NEURAL CONTENT BASED COLLABORATIVE FILTERING FOR RECOMMENDATION SYSTEMS

Prateek Verma
CCRMA
Stanford University
prateekv@ccrma.stanford.edu

Jonathan Berger
CCRMA
Stanford University
brg@ccrma.stanford.edu

EXTENDED ABSTRACT

Recommendations systems are at the core of a number of applications and interactions among people every day. They are used in diverse fields such as recommending music, images, videos, text to everyday products that we buy/sell and even in areas such as search. It combines a variety to disciplines like machine learning, statistics combined with expertise in domain specific fields like computer vision, music and audio processing, natural language processing etc. The field is well researched, with renewed interest, due to the success of the Music Genome project [1], Netflix prize [4], and more recently with advent of deep neural networks. The approaches are loosely based on collaborative filtering and content based filtering. A collaborative filtering approach models interactions between a subset of users-items in order to recommend unknown user-item entries via an interaction matrix. There are a variety of approaches plausible in order to get the missing entries of user-item matrix, e.g. matrix factorization, clustering models and more recently a neural network based in-filling algorithm [5]. On the other hand, content based filtering relies on various characteristics of the desired item, and does not take into account user behaviour. More recently, neural embeddings have been found to be very useful in recommendations systems.

In this paper, we propose to combine these two approaches. The main contribution of the paper is an algorithm which proposes a content based collaborative filtering approach in which the user-item matrix is replaced by user-filter matrix based on activations of a deep neural network. To the best of our knowledge the idea of user-filter activation matrix has not been proposed before. For our proposed algorithm, we focus on music and audio recommendation. Almost identical neural architectures have been used in various domains such as audio, vision [7], natural language processing [3]. It is easy to see that the current approach is applicable in several different areas. It has been shown in domains such as audio/images, that the family of convolutional models learn various hierarchies of features. For the case of music, the initial layers learn lower level features such as pitch, harmonies, vibrato, onsets etc. The higher layers combine these low level activations for understanding higher level concepts depending on the problem of interest like language, genre, artist, melodies etc. For the constitution of any song, both the high level and the low level features are important. These features give us far greater information than categorical description of items (song).

To give an example to further motivate this idea, a jazz tune will invoke particular filters fired for 'high level' categories (eg genre (jazz), timbre (piano, acoustics bass, sax, etc). There is a stronger likelihood that a different user liking sax, bass and piano in popular music would listen to jazz than to, say, Indian classical Music. Several such arguments can be made in regard to combinations of low level and high level features, supporting our argument, and advantages of moving from user-item matrix to user-filter matrix. The current user-item interaction matrix fails to capture such dependencies, and relies on powerful algorithms to learn them.

For our case, we train a variant of state of the art DenseNet [6] architecture on balanced Audio Set [2], giving rise to a total of 1316 convolutional channels which the input is progressively downsampled and reduced in size from. The input to our system is an audio file of $3s$ duration with the targets being 527 categories. As motivated before, we replace the traditional user-item matrix with that of user-filter (activation) matrix. The values in the matrix are categorical, indicating the presence/absence of a musical attribute modelled by

convolutional filter activation. The continuous filter activations are discretized using mean, μ and variance, σ across all the activations present in the dataset for a single filter. As a sample rule, an activation greater than μ + 2σ can be assigned 1, with values less than μ - 2σ assigned -1 indicative of a strong dislike, with the rest of the entries as unfilled. The values consisting of a user-filter pair gets assigned based upon the average of all the convolutional filter activations of all the audio pieces the user has interacted with. A single song would have several such patches of $3s$ duration. The final value of a user-filter interaction matrix would simply be an average across all the patches a user has interacted with.

Mathematically, let $\mathcal{M}$ be the user-filter collaborative matrix, then each entry of the interaction matrix, m_{uf}, is the value corresponding to user u and filter index f. For a total of N interactions for a user u,

$$m_{uf} = \frac{1}{N} \sum_{n=1}^{N} W_n^{lk}$$

where W_n^{lk} is the activation of the convolutional filter k in the l th layer of a deep neural net architecture (assigned an index f), with a $3s$ input audio, with N being the total number of interactions the user u had. The activations of different filters, are already normalized as we discretized the scores to $[-1, 1]$ using first order statistics. The missing user-filter entries can be inferred using a wide body of existing literature like [4]. For getting the users, a song back, we first store the sorted lists of songs according to the filter activation for *each* filter. We then recommend the song that the user has not interacted with that activates the predicted positive entry for the filter index in our interaction matrix.

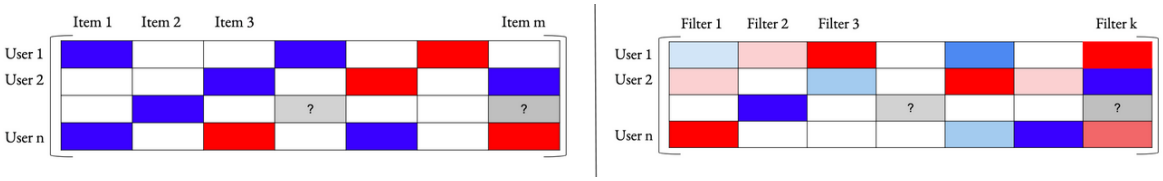

Figure 1. Comparisons of traditional user-item matrix and our proposed user-filter interaction matrix. Blue and Red refer to likes-dislikes, with the same being modelled using μ-σ statistics, and stored as continuous numbers between $[-1, 1]$ for user-filter matrix. The shades of blue-red represent such values, and white being unfilled entries.

REFERENCES

[1] Michael Castelluccio. The music genome project. *Strategic Finance*, pages 57–59, 2006.

[2] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[3] Michelle Guo, Albert Haque, and Prateek Verma. End-to-end spoken language translation. *arXiv preprint arXiv:1904.10760*, 2019.

[4] Blake Hallinan and Ted Striphas. Recommended for you: The netflix prize and the production of algorithmic culture. *New media & society*, 18(1):117–137, 2016.

[5] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[7] Prateek Verma and Julius O Smith. Neural style transfer for audio spectograms. *arXiv preprint arXiv:1801.01589*, 2018.