LEARNING THE SEMANTICS OF MULTIMEDIA CONTENT WITH APPLICATION TO WEB IMAGE RETRIEVAL AND CLASSIFICATION

Alexei Vinokourov, David R. Hardoon and John Shawe-Taylor

Dept. Computer Science, Royal Holloway, University of London, Egham, TW20 0EX, UK
{alexei, davidh, john}@cs.rhul.ac.uk

ABSTRACT

We use kernel Canonical Correlation Analysis to learn a semantic representation of Web images and their associated text. This representation is used in two applications. In first application we consider classification of images into one of three categories. We use SVM in the semantic space and compare against the SVM on raw data and against previously published results using ICA. In the second application we retrieve images based only on their content from a text query. The semantic space provides a common representation and enables a comparison between the text and image. We compare against a standard cross-representation retrieval technique known as the Generalised Vector Space Model.

1. INTRODUCTION AND PREVIOUS WORK

With increasingly vast amount of multimedia content available in digital form on- and offline, it has become crucially important to be able to process large amounts of mixed text / image / video information. A separate image / text data processing has been extensively explored in literature though only few works appear to have taken an advantage of combined data analysis. In this work we suggest a novel approach in this area.

The best way to present the method being suggested in this study is, probably, to start from the bunch of approaches to semantically analyse text and/or image data. An opening accord in semantic text processing has been played by latent semantic indexing work [3] closely related to Principal Component Analysis (PCA). A number of follow-up methods also have emerged, for example, Latent Semantic Kernel (LSK) [2]. The sparsity of text data, however, turned out to be an obstacle to further development of the approach. It has been realised that text documents in a collection can be considered to be generated by a mixture of a few independent sources, similarly in a way to newswire articles that come to a news agency from diverse sources. A pioneering work on analysis of text from this prospective has demonstrated the success of the approach [4]. Another landmark work in this area exploits Nonnegative Matrix Factorisation (NMF), closely related to ICA, to analyse text and images content separately [6]. Naive Bayes and extending it Multinomial Hierarchical ASymmetric Analysis (MASHA) also have proved to be quite successful approaches in text as an ICA-related alternative to LSI/PCA [7].

However, in [5] it is shown that in combination the semantic multimedia content analysis of different types of data (text and images) can bring more advantage than the analysis of each type of the data separately for, particularly, data categorisation problem.

We have analysed previously the semantics of text using different sources of semantically the same content - original text in English and its translation in French and we have shown that by finding canonical correlations between originals and their translations one can extract the translation-invariant semantics of the text [9]. It is therefore natural to apply the same method, named kernel Canonical Correlation Analysis (KCCA) to find correlations between web images and attached text. In this work the set of canonical correlation directions, comprising the information about semantics of the text, is used not only to improve categorisation of the images but also to retrieve them according to text queries.

In Section 2 we give the description of the algorithm, which, in contrast to previously published work [9], is extended to show connection between KCCA via incomplete Cholesky decomposition and Gram-Schmidt procedure used in Latent Semantic Kernel [2]. In Section 3.1 we apply KCCA to classify images and attached text and compare it with plain Support Vector Machine (SVM) classification, and in Section 3.2 we explore image retrieval using projection of text queries and images from a test collection onto KCCA components.

2. ALGORITHM DESCRIPTION

Using the kernel-cca algorithm [1] [9] we try to obtain a standard eigenproblem for the kernel mapping of the text and image kernels. We use kernel-cca with a control on the flexibility of the projection mappings by convexly combining two utmost cases in the denominator:

$$\rho = \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{\alpha' K_x^2 \alpha \cdot \beta' K_y^2 \beta}}$$

$$\begin{aligned}\rho &= \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa \|\alpha\|^2) \cdot (\beta' K_y^2 \beta + \kappa \|\beta\|^2)}} \\ &= \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa \alpha' K_x \alpha) \cdot (\beta' K_y^2 \beta + \kappa \beta' K_y \beta)}}\end{aligned}$$

constricting $(\alpha' K_x^2 \alpha + \kappa \alpha' K_x \alpha) = 1$ and $(\beta' K_y^2 \beta + \kappa \beta' K_y \beta) = 1$. The corresponding Lagrangain is

$$\begin{aligned}L(\lambda_\alpha, \lambda_\beta, \alpha, \beta) &= \alpha' K_x K_y \beta \\ &\quad -\frac{\lambda_\alpha}{2}(\alpha' K_x^2 \alpha + \kappa \alpha' K_x \alpha - 1) \\ &\quad -\frac{\lambda_\beta}{2}(\beta' K_y^2 \beta + \kappa \beta' K_y \beta - 1)\end{aligned}$$

the derivatives are

$$\frac{\partial f}{\partial \alpha} = K_x K_y \beta - \lambda_\alpha (K_x^2 \alpha + \kappa K_x \alpha) \qquad (1)$$

$$\frac{\partial f}{\partial \beta} = K_y K_x \alpha - \lambda_\beta (K_y^2 \beta + \kappa K_y \beta) \qquad (2)$$

we can show that

$$\begin{aligned}\alpha' K_x K_y \beta - \lambda_\alpha \alpha' (K_x^2 \alpha + \kappa K_x \alpha) &= 0 \\ \beta' K_y K_x \alpha - \lambda_\beta \beta' (K_y^2 \beta + \kappa K_y \beta) &= 0 \\ \lambda_\alpha (\alpha' K_x^2 \alpha + \kappa \alpha' K_x \alpha) - \lambda_\beta (\beta' K_y^2 \beta + \kappa \beta' K_y \beta) &= 0\end{aligned}$$

$$\lambda_\alpha = \lambda_\beta$$

Where the case is that K_x and K_y are not full rank matrices, we explore the Gram-Schmidt decomposition algorithm, which is described in [2], as it is also known as the incomplete Cholesky decomposition. Complete decomposition of a kernel matrix is an expensive step and should be avoided with real world data. We slightly modify the Gram-Schmidt algorithm so it will use a precision parameter as a stopping criterion as shown in [1].
The projection is built up as the span of a subset of the projections of a set of m training examples. These are selected by performing a Gram-Schmidt orthogonalisation of the training vectors in the feature space.

Given a kernel K and precision parameter η:

Initializations:

```
$m =$ size of $K$
$size$ and $index$ are a vector with the same length as $k$
$feat$ a zeros matrix equal to the size of $k$
for $i = 1$ to $m$ do
    norm2[$i$] $= K_{ii}$;
```

Algorithm:

```
$j = 1$;
while $\sum_{i=j}^{m} norm2[i] > \eta$ do
    $i_j = argmax_i$(norm2[$i$]);
    index[$j$] $= i_j$;
    size[$j$] $= \sqrt{norm2[i_j]}$;
    for $i = 1$ to $m$ do
        feat[$i, j$] $= \frac{\left(K_{i,i_j} - \sum_{t=1}^{j-1} feat[i,t] \cdot feat[i_j,t]\right)}{size[j]}$;
        norm2[$i$] $=$ norm2[$i$]$-$ feat$(i, j) \cdot$ feat$(i, j)$;
    end;
    $j = j + 1$;
end;
return feat[$i, j$] as the $j-$th feature of input $i$;
```

Output:

```
Features satisfying $\|K - feat \cdot feat'\| \leq \eta$
```

To classify a new example:

```
for $j = 1$ to $T$
    newfeat[$j$] = $(K_{i,i_j} - \sum_{t=1}^{j-1}$ newfeat[$j, t$] $\cdot$ feat[$i_j, t$]$)/$
                    $/$size[$j$];
end;
```

Setting via Gram-Schmidt decomposition

$$\begin{aligned}K_x &\cong R_x R_x' \\ K_y &\cong R_x R_x'\end{aligned}$$

we can rewrite equation 1 and 2 as

$$\begin{aligned}R_x R_x' R_y R_y' \beta - \lambda (R_x R_x' R_x R_x' + \kappa R_x R_x') \alpha &= 0 \\ R_y R_y' R_x R_x' \alpha - \lambda (R_y R_y' R_y R_y' + \kappa R_y R_y') \beta &= 0 \\ R_x' R_x R_x' R_y R_y' \beta - \lambda R_x' (R_x R_x' R_x R_x' + \kappa R_x R_x') \alpha &= 0 \\ R_y' R_y R_y' R_x R_x' \alpha - \lambda R_y' (R_y R_y' R_y R_y' + \kappa R_y R_y') \beta &= 0\end{aligned}$$

setting

$$\begin{aligned}Z_{xx} &= R_x' R_x \\ Z_{yy} &= R_y' R_y \\ Z_{xy} &= R_x' R_y \\ Z_{yx} &= R_y' R_x \\ \tilde{\alpha} &= R_x' \alpha \\ \tilde{\beta} &= R_y' \beta\end{aligned}$$

we rewrite the equation in the following manner where the Z matrices are invertable

$$\begin{aligned}Z_{xx} Z_{xy} \tilde{\beta} - \lambda Z_{xx} (Z_{xx} + \kappa I) \tilde{\alpha} &= 0 \\ Z_{yy} Z_{yx} \tilde{\alpha} - \lambda Z_{yy} (Z_{yy} + \kappa I) \tilde{\beta} &= 0\end{aligned}$$

$$\begin{aligned}\tilde{\beta} &= \frac{(Z_{yy}+\kappa I)^{-1} Z_{yy}^{-1} Z_{yy} Z_{yx}\tilde{\alpha}}{\lambda} \\ \tilde{\beta} &= \frac{(Z_{yy}+\kappa I)^{-1} Z_{yx}\tilde{\alpha}}{\lambda}\end{aligned}$$

$$\begin{aligned}Z_{xx}Z_{xy}(Z_{yy}+\kappa I)^{-1}Z_{yx}\tilde{\alpha} &= \lambda^2 Z_{xx}(Z_{xx}+\kappa I)\tilde{\alpha} \\ Z_{xy}(Z_{yy}+\kappa I)^{-1}Z_{yx}\tilde{\alpha} &= \lambda^2 (Z_{xx}+\kappa I)\tilde{\alpha}\end{aligned}$$

Performing a complete Cholesky decomposition on $(Z_{xx} + \kappa I) = SS'$ and setting $\hat{\alpha} = S' \cdot \tilde{\alpha}$

$$S^{-1}Z_{xy}(Z_{yy}+\kappa I)^{-1}Z_{yx}S^{-1'}\hat{\alpha} = \lambda^2\hat{\alpha}$$

which is the $Ax = \lambda x$ eigenproblem.

3. EXPERIMENTS

3.1. Learning multimedia content semantics

In the following application the problem of learning semantics of multimedia content by combining image and text data is addressed. The synthesis is addressed by the kernel Canonical Correlation Analysis described in Section 2. The learnt semantics is used for classification of images and their assigned text labels. The reported results show substancially better classification than for text or image taken alone that clearly demonstrates the ability of the method to learn the connection between the two different representations of multimedia Web content. The results also demonstrate an improvement over plain SVM classification in the combined Vector Space Model.

The semantic representation is obtained by projecting data onto the semantic space spanned by vector solutions of the KCCA problem. For example, to process an unseen data point x we expand x into the vector representation for medium (image or text) space to get $\widetilde{x}$ and then project it onto the d canonical $\mathcal{F}$-correlation components: $[x] = \alpha^T Z^T \widetilde{x}$ using the appropriate vector of the appropriate medium, where α is $N \times d$ matrix whose columns are the first solutions of the KCCA problem for the given medium space sorted by eigenvalue in descending order. Here we assumed that $(\Phi(z), \Phi(\widetilde{q}))$ is simply $z^T\widetilde{x}$ where Z is the training corpus in the given space. The so-called semantics $W = Z\alpha$ can be exported and used in, for example, Support Vector Machine classification. We compute the new kernel which is the inner product of the projected data:

$$K'(x_i, x_j) = x_i^T W W^T x_j \qquad (3)$$

The multimedia image-text Web database was kindly provided by the authors of [5]. The data was divided into three classes (yahoo categories) - Sport, Aviation and Paintball - 400 records each and consisted of jpeg images retrieved from the Internet with attached text. We randomly split each class into two halves which were used as training and test data accordingly. The extracted features of the data were used the same as in [5]: image HSV colour, image Gabor texture and term frequencies in text. The results of SVM classification without semantic projection are presented in Table 1 when classification error for the SVM with semantic kernel is given in Table 1. Using cross validation with small random subsamples of 50 points per each class the optimal set of parameters was found to be as the following: KCCA regularization parameter $\kappa = 1.5$ and SVM generalization trade-off parameter $C = 1$ (the input data was normalised). The best results were found with Gaussian kernel, parameter σ equal average distance between data points in the train corpus d, and 30 semantic vectors.

K	30	60
ICA	3%	
plain SVM	2.13%±0.3%	
KCCA-SVM $\sigma = d$	**0.4**%±0.3%	1.4%±0.4%
KCCA-SVM $\sigma = d/2$	1.9%±0.3%	1.8%±0.4%
KCCA-SVM $\sigma = d/5$	4.3%±0.3%	3.4%±0.8%

Table 1. SVM classification error with and without semantic projection averaged over all three classes - Sport, Aviation and Paintball - and 10 runs.

3.2. Image retrieval using text queries

We next tested the use of the derived semantic space in an image retrieval task that uses only image content the aim is to allow retrieval of images from a text query but without reference to any labelling associated with the image. This can be viewed as a cross-modal retrieval task. We used the combined image and text database, described above, where we are trying to facilitate mate retrieval on a test set.

We set the value of κ for the regularization by running the kernel-cca with the association between image and text randomized. Let $\lambda(\kappa)$ be the spectrum without randomization and $\lambda_R(\kappa)$ be the spectrum with randomization (by spectrum it is meant that the vector whose entries are the eigenvalues). We expect for $\kappa = 0$ that we may have $\lambda(\kappa) = \lambda_R(\kappa) = j$ the all ones vector, since it is very possible that the examples are linearly independent. Though in practice only 50% of the examples are linearly independent but this does not effect the method of selection of κ . We choose κ so that the κ for which the difference between the spectrum of the randomized set is maximally different (in the two norm) from the true spectrum.

$$\kappa = argmax \|\lambda_R(\kappa) - \lambda(\kappa)\|$$

We find that $\kappa = 7$ and set the Gram-Schmidt precision parameter $\eta = 0.5$.

Dividing the overall examples into 50:50, in a random manner, we obtain 600 training examples and 600 testing examples. Training on the training examples to obtain $\tilde{\alpha}$ for the image kernel and $\tilde{\beta}$ for the text kernel.

To perform the test image retrieval we compute the features of the images and text query using the Gram-Schmidt algorithm. Once we have obtained the features for the test query (text) and test images we project them into the semantic feature space using $\tilde{\beta}$ and $\tilde{\alpha}$ respectively. Now we can compare them using an inner product of the semantic feature vector. The higher the value of the inner product, the more similar the two objects are. Hence, we retrieve the images whoe inner products with the test query are highest.

In the experiments we used the first 150 $\tilde{\alpha}$ eigenvectors and $\tilde{\beta}$ eigenvectors (corresponding to the largest eigenvalues). We computed the 10 and 30 images for which their semantic feature vector has the closest inner product with the semantic feature vector of the chosen text. A successful match is considered if the image that actually matched the chosen text is contained in this set.

Image Set	GVSM success	kernel-cca success
10	9.5%	36.34%
30	22.34%	50.67%

Table 2. Success cross-results between kernel-cca & generalised vector space. (Linear kernel for image colour)

Image Set	GVSM success	kernel-cca success
10	8%	59.5%
30	19%	69%

Table 3. Success cross-results between kernel-cca & generalised vector space. (Gaussian kernel for image colour)

We compared the performance of our method with a retrieval technique based on the Generalized Vector Space Model. This uses as a semantic feature vector the vector of inner products between either a text query and each training label or test iamge and each training image.

As shown in Tables 2 and 3 we compare the performance of the kernel-cca algorithm and generalised vector space model, where in Table 2 we use a linear kernel as above for the image colour while in Table 3 we use a Gaussian kernel wtih σ = max distance/20 . In both cases the kernel CCA method sharply outperforms GVSM. It is also clear that the Gaussian kernel gives significally better results for the KCCA though for GVSM it reduces the performance.

4. CONCLUSIONS AND FUTURE WORK

It has been demonstrated that the image / text classification is best with KCCA representation on combined data compared to plain SVM and non-combined data. It has also been shown that using this representation one can retrieve images with a mush better accuracy than with the generalised vector space model. In future we will extend our experiments on other data collections. We will also use KCCA to derive text features from image and vice versa.

5. REFERENCES

[1] F. R. Bach and M. I. Jordan. Kernel indepedendent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[2] Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. Latent semantic kernels. *Journal of Intelligent Information Systems*, 18(2/3):127–152, 2002. Special Issue on Automated Text Categorization.

[3] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[4] T. Kolenda, L. Hansen, and S. Sigurdsson. *Advances in Independent Component Analysis*, chapter Independent Components in Text, pages 229–250. Springer-Verlag, 2000.

[5] T. Kolenda, L. K. Hansen, J. Larsen, and O. Winther. Independent component analysis for understanding multimedia content. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, and S. Douglas, editors, *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, pages 757–766, Piscataway, New Jersey, 2002. IEEE Press. Martigny, Valais, Switzerland, Sept. 4-6, 2002.

[6] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, pages 788–791, 1999.

[7] Alexei Vinokourov and Mark Girolami. A probabilistic framework for the hierarchic organization and classification of document collections. *Journal of Intelligent Information Systems*, 18(2/3):153–172, 2002. Special Issue on Automated Text Categorisation.

[8] Alexei Vinokourov, David R. Hardoon, and John Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. In *Submitted to Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, Nara, Japan, 2003.

[9] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances of Neural Information Processing Systems 15 (to appear)*, 2002.