

REVISITING THE PROBLEM OF AUDIO-BASED HIT SONG PREDICTION USING CONVOLUTIONAL NEURAL NETWORKS

Li-Chia Yang*, Szu-Yu Chou*, Jen-Yu Liu*, Yi-Hsuan Yang*, Yi-An Chen†

*Research Center for Information Technology Innovation, Academia Sinica, Taiwan

†Machine Learning Research Team, KKBOX Inc., Taiwan

ABSTRACT

Being able to predict whether a song can be a hit has important applications in the music industry. Although it is true that the popularity of a song can be greatly affected by external factors such as social and commercial influences, to which degree audio features computed from musical signals (whom we regard as internal factors) can predict song popularity is an interesting research question on its own. Motivated by the recent success of deep learning techniques, we attempt to extend previous work on hit song prediction by jointly learning the audio features and prediction models using deep learning. Specifically, we experiment with a convolutional neural network model that takes the primitive mel-spectrogram as the input for feature learning, a more advanced JYnet model that uses an external song dataset for supervised pre-training and auto-tagging, and the combination of these two models. We also consider the inception model to characterize audio information in different scales. Our experiments suggest that deep structures are indeed more accurate than shallow structures in predicting the popularity of either Chinese or Western Pop songs in Taiwan. We also use the tags predicted by JYnet to gain insights into the result of different models.

Index Terms— Hit song prediction, deep learning, convolutional neural network, music tags, cultural factors

1. INTRODUCTION

The popularity of a song can be measured *a posteriori* according to statistics such as the number of digital downloads, playcounts, listeners, or whether the song has been listed in the Billboard Chart once or multiple times. However, for music producers and artists, it would be more interesting if song popularity can be predicted *a priori* before the song is actually released. For music streaming service providers, an automatic function to identify emerging trends or to discover potentially interesting but not-yet-popular artists is desirable to address the so-called “long tail” of music listening [1]. In academia, researchers are also interested in understanding the

factors that make a song popular [2,3]. This can be formulated as a pattern recognition problem, where the task is to generalize observed association between song popularity measurements and feature representation characterizing the songs in the training data to unseen songs [4].

Our literature survey shows that this *automatic hit song prediction* task has been approached using mainly two different information sources: 1) *internal factors* directly relating to the content of the songs, including different aspects of audio properties, song lyrics, and the artists; 2) *external factors* encompassing social and commercial influences (e.g. concurrent social events, promotions or album cover design).

The majority of previous work on the internal factors of song popularity are concerned with the audio properties of music. The early work of Dhanaraj and Logan [4] used support vector machine to classify whether a song will appear in music charts based on latent topic features computed from audio Mel-frequency cepstral coefficients (MFCC) and song lyrics. Following this work, Pachet *et al.* [5] employed a large number of audio features commonly used in music information retrieval (MIR) research and concluded that the features they used are not informative enough to predict hits, claiming that hit song science is not yet a science. Ni *et al.* [6] took a more optimistic stand, showing that certain audio features such as tempo, duration, loudness and harmonic simplicity correlate well with the evolution of musical trends. However, their work analyzes the evolution of hit songs [7–9], rather than discriminates hits from non-hits. Fan *et al.* [10] performed audio-based hit song prediction of music charts in mainland China and UK and found that Chinese hit song prediction is more accurate than the UK version. Purely lyric-based hit song prediction was relatively unexplored, except for the work presented by Singhi and Brown [11].

On the other hand, on external factors, Salganik *et al.* [12] showed that the song itself has relatively minor role than the social influences for deciding whether a song can be a hit. Zangerla *et al.* [13] used Twitter posts to predict future charts and found that Twitter posts are helpful when the music charts of the recent past are available.

To our best knowledge, despite its recent success in various pattern recognition problems, deep learning techniques have not been employed for hit song prediction. In particular,

This work was partially supported by the Ministry of Science and Technology of Taiwan under Contracts 104-2221-E-001-029-MY3 and 105-2221-E-001-019-MY2.

in speech and music signal processing, convolutional neural network (CNN) models have exhibited remarkable strength in learning task-specific audio features directly from data, outperforming models based on hand-crafted audio features in many prediction tasks [14–16].

We are therefore motivated to extend previous work on audio-based hit song prediction by using state-of-the-art CNN-based models, using either the primitive, low-level mel-spectrogram directly as the input for feature learning, or a more advanced setting [17] that exploits an external music auto-tagging dataset [18] for extracting high-level audio features. Moreover, instead of using music charts, we use a collection of user listening data from KKBOX Inc., a leading music streaming service provider in East Asia. We formulate hit song prediction as a regression problem and test how we can predict the popularity of Chinese and Western Pop music among Taiwanese KKBOX users, whose mother tongue is Mandarin. Therefore, in addition to testing whether deep models outperform shallow models in hit song prediction, we also investigate how the culture origin of songs affects the performance of different CNN models.

2. DATASET

Because we are interested in discriminating hits and non-hits, we find it informative to use the playcounts a song receives over a period of time from streaming services to define song popularity and formulate a regression problem to predict song popularity. In collaboration with KKBOX Inc., we obtain a subset of user listening records contributed by Taiwanese listeners over a time horizon of one year, from Oct. 2012 to Sep. 2013, involving the playcounts of close to 30K users for around 125K songs. Based on the language metadata provided by KKBOX, we compile a *Mandarin* subset featuring Chinese Pop songs and a *Western* subset comprising of songs sung mainly in English. There are more songs in the Western subset but the Mandarin songs receive more playcounts on average, for Mandarin is the mother tongue of Taiwanese.

The following steps are taken to gain insights into the data and for data pre-processing. First, as the songs in our dataset are released in different times, we need to check whether we have to compensate for this bias, for intuitively songs released earlier can solicit more playcounts. We plot in Fig. 1 the average playcounts of songs released in different time periods, where Q1 denotes the first three months starting from Oct. 2012 and –Q1 the most recent three months before Oct. 2012, etc. The y-axis is in log scale but the actual values are obscured due to a confidentiality agreement with KKBOX. From the dash lines we see that the average playcounts from different time periods seem to be within a moderate range in the log scale for both subsets, exempting the need to compensate for the time bias by further operations.

Second, we define the *hit score* of a song according to the multiplication of its playcount in log scale and the number of

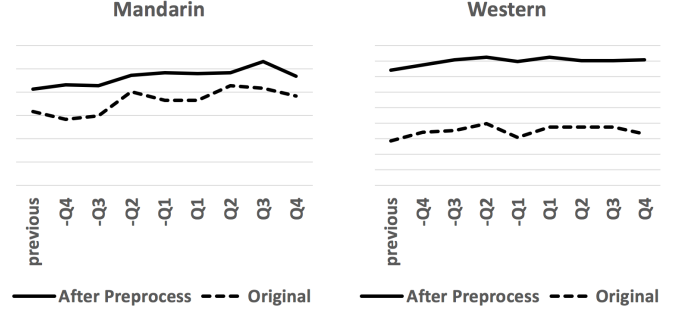


Fig. 1. The average playcounts (in log scale) of songs released in different time periods.

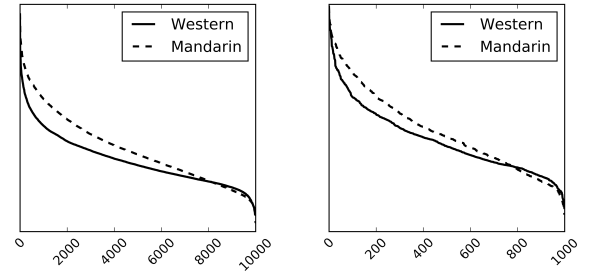


Fig. 2. The distribution of hit scores (see Section 2 for definition) in the (left) whole and (right) test sets.

users (also in log scale) who have listened to the song. We opt for not using the playcounts only to measure song popularity because it is possible that the playcount of a song is contributed by only a very small number of users.

Third, to make our experimental results on the two subsets comparable, we sample the same amount of 10K songs in our experiment for both subsets. These songs are those with the highest playcounts within the subset. It can be seen from Fig. 2 that the distributions of hit scores of the sampled songs are similar. The solid lines in Fig. 1 show that after this sampling the time bias among the sampled songs remains moderate.

Finally, we randomly split the songs to have 8K, 1K, and 1K songs as the training, validation, and test data for each of the subsets. Although it may be more interesting to split the songs according to their release dates so as to ‘learn from the past and predict the future,’ we leave this as a future work. Our focus here is to study whether deep models perform better than shallow models in audio-based hit song prediction.

The scale and the time span of the dataset are deemed appropriate for this study. Unlike previous work on musical trend analysis that may involve more than ten years’ worth of data (e.g. [6], [19]), for the purpose of our work we want to avoid changes in public music tastes and therefore it is better to use listening records collected within a year.

3. METHODS

We formulate hit song prediction as a regression problem and train either shallow or deep neural network models for predicting the hit scores. Given the audio representation \mathbf{x}_n for each song n in the training set, the objective is to optimize the parameters Θ of our model $f(\cdot)$ by minimizing the squared error between the ground truth y_n and our estimate, expressed as $\min_{\Theta} \sum_n \|y_n - f_{\Theta}(\mathbf{x}_n)\|_2^2$. As described below, a total number of six methods are considered, All of them are implemented based on the lasagne library [20], and the model settings such as learning rate update strategy, dropout rate, and numbers of feature maps per layer are empirically tuned by using the validation set.

3.1. Method 1 (m1): LR

As the simplest method, we compute 128-bin log-scaled mel-spectrograms [21] from the audio signals and take the mean and standard deviation over time, leading to a 256-dim feature vector per song. The feature vectors are used as the input to a single-layer shallow neural network model, which is effectively a linear regression (LR) model. The mel-spectrograms are computed by short-time Fourier transform with 4,096-sample, half-overlapping Hanning windows, from the middle 60-second segment of each song, which is sampled at 22 kHz. In lasagne, we can implement the LR model by a dense layer.

3.2. Method 2 (m2): CNN

Going deeper, we use the mel-spectrograms directly as the input, which is a 128 by 646 matrix for there are 646 frames per song, to a CNN model. Our CNN model consists of two early convolutional layers, with respectively 128-by-4 and 1-by-4 convolutional kernels, and three late convolutional layers, which all has 1-by-1 convolutional kernels. Unlike usual CNN models, we do not use fully connected layers in the latter half of our model for such *fully convolutional* model has been shown more effective for music [14, 17, 22].

3.3. Method 3 (m3): inception CNN

The idea of *inception* was introduced in GoogLeNet for visual problems [23]. It uses multi-scale kernels to learn features. We make an audio version of it by adding two more parallel early convolutional layers with different sizes: 132-by-8 and 140-by-16, as illustrated in the bottom-right corner of Fig. 3. To combine the output of these three kernels by concatenation, the input mel-spectrogram needs to be zero-padded.

3.4. Method 4 (m4): JYnet (a CNN model) + LR

While generic audio features such as mel-spectrogram may be too primitive to predict hits, we employ a state-of-art music auto-tagging system referred to as the JYnet [17] to compute

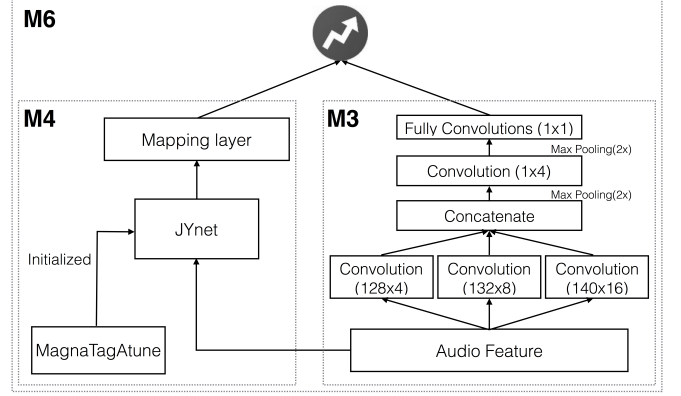


Fig. 3. Architecture of the investigated CNN models.

high-level tag-based features. JYnet is another CNN model that also takes the 128-bin log-scaled mel-spectrograms as the input, but the model is trained to make tag prediction using the MagnaTagATune dataset [18]. The output is the activation scores of 50 music tags, including genres, instruments, and other performing related tags such as male vocal, female vocal, fast and slow. From the output of JYnet (i.e. 50-dim tag-based features), we learn another LR model for predicting hit scores, as illustrated in the bottom-left corner of Fig. 3.

3.5. Methods 5 and 6 (m5 & m6): Joint Training

We also try to combine (m4) with (m2) or (m3) to exploit information in both the mel-spectrograms and tags, leading to (m5) and (m6). Instead of simply combining the results of the two models $f_{\Theta_1}(\cdot)$ and $f_{\Theta_2}(\cdot)$ being combined, we add another layer on top of them for joint training, as illustrated in Fig. 3. The learning objective becomes:

$$\min_{w, \Theta_1, \Theta_2} \sum_n \|y_n - w f_{\Theta_1}(\mathbf{x}_n) - (1 - w) f_{\Theta_2}(\mathbf{x}_n)\|_2^2, \quad (1)$$

where w determines their relative weight. In this way, we can optimize the model parameters of both models jointly. However, when method 4 is used in joint training we only update the parameters of its LR part, as JYnet is treated as an external, pre-trained model in our implementation.

4. EXPERIMENTAL RESULTS

We train and evaluate the two data subsets separately. For evaluation, the following four metrics are considered:

- Recall@100: Treating the 100 songs (i.e. 10%) with the highest hit scores among the 1,000 test songs as the hit songs, we rank all the test songs in descending order of the predicted hit scores and count the number of hit songs that occur in the top 100 of the resulting ranking.
- nDCG@100: normalized discounted cumulative gain (nDCG) is another popular measure used in ranking

Table 1. Accuracy of Hit Song Prediction

Method	Mandarin subset				Western subset			
	recall	nDCG	Kendall	Spearman	recall	nDCG	Kendall	Spearman
(m1) audio+LR	0.1900	0.1997	0.1679	0.2480	0.1400	0.1271	0.0674	0.1002
(m2) audio+CNN	0.2300	0.2334	0.1806	0.2678	0.1300	0.1294	0.1031	0.1564
(m3) audio+inception CNN	0.2500	0.2369	0.2286	0.3374	0.1800	0.1989	0.1093	0.1636
(m4) tag+LR	0.2400	0.2372	0.1671	0.2473	0.2000	0.1774	0.0918	0.1372
(m5) (m2)+(m4)	0.2500	0.2558	0.2018	0.2971	0.1800	0.1791	0.1300	0.1941
(m6) (m3)+(m4)	0.3000	0.2927	0.2665	0.3894	0.2100	0.2413	0.1341	0.1996

problems [24]. It is computed in a way similar to recall@100, but the positions of recalled hit songs in the ranking list are taken into account.

- Kendall’s τ : we directly compare the ground truth and predicted rankings of the test songs in hit scores (without defining which songs are hit songs) and compute a value that is based on the number of correctly and incorrectly ranked pairs [25].
- Spearman’s ρ : the rank correlation coefficient (considering the relative rankings but not the actual hit scores) between the ground truth and predicted rankings.

The result is shown in Table 1, which is obtained by averaging the result of 10 repetition of each method. The following observations can be made. First, by comparing the result of (m1), (m2) and (m3), we see that better result in most of the four metrics is obtained by using deeper and more complicated models for both subsets. This suggests the effectiveness of deep structures for this task. Furthermore, by comparing the result of the two subsets, we see that audio-based hit song prediction is easier for the Mandarin subset, confirming the findings of Fan *et al.* [10].

Second, as both (m1) and (m4) use LR for prediction, by comparing their result we see that the tag-based method (m4) outperforms the simple audio-based method (m1) in all the four metrics for the Western subset, demonstrating the effectiveness of the JYnet tags. This is however not the case for the Mandarin subset for Kendall’s τ and Spearman’s ρ .

Third, from the result of (m5) and (m6), we see that the joint learning structure can further improve the result for both subsets. The best result is obtained by (m6) in all metrics.

To gain insights, we employ JYnet to assign genre labels to all the test songs and examine the distribution of genres in the top-50 hit songs determined by either automatic models or the ground truth. For each song, we pick the genre label that has the strongest activation as predicted by JYnet. The resulting genre distributions are shown in Fig. 4. We see from the result of ground truth that the Western hits have more diverse genres. The predominance of ‘Pop’ songs in the Mandarin subset might explain why 1) hit song prediction in this subset is easier and 2) (m4) alone cannot improve τ and ρ . Moreover, for the Western subset, we see that the genre distribution of (m4) is more diverse than that of (m3), despite that

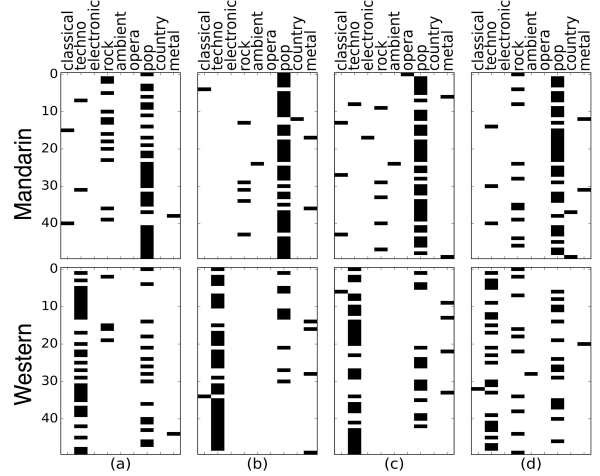


Fig. 4. The predominate tags (predicted by JYnet) for the top-50 hit songs determined by different methods for the (top) Mandarin and (bottom) Western subsets. From left to right: (a) the tag-based model (m4), (b) the audio-based model (m3), (c) the hybrid model (m6), and (d) the ground truth.

(m3) achieves slightly higher nDCG and Spearman’s ρ . This might imply that the ability to match the genre distribution of the ground truth is another important performance indicator.

5. CONCLUSION

In this paper, we have introduced state-of-the-art deep learning techniques to the audio-based hit song prediction problem. Instead of aiming at classifying hits from non-hits, we formulate it as a regression problem. Evaluations on the listening data of Taiwanese users of a streaming company called KKBOX confirms the superiority of deep structures over shallow structures in predicting song popularity. Deep structures are in particular important for Western songs, as simple shallow models may not capture the rich acoustic and genre diversity exhibited in Western hits. For future work, we hope to understand what our neural network models actually learn, to compare against more existing methods (preferably using the same datasets), and to investigate whether our models can predict future charts or emerging trends.

6. REFERENCES

- [1] H. Silk, R. Santos-Rodriguez, C. Mesnage, T. De Bie, and M. McVicar, “Data science for the detection of emerging music styles,” *EPSRC*, pp. 4–6, 2014.
- [2] S. McClary, *Studying popular music*, vol. 10, 1991.
- [3] P. D. Lopes, “Innovation and diversity in the popular music industry, 1969 to 1990,” *American Sociological Review*, vol. 57, no. 1, pp. 56, 1992.
- [4] R. Dhanaraj and B. Logan, “Automatic prediction of hit songs,” in *Proceedings of International Society for Music Information Retrieval*, pp. 11–15, 2005.
- [5] F. Pachet and P. Roy, “Hit song science is not yet a science,” in *Proceedings of International Society for Music Information Retrieval*, pp. 355–360, 2008.
- [6] Y. Ni and R. Santos-Rodriguez, “Hit song science once again a science,” *International Workshop on Machine Learning and Music*, pp. 2–3, 2011.
- [7] R. M. MacCallum, M. Mauch, A. Burt, and A. M. Leroi, “Evolution of music by public choice,” in *Proceedings of the National Academy of Sciences*, vol. 109, no. 30, pp. 12081–12086, 2012.
- [8] M. Mauch, R. M. MacCallum, M. Levy, and A. M. Leroi, “The evolution of popular music: USA 1960–2010,” *Royal Society Open Science*, vol. 2, no. 5, pp. 150081, 2015.
- [9] J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J. L. Arcos, “Measuring the evolution of contemporary western popular music,” *Scientific Reports*, vol. 2, pp. 1–6, 2012.
- [10] J. Fan and M. Casey, “Study of Chinese and UK hit songs prediction,” *10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pp. 640–652, 2013.
- [11] A. Singhi and D. G. Brown, “Hit song detection using lyric features alone,” in *Proceedings of International Society for Music Information Retrieval*, 2014.
- [12] M. J. Salganik, P. S. Dodds, and D. J. Watts, “Experimental study of inequality and cultural market,” *Science*, vol. 311, no. 5762, pp. 854–856, 2006.
- [13] E. Zangerle, M. Pichl, B. Hupfaut, and G. Specht, “Can microblogs predict music charts ? an analysis of the relationship between # nowplaying tweets and music charts,” in *Proceedings of International Society for Music Information Retrieval*, 2016.
- [14] K. Choi, G. Fazekas, and M. Sandler, “Automatic tagging using deep convolutional neural networks,” in *Proceedings of International Society for Music Information Retrieval*, 2016.
- [15] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1533–1545, 2014.
- [16] S. Dieleman and B. Schrauwen, “End-to-end learning for music audio,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 6964–6968.
- [17] J.-y. Liu and Y.-h. Yang, “Event localization in Music auto-tagging,” in *Proceedings of the 24th ACM international conference on Multimedia*, 2016.
- [18] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, “Evaluation of algorithms using games: the case of music tagging,” in *Proceedings of International Society for Music Information Retrieval*, 2009.
- [19] S. Kinoshita, T. Ogawa, and M. Haseyama, “Popular music estimation based on topic model using time information and audio features,” *IEEE 3rd Global Conference on Consumer Electronics (GCCE)*, pp. 102–103, 2014.
- [20] “lasagne,” [online] <https://lasagne.readthedocs.org/en/latest/>.
- [21] S. Dieleman and B. Schrauwen, “Multiscale approaches to music audio feature learning,” in *Proceedings of International Society for Music Information Retrieval*, pp. 116–121, 2013.
- [22] E. Shelhamer, J. Long, and T. Daeedll, “Fully convolutional networks for semantic segmentation,” *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [23] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet, “Going deeper with convolutions,” *arXiv preprint arXiv: 1409.4842*, 2014.
- [24] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, and W. Chen, “A theoretical analysis of NDCG ranking measures,” in *Proceedings of the Annual Conference on Learning Theory*, pp. 1–30, 2013.
- [25] M. Kendall and J. D. Gibbons, *Rank correlation methods*, vol. 3, Oxford University Press, 1990.