HyperLearn: A Distributed Approach for Representation Learning in Datasets With Many Modalities

Arya, D.; Rudinac, S.; Worring, M.

DOI
10.1145/3343031.3350572

Publication date
2019

Document Version
Final published version

Published in
MM'19

HyperLearn: A Distributed Approach for Representation Learning in Datasets With Many Modalities

Devanshu Arya
University of Amsterdam
Amsterdam, The Netherlands
d.arya@uva.nl

Stevan Rudinac
University of Amsterdam
Amsterdam, The Netherlands
s.rudinac@uva.nl

Marcel Worring
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

ABSTRACT

Multimodal datasets contain an enormous amount of relational information, which grows exponentially with the introduction of new modalities. Learning representations in such a scenario is inherently complex due to the presence of multiple heterogeneous information channels. These channels can encode both (a) inter-relations between the items of different modalities and (b) intra-relations between the items of the same modality. Encoding multimedia items into a continuous low-dimensional semantic space such that both types of relations are captured and preserved is extremely challenging, especially if the goal is a unified end-to-end learning framework. The two key challenges that need to be addressed are: 1) the framework must be able to merge complex intra and inter relations without losing any valuable information and 2) the learning model should be invariant to the addition of new and potentially very different modalities. In this paper, we propose a flexible framework which can scale to data streams from many modalities. To that end we introduce a hypergraph-based model for data representation and deploy Graph Convolutional Networks to fuse relational information within and across modalities. Our approach provides an efficient solution for distributing otherwise extremely computationally expensive or even unfeasible training processes across multiple-GPUs, without any sacrifices in accuracy. Moreover, adding new modalities to our model requires only an additional GPU unit keeping the computational time unchanged, which brings representation learning to truly multimodal datasets. We demonstrate the feasibility of our approach in the experiments on multimedia datasets featuring second, third and fourth order relations.

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; *Distributed artificial intelligence*; *Multi-task learning*.

KEYWORDS

Multimodal Representation Learning; Hypergraph; Tensor Factorization; Geometric Deep Learning; Highly Multimodal Datasets

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350572

ACM Reference Format:
Devanshu Arya, Stevan Rudinac, and Marcel Worring. 2019. HyperLearn: A Distributed Approach for Representation Learning in Datasets With Many Modalities. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350572

1 INTRODUCTION

The field of multimedia has been slowly, but steadily growing beyond simple combining of diverse modalities, such as text, audio and video, to modeling their complex relations and interactions. These relations are commonly perceived, and therefore, modelled as only pair-wise connections between two items, which is a major drawback in the majority of the existing techniques. Going beyond pair-wise connections to encode higher-order relations can not only discover complex inter-dependencies between items but also help in removing ambiguous relations. For instance: in the task of social image-tag refinement, conventional approaches focus on exploiting the pairwise tag-image relations, without considering the user information which has been proven extremely useful in resolving tag ambiguities and closing the semantic gap between visual representation and semantic meaning [8, 23, 43, 44]. It is hence an interesting, but far more challenging problem in multimedia to exploit and learn higher-order relations to be able to (a) learn a better representation for each item, (b) improve pairwise retrieval tasks and (c) discover far more complex relations which can be ternary (3^{rd} order), quaternary (4^{th} order), quinary (5^{th} order) or even beyond. As examples, figure 1 shows the importance of modeling higher-order relations in social networks and in artistic analysis respectively. In the upper example from Figure 1, textual annotations and information about user demographics is utilized for disambiguation between landmarks with very similar visual appearance. Similarly, the second example illustrates quaternary relations formed by the artworks, media, artists and the time-frame in which they were active. Capturing such complex relations is of utmost importance in a number of tasks performed by the domain experts, such as author attribution, influence and appreciation analysis.

Learning representations in multimodal datasets is an extremely complex task due to the enormous amount of relational information available. At the same time most of these relations have an innate property of 'homophily', which is the fact that similarity breeds connections. Exploiting this property of similarities can help in immensely simplifying the understanding of these relations. These similarities can be derived from both intra relations between items of the same modality and inter-relations between items across

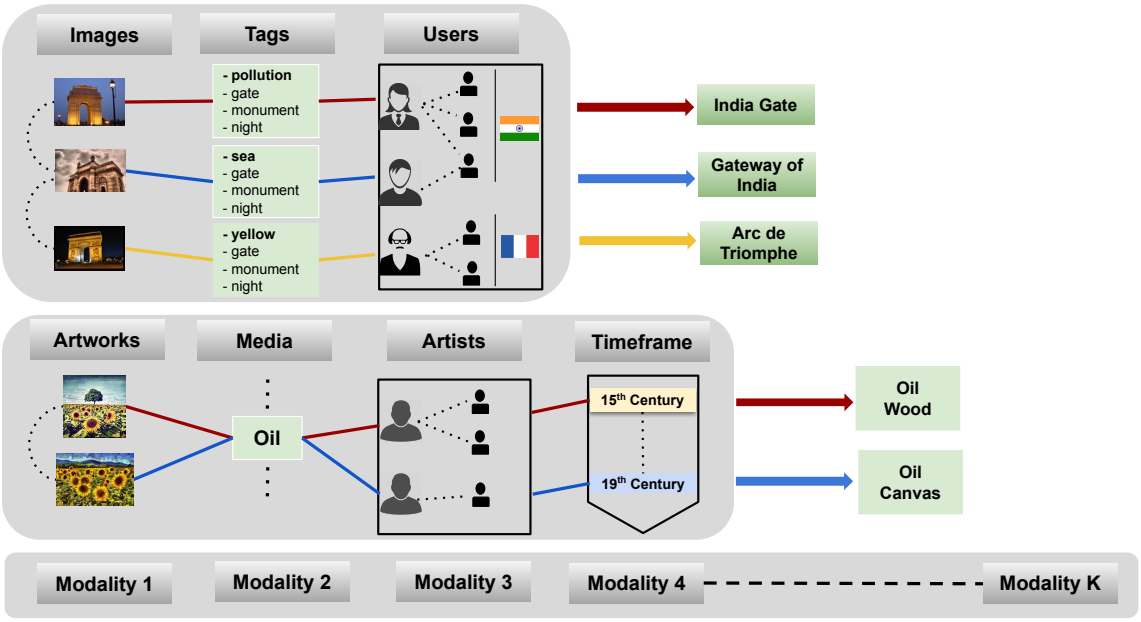


Figure 1: Example showing importance of capturing ternary relations (images-tags-users) in a social network dataset and quaternary relations (artworks-media-artists-timeframe) in artistic dataset. HyperLearn exploits such relations to learn complex representations for each modality. At the same time, HyperLearn provides a distributed learning approach, which makes it scalable to datasets with many modalities

different modalities. Unifying the two types of relations in a complementary manner has the potential to bolster the performance of practically any multimedia task. Thus, in this work we propose an efficient learning framework that can merge information generated by both intra as well as inter-relations in datasets with many modalities. We conjecture that such an approach can pave the way for a generic methodology for learning representations by exploiting higher-order relations. At the same time, we introduce an approach that makes our framework scale to multiple modalities.

We focus on learning a low-dimensional representation for each multimodal item using an unsupervised framework. The unsupervised methods utilize relational information both within as well as across modalities to learn common representations for a given multimodal dataset. The co-occurrence information simply means that two items from different modalities are semantically similar if they co-exist in a multimedia collection. For example, the textual description of a video often describes the events shown in the visual channel. Many of the multimedia tasks revolve around this compact latent representation of each multimodal entity [31, 32]. The major challenge lies in bridging the learning gap between the two types of relations in a way that they can be semantically complementary in describing similar concepts. Learning representation is usually extremely expensive, both in computational time and required storage as even a relatively small multimedia collection normally contains a multitude of complex relations.

Handling a large amount of relations requires a framework with a flexible approach to training across multiple pipelines. Most of the existing algorithms fail in parallelizing their framework into separate pipelines [24, 50], resulting in large time and memory consumption. Thus, in the proposed framework we can parallelize the training process for different modalities into separate pipelines, each requiring just an additional GPU core. By doing so, we facilitate joint multimodal representation learning on highly heterogeneous multimedia collections containing an arbitrarily large number of modalities, effectively hitting an elusive target sought after since the early days of multimedia research. The points below highlight the contributions of this paper:

- We address the challenging problem of multimodal representation learning by proposing HyperLearn, an unsupervised framework capable of jointly modeling relations between the items of the same modality, as well as across different modalities.
- Based on the concept of geometric deep learning on hypergraphs, our HyperLearn framework is effective in extracting higher-order relations in multimodal datasets.
- In order to reduce prohibitively high computational costs associated with multimodal representation learning, in this work we propose a distributed learning approach, which can be parallelized across multiple GPUs without harming the accuracy. Moreover, introducing a new modality into HyperLearn framework requires only an additional GPU, which makes it scalable to datasets with many modalities.
- Extensive experimentation shows that our approach is task-independent, with a potential for deployment in a variety of applications and multimedia collections.

2 RELATED WORK

The core challenge in multimodal learning revolves around learning representations that can process and relate information from multiple heterogeneous modalities. Most of existing multimodal representation learning methods can be split into two broad categories – multimodal network embeddings and tensor factorization-based latent representation learning. In this section we reflect on the representative approaches from these two categories. Since, in this work we extend the notion of graph convolution networks for multimodal datasets, we also touch upon some of the existing techniques that aim to deploy deep learning on graphs.

2.1 Multimodal Network Embedding

A common strategy for representation learning is to project different modalities together into a joint feature space. Traditional methods [27, 33, 45] focus on generating node embeddings by constructing an affinity graph on the nodes and then finding the leading eigenvectors for representing each node. With the advent of deep learning, neural networks have become a popular way to construct combined representations. They owe their popularity to the ability to jointly learn high-quality representations in an end-to-end manner. For example, Srivastava and Salakhutdinov proposed an approach for learning higher-level representation for multiple modalities, such as images and texts using Deep Boltzmann Machines (DBM) [38]. Since then a large number of multimodal representation learning methods based on deep learning have been proposed. Some of these methods attempt to learn a multimodal network embedding by combining the content and link information [7, 12, 20, 24, 42, 50, 51]. Other set of methods focuses on modeling the correlation between multiple modalities to learn a shared representation of multimedia items. An example of such coordinated representation is Deep Canonical Correlation Analysis (DCCA) that aims to find a non-linear mapping that maximizes the correlation between the mapped vectors from the two modalities [49]. Ambiguities often occur while using network embedding methods to learn multimodal relations due to sub-optimal usage of available information. This is mostly because these methods assume relations between items to be pairwise which often leads to loss of information [1, 6, 19].

2.2 Tensor Factorization Based Latent Representation Learning

Decoupling a multidimensional tensor into its factor matrices has been proven successful in unraveling latent representations of their components in an unsupervised manner [17, 18, 30]. Most existing approaches aim to embed both entities and relations into a low-dimensional space for tasks such as link prediction [46], reasoning in knowledge bases [37] or multi-label classification problems [25]. Recent methods on social image understanding incorporate user information as the third modality for tag based image retrieval and image-tag refinement problems [41, 43, 44]. Even though most of these approaches are suitable for large datasets, one of the main disadvantages of using a factorization based model is the lack of flexibility when scaling to highly multidimensional datasets. Additionally, most of the tensor decomposition methods are based on the optimization with a least squared criterion, which severely lacks robustness to outliers [15].

In this work, we first overcome the issues of network embedding methods by using a hypergraph-based learning method. Secondly, we introduce a scalable approach to tensor decomposition for scaling representation learning to many modalities. Finally, we can combine the advantages of rich information from network structure with the unsupervised nature of tensor decomposition in one single end-to-end framework.

2.3 Geometric Deep Learning on graphs

Geometric deep learning [4] brings the algorithms that can help learn from non-euclidean data like graphs and 3D objects by proposing an ordering of mathematical operators that is different from common convolutional networks. The aim of Geometric Deep Learning is to process signals defined on the vertices of an undirected graph $\mathbb{G}(V, E, W)$, where V is the set of vertices, E is set of edges, and $W \in \mathbb{R}^{|V|\times|V|}$ is the adjacency matrix. Following [9, 36], spectral domain convolution of signals x and y defined on the vertices of a graph is formulated as:

$$x \circledast y = \Phi(\Phi^T x).(\Phi^T y) = \Phi(\mathcal{F}(x).\mathcal{F}(y)) \tag{1}$$

Here, $\Phi^T x$ corresponds to Graph Fourier Transform and $\mathcal{F}(.)$ represents Fourier Transform; the eigen functions Φ of the graph laplacian play the role of Fourier modes; the corresponding eigenvalues Λ of the graph laplacian are identified as the frequencies of the graph. Recent applications of graph convolutional networks range from computer graphics [3] to chemistry [10]. The spectral graph convolutional neural networks (*GCN*), originally proposed in [5] and extended in [9] were proven effective in classification of handwritten digits and news texts. A simplification of the *GCN* formulation was proposed in [16] for semi-supervised classification of nodes in a graph. In the computer vision community, *GCN* has been extended to describe shapes in different human poses [26], perform action detection in videos [47] and for image and 3D shape analysis [29]. However, in the multimedia field there have been considerably less examples of using deep learning on graphs for modeling highly multimodal datasets with [1, 34] as notable exceptions.

In this paper, we propose an approach that introduces the application of graph convolutional networks on multimodal datasets. We deploy Multi-Graph Convolution Network (MGCNN) originally proposed by [29] for the matrix completion task using row and column graphs as auxiliary information. It aims at extracting spatial features from a matrix by using information from both the row and column graphs. For a matrix $X \in \mathbb{R}^{N_1\times N_2}$, MGCNN is given by

$$\widetilde{X} = \sum_{j,j'=0}^{q} \theta_{jj'} T_j(\mathbb{L}_r) X T_{j'}(\mathbb{L}_c) \tag{2}$$

where, $\Theta = \theta_{jj'}$ is $(q+1)\times(q+1)$ dimensional matrix which represents the coefficients of the filters, $T_j(.)$ denotes the Chebyshev polynomial of degree j and $\mathbb{L}_r, \mathbb{L}_c$ are the row and column Graph Laplacians respectively. Using Equation 2 as the convolutional layer of MGCNN, it produces q output channels ($N_1 \times N_2 \times q$) for matrix $X \in \mathbb{R}^{N_1\times N_2}$ with a single input channel. In this way, one can extract q dimensional features for each item in matrix X by combining

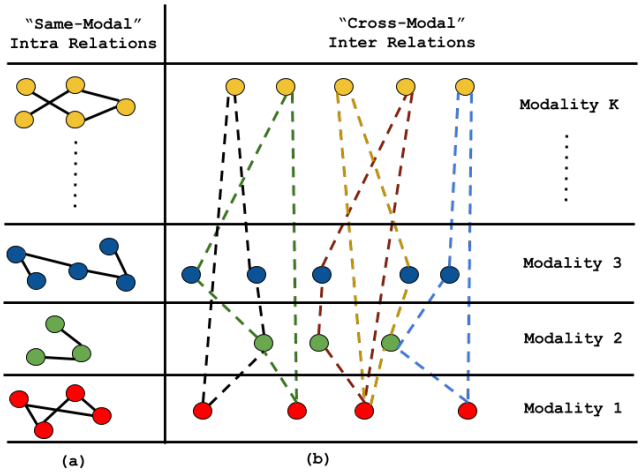


Figure 2: (a) Pair-wise relationship among the items of the same modality in K-modal data. (b) Complex higher-order heterogeneous relationships between entities of different modalities using a Hypergraph representation.

information from row and column graphs, which can correspond to e.g. individual modalities.

3 THE PROPOSED FRAMEWORK

In this section, we propose a novel distributed learning framework that can simultaneously exploit both intra and inter-relations in multimodal datasets. We depict these inter-relations on a hypergraph and conjecture that this way of representing higher-order relations reduces any loss of information contained within the multimodal network structure [1, 19, 52]. Mathematically, a hypergraph is depicted by its adjacency tensor [2]. A simple tensor factorization on this adjacency tensor can disentangle modalities into their compact representations. However, this kind of representation lacks information from the intra-relation of items belonging to the same modality. Subsequently, we therefore incorporate intra-relations among entities as auxiliary information to facilitate flow of within-modal relationship information.

3.1 Notations

We use boldface underlined letters, such as $\underline{\boldsymbol{X}}$, to denote tensors and simple upper case letters, such as U, to denote matrices. Let $\odot$ represent the "Khatri-Rao" product [14] defined as

$$U \odot V = (U_{ij} \otimes V_{ij})_{ij} \tag{3}$$

where, $U \in \mathbb{R}^{L \times R}$ and $V \in \mathbb{R}^{M \times R}$ are arbitrary matrices and $\otimes$ is the Kronecker Product. The resulting matrix $U \odot V$ is an expanded matrix of dimension $LM \times R$ on the columns of U and V.

3.2 Representing Cross-Modal Inter-Relations using Hypergraphs

Hypergraphs have been proven extremely efficient in depicting higher-order and heterogeneous relations. A hypergraph is the most efficient way to represent complex relationships between a multitude of diverse entities, as it minimizes any loss of available information [1, 6, 48]. Given multimodal data, we construct a unified hypergraph $\mathcal{H}(V, E)$ by building hyperedges (E) around each of the individual multimodal items which are represented on a set of nodes (V). These hyperedges correspond to the cross-modal relations between items of different modalities as illustrated in Figure 2.

A more formalised mathematical interpretation of this unified hypergraph is given by its adjacency tensor $\underline{\boldsymbol{X}}$, where the number of components of the tensor is equal to the number of modalities in the hypergraph. Further, each hyperedge corresponds to an entry in the tensor whose value are the weights of the hyperedge. For simplicity, in this work we focus on unweighted hypergraphs.

Thus, a multimodal data with K modalities is depicted on a tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{N_1 \times N_2 \ldots . \times N_K}$, where each component N_θ $(1 \leq \theta \leq K)$ of this tensor represents one of the K heterogeneous modalities. A single element $\underline{\boldsymbol{x}}$ of $\underline{\boldsymbol{X}}$ is addressed by providing its precise position by a series of indices $n_1, n_2, .., n_K$ i.e.

$$\underline{\boldsymbol{x}}_{n_1 n_2 .. n_K} \equiv \underline{\boldsymbol{X}}_{n_1 n_2 .. n_K}; \quad 1 \leq n_1 \leq |N_1|, ..., 1 \leq n_K \leq |N_K| \tag{4}$$

Further, a hyperedge around a set of nodes can be represented as binary values such that $\underline{\boldsymbol{x}}_{n_1 n_2 .. n_K} = 1$ if the relation $(n_1, n_2, ..., n_K)$ is known i.e. if there exists a mutual relation between the K modalities for that instance. For example, in the social network use case, with a possible corresponding image-tag-user associated tensor $\underline{\boldsymbol{X}} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$, the images (n_1) are represented on rows, users (n_2) on columns and tags (n_3) on tubes. If the l^{th} image uploaded by the m^{th} user is annotated with the n^{th} tag, then $\underline{\boldsymbol{x}}_{lmn} = 1$ and 0 otherwise.

3.3 Representing Intra-Relations Between the Items of the Same Modality

Relationships between items of the same modality are dependent on the nature/properties of the modality. For instance, relationships between users in a social network is defined based on their common interests. To make our framework flexible, each modality $(\theta; 1 \leq \theta \leq K)$ is represented on a separate graph $\boldsymbol{G_\theta}$ whose connections can be defined independently. For example: relations among images can be established based on their visual features, for tags it can be calculated based on their co-occurrence and for users it can very well be based on their mutual likes/dislikes. We denote the adjacency matrix of $\boldsymbol{G_\theta}$ by Λ_θ where each of its entries $\Lambda_\theta^{i,j} = 1$, if there exists a relation between the i^{th} and j^{th} element and 0 otherwise. The corresponding normalized graph laplacians $(\mathbb{L}_\theta)$ are given by

$$\mathbb{L}_\theta = D_\theta^{\frac{1}{2}} \Lambda_\theta D_\theta^{-\frac{1}{2}} \tag{5}$$

where, $D_\theta = diag(\sum_{j \neq i} \Lambda_\theta^{i,j})$ is known as the degree matrix.

3.4 Combined Inter-Intra Relational Feature Extraction

Tensor $\underline{\boldsymbol{X}}$ can be factorized using Candecomp/Parafac(CP) - decomposition [11] which decomposes a tensor into a sum of outer products of vectors $(a_r^{(\theta)})$.

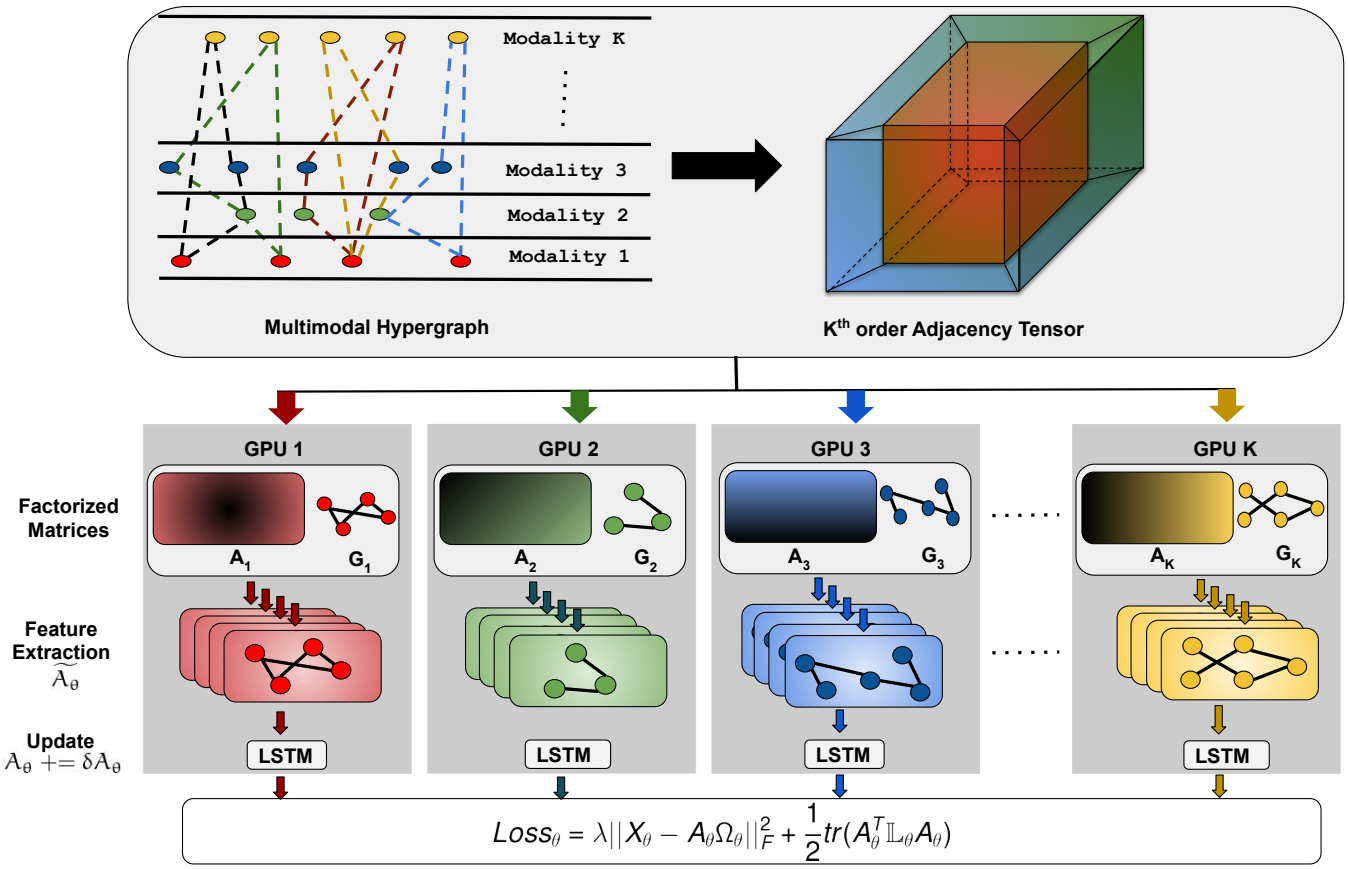


Figure 3: Proposed HyperLearn framework deployed on K modalities with a distributed learning approach

The CP-decomposition $\widetilde{\underline{X}}$ of $\underline{X}$ is defined as

$$\widetilde{\underline{X}} = \sum_{r=1}^{R} a_r^{(1)} \circ a_r^{(2)} \circ \ldots \circ a_r^{(K)} \tag{6}$$

$$= \boldsymbol{I} \times_1 A_1 \times_2 A_2 \times_3 \ldots \times_K A_K \tag{7}$$

where $\circ$ is the outer product and $\times_i$ represents mode-i multiplication (Tensor matrix product). Matrices $A_\theta \in \mathbb{R}^{|N_\theta| \times R}$ are called factor matrices of rank R and $\boldsymbol{I}$ is an R^{th} order identity tensor. Matrices A_θ are essentially the latent lower dimensional representations for each of the N_θ components of the tensor and therefore, for each of the K modalities.

Subsequently, we introduce an approach that can learn robust representations A_θ by combining intra relational information. We extract spatial features that merge information from each of the graphs $\boldsymbol{G_\theta}$ with the latent representation matrices A_θ using Multi-Graph Convolutional Network (MGCNN) layers given by

$$\widetilde{A_\theta} = \sum_{j,j'=0}^{q} \theta_{jj'} T_j(\mathbb{L}_\theta) A_\theta \tag{8}$$

where, the output $\widetilde{A_\theta} \in \mathbb{R}^{|N_\theta| \times R \times q}$ has q output channels. Similar to [29], we use an *LSTM* to implement the feature diffusion process which essentially iteratively predicts accurate changes δA_θ for the matrix A_θ. Due to its ability to keep long-term internal states, this *LSTM* architecture is highly efficient in learning complex non-linear diffusion processes.

3.5 Loss Function Incorporating Cross-Modality Inter-Relations and Within-Modality Intra Relations

In standard CP decomposition of a tensor, its factor matrices are approximated by finding a solution to the following equation

$$\min_{A_1,..,A_K} ||\underline{X} - (\boldsymbol{I} \times_1 A_1 \times_2 A_2 \times_3 \ldots \times_K A_K)||_F^2 \tag{9}$$

This equation essentially tries to find low dimensional factor matrices A_θ such that their combination is as close as possible to the original tensor $\underline{X}$. Further, to add relational information among items within each of these A_θ, we extend the "within-mode" regularization term introduced in [21] for matrices and [30] for third order tensors to generic K^{th} order tensors. The basic idea is to add a regularization term to Equation 9 such that it can force two similar objects in each modality to have similar factors, so that they operate similarly. Thus, the combined loss function is given

Table 1: Table showing the total number of intra and inter-relations between items on MovieLens, MIR Flickr and OmniArt datasets.

Movie Lens	$\mathscr{R}$(U) 12,594	$\mathscr{R}$(M) 28,928	$\mathscr{R}$(U-M) 100,000		
MIR Flickr	$\mathscr{R}$(I) 93,695,167	$\mathscr{R}$(T) 25,170	$\mathscr{R}$(U) 9,900,716	$\mathscr{R}$(I-T-U) 48,760	
Omni Art	$\mathscr{R}$(I) 4,628,009	$\mathscr{R}$(A) 849,482	$\mathscr{R}$(M) 21,178	$\mathscr{R}(T_f)$ 144	$\mathscr{R}$(I-A-M-T_f) 28,399

by:

$$\min_{A_1,..,A_K} \frac{1}{2}(tr \sum_{\theta=1}^{K} A_\theta^T \mathbb{L}_\theta A_\theta) + \lambda ||\underline{X} - (\boldsymbol{I} \times_1 A_1 \times_2 .. \times_K A_K)||_F^2 \quad (10)$$

where, $tr(.)$ returns the trace of a matrix. In Equation 10, the first term ensures closeness between items of the same modalities and the second term consolidates the relative similarities between items across modalities. Minimizing Equation 10 is a non-convex optimization problem for a set of variables $A_1, .., A_K$. Apart from being an NP-hard problem, computationally it is also expensive to perform even simple operations like element wise product on a K^{th} order tensor. To get a more robust solution, we introduce an alternating method to tensor decomposition similar to [15, 17]. The key insight of such a method is to iteratively solve one of the K components of the tensor while keeping the rest fixed. We exploit this kind of alternating optimization solution to parallelize our framework across multiple GPUs, by placing each modality on one of them. This creates an independent pipeline for all of the K modalities as shown in Figure 3 which summarizes our distributed learning framework for multimodal datasets.

3.6 Distributed Training Approach for Learning Latent Representations

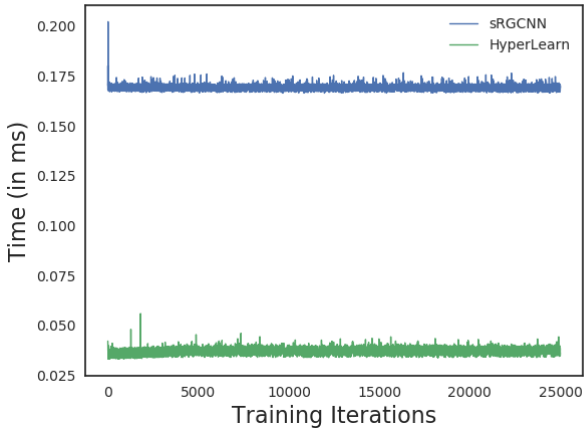
The separable feature extraction process for each modality makes our methodology unique and scalable to multiple modalities. These separate pipelines are combined by a joint loss function. Consider solving Equation 10 by keeping all other components except N_{θ_0} as constant. Since, all but one component of the tensor is a variable, unfolding original tensor $\underline{X}$ into a matrix along the N_{θ_0} component results in matrix $X_\theta^{(0)}$ with dimensions $|N_{\theta_0}| \times |N_1 N_2 ..... N_\theta|$ (where $1 \leq \theta \leq K \quad s.t. \quad \theta \neq \theta_0$). So, the loss function in Equation 10 can be rewritten for each of the K components (N_θ) as

$$Loss_\theta = \lambda ||X_\theta - A_\theta \Omega_\theta||_F^2 + \frac{1}{2} tr(A_\theta^T \mathbb{L}_\theta A_\theta) \quad (11)$$

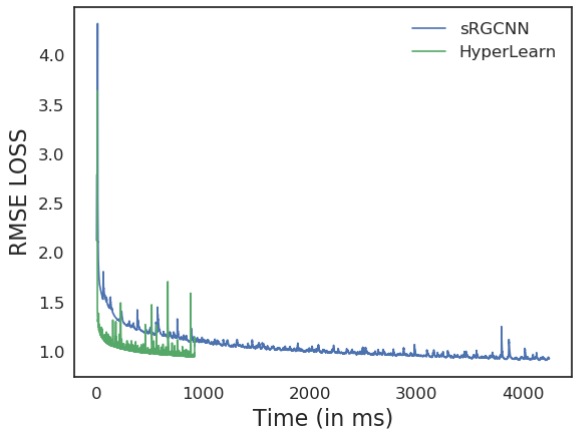
, where $\Omega_\theta = A_1 \odot A_2 \odot .. \odot A_{\theta-1} \odot A_{\theta+1} .. \odot A_K$ and $\odot$ represents the "Khatri-Rao" product.

4 EXPERIMENTS

We start our experimental evaluation showing the performance of our approach on a 2-dimensional standard matrix completion task and then extend it to 3 and 4 dimensional cases. For 2D, 3D and 4D case, we use MovieLens [28], MIR Flickr [13] and OmniArt [39] datasets respectively. We conjecture that our framework can be generalized to datasets with even more modalities. Table 1 summarizes the number of inter and intra relations for the three above mentioned cases. Here, $\mathscr{R}(.)$ represents the number of relations. As seen from the table, even relatively small datasets feature a multitude of relations, which makes learning them even more challenging.



(a) Time (in ms) taken for each training iteration



(b) Convergence rate of RMSE Loss over time

Figure 4: Illustration of the convergence rate of HyperLearn against sRGCNN. Our method clearly requires a much lower training time per iteration and also converges much faster than sRGCNN.

4.1 Task 1: Matrix Completion on Graphs

We show the computational advantage of our approach against a matrix completion method that makes use of side information as a baseline. For this, we use the standard MovieLens 100K dataset [28], which consists of 100,000 ratings on a scale of 0 to 5 corresponding to 943 users (U) and for 1,682 movies (M). We follow the experimental setup of Monti et.al. [29] for constructing the respective user

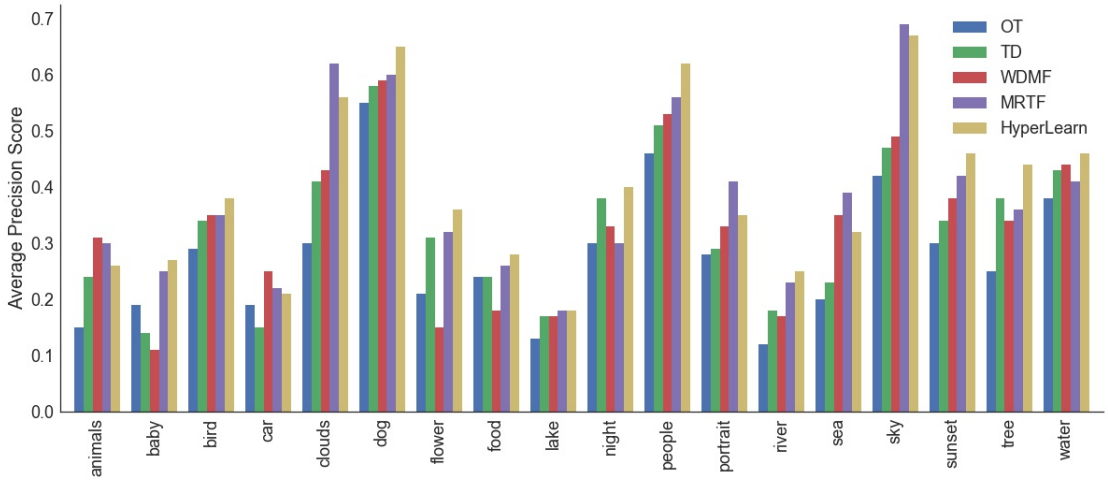


Figure 5: Detailed performance comparison in terms of Average Precision over 18 concepts on the MIRFlickr dataset

and movie intra-relation graphs as unweighted 10-nearest neighbor graphs.

We compare the performance of HyperLearn with separable Recurrent Graph Convolutional Networks (sRGCNN) as proposed in [29]. As can be seen from Figure 4, our approach attains comparable performance to the state of the art alternative, while being much faster. The feature extraction approach alternating between movie and item graphs reduces the time complexity (although not linearly) considerably as can be seen from Figure 4(a) which in turn increases the rate of convergence for the algorithm as depicted in Figure 4(b). However, due to continuous alternating loss calculations, sometimes the back-propagated gradients tend to get biased towards one of the modalities resulting in some higher peaks for HyperLearn in Figure 4(b).

4.2 Task 2: Social Image Understanding

In this experiment, we test the performance of our model on a 3rd order multimodal relational dataset. We apply our method to uncover latent image representations by jointly exploring user-provided tagging information, visual features of images and user demographics. We conduct experiments on the social image dataset: MIR Flickr [13]. The MIR Flickr dataset consists of 25,000 images (I) from Flickr posted by 6,386 users (U) with over 50,000 user-provided tags (T) in total. Some tags are obviously noisy and should be removed. Tags appearing at least 50 times are kept and the remaining ones are removed as in [23, 44]. To include user information, we crawl the groups joined by each user through the Flickr API. Some images have broken links, or are deleted by their users. We remove such images from our dataset which leaves us with 15,662 images, 6,618 users and 315 tags. The dataset also provides manually-created ground truth image annotations at the semantic concept level. For this filtered dataset, there are 18 unique concepts such as animals, bird, sky etc. for the images which we adopt to evaluate the performance. We create an intra-relation graph for images by taking 10-nearest neighbors based on their widely used standard SIFT features. For users, we create edges between them if they joined the same groups and for tags a graph is created based on their co-occurrence.

To empirically evaluate the effectiveness of our proposed method, we present the performance of the latent representation of images in classifying them into 18 concepts. We compare our model with the following methods:

- **OT**: The user-provided tags from Flickr as baseline.
- **TD**: The conventional CANDECOMP/PARAFAC (CP) tensor decomposition [11]
- **WDMF**: Weakly-supervised Deep Matrix Factorization for Social Image Understanding [22]
- **MRTF**: Multi-correlation Regularized Tensor Factorization approach [35]

Table 2: Comparison of the training times (in hours) on MIR Flickr dataset

Model	Training Time (in hours)
WDMF	4.2 ± 0.4
MRTF	2.7 ± 0.3
HyperLearn	1.8 ± 0.3

These methods, to the best of our knowledge, cannot provide the flexibility of performing distributed training for each modality using multiple GPUs. We report Average Precision (AP) scores for comparing our HyperLearn approach against all of these methods. Average Precision (AP) is the standard measure used for multi-label classification benchmarks. It corresponds to the average of the precision at each position where a relevant image appears. Figure 5 shows the comparative performance for all the 18 concepts. We also compare HyperLearn with MRTF and WDMF in terms of the training time and report the results in Table 2. As can be seen from this table, HyperLearn executes faster than MRTF and WDMF while

its performance is at par or even better for most of the concepts in the multi-label classification task shown in Figure 5.

Through this experiment we show that - (a) the performance of our approach is at par with the existing methods in understanding social image relationships (b) by introducing a distributed approach we can cut down training time of the model significantly.

4.3 Task 3: OmniArt

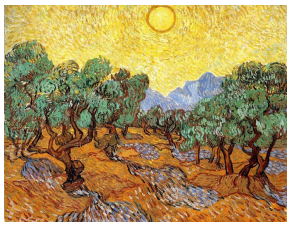
In the last experiment, we show the performance of our model in learning relations that go beyond 3rd order of connections. For this we require a highly multimodal dataset containing complex relations that are hard to interpret. One such dataset is OmniArt, a large-scale artistic benchmark collection consisting of digitized artworks from museums across the world [39, 40]. OmniArt comprises millions of artworks coupled with extensive metadata such as genre, school, material information, creation period and dominant color. This makes the dataset extremely multi-relational and, at the same time, very challenging to perform learning tasks.

For the purpose of comparison with related work, we first perform the artist attribution task in which we attempt to determine the creator of an artwork based on his/her inter-relations with artworks, media (e.g., oil, watercolor, canvas etc.) and creation period (timeframe), along with their intra-relations. To this end we select artworks corresponding to the most common artists in the collection. Considering each of these data streams - artworks (I), artist (A), media (M) and their timeframe (T_f) in centuries as a separate modality, we create the inter-relation hypergraph between them. Subsequently, intra-relation graphs are created for each of the 4 modalities in the following way:

- G_I: Based on color palettes similarity
- G_A: Based on the schools the artist belongs to
- G_M: Based on the co-occurrence in all artworks
- G_{T_f}: Based on the style and genre prevalent in that century

We take a sub-sample of the OmniArt dataset consisting of 10,000 artworks from 2,776 artists in the time period ranging all the way from 8th to 20th century along with 63 prominent media types. On this sampled dataset, we achieve an accuracy of 61.7% for the artist attribution task. The performance of our model is at par with the benchmark accuracy of 64.5% [39]. In addition, we conjecture that Hypergraph has an important advantage – the ability to learn even higher order relations, i.e. 5th, 6th and beyond, something that we intend proving in future work.

In the particular case of OmniArt, such higher-order relations would include information about e.g., artist, school, timeframe, medium, dominant colour use, semantics and (implicit) social network. For example, Figure 6 shows the well-known "Olive Trees with Yellow Sky and Sun" painted by Vincent van Gogh in 1889 and Claude Monet's masterpiece "Marine View with a Sunset" from 1875. As nicely portrayed by these two examples, while the two artists exhibit many stylistic similarities, sharing motives and a time period, their materialization is very different. Influenced by Monet, Van Gogh changed both his colour palette and coarseness of brushstrokes, so technically, his work became closer to the French Impressionism. Detecting "tipping points" in the artist's opus would require multimedia representations capable of capturing information about e.g. colour, texture and semantic concepts depicted in the paintings, but also information about school, social network, relevant locations and timeframe and historical context. We believe that our proposed framework is a significant and brave step forward in ultimately deploying multimedia analysis for solving such complex tasks.



(a) Vincent van Gogh – Olive Trees with Yellow Sky and Sun, 1889

(b) Claude Monet – Marine View with a Sunset, 1875

Figure 6: Van Gogh (a) and Monet (b) have many stylistic similarities, but their materialization is different. Capturing their similarities, differences and influences requires the ability to model higher-order relations.

5 CONCLUSION AND FUTURE WORK

In this paper we propose HyperLearn, a hypergraph-based framework for learning complex higher-order relationships in multimedia datasets. The proposed distributed training approach makes this framework scalable to many modalities. We demonstrate benefits of our approach with regards to both performance and computational time through extensive experimentation on MovieLens and MIR-Flickr datasets with 2 and 3 modalities respectively. To show the flexibility of HyperLearn in encoding a larger number modalities, we perform experiments on 4th order relations from the OmniArt dataset. In conclusion, on the examples of very different datasets, domains and use cases, we demonstrate that HyperLearn can be extremely useful in learning representations that can capture complex higher order relations within and across multiple modalities. For future work we plan to test the approach on even higher number of heterogeneous modalities and further extend this approach to much larger datasets by solving sub-tensors derived from slicing hypergraph into multiple smaller hypergraphs.

6 ACKNOWLEDGMENT

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 700381 (ASGARD project).

REFERENCES

[1] Devanshu Arya and Marcel Worring. 2018. Exploiting Relational Information in Social Networks using Geometric Deep Learning on Hypergraphs. In *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*. ACM, 117–125.
[2] Anirban Banerjee, Arnab Char, and Bibhash Mondal. 2017. Spectra of general hypergraphs. *Linear Algebra Appl.* 518 (2017), 14–30.
[3] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. 2016. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*. 3189–3197.
[4] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. 2017. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* 34, 4 (2017), 18–42.

[5] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. 2014. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014.*

[6] Jiajun Bu, Shulong Tan, Chun Chen, Can Wang, Hao Wu, Lijun Zhang, and Xiaofei He. 2010. Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 391–400.

[7] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C Aggarwal, and Thomas S Huang. 2015. Heterogeneous network embedding via deep architectures. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 119–128.

[8] Peng Cui, Shao-Wei Liu, Wen-Wu Zhu, Huan-Bo Luan, Tat-Seng Chua, and Shi-Qiang Yang. 2014. Social-sensed image search. *ACM Transactions on Information Systems (TOIS)* 32, 2 (2014), 8.

[9] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*. 3844–3852.

[10] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*. 2224–2232.

[11] Richard A Harshman et al. 1970. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis. (1970).

[12] Feiran Huang, Xiaoming Zhang, Chaozhuo Li, Zhoujun Li, Yueying He, and Zhonghua Zhao. 2018. Multimodal network embedding via attention based multi-view variational autoencoder. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, 108–116.

[13] Mark J Huiskes and Michael S Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 39–43.

[14] CG Khatri and C Radhakrishna Rao. 1968. Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhyā: The Indian Journal of Statistics, Series A* (1968), 167–180.

[15] Hyon-Jung Kim, Esa Ollila, Visa Koivunen, and Christophe Croux. 2013. Robust and sparse estimation of tensor decompositions. In *2013 IEEE Global Conference on Signal and Information Processing*. IEEE, 965–968.

[16] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *Proceedings of the International Conference on Learning Representations* (2017).

[17] Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM review* 51, 3 (2009), 455–500.

[18] Timothee Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical Tensor Decomposition for Knowledge Base Completion. In *International Conference on Machine Learning*. 2869–2878.

[19] Dong Li, Zhiming Xu, Sheng Li, and Xin Sun. 2013. Link prediction in social networks based on hypergraph. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 41–42.

[20] Hang Li, Haozheng Wang, Zhenglu Yang, and Masato Odagaki. 2017. Variation autoencoder based network representation learning for classification. In *Proceedings of ACL 2017, Student Research Workshop*. 56–61.

[21] Wu-Jun Li and Dit-Yan Yeung. 2009. Relation regularized matrix factorization. In *Twenty-First International Joint Conference on Artificial Intelligence*.

[22] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. 2016. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys (CSUR)* 49, 1 (2016), 14.

[23] Zechao Li and Jinhui Tang. 2016. Weakly supervised deep matrix factorization for social image understanding. *IEEE Transactions on Image Processing* 26, 1 (2016), 276–288.

[24] Zechao Li, Jinhui Tang, and Tao Mei. 2018. Deep collaborative embedding for social image understanding. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[25] Koji Maruhashi, Masaru Todoriki, Takuya Ohwa, Keisuke Goto, Yu Hasegawa, Hiroya Inakoshi, and Hirokazu Anai. 2018. Learning multi-way relations via tensor decomposition with neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[26] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. 2015. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*. 37–45.

[27] Julian McAuley and Jure Leskovec. 2012. Image labeling on a network: using social-network metadata for image classification. In *European conference on computer vision*. Springer, 828–841.

[28] Bradley N Miller, Istvan Albert, Shyong K Lam, Joseph A Konstan, and John Riedl. 2003. MovieLens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 263–266.

[29] Federico Monti, Michael Bronstein, and Xavier Bresson. 2017. Geometric matrix completion with recurrent multi-graph neural networks. In *Advances in Neural Information Processing Systems*. 3697–3707.

[30] Atsuhiro Narita, Kohei Hayashi, Ryota Tomioka, and Hisashi Kashima. 2012. Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery* 25, 2 (2012), 298–324.

[31] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.

[32] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 251–260.

[33] Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* 290, 5500 (2000), 2323–2326.

[34] Stevan Rudinac, Iva Gornishka, and Marcel Worring. 2017. Multimodal Classification of Violent Online Political Extremism Content with Graph Convolutional Networks. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. ACM, 245–252.

[35] Jitao Sang, Jing Liu, and Changsheng Xu. 2011. Exploiting user information for image tag refinement. In *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 1129–1132.

[36] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine, vol. 30, no. 3, pp. 83 – 98* (2013).

[37] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*. 926–934.

[38] Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[39] Gjorgji Strezoski and Marcel Worring. 2017. Omniart: multi-task deep learning for artistic data analysis. *arXiv preprint arXiv:1708.00684* (2017).

[40] Gjorgji Strezoski and Marcel Worring. 2018. OmniArt: A Large-scale Artistic Benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 4 (2018), 88.

[41] Jinhui Tang, Zechao Li, Meng Wang, and Ruizhen Zhao. 2015. Neighborhood discriminant hashing for large-scale image retrieval. *IEEE Transactions on Image Processing* 24, 9 (2015), 2827–2840.

[42] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[43] Jinhui Tang, Xiangbo Shu, Zechao Li, Yu-Gang Jiang, and Qi Tian. 2019. Social Anchor-Unit Graph Regularized Tensor Completion for Large-Scale Image Retagging. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[44] Jinhui Tang, Xiangbo Shu, Guo-Jun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. 2017. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE transactions on pattern analysis and machine intelligence* 39, 8 (2017), 1662–1674.

[45] Joshua B Tenenbaum, Vin De Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290, 5500 (2000), 2319–2323.

[46] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*. 2071–2080.

[47] Xiaolong Wang and Abhinav Gupta. 2018. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 399–417.

[48] Michael M Wolf, Alicia M Klinvex, and Daniel M Dunlavy. 2016. Advantages to modeling relational data using hypergraphs versus graphs. In *2016 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 1–7.

[49] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3441–3450.

[50] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Chang. 2015. Network representation learning with rich text information. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[51] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2017. User Profile Preserving Social Network Embedding.. In *IJCAI*. 3378–3384.

[52] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. 2007. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in neural information processing systems*. 1601–1608.