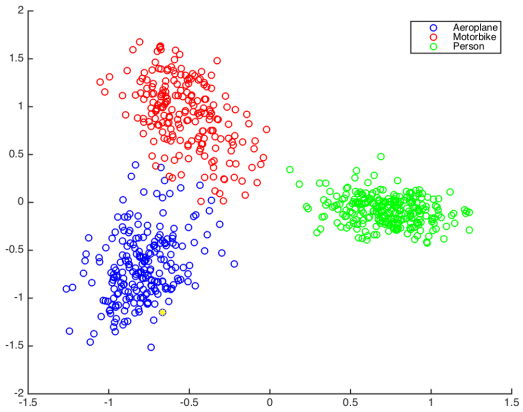
(a) Canonical variates along the *first two* directions; 'visual'.



(b) Canonical variates along the *first two* directions; 'text'.



(a) Canonical variates along directions 2 and 3 : $W_m(:,2)$ and $W_m(:,3)$; 'visual'.



(b) Canonical variates along directions 2 and 3 : $W_m(:,2)$ and $W_m(:,3)$; 'text'.

1.2 Dataset: INRIA-webqueries

We will see in this section the characteristics of the dataset: INRIA-webqueries is composed of 71,748 images, each image is related to one xml file containing information of the web page from which the image is downloaded. The dataset is originated from 353 web search queries that represent various semantics such as tourist attractions, machines and objects, stars, sports, logos etc. Each of the 353 classes is comprised of about 200 ~ 300 images, some of which being irrelevant to the semantic, which is one source of the noise.

The **metadata** of each image is taken from the webpage and contains notably the title of the page and the texts surrounding the image (before and after it). We only use the text in these three fields to get text descriptions for each image. However, a non-negligeable proportion of these texts (not to mention those of other fields such as <imageURL> or

<referer>) are irrelevant to the image content. As a result, INRIA-webqueries' text view has a source with greater noise compared to other datasets such as Pascal1K, Flickr8K, Flickr30K etc.

1.2.1 Visual feature

The visual view is represented by CNN features of images using **Matconvnet**. The feature vector of the 20th and last layers of dimensions 4096 and 1000 are both tested as the visual feature.

We note that certain images (4.74% of all) do not have valid CNN representations due to their image file property (e.g. some of the images are actually in format **gif**). On the other hand, a very small percentage(0.51%) of images do not have text in neither of the three fields(<ptitle>, <before> and <after>).

Nevertheless, as these problems only represent 5.2% of the total dataset, in the whole project we only consider the images that have valid CNN features and text descriptions at the same time, which represent 94.8% of the dataset.

The following example shows the content of the dataset.

metadata file	metadata/query_150_document_0_textmeta.xml	<?xml version="1.0" encoding="utf-8"?> <documentmeta> <concept> trompette </concept> <rank> 1 </rank> <language> en </language> <referer> http://allthingschill.com/wordpress/archives/2004/11/wallpapers/ </referer> <imageUrl> http://allthingschill.com/img/wallpaper/trompette2.jpg </imageUrl> <before> 1024 x 768 Au Revoir Trompette </before> <after> </after> <ptitle> All Things Chill » Blog Archive » Wallpapers </ptitle> <alt> Trompette </alt> </documentmeta>
thumbnail file	images/query_150_document_0_imagethumb.jpg	

Figure 1.3: For the text features, only text between (<ptitle>, </ptitle>), (<before>, </before>) and ((<after>, </after>)) are extracted.

1.2.2 Textual feature

The text view is represented by vectors constructed from those of texts between <ptitle>, <before> and <after>; the vectors of these words come from a word-vector dictionary trained on the entire metadata using **word2vec**.

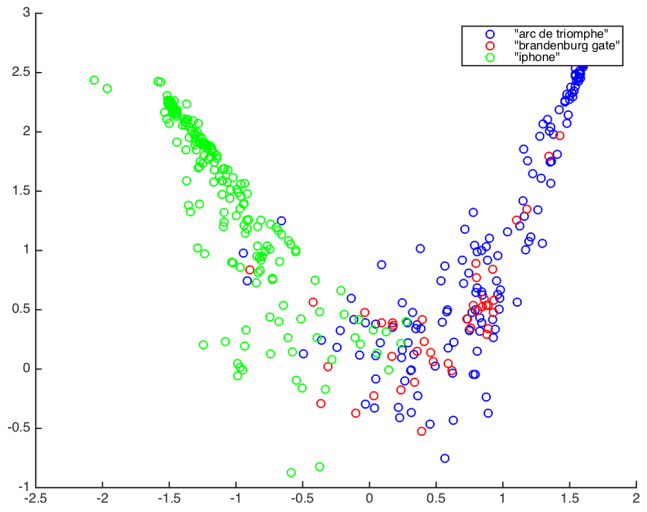
Precisely, for the nature of **metadata**, we only extract texts between (<ptitle>, </ptitle>), (<before>, </before>) and ((<after>, </after>)), `sed` and `awk` are used for this extraction. As a result, the entire text `inria_text_all.txt` of the metadata is the concatenation of all images' three text fields. We put this text file to train the dictionary with **word2vec**. The dictionary `dictionary_inriaPBA.txt` contains 22,291 word(s)-vector correspondances. The word vector is 200−dimensional.

We notice that two words that have similar concepts or originated from the same context (in the original text) will lead to large cosine value between their vectors, thus the word2vec features provide useful and relevant representation for the text view.

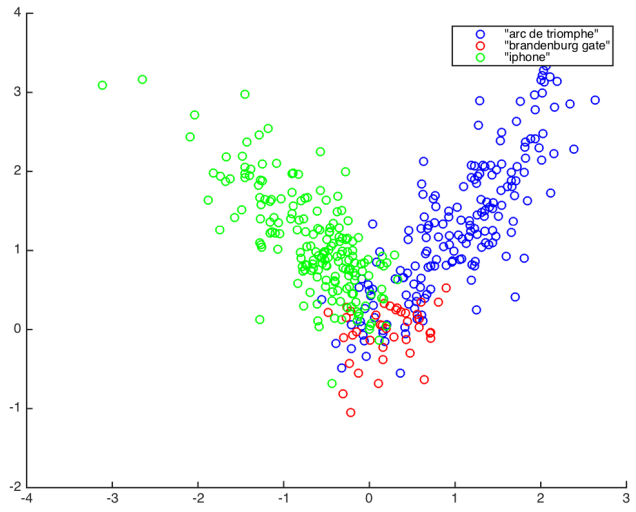
As the textual feature depends almost totally on the word-vector dictionary trained by **word2vec**, which has two versions, one with only words and very few word combinaisons, and the other provides vectors for almost all highly frequent or correlated word combinations, such as *"arc_de_triomphe", "mont_blanc", "september_11th"* etc. Therefore, two approaches of extracting textual features are tried with respect to these two versions of dictionaries.

1.3 Training and test

We take 4/5 of the total dataset as the training data and obtained the matrices containing (w_x^k) and (w_t^k), now we shall visualize the real canonical variates. There are $54,211$ training images/texts of 300 classes, we only show the data points of three classes of images/texts.



(a) Canonical variates along directions 1 and 2: $W_m(:,1)$ and $W_m(:,2)$; 'visual'.



(b) Canonical variates along directions 1 and 2: $W_m(:,1)$ and $W_m(:,2)$; 'text'.

Figure 1.4: We notice that the projected points of *arc de triomphe* and *brandenburg gate* are difficult to be distinguished in the 'visual' view and that is easier in the 'textual' view. In both views, the clusters of *arc de triomphe* and *iphone* have two different directions.

One important parameter for the retrieval tasks is the number d of top ranked eigenvectors $(w_m^k)_{k=1,..,d} = W_m$, for $m = x$ (visual) and t (textual) . The cosine similarity measure in the latent space depends much on this dimensionality:

$$S(x,t) = \frac{(W_x^\top \varphi(x))^\top (W_t^\top \varphi(t))}{\|W_x^\top \varphi(x)\|_2 \|W_t^\top \varphi(t)\|_2}.$$

We notice that it is equivalent to the similarity meaure defined in (*Gong et al.*, 2014) in case the p-th power is set to 1. The measure S between the visual view x and the textual view t will be extended naturally to that within the visual view (for image-to-image retrival).

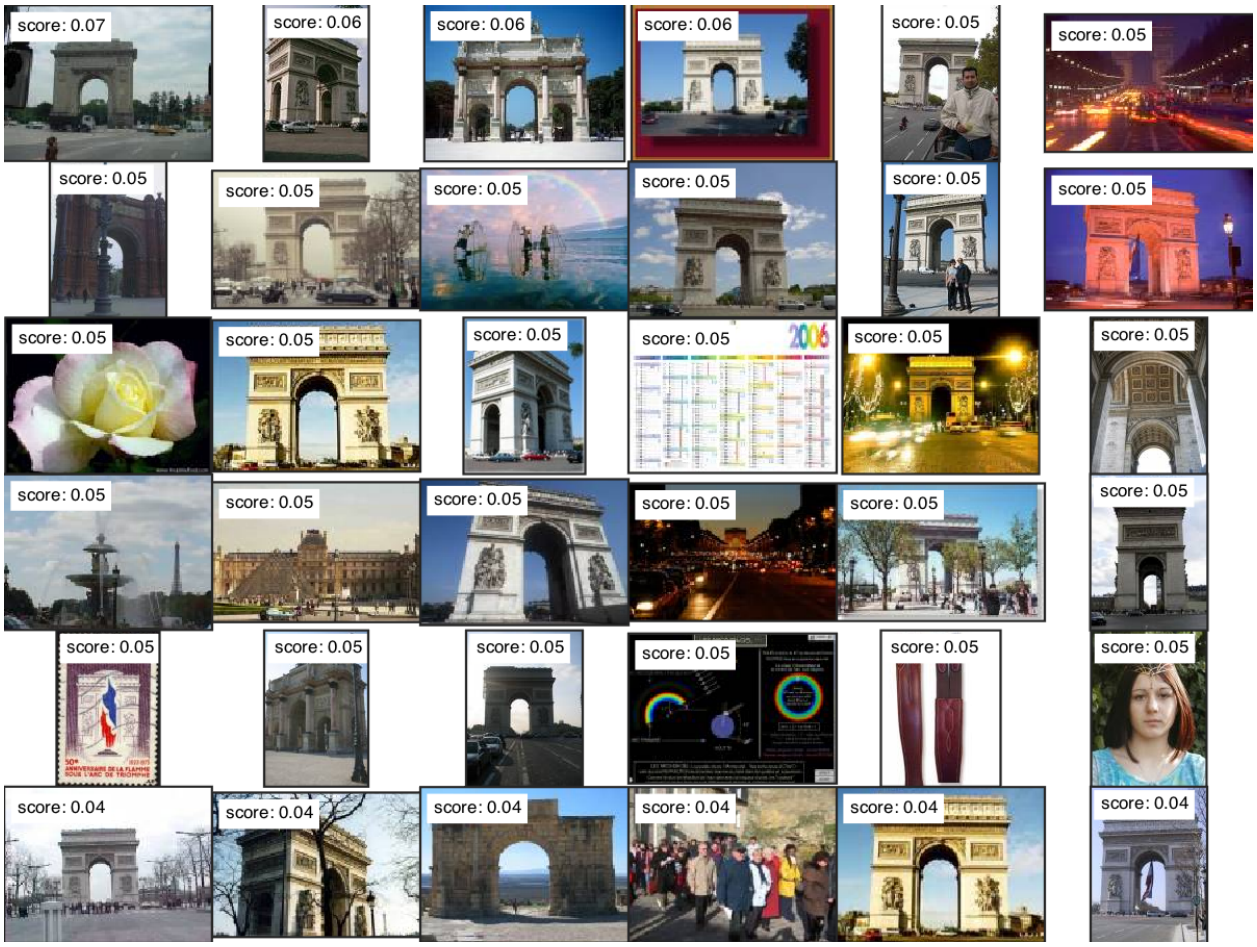
We visualize some of the image-text retrieval examples on test data using the above similarity measure.

1.3.1 Image-to-text search on test data

We carry out the image-to-text search on the remaining test data and visualize the results of several images

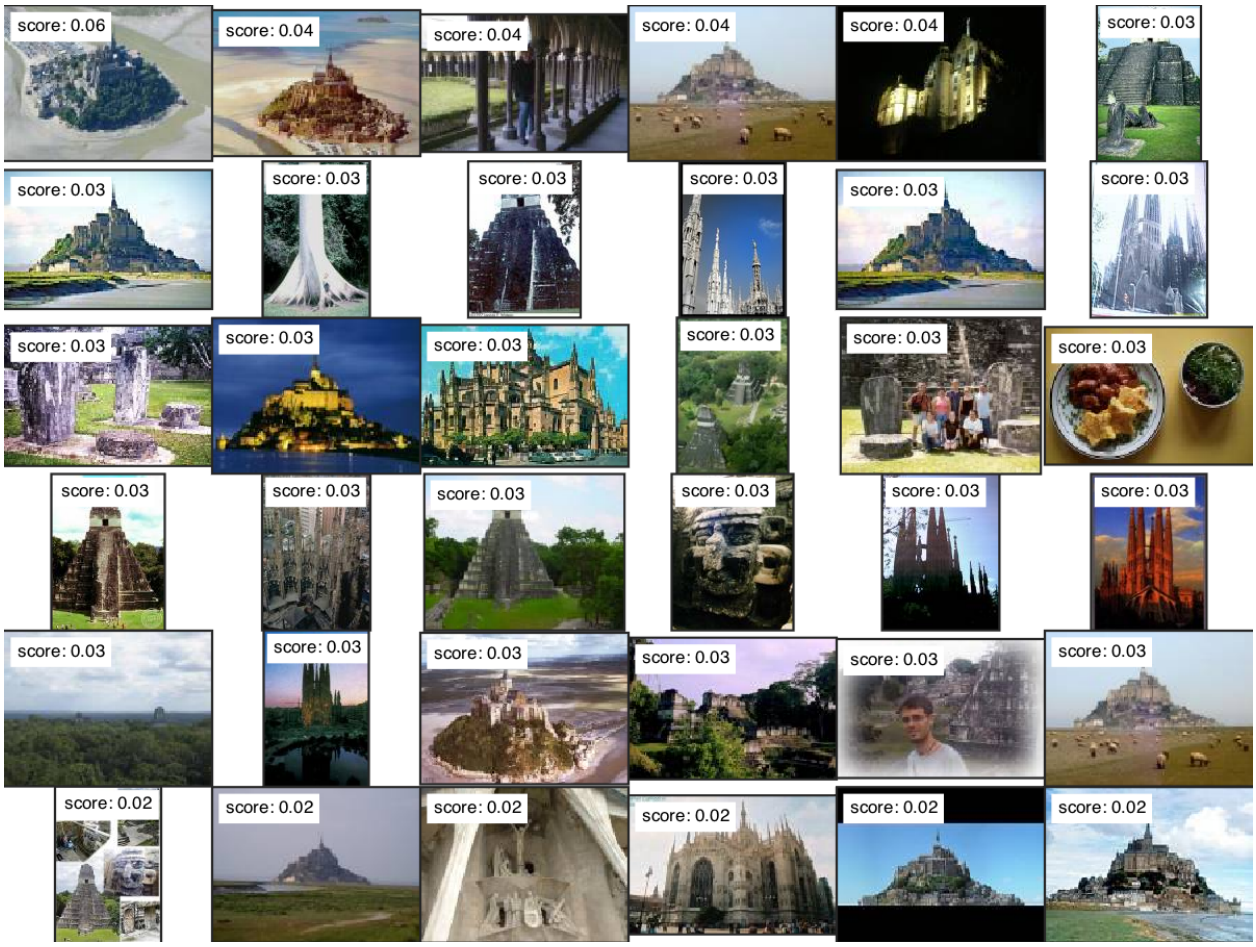
Figure 1.5: Five examples of image-to-text search on test data: the words correspond to the trained text vectors that have the highest similarities with the test image's visual feature; the scores is the count of occurrences of such each word.

For three images in the figure (1.5) and we visualize the images in training set that correspond to the top ranked texts:




score: 0.06
score: 0.06
score: 0.05
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
Click to view the W960i
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
score: 0.04
Say hello to everyone.
score: 0.04
score: 0.04
score: 0.04
(Click to enlarge)
score: 0.04
score: 0.04
score: 0.04
score: 0.04
MOTOROLA
score: 0.03
score: 0.03
score: 0.03

Figure 1.6

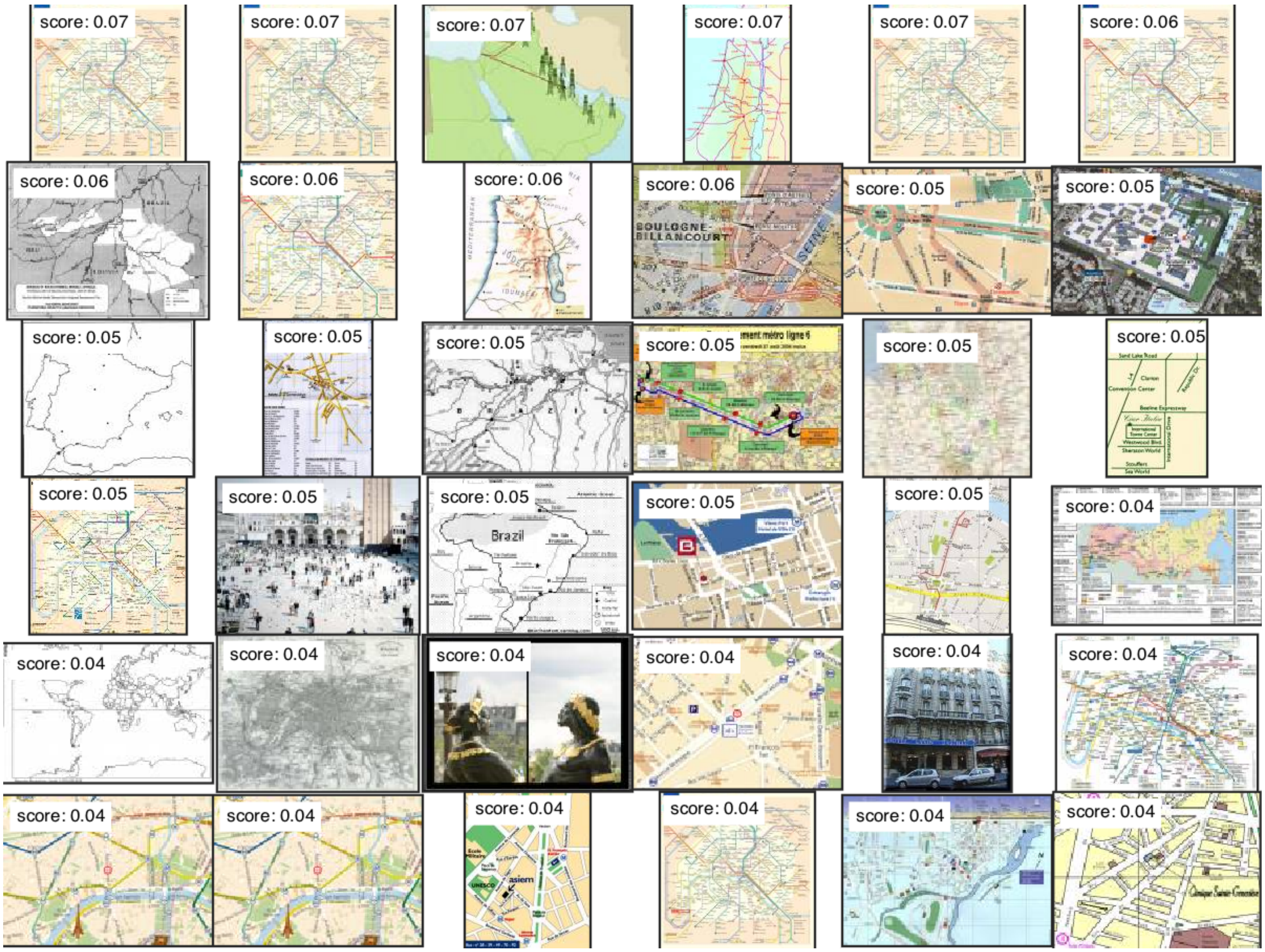


1.3.2 Text-to-image search on test data

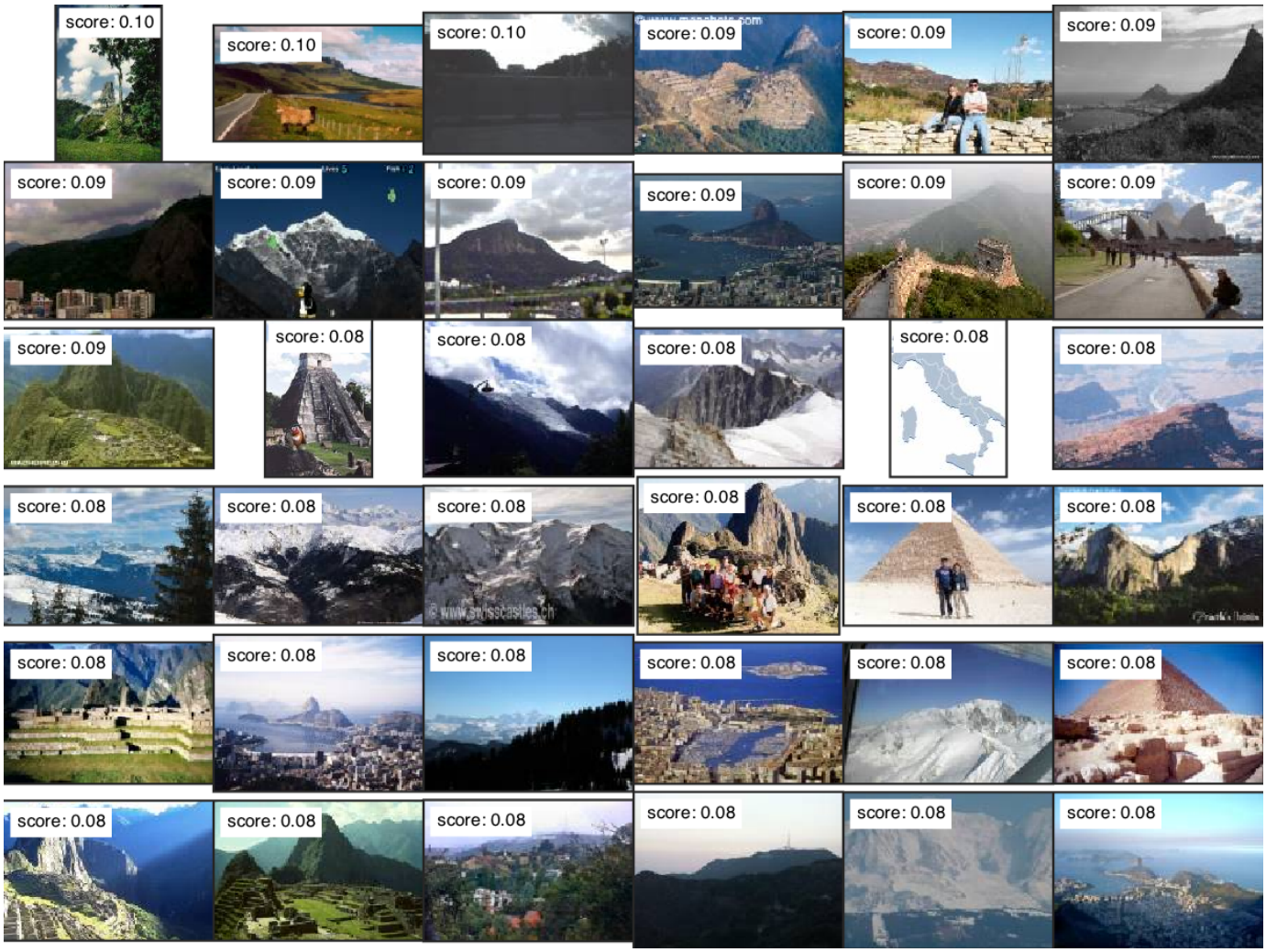
We carry out the text-to-image search on the same test data and visualize the top ranked images:

By typing the follwoing texts, we have:

" metro paris plan":



And "mont blanc":



1.4 Evaluations and conclusion

Bibliography

Bach, F. R., and M. I. Jordan (2005), A probabilistic interpretation of canonical correlation analysis, *Tech. rep.* *Cited on page 2*

Gong, Y., Q. Ke, M. Isard, and S. Lazebnik (2014), A multi-view embedding space for modeling internet images, tags, and their semantics, *Int. J. Comput. Vision*, *106*(2), 210–233, doi:10.1007/s11263-013-0658-4. *Cited on page 6*

Hardoon, D. R., S. Szedmak, O. Szedmak, and J. Shawe-taylor (2007), Canonical correlation analysis; an overview with application to learning methods, *Tech. rep.* *Cited on page 1*