Latent Factor Representations for Cold-Start Video Recommendation

Sujoy Roy
SAP Innovation Center Network
Singapore
sujoy.roy@sap.com

Sharath Chandra Guntuku
School of Computer Engineering
Nanyang Technological University
Singapore
sharathc001@e.ntu.edu.sg

ABSTRACT

Recommending items that have rarely/never been viewed by users is a bottleneck for collaborative filtering (CF) based recommendation algorithms. To alleviate this problem, item content representation (mostly in textual form) has been used as auxiliary information for learning latent factor representations. In this work we present a novel method for learning latent factor representation for videos based on modelling the emotional connection between user and item. First of all we present a comparative analysis of state-of-the art emotion modelling approaches that brings out a surprising finding regarding the efficacy of latent factor representations in modelling emotion in video content. Based on this finding we present a method visual-CLiMF for learning latent factor representations for cold start videos based on implicit feedback. Visual-CLiMF is based on the popular collaborative less-is-more approach but demonstrates how emotional aspects of items could be used as auxiliary information to improve MRR performance. Experiments on a new data set and the Amazon products data set demonstrate the effectiveness of visual-CLiMF which outperforms existing CF methods with or without content information.

Keywords

Affective Computing, Personality, Likes, Videos, Emotion Prediction

1. INTRODUCTION

Matrix Factorization based collaborative filtering approaches are widely used in modern recommender systems. One of the pain points of this approach is when the collaborative information is very sparse (*cold start*) and in many cases not available (*very cold start scenario*). This affects recommendation performance. To address this issue, several methods have been proposed that add content or contextual information as auxiliary information with the collaborative information to show performance improvements. How best to *represent* the content and contextual information and *combine* it with the collaborative information depends on the nature of the content and context and clearly there is no one approach that fits all types of contents. In this work the content type is videos for a video recommendation scenario where there is not much contextual information available.

Content Representation. Our investigations into finding a good content representation is inspired by the observation that users like videos because they find some kind of emotional connect with it. Therefore the way users and videos are represented has to carry these emotive factor information. Hence in this work, as a basis for recommending videos, we investigate the emotive factors that influence users to 'like' videos [19]. *This involves learning representations that best capture the expected emotional response of users towards a video.* These representations could be used to either (1) build discriminative emotion models that can predict users emotional response to videos or (2) infer users preference for a video. Generating discriminative representations that can effectively model emotions faces the following challenges.

A good feature engineering based approach needs to incorporate knowledge of how contents influence users. This is a challenge because of the abstract and subjective nature of the problem and also the dimensions of emotions [15]. While sometimes it may seem relatively easier to pinpoint the emotion evoked by some videos, it is very hard to do so for most. For example, it is difficult to exactly ascertain the emotion experienced where there is no strong positive or negative valence associated with the emotion [20]. This makes creating labeled datasets with explicit labeling sometimes unreliable. Therefore most existing bigger datasets for emotion modeling are based on highly positive or highly negative emotion categories. They are created by using high valence terms as query to search for images/videos that might have such terms as tags. But this has its issues. For example, the term "angry" might return "angry bird video" which may not have to do anything with the emotion[9]. Several such issues are discussed in [23]. This points to the challenge of getting supervisory information to model emotions. Learning representations using deep networks as an alternative to feature engineering requires a lot of labelled training data

RecSys '16, September 15-19, 2016, Boston , MA, USA
© 2016 ACM. ISBN 978-1-4503-4035-9/16/09. . . $15.00
DOI: http://dx.doi.org/10.1145/2959100.2959172

and hence is limited by the lack of availability of supervisory information [14].

In this work we present an alternative to address the above challenge by demonstrating the possibility of using the implicit feedback ("likes") received from users as a potential source of supervisory signal that can help build better emotion recognition models that can in-turn help serve better recommendations. Note that implicit feedback on user preference for videos is usually available in plenty in the form of usage history. The intuition is that, latent factors learnt from factorizing user-item interaction data, carries information that captures the emotive factors of an item that makes a user "like" it.

To verify the intuition, first of all we present a comparative analysis of state of the art representations for modelling emotion from visual data vs. their latent factor representations derived from collaborative information through matrix factorization. The analysis establishes a novel finding about the efficacy of latent factor representation in modelling emotional response to videos. Latent factor representations are shown to be more discriminative in modelling specific emotions. This addresses the issue of finding an appropriate representation that carries emotive factors.

Combining Collaborative with Content Info. Latent factor representations are derived using factorization of collaborative data. Unfortunately cold start videos do not have sufficient collaborative information to have a discriminative latent factor representation. To deal with this problem, next, we present visual-CLiMF, a pipeline for learning a latent factor representation from a combination of content and collaborative information. To ensure that emotive factors are part of such learned latent representations, visual-CLiMF learns a transformation from existing emotion based representations to the latent factor representation space, based on the objective of maximizing reciprocal rank performance. This enables inferring latent factor representations for even videos that have no collaborative information. Note that we still need collaborative information during the transformation learning step. Our proposed matrix factorization approach is based on the less-is-more concept[21] of maximizing reciprocal rank (MRR) metric, although the concept could be extended to optimizing other evaluation metrics like area under the curve (AUC) [17] etc. Our experimental evaluations and analysis demonstrate the efficacy of the proposed approach in very cold start video recommendation. Figure 1 depicts an illustration of our proposed latent factor learning framework.

2. RELATED WORK

The motivation for this work rests on the intuition that latent factor representations from matrix factorization carry valuable information about the emotional factors that connect users to items. In this work we validate this intuition. Hence studying these latent factors is interesting both from the perspective of video emotion modelling as well as building effective recommender systems. In this section we present related work from both perspectives.

The discriminative information carried by the latent factors are clearly a function of the amount of collaborative information available. Matrix factorization has been very successfully applied to explicit and dense user-item collaborative information [22]. In most practical situations the collaborative information is implicit (positive-only feedback) and sparse which leads to reduced recommendation performance. So although there have been several works combining explicit collaborative information with content information [5, 18], works dealing with implicit feedback are only relevant for this work.

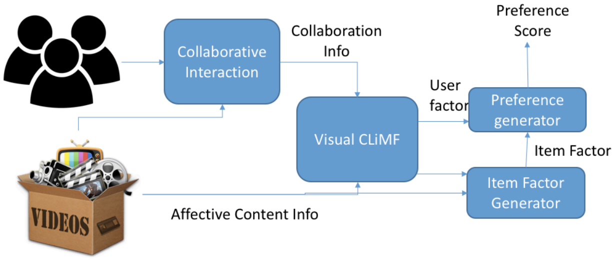

Figure 1: A flow diagram of the interactions in the proposed visual-CLiMF approach.

To deal with implicit feedback several works propose a numerical optimization framework that (1) *point-wise* associate confidence levels to preference values (with "0" preference for non-observed associations)[8], (2) considers the *pair-wise* comparison between preferred and a selection of non-preferred instances[17] and (3) considers only the preferred items[21]. The recommendation problem in the last two cases is formulated as a "learning to rank" problem where the objective is a smoothed version of some performance metric like AUC, MRR etc. In this work we consider the proposal in [21] as a basis to dealing with implicit feedback.

To deal with sparsity, content (visual[7], textual[25, 26]) and contextual[1] auxiliary information has been used to show improvements in recommendation performance. However, it is not clear how the syntactic information extracted from the content translates to high-level semantic information that can contribute to improving preference prediction. Some of these works aim to learn deep semantic representations[26] (for textual data). But it is not clear why the latent factor representations improved performance. This forms the motivation for studying the abstract/emotional factors carried by the latent factor representations, based on the intuition that latent factors carry some form of information regarding the emotional connect that users feel for items [10, 11].

If we consider the latent factor representations to be carrying some form of emotion response information, we also wish to study if they can help model emotional response to content, based on syntactic and mid-level semantic information. Several approaches have been proposed for modeling emotional response to content (textual[3], audio[4], images[2], videos[24]). But they all face some fundamental challenges. There are very few labelled data sets for emotion modelling and that too very small [27, 6] because manual labelling is expensive and the abstract nature of the labels leads to unreliable labels. Hence most labelled data sets are available for very positive or negative emotional responses. Also if they are not manually created they are created by querying for contents with very positive or negative emotional terms associated with them [9].

For the purpose of completeness we discuss some related works that propose visual information based feature representations for emotion modeling. A set of low-level features based on research in affective computing was proposed in [12]. As an investigation into semantic features [2] presents emotion modeling pipeline based on a label set of 2089 adjective noun pairs that are either very positive or very negative in nature. A detailed analysis of affective feature representations applied to video emotion modeling was presented in [14]. In [16] the authors present a comparative analysis of the efficacy of feature representations learnt by the above feature engineering approaches compared to representations learnt by the use of deep non-linear networks. Although the deep non-linear representations are shown[16] to be better than traditional feature engineering based approaches, learning deep networks without overfitting require a lot of training data which is usually not available in emotion modeling scenarios. Hence the improvements reported is not significant enough. On the other hand using off the shelf deep feature representations that have been proven to be extremely useful for detecting visual concepts in images/videos do not generalize well for emotion modeling (as shown in [16] and our own experiments). Hence the improvements reported are not significant enough.

From the perspective of related work, this work presents a novel method of combining auxiliary content information with collaborative information in addressing the cold start problem in video recommendation. In the process we also present a comparative validation of the efficacy of matrix factor based latent factor representations in modelling emotional responses to videos. This is also a novel result that we believe will be of interest to the research community.

3. PROBLEM FORMULATION

For a collection of videos $\mathcal{V}$ and a set of users $\mathcal{U}$ we are given $\mathcal{V}_i^+ \subseteq \mathcal{V}$, which is a set of videos liked by a user i. We are also given a collection $\mathcal{C} = \{C_j : 1 \leq j \leq n\}$, where C_j is the content representation for video j. The problem is to generate a ranked personalized recommendation of videos $\mathcal{V}_i^-$ not in $\mathcal{V}_i^+$, i.e., $\mathcal{V}_i^- \in \mathcal{V} \setminus \mathcal{V}_i^+$ for user i. The performance of the ranking is measured by reciprocal rank. And the method should also work for recommending those items that do not carry any collaborative information.

4. METHODOLOGY

This work proposes a method for combining content and collaborative information to learn latent factor representations for users and items. To investigate what kind of latent factor representations should we learn that can alleviate the cold start problem, we first study the nature of latent factor representations and what kind of information they carry. Towards this goal, we start with the intuition that out of the many semantic factors that they carry, one of the most important ones are the factors that emotionally connect the user to the item. To investigate the discriminative ability of the emotional information they carry, we study their effiacy in modeling desired emotional responses.

4.1 Emotion Modeling

Emotion modeling refers to the problem of automatically estimating the expected emotional response that a content will receive from users, as opposed to a subjective notion of response. To understand the nature of latent factors we benchmark three existing visual feature representations for emotion modeling, namely *affective* features[12] (114 dimensions), *sentibank* features[2] (2089 dimensions, PCA version - 312 dimensions), *hybrid CNN*[1] features[28] (4096 dimensions - PCA version 312 dimensions), with the MF-based latent factor[21] representations. Each video in our data set (to be described in Section 5) is labeled with 9 emotion labels - Amusement, Anger, Disgust, Fear, Interest, Joy, Sadness, Surprise and Tension. We learn a multi-label SVM classifer for each feature representation over the 9 emotion labels to evaluate the discriminative nature of the competing representations. That is, the feature representation with the minimum hamming loss (metric for evaluating multi-label multi-class classification) is the most suitable for emotion modeling. Figure 2 presents the hamming loss for each feature representation under multi-label multi-class classification. Note that the latent factor representations out-perform other feature representations by a significant margin (7-8% improvement over second best feature - affective features).

Figure 3 depicts the change in hamming loss with change in size of latent factor representation. The hamming loss starts high for low dimensional factors but becomes stable with higher dimensions. This verifies the fact that the hamming loss performance is not a function of the size of the latent factor representation. Note that the size of the other feature representations are fixed as reported in existing literature. This establishes the fairness of the comparison in terms of feature dimensions chosen.

This interesting result points to two facts. (1) Latent factor representations indeed carry some form of emotion response related information. (2) The emotional information captured by the latent factors are more discriminative in modeling emotion than the high level semantic reprresentations proposed in existing literature for affective analysis of visual content.

This leads to the next questions. (1) How can we generate such latent factors for contents that do not have collaborative information? What it has to do with the fact that latent factors model emotion well? (2) How do these latent factor representations compare with other MF approaches that use content information (specifically visual)? These questions are answered in Sections 4.2 and 5.

4.2 Visual-CLiMF

In this section we present the visual-CLiMF approach

[1] Convolutional Neural Networks (CNN) trained on 1183 categories (205 scene categories from Places dataset and 918 categories from the ILSVRC2012 training dataset from ImageNet)

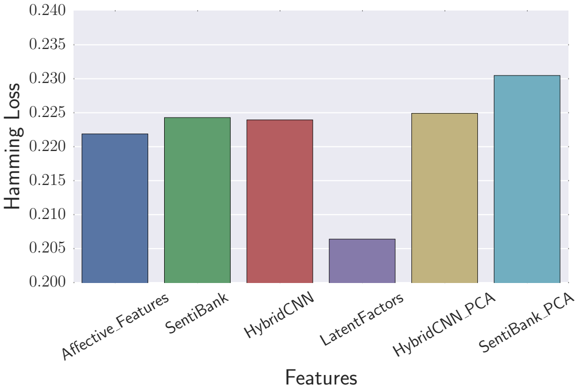

Figure 2: Performance of different features at modelling emotions in clips: Latent factors achieve lower hamming loss when compared to other features in multi-label classification followed by affective features, Sentibank and HyrbidCNN features.

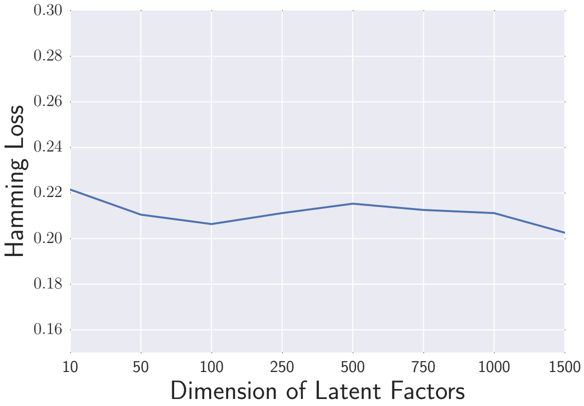

Figure 3: Performance of Latent Factors at predicting emotion distributions. Hamming Loss is seen to be stable across varying latent factor dimensions.

that combines collaborative and content information to generate latent factors for contents that do not contain collaborative information. The goal is to improve ranking of personalized recommendations for users. The model parameters are learned by directly maximizing the Mean Reciprocal Rank (MRR) which measures the performance of top-k recommendations.

The reciprocal rank of a list of N recommended items for user i is formulated as

$$RR_i = \sum_{j}^{N} \frac{Y_{ij}}{R_{ij}} \prod_{k=1}^{N} (1 - Y_{ik}\mathbb{1}(R_{ik} < R_{ij})), \qquad (1)$$

where R_{ij} denotes the rank of item j in the ranked list (in descending order) of item for user i. Y_{ij} denotes the binary relevance score of item j to user i, i.e., $Y_{ij} = 1$ if item j is relevant to user i else 0. $\mathbb{1}(x)$ is an indicator function that is equal to 1, if x is true, otherwise 0. Note that the ranking of relevant items is a non-smooth function of the relevance scores. Thus RR_i is also non-smooth. To deal with this problem a smoothed version of RR_i was presented [21] by basically approximating $\frac{1}{R_{ij}} \approx g(f_{ij})$ and $\mathbb{1}(R_{ik} < R_{ij}) \approx g(f_{ik} - f_{ij})$ where $f_{ij} = U_i^T V_j$.

A method for optimizing a lower bound of the smoothed reciprocal rank was presented in [21] to generate latent factor representations U_i and V_j for user i and item j respectively. To leverage on content information we represent the latent factor representation V_j for item j (which is a m-dimensional vector) as a transformation T of the content representation C_j (which is a n-dimensional vector)

$$V_i = \mathbf{T}\ C_j \qquad (2)$$

where T is a $m \times n$ matrix.

This leads to the following objective function.

$$\begin{aligned} F(U, \mathbf{T}) = \sum_{i=1}^{M} \sum_{j=1}^{N} Y_{ij} [ln\ g(U_i^T \mathbf{T} C_j) \\ + \sum_{k=1}^{N} ln(1 - Y_{ik} g(U_i^T \mathbf{T} C_k - U_i^T \mathbf{T} C_j))] \\ - \frac{\lambda}{2} (||U||_F^2 + ||\mathbf{T}||_F^2) \qquad (3) \end{aligned}$$

where λ is the regularization coefficient, $||U||_F^2$ is the Frobenius norm of U and $||\mathbf{T}||$ is the Frobenius form of $\mathbf{T}$.

Stochastic gradient ascend is used to maximize the objective function in Eq. (3). For gradient of the objective for user i with respect to U_i can be computed as below:

$$\begin{aligned} \frac{\delta F}{\delta U_i} = \sum_{j=1}^{N} Y_{ij} [g(-f_{ij}) V_j \\ + \sum_{k=1}^{N} \frac{Y_{ik} g'(f_{ik} - f_{ij})}{1 - Y_{ik} g(f_{ik} - f_{ij})} (V_j - V_k)] - \lambda U_i \end{aligned} \qquad (4)$$

The gradient of the objective with respect to $\mathbf{T}$ for every user i and item j is

$$\begin{aligned} \left(\frac{\delta F}{\delta \mathbf{T}}\right)_j &= \sum_{m,n} \mathbf{E}_{m,n} \frac{\delta F}{\delta T_{mn}} \\ &= Y_{ij} \frac{g'(U_i \mathbf{T} C_j)\ U_i \otimes C_j}{g(U_i \mathbf{T} C_j)} \\ &+ \sum_{k=1}^{N} \frac{Y_{ik} g'(U_i^T \mathbf{T} C_j)(U_i \otimes C_j - U_i \otimes C_k)}{1 - Y_{ik} g(U_i^T \mathbf{T} C_j)} \\ &+ \lambda \mathbf{T}, \end{aligned} \qquad (5)$$

where $\mathbf{E}_{m,n}$ is the elementary matrix of order $(m \times n)$, $g'(x)$ is the derivative of $g(x)$ and $\otimes$ is the outer product.

The learning algorithm is stochastic gradient ascend and is presented in Algorithm 1. Note that the difference with [21] is that here we learn U_i and $\mathbf{T}$ (and hence V_j) based on content information C_j and collaborative information Y_{ij}. In [21] only the collaborative information Y_{ij} is used to learn U_i and V_j. Once we know $\mathbf{T}$, given any new content C, we could estimate the latent factor representation $V = \mathbf{T}C$.

Table 1: Performance comparison of visual CLiMF and baselines (original CLiMF and Popularity Ranking) using different feature representations.

Split	Given 3			Given 5			Given 7			Given 9		
Feature	*Affective*	*Sentibank*	*HybridCNN*	*Affective*	*Sentibank*	*HybridCNN*	*Affective*	*Sentibank*	*HybridCNN*	*Affective*	*Sentibank*	*HybridCNN*
Test MRR (Content-based)	**0.2718**	**0.2256**	**0.238**	**0.2789**	**0.1884**	**0.2741**	**0.3042**	**0.1327**	**0.2293**	**0.2713**	**0.1157**	**0.1919**
Test MRR (CLiMF)	0.2156			0.2526			0.2286			0.1996		
Test MRR (Popularity)	0.2029			0.2009			0.2058			0.1362		

The relevance score of an item V_j for user U_i is computed as the dot product $r_{ij} =< U_i, V_j >$ that returns a scalar value. The recommendation process involves ranking (sorting in descending order) all items V_j according to their score r_{ij} for user U_i.

Algorithm 1 Algorithm: Learning Latent Representation

```
Input: Training set $Y$ , regularization parameter $\lambda$ learning
    rate $\gamma$, and the maximal number of iterations $itermax$
Output: The learned latent factors $U$, $\mathbf{T}$
Initialize $U^(0)$ and $\mathbf{T}^(0)$ with random values
Initialize $t = 0$;
repeat
    for $i \leftarrow 1, M$ do
        $U^{(t+1)} \leftarrow U^{(t)} + \gamma \frac{\delta F}{\delta U_i}$
        for $j \leftarrow 1, N$ do
            $\mathbf{T}^{(t+1)} \leftarrow \mathbf{T}^{(t)} + \gamma \left(\frac{\delta F}{\delta \mathbf{T}}\right)_j$
    $t \leftarrow t + 1$
until $t \geq itermax$
$U = U^{(t)}$ and $\mathbf{T} = \mathbf{T}^{(t)}$
```

5. EXPERIMENTS

5.1 Dataset

We evaluate visual-CLiMF with existing works on two data sets. The first data set is called **Video Emotion data set**, which is a new data set that we present for this work[2]. The second data set is the Amazon product data set[13].

Video Emotion Data Set:. We present a new data set that contains 323 video clips, rated by 111 users with a total of 1917 ratings. The videos are general in nature and include movie/TV series clips and so on. Each video is of an average duration of 3 minutes. Each user rated on an average of 17.27 videos. Each video is also labelled (by several volunteers based on majority voting) to belong to one or more of the 9 emotion categories - Amusement, Anger, Disgust, Fear, Interest, Joy, Sadness, Surprise and Tension. Each emotion is also rated by a valence score between 0-3. Apart from these item related information we also collected a few user related information like their answer to personality modelling questions, their age, gender etc. The user related information is not relevant for this work. All this information and ratings were collected using Amazon Mechanical Turk.

Amazon Product Data Set:. This dataset contains product reviews and metadata from Amazon. This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). For this work we only look at the "Cell Phones and Accessories" product category. We take the review data and the image features (which are basically 4096 dimensional CNN features of pre-trained network fo visual concept detection) of the product images.

Table 2: Performance comparison for cold start items (< 5 likes) for given 3. Note how visual-CLiMF outperforms original CLiMF.

Feature	*Affective*	*Sentibank*	*HybridCNN*
Test MRR (Content-based)	0.2033	0.1464	0.1377
Test MRR (CLiMF)	0.1319		
Test MRR (Popularity)	0.0973		

In this data set we look at the collaborative information between 66646 users and 118314 items (products). The rating of a product is in the range 0 to 5. Note that unlike an implicit data set, we do know that the user has given the item an explicit rating. If the rating is greater than equal to 3 we bucket it as a "like" and if the rating is less than 3 then either the user has not seen it or does not like it. We do this form of explicit to implicit feedback mapping to prepare the data to be used by visual-CLiMF which only sees implicit feedback. This is also used to compare performance with VBPR[7] (note: it is not clear to us if this is how implicit feedback is extracted in [7]).

5.2 Comparison: Only Collaborative

In this section we consider two baseline methods for ranking items based on only collaborative information. They are (1) original CLiMF[21] method to get personalized ranked item list and (2) popularity based approach which gives non-personalized ranked list. Both these methods do not consider any content information in the representation learning process. Our evaluation is performed on both data sets presented earlier.

We divide each users "liked" items into a training set and a test set where"Given n" refers to "n" items being considered as part of the training set $\mathcal{V}_i^+$ and the rest is taken as the test set $\mathcal{V}_i^-$. For CLiMF we first learn the latent factor representations for both users U_i and items V_j based on the training data with hyper-parameters $\lambda = 0.001$ and $\gamma = 0.0005$ (determined based on experimentation). Next we generate a ranked list of all the items for each user i (by sorting $< U_i, V_j >$) and compute the MRR for all the test items $\mathcal{V}_i^-$. For visual-CLiMF we follow the steps outined in Algorithm 1 with the same values of λ and γ to learn the user latent factors U and the transformation matrix $\mathbf{T}$ over the training set. Then we generate the item latent representation based on Equation 2. Next, a ranked list of items is generated and MRR computed for the test items.

[2] Data set will be made available upon request

Table 3: Performance of visual CLiMF compared to VBPR on Amazon Phones Dataset based on AUC.

Amazon Phones Dataset	Test AUC
visual CLiMF	0.4064
CLiMF	0.4035
VBPR	0.2527 +/- .38

Table 1 depicts the performance comparison of visual-CLiMF with the original CLiMF and popularity based methods on the Video Emotions dataset. Visual-CLiMF outperforms both the personalized (original CLiMF) and non-personalized (Popularity) recommendation approaches. Of all the feature representations, *affective features* proposed in [12] seem to perform the best and provide significant performance gains. The performance is also quite consistent across different sizes of training data. This seems to be consistent with the observation we made in Figure 2 where the affective features proposed in [12] are next best in emotion modeling in terms of hamming score. This shows how using emotion features to represent content is the key, and latent factors seem to be the best in doing that.

Table 4 depicts the performance comparison of visual CLiMF with the original CLiMF and popularity based methods on the Amazon products dataset, specifically the Cell phones and accessories category. For this data set also there is slight improvement in MRR performance compared to original CLiMF and popularity based approach. This improvement could be due to the following reasons. (1) People buy cell phones based on specifications, functionalities, sometimes design and also brand name. This means the affective features derived from that kind of metadata would probably be more useful and this is a subject of future work. (2) Also the CNN features available as part of the data set are good for visual concept detection but not for emotion modeling. In fact we do see for this data set that popularity based approaches have comparable MRR performance. This confirms the fact that the effectiveness in adding auxiliary information into the recommendation process depends on the usefulness of that information in capturing why users will want that item. In fact collaborative information seems to be more relevant for this data set.

We also conducted experiments to evaluate the performance of visual CLiMF in ranking cold start videos (< 5 likes) - depicted in Table 2. Note that for all the feature representations visual CLiMF significantly outperforms the original CLiMF and popularity based recommender. In fact comparing Table 1 and Table 2, we observe that the performance for cold start videos is comparable to the original CLiMF applied to the users who have more "likes". This is a promising observation.

MRR Effectiveness:. Figure 4 depicts the change in MRR performance with iterations for visual-CLiMF and original CLiMF on test data and training CLiMF. Note that these observations follow the less-is-more flavour of the CLiMF approach. The MRR performance increases with increase in iterations.

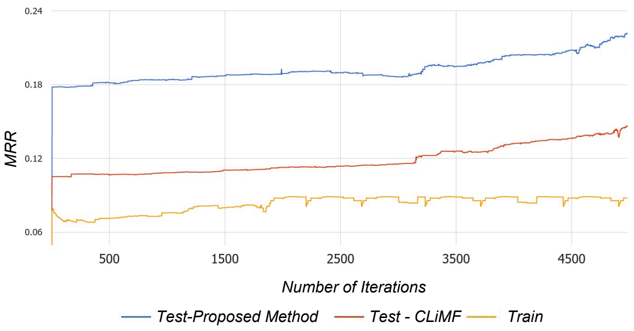

Figure 4: Effectiveness of the learning algorithm under different iterations for visual-CLiMF and original CLiMF under the Given-5

5.3 Comparison: Visual + Collaborative

There have been several works that have looked at adding visual information to collaborative information. Visual Bayesian Personalized Ranking(VBPR)[7] is one such recent work that has looked at integrating the visual information into the latent factor learning process. Few of the significant difference of visual CLiMF with VBPR is that both optimize different performance metrics. VBPR optimizes AUC whereas CLiMF optimizes MRR. VBPR uses visual concept descriptors (CNN) to represent content information. In fact these CNN features constitute a part of the item factor representation. In contrast visual CLiMF learns a latent representation from emotionally relevant representations which is also verified to work well based on results reported in earlier sections. Although visual CLiMF combines visual with collaborative information we also report results for very cold start situation where items have no collaborative information. Our understanding is that VBPR is not meant for those situations.

Table 3 presents a comparison of visual CLiMF, original CLiMF and VBPR based on AUC performance metric[3]. Note that although our method is optimzed for MRR, both CLiMF and visual CLiMF outperform VBPR (mean performance value). Also note that the AUC result for VBPR has a large enough standard deviation which makes the reported accuracy unstable. We have only tested VBPR on the Amazon products data set for fair comparison.

5.4 Visual Features

In this section we address two scenarios/questions. (1) What if we replace the latent factor V_j by their content representation C_j to compute the preference score $U_i^T C_j$ instead of $U_i^T V_j$ in generating a ranked list of items. This is to test whether we even need a transformation matrix $\mathbf{T}$ that we claim is learning something meaningful. Note that in this scenario, based on our observation that latent factors carry emotional information, the user factor U_i does carry some collaboratively shared emotional information about items that he may not have watched or liked. (2) The second question we address is what about items that have never been seen by any user. This is the

[3] Using code from authors of [7]

Table 4: Performance comparison of visual CLiMF and two baselines (i.e., original CLiMF and Popularity Ranking) for Amazon Phones Dataset based on MRR. ImageNet CNN features were used for visual CLiMF

Amazon Phones Dataset	*Given 3*	*Given5*
Test MRR (Visual CLiMF)	**0.0310**	**0.0331**
Test MRR (CLiMF)	0.0287	0.0303
PopRec	0.0282	0.0231

Table 5: MRR performance comparison by replacing latent representation with affective features [12]. This shows that transformation T learnt does improve performance over replacing the latent factors with the content representation.

Split	*visual-CLiMF (MRR)*	*Affective (MRR)*
Given 3	0.2691	0.2362
Given 5	0.2480	0.1886
Given 7	0.3102	0.2236
Given 9	0.2833	0.2032

very cold start problem where no collaborative information is available.

Scenario 1. We use *affective features* [12] to represent C_j and use it in place of V_j to generate $U_i^T V_j$. We chose affective features because they seem to perform second best after latent factor representations in modeling emotions. This would also tell us how consistent is the conclusion that indeed latent factors should carry some form of affective/emotional information to improve recommendation performance.

Table 5 reports the MRR scores obtained by using the affective feature representation [12] in place of the item latent factors. When we compare this to Table 1 we make some interesting observations. Note that using the *affective features* even outperform the original CLiMF approach in terms of MRR metric. This also demonstrates the fact that the transformation matrix **T** indeed assists in learning latent factors that carry emotional information that helps improve recommendation performance, which are better than directly replacing V_j with a content representation C_j or a purely collaborative representation for an item as in the original CLiMF method.

Scenario 2. Here we are concerned about a much tougher situation where we wish to apply visual-CLiMF to items which no one has watched - the *very cold start* problem. Table 6 depicts the MRR results for items that were never seen before and hence do not have any collaborative information. Note that for these items none of the existing methods that combine collaborative and content information will work as there is no collaboative information for these items. Hence there is no comparitive analysis with CF approaches.

Compared to Table 5 there is clearly a performance drop but none the less we do see how visual-CLiMF delivers at least 1 relevant item in the top-10 items returned in the ranked list of items. This is significant because the item is new and has not been seen by anyone.

Table 6: Performance of visual-CLiMF for very cold start items in the Video Emotion data set. That is, these items are not present in any $\mathcal{V}_i^+$.

% of items	*visual-CLiMF (MRR)*	*Affective (MRR)*
7%	0.0886	0.0513
27%	0.0498	0.0369
57%	0.0181	0.01808

6. CONCLUSION

Emotional factors are important in connecting users to items. In this work we presented some interesting findings about how to interpret latent factors in terms of the emotional elements they carry. In fact we show how latent factors that are generated from implicit feedback actually address the problem of lack of labelled data for modelling emotion categories. Now, we can learn better emotion models from implicit feedback. Then we show how this knowledge can be exploited to improve MRR performance of recommender systems. We propose visual-CLiMF that builds on the original CLiMF approach by combining visual with collaborative information to learn latent factors that give better MRR performance. In the process we also present our investigations in dealing with recommendation of *very cold start* videos to users - videos that have not been seen by any user.

7. ACKNOWLEDGEMENTS

Sujoy Roy's research is partially funded by the Economic Development Board and National Research Foundation, Singapore.

8. REFERENCES

[1] L. Baltrunas, B. Ludwig, and F. Ricci. Matrix factorization techniques for context aware recommendation. In *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, pages 301–304. ACM, 2011.

[2] D. Borth, T. Chen, R. Ji, and S.-F. Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 459–460. ACM, 2013.

[3] E. Cambria, J. Fu, F. Bisio, and S. Poria. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *AAAI*, pages 508–514, 2015.

[4] A. K. Elshenawy, S. Carter, and D. Braga. ItâĂŹs not just what you say, but how you say it: Muiltimodal sentiment analysis via crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[5] P. Forbes and M. Zhu. Content-boosted matrix factorization for recommender systems: Experiments with recipe recommendation. In *Proceedings of the*

Fifth ACM Conference on Recommender Systems, RecSys '11, pages 261–264, 2011.

[6] S. C. Guntuku, M. J. Scott, G. Ghinea, and W. Lin. Personality, culture, and system factors-impact on affective response to multimedia. *arXiv preprint arXiv:1606.06873*, 2016.

[7] R. He and J. McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1510.01784*, 2015.

[8] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. Ieee, 2008.

[9] Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[10] J. Kim, J.-B. Yoo, H. Lim, H. Qiu, Z. Kozareva, and A. Galstyan. Sentiment prediction using collaborative filtering. In *ICWSM*, 2013.

[11] F. Li, S. Wang, S. Liu, and M. Zhang. Suit: A supervised user-item based topic model for sentiment analysis. In *AAAI*, pages 1636–1642, 2014.

[12] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, pages 83–92. ACM, 2010.

[13] J. McAuley, R. Pandey, and J. Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 785–794, New York, NY, USA, 2015. ACM.

[14] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2015.

[15] R. Plutchik. *The emotions*. University Press of America, 1991.

[16] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[17] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press, 2009.

[18] R. Ronen, N. Koenigstein, E. Ziklik, and N. Nice. Selecting content-based features for collaborative filtering recommenders. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 407–410. ACM, 2013.

[19] M. Sakaki, K. Niki, and M. Mather. Beyond arousal and valence: The importance of the biological versus social relevance of emotional stimuli. *Cognitive, Affective, & Behavioral Neuroscience*, 12(1):115–139, 2012.

[20] K. R. Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[21] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, N. Oliver, and A. Hanjalic. Climf: learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 139–146. ACM, 2012.

[22] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 47(1):3, 2014.

[23] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic. Corpus development for affective video indexing. *Multimedia, IEEE Transactions on*, 16(4):1075–1089, 2014.

[24] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *Affective Computing, IEEE Transactions on*, 3(2):211–223, 2012.

[25] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. ACM.

[26] H. Wang, N. Wang, and D. Yeung. Collaborative deep learning for recommender systems. *CoRR*, abs/1409.2944, 2014.

[27] S. Wang and Q. Ji. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, 6(4):410–430, 2015.

[28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.