Deep Correlation for Matching Images and Text

Fei Yan, Krystian Mikolajczyk
Centre for Vision, Speech and Signal Processing, University of Surrey.

Given two sets of m random vectors $X \in \mathbb{R}^{d_x \times m}$ and $Y \in \mathbb{R}^{d_y \times m}$, let their covariances be Σ_{xx} and Σ_{yy} respectively, and let the cross covariance be Σ_{xy}. Canonical correlation analysis (CCA) seeks pairs of linear projections that maximise the correlation of the two views:

$$\begin{aligned}(\mathbf{w}_x^*, \mathbf{w}_y^*) &= \underset{\mathbf{w}_x, \mathbf{w}_y}{\operatorname{argmax}} \ \operatorname{corr}(\mathbf{w}_x^T X, \mathbf{w}_y^T Y) \\ &= \underset{\mathbf{w}_x, \mathbf{w}_y}{\operatorname{argmax}} \ \frac{\mathbf{w}_x^T \Sigma_{xy} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \Sigma_{xx} \mathbf{w}_x \mathbf{w}_y^T \Sigma_{yy} \mathbf{w}_y}} \end{aligned} \quad (1)$$

Define $T = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2}$, and let the singular value decomposition (SVD) of T be $T = UDV^T$. It is shown in [1] that the gradient of the total correlation with respect to X is given by:

$$\frac{\partial \operatorname{corr}(X,Y)}{\partial X} = \frac{1}{m-1}(2\nabla_{xx}\bar{X} + \nabla_{xy}\bar{Y}) \quad (2)$$

where

$$\nabla_{xx} = -\tfrac{1}{2}\Sigma_{xx}^{-1/2} U D U^T \Sigma_{xx}^{-1/2} \quad (3)$$

$$\nabla_{xy} = \Sigma_{xx}^{-1/2} U V^T \Sigma_{yy}^{-1/2} \quad (4)$$

and $\frac{\partial \operatorname{corr}(X,Y)}{\partial Y}$ has a symmetric form.

Using the insight that the the gradient of the correlation sought in CCA can be computed in Eq. (2), deep canonical correlation analysis (DCCA) [1] propagates the gradient along the two branches of a deep neural network, achieving end-to-end learning.

In [1] DCCA is applied to medium-sized problems where d_x and d_y are in the order of 10^1. Moreover, it is evaluated in terms of total correlation obtained in the learnt latent space, which is not the final goal of real-world applications. In this paper, we employ DCCA to learn a latent space for matching images and text, where the number of features required to encode the rich information is in the order of 10^3 [3, 4]. To this end, we propose to process image and text in the two branches of the network, and carefully address complexity and overfitting issues.

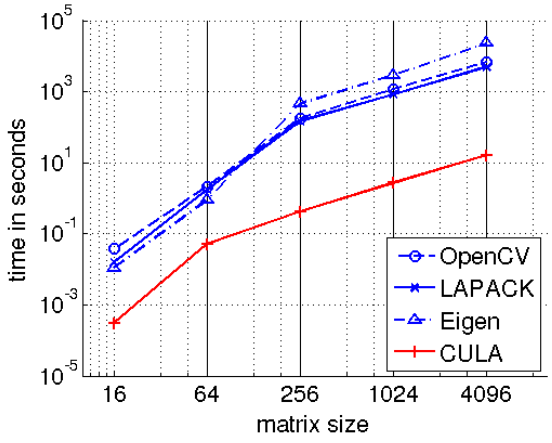

Figure 1: Log-log plot of speed of SVD solvers. The GPU based CULA solver is two to three orders of magnitude faster than CPU based ones when matrix is 4096×4096.

We implement the CCA loss layer on a GPU with the CUBLAS and CULA libraries. For comparison we also implement the layer using several CPU based linear algebra libraries. Figure 1 compares the time needed to solve an SVD with various libraries, where OpenCV, LAPACK, Eigen are CPU based and CULA is GPU based. When d is in the order of 10^3, CULA is typically two to three orders of magnitude faster. For example, when $d = 4096$, CULA takes only 16.3 seconds, while LAPACK, OpenCV and Eigen take 4922.6, 6714.5 and 22971.1 seconds, respectively.

Other linear operations such as matrix multiplication and the Cholesky decomposition required for matrix inversion also get a significant speedup with CULA and CUBLAS. Overall each iteration for a batch of size $m = 100$ takes approximately 26.5 seconds to complete. Since typically thousands of iterations are needed for the network to converge, it is clear that the migration from CPU to GPU is a crucial step for DCCA to be practically employed in our problems.

We evaluate the DCCA learning scheme on three image-text parallel datasets, namely Flickr8K, Flickr30K, and IAPR TC-12. We follow the common practice on these datasets and compute the average recall of the gold item at position 1, 5, 10 of the ranked list (R@1, R@5, R@10), and the median rank (MR) of the gold item, for both image annotation and image retrieval tasks. Results for the image annotation task on Flicker30K dataset are shown in Table 1.

		R@1	R@5	R@10	MR
Protocol I	transfer CCA [3]			**32.8** / 32.4	
	DCCA			32.5	
Protocol II	DeViSE [2]	4.5	18.1	29.2	26
	SDT-RNN [6]	9.6	29.8	41.1	16
	Deep Fragment [5]	16.4	**40.2**	**54.7**	**8**
	DCCA	**16.7**	39.3	52.9	**8**
Protocol III	DCCA	27.9	56.9	68.2	4

Table 1: Performance on Flickr30K: image annotation

The results in Table 1 indicate that under protocol I the transfer CCA in [3] achieves an R@10 score of 32.8 and 32.4 when using Flickr1M and SBU1M as additional training data respectively, where Flickr1M and SBU1M each contain 1 million image-caption pairs. Our method has an R@10 score of 32.5, which is on par with [3] but does not use additional data for training. On the other hand, when no extra data is used for training, the performance of the proposed learning scheme is comparable to that of [5], which is the state of the art on this dataset under protocol II.

Acknowledgement This work has been supported by EU Chist-Era EPSRC EP/K01904X/1 Visual Sense project.

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

[2] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[3] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014.

[4] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[5] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014.

[6] R. Socher, A. Karpathy, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. In *ACL*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.