Extracting Highly Effective Features for Supervised Learning via Simultaneous Tensor Factorization

Sunny Verma[†], **Wei Liu**[†], **Chen Wang**[‡], **and Liming Zhu**[‡]
[†]Advanced Analytics Institute, University of Technology Sydney, Sydney, Australia
[‡]Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, Australia
Sunny.Verma@student.uts.edu.au, Wei.Liu@uts.edu.au, Chen.Wang@data61.csiro.au, Liming.Zhu@data61.csiro.au

Abstract

Real world data is usually generated over multiple time periods associated with multiple labels, which can be represented as multiple labeled tensor sequences. These sequences are linked together, sharing some *common* features while exhibiting their own *unique* features. Conventional tensor factorization techniques are limited to extract either *common* or *unique* features, but not both simultaneously. However, both types of these features are important in many machine learning systems as they inherently affect the systems' performance. In this paper, we propose a novel supervised tensor factorization technique which simultaneously extracts ordered *common* and *unique* features. To test the effectiveness of the proposed method we utilize CIFAR-10 and ETH-80 databases. Images belonging to each category in these databases shares similar representations but vary in texture, scale, color, and rotation making it suitable to test our method. Classification results using features extracted by our method achieves significantly better performance over other factorization methods on these databases, illustrating the effectiveness of the proposed technique.

Introduction and Motivation

In the real world, data is often acquired as a sequence of matrices rather than a single matrix. These matrices can be represented as multiple labeled tensor (multidimensional arrays) sequences (Lahat, Adali, and Jutten 2015). Due to the underlying data generation mechanism, these sequences are naturally linked together and share some *common* features. While at the same time, they also exhibit their own *unique* features. When one is faced with the scenario of extracting features from these multiple labeled tensors sequences, two common approaches are followed: a) concatenate all tensor instances and factorize them together, or b) factorize each tensor instance individually. However, both these approaches suffer information loss. The former approach is limited to extract *common* features, suffering the loss of unique features. While, the latter approach is limited to extract *unique* features, suffering the loss of common features. This is because conventional tensor factorization techniques are unsupervised i.e., tensor instances are factorized without considering its label (category). Hence, one is only able to extract either *common* or *unique* features, but not both simultaneously from multiple labeled tensor sequences.

To overcome this limitation of information loss in conventional factorization techniques, we propose a novel supervised tensor factorization technique called **C**ommon and **U**nique **T**ensor **F**actorization (**CUTF**). The proposed technique is able to simultaneously extract *common* and *unique* features from multiple labeled tensor sequences. Furthermore, features extracted using **CUTF** are ordered by their singular value significance, enabling re-utilization of these features in various machine learning tasks like classification, clustering, recommender systems, data fusion etc.

Related Work and Tensor Notations

Tensors are higher order generalizations of matrices denoted in this paper by boldface Euler script letters $\mathcal{X}$,$\mathcal{Y}$,$\mathcal{Z}$. The number of modes also called order of tensor is equal to the number of dimensions of tensor $\mathcal{X}$. Similar to matrix rows and columns, fibers are defined by fixing all except one index for tensors. Matrix column is a mode-1 fiber, matrix row is a mode-2 fiber. Mode-3 fiber of tensor is called as tube fiber. An N mode tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \ldots \times I_N}$ can be unfolded as a matrix in mode-*n* by rearranging mode-*n* fibers as columns of matrix denoted in boldface capital letter $\mathbf{X}_{(n)}$. For a detailed review on tensors and related factorization literature, refer to Kolda and Bader.

Acar, Kolda, and Dunlavy were the first to propose extraction of *common* features shared among multiple data sources. In their work, the authors proposed joint factorization of a tensor with a matrix sharing *common* features on a single identical mode - coupled matrix and tensor factorization (**CMTF**) (Acar, Kolda, and Dunlavy 2011). However, in their work, the authors did not addressed these two issues: 1) how to extract *common* features from more than one identical mode, and more importantly 2) how to extract *unique* features from the same shared identical mode. These challenging issues are addressed by our method introduced in the next section.

Table1 compares the strengths and weaknesses between **CMTF** and **CUTF**.

Proposed CUTF

CUTF is developed using Higher-Order Orthogonal Iteration (*HOOI*) technique (Liu et al. 2015). *HOOI* is a

Table 1: Tensor Decomposition Techniques

Factorization Techniques	Properties		
	Unique	*Common*	*Ordered*
CMTF		✓	
CUTF (*Proposed*)	✓	✓	✓

generalization of matrix *SVD* technique and is developed for factorizing single tensors. It decomposes tensor $\mathcal{X} \approx [\![\mathcal{G}; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, ..., \mathbf{A}^{(N)}]\!]$, mathematically $\mathcal{G} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} ... \times_N \mathbf{A}^{(N)}$ representing sequential multiplication of a tensor with a matrix in i^{th}-mode ($1 \leq i \leq N$). Here, $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 ... \times R_N}$ can be thought as compressed version of $\mathcal{X}$ and $\mathbf{A}^{(i)s} \in \mathbb{R}^{I_i \times R_i}$ represents low-rank factor matrices of i^{th}-mode in tensor $\mathcal{X}$. *HOOI* guarantees best rank-$(R_1, R_2, ..., R_N)$ approximation of tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 ... \times I_N}$ (Liu et al. 2015).

Without loss of generality, we focus on two 3-mode tensors $\mathcal{X}$ and $\mathcal{Y} \in$ Class $[+1, -1]$. Assume that the tensors share common features only in their first mode. Denote $\mathbf{W}$ as the *common* features shared among these two tensors and, denote $\mathbf{V}$ and $\mathbf{S}$ as the remaining unique features of tensors in the same mode and denote $\mathbf{U}^i$ and $\mathbf{K}^i$ as the factors of other modes of $\mathcal{X}$ and $\mathcal{Y}$. Simply, $(\mathbf{W}|\mathbf{V} \simeq U^1)$ and $(\mathbf{W}|\mathbf{S} \simeq K^1)$ represents factor matrices of the tensors $\mathcal{X}$ and $\mathcal{Y}$ in their first mode respectively. Our objective is to jointly factorize $\mathcal{X}$ and $\mathcal{Y}$ to obtain their low rank approximations, simultaneously extracting their common $(\mathbf{W})$ and unique $(\mathbf{V}, \mathbf{S})$ features from their first mode: Fig-1 illustrates comparison between conventional factorization technique and proposed **CUTF**. The complete procedure of solving our objective function is presented in Algorithm 1.

$$obj = \min \left[\, \| \, \mathcal{X} - [\![\mathcal{G}_{\mathcal{X}}; (\mathbf{W}|\mathbf{V}), \mathbf{U}^{(2)}, \mathbf{U}^{(3)}]\!] \, \|_F + \| \, \mathcal{Y} - [\![\mathcal{G}_{\mathcal{Y}}; (\mathbf{W}|\mathbf{S}), \mathbf{K}^{(2)}, \mathbf{K}^{(3)}]\!] \, \|_F \, \right].$$

Experiments and Analysis

In this section we evaluate the performance of proposed **CUTF** on two different datasets: 1) CIFAR-10 (Krizhevsky and Hinton 2009) and 2) ETH-80 (Leibe and Schiele 2003). Three different feature sets: 1) *common* (*Com*), 2) *unique* (*Unq*) and, 3) both *common* and *unique* (e.g., **CUTF**) are extracted from each of the dataset (note that the *Com* is the same as the **CMTF** and *Unq* is the same as *HOOI)*. These features are classified using Logistic Regression (LR), SVM with polynomial kernel (SVM-Poly), SVM without Kernel (SVM), and K-nearest neighbor (KNN).

CUTF is implemented using Matlab tensor toolbox (Bader, Kolda, and others 2015) and the classifiers are Matlab R2016a inbuilt subroutines. To validate the superiority of **CUTF**, Friedman tests (Demšar 2006) were also performed on the classification results. Friedman test is a non-parametric statical method to validate the claim of significance of multiple classifiers on multiple datasets. The hypothesis "there is no significant difference between the two factorization methods" is rejected if p-values obtained through Friedman test are lower than 0.05.

Algorithm 1 **C**ommon and **U**nique **T**ensor **F**actorization

In $(\mathcal{X}, \mathcal{Y}, \mathbf{R_1}, \mathbf{R_2}, \mathbf{R_3}, \textit{MaxIter})$
Out $[\mathcal{G}_{\mathcal{X}}, (\mathbf{W}|\mathbf{V}), \mathbf{U}^{(2)}, \mathbf{U}^{(3)}, \mathcal{G}_{\mathcal{Y}}, (\mathbf{W}|\mathbf{S}), \mathbf{K}^{(2)}, \mathbf{K}^{(3)}]$
$\mathbf{U}^{(i)} \leftarrow \mathbf{R}_i$ left singular vectors of $\mathbf{X}_{(i)}$ $i = 2, 3$
$\mathbf{K}^{(i)} \leftarrow \mathbf{R}_i$ left singular vectors of $\mathbf{Y}_{(i)}$ $i = 2, 3$
$\mathbf{W} \leftarrow \lceil \mathbf{R}_1/2 \rceil$ left singular vectors of $[\mathbf{X}_{(1)} \mathbf{Y}_{(1)}]$
$\mathbf{V} \leftarrow \lfloor \mathbf{R}_1/2 \rfloor$ left singular vectors of $\mathbf{X}_{(1)}$
$\mathbf{S} \leftarrow \lfloor \mathbf{R}_1/2 \rfloor$ left singular vectors of $\mathbf{Y}_{(1)}$
$\mathcal{G}_{\mathcal{X}} \leftarrow [\![\mathcal{X}; (\mathbf{W}|\mathbf{V})^T, (\mathbf{U}^{(2)})^T, (\mathbf{U}^{(3)})^T]\!]$
$\mathcal{G}_{\mathcal{Y}} \leftarrow [\![\mathcal{Y}; (\mathbf{W}|\mathbf{S})^T, (\mathbf{K}^{(2)})^T, (\mathbf{K}^{(3)})^T]\!]$
while *obj* converges or *MaxIter* exhausted **do**
  for $i = 2, 3$ **do**
    $\mathcal{M} \leftarrow [\![\mathcal{G}_{\mathcal{X}}; (\mathbf{W}|\mathbf{V})^T, (\mathbf{U}^{(j)})^T]\!]$
    $\mathbf{U}^{(j)} \leftarrow \mathbf{R}_j$ left singular vectors of $\mathbf{M}_{(j)}$
    $\mathcal{N} \leftarrow [\![\mathcal{G}_{\mathcal{Y}}; (\mathbf{W}|\mathbf{S})^T, (\mathbf{K}^{(j)})^T]\!]$
    $\mathbf{K}^{(j)} \leftarrow \mathbf{R}_j$ left singular vectors of $\mathbf{N}_{(j)}$
    where $j \in [2, 3]$ & $j \neq i$
  end for
  $\mathcal{M} \leftarrow [\![\mathcal{G}_{\mathcal{X}}; (\mathbf{U}^{(2)})^T, (\mathbf{U}^{(3)})^T]\!]$
  $\mathcal{N} \leftarrow [\![\mathcal{G}_{\mathcal{Y}}; (\mathbf{K}^{(2)})^T, (\mathbf{K}^{(3)})^T]\!]$
  $\mathbf{W} \leftarrow \lceil \mathbf{R}_1/2 \rceil$ left singular vectors of $[\mathbf{M}_{(1)} \; \mathbf{N}_{(1)}]$
  $\mathbf{V} \leftarrow \lfloor \mathbf{R}_1/2 \rfloor$ left singular vectors of $\mathbf{M}_{(1)}$
  $\mathbf{S} \leftarrow \lfloor \mathbf{R}_1/2 \rfloor$ left singular vectors of $\mathbf{N}_{(1)}$
  $\mathcal{G}_{\mathcal{X}} \leftarrow [\![\mathcal{X}; (\mathbf{W}|\mathbf{V})^T, (\mathbf{U}^{(2)})^T, (\mathbf{U}^{(3)})^T]\!]$
  $\mathcal{G}_{\mathcal{Y}} \leftarrow [\![\mathcal{Y}; (\mathbf{W}|\mathbf{S})^T, (\mathbf{K}^{(2)})^T, (\mathbf{K}^{(3)})^T]\!]$
end while

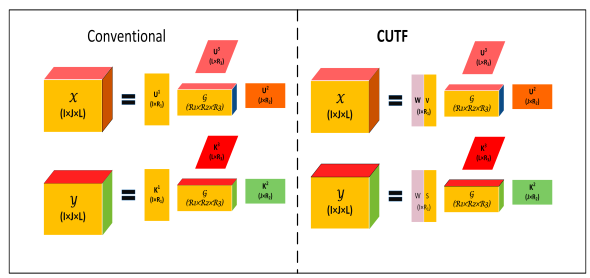

Figure 1: Conventional factorization method vs **CUTF**

CIFAR-10

CIFAR-10 dataset consists of 60,000 RGB images of 32×32 pixels equally divided among 10 categories. The first 5000 images of each category are for testing while the last 1000 images are for testing. Images were pre-processed using global contrast normalization and ZCA whitening (Goodfellow et al. 2013). For each class label, we build a tensor of 4 modes: $RowPixels \times ColumnPixels \times color \times position$. We randomly chose multiple pairs of binary categories from the database and extracted feature using three different methods sets: 1) *Com* , 2) *Unq* and, 3) **CUTF**.

Classification results using LR and SVM-Poly and reported in Table 2, while results obtained using SVM and KNN are reported in Table 3. Also on the bottom of the tables, p-values obtained through Friedman test are reported. These low p-values illustrate the statistical significance of our technique. Moreover, Fig-2 compares the accuracies obtained through SVM-Poly on different factorization ranks, which demonstrates the advantages of the proposed method.

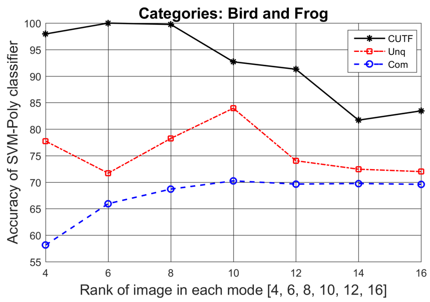

Figure 2: Accuracy comparison by factorization ranks

Table 2: Comparisons of different factorization methods

Categories	Accuracy from LR			Accuracy from SVM-Poly		
	Com	*Unq*	**CUTF**	*Com*	*Unq*	**CUTF**
Cat - Dog	.571	.625	**.648**	.535	.787	**.997**
Cat - Bird	.609	.612	**.631**	.539	.722	**.994**
Bird - Frog	.672	.671	**.682**	.687	.782	**.997**
Dog - Frog	.669	.668	**.687**	.675	.858	**.996**
Ship - Truck	.715	.870	**.879**	.753	.950	**.987**
Horse - Dog	.649	.670	**.693**	.704	.808	**.999**
Dog - Deer	.583	.649	**.686**	.655	.841	**.999**
Bird - Horse	.629	**.699**	.679	.669	.760	**.993**
Deer - Horse	.637	.682	**.707**	.678	.843	**.995**
Plane - Ship	.682	.825	**.843**	.695	.801	**.992**
Mobile - Ship	.719	.851	**.864**	.760	.917	**.990**
Mobile - Plane	.737	.777	**.825**	.771	.926	**.985**
Truck - Mobile	.661	.854	**.899**	.685	.945	**.993**
Friedman tests	.0003	.0023	Base	.0003	.0003	Base

ETH-80

ETH-80 dataset consists of $3,280$ RGB images of 128×128 pixels grouped among 8 categories. Each category is further subdivided into 10 objects (similar objects are grouped into a single category) with 41 views each object, equally spaced over the viewing hemisphere. Objects within the same category shares *common* features and also have their own unique features, making this database suitable to evaluate the performance of proposed **CUTF**. Images were converted into grayscale and 5 modes: $RowPixels \times ColumnPixels \times view \times object \times categoryt$ tensor was built. Features were extracted using the three methods : $1)$ Com , $2)$ Unq and, $3)$ **CUTF** and, extracted features were classified using classifiers- LR, SVM-Poly, SVM, and KNN.

The objects within each category were classified using two different protocols. Protocol: $a)$ 20% training data and 80% testing data and $b)$ 40% training data and 60% testing data. Training and testing data was randomly selected for each category and mean of 10 fold classification accuracies are reported in Table 4.

Conclusion and Future Work

In this research, we have proposed a novel supervised tensor factorization technique, which simultaneously extracts *common* and *unique* features. These features are ordered by their singular value significance with respect to multiple labeled tensor sequences. Experiments reported in this paper demonstrate huge potential of simultaneously extracting *common* and *unique* features. Our future work includes extending the proposed **CUTF** for sparse tensor factorizations.

Table 3: Comparisons of different factorization methods

Categories	Accuracy from KNN			Accuracy from SVM		
	Com	*Unq*	**CUTF**	*Com*	*Unq*	**CUTF**
Cat - Dog	.572	.764	**.999**	.574	.637	**.756**
Cat - Bird	.605	.636	**.996**	.610	.614	**.749**
Bird - Frog	.694	.692	**.922**	.675	.694	**.789**
Dog - Frog	.633	.840	**.993**	.665	.670	**.783**
Ship - Truck	.722	.928	**.982**	.724	.876	**.92**
Horse - Dog	.681	.837	**.999**	.662	.683	**.774**
Dog - Deer	.637	.841	**.978**	.581	.648	**.766**
Bird - Horse	.664	.834	**.990**	.625	.702	**.776**
Deer - Horse	.691	.908	**.995**	.643	.711	**.722**
Plane - Ship	.682	.832	**.993**	.684	.865	**.888**
Mobile - Ship	.733	.924	**.997**	.72	.746	**.876**
Mobile - Plane	.762	.909	**.974**	.74	.798	**.893**
Truck - Mobile	.656	.944	**.999**	.663	.844	**.927**
Friedman tests	.0003	.0003	Base	.0003	.0003	Base

Table 4: 10-fold Classification on ETH-80 Dataset

Classifiers	20 % Training Data			40 % Training Data		
	Com	*unq*	**CUTF**	*Com*	*Unq*	**CUTF**
SVM-Poly	.670	.665	**.862**	.741	.808	**.917**
SVM	.630	.619	**.845**	.683	.706	**.901**
LR	.381	**.484**	.440	**.562**	.426	.511
KNN	.678	.686	**.877**	.744	.801	**.927**

References

Acar, E.; Kolda, T. G.; and Dunlavy, D. M. 2011. All-at-once optimization for coupled matrix and tensor factorizations. *arXiv preprint arXiv:1105.3422*.

Bader, B. W.; Kolda, T. G.; et al. 2015. Matlab tensor toolbox version 2.6. Online, http://www.sandia.gov/ tgkolda/TensorToolbox/.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* 7(Jan):1–30.

Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A. C.; and Bengio, Y. 2013. Maxout networks. *ICML (3)* 28:1319–1327.

Kolda, T. G., and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM review* 51(3):455–500.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*.

Lahat, D.; Adali, T.; and Jutten, C. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE* 103(9):1449–1477.

Leibe, B., and Schiele, B. 2003. Analyzing appearance and contour based methods for object categorization. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, II–409. IEEE.

Liu, Y.; Shang, F.; Fan, W.; Cheng, J.; and Cheng, H. 2015. Generalized higher-order orthogonal iteration for tensor decomposition and completion. In *Advances in Neural Information Processing Systems (NIPS)*, 1763–1771.