Applications of Deep Learning Network on Audio and Music Problems

Kanishka Tyagi, Student Member, *IEEE*,
Supervisor : Dr.Kyogu Lee

***Abstract*— An efficient pipeline procedure for designing unsupervised deep learning algorithms, multi label multi layer perceptron and stacked kernel classifiers is presented. By combing the deep learner with the kernel method we make optimal use of the localization property of the kernel methods as well as the distributed representation of the deep learners. Our models are tested on the three problems namely Multi label music tag classification, Audio scene classification and Bird audio classification. We analyze the experimental results that establish a relationship between the features captured by the proposed models and the musical and spectral dynamics of the music and audio signals.**

***Index Terms*—multi label, multi instance, sparse autoencoder, restricted Boltzmann machine, multi kernel.**

I. Introduction

DEEP LEARNING features are extremely useful in capturing various intrinsic details of the problem in hand. In this report we first explain the multi label music tag classification problem that essentially encompass all the basic formulations and building blocks required, in order to explain the remaining two problems.

II. Problem 1: Music Tag Classification

In the field of pattern recognition feature extraction techniques are usually task specific. We need to automatically extract useful features without being entirely dependent on the specific domain. For the last decade, kernel machines (SVM's) have been used for many speech, audio, image applications [1],[2],[3]. The performance of these algorithms is limited by the computational complexity for large problems where dimensionality of the input space is high. Obtaining high level task independent features has been a challenge in the machine learning community for decades.

In [4], a scalable Gaussian mixture model (GMM) based annotation and retrieval system is proposed. Its high dependence on the training data leads to the major problem of generating new music data. It's not only expensive but also requires a physiological setting to record clean data. [5] uses an approach where it adopts dynamic texture mixture (DTM) model to propose a hierarchical EM algorithm called HEM-DTM. It models the temporal dynamics and timbral contents and does not refer to a song as "bag of features.

Recent work by [6] Deep belief network (DBN) or deep learning has been actively used for feature learning in an unsupervised pre-training setting [26],[27]. Low level acoustic features that are obtained using conventional techniques can be fed into a DBN to develop more high level features that can better represent music signal. Can we learn the task independent high level features and use them to better represent the music signal ? We investigate this question in the present paper.

In order to train the DBN, [7] propose a fast greedy algorithm to train these unsupervised models and stacked on top, with a supervised algorithms such as softmax, SVM. However stacking the supervised algorithm and using them as prior is cumbersome. Whether the stacked classifier used after the pre-training is useful or does it have any effect on performance, is not well investigated. In [8] local kernels relying solely on smoothness prior leads to high sensitivity for curse of dimensionality. In this paper, we use support vector machine (SVM) [9] as our discriminative kernel based method. In literature, multi kernel has been used extensively in image classification [10] and two stage multi kernel learning [11].

A. *HOW IT IS DIFFERENT FROM PRIOR WORK* ?

In audio and music processing, classification problems using deep learning are mostly based on a routine method of extracting the features using deep learning and then stacking up with a classifier to get the results. [4] proposes a system in a multi class scenario that annotate a novel song by modeling heterogeneous dataset of music and words. However multi word queries are not considered. Usually most of the music tagging problems uses timbral and temporal features. [12] although this work has been automated in [13], using genetic algorithms but it's used for single class genre classification problem . DBN has been applied in [14]. In [15], DBN's are used in convolutional neural nets setting for genre and artist classification. None of these approaches uses the music vocal and timbre information to solve the task.

This work was supported by IEEE CIS Walter Karplus Summer Research Grant-2013 and Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea.

Kanishka Tyagi is with the Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX, 76013, USA (e-mail: kanishka.tyagi@mavs.uta.edu).

Kyogu Lee is with the Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea (e-mail: kglee@snu.ac.kr).

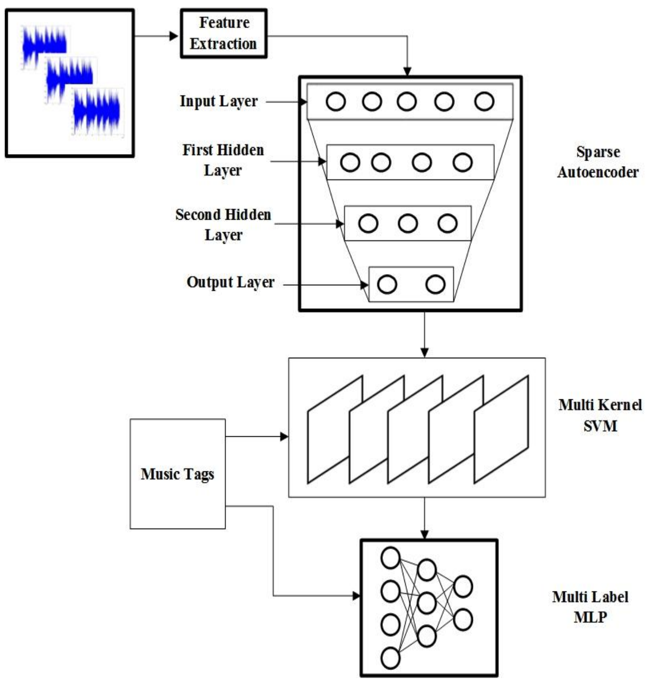

Figure 1. Overall proposed method. The neurons in SAE and MLP serves for illustrative purposes and does not imply the actual number.

In this report, we develop a streamline approach to extract the features that captures the vocal dynamics of music from a stacked sparse auto encoder (SAE) and use a supervised multi kernel and multi label MLP [16] learning to classify the music tags. The multi-kernel classifier uses appropriate priors [17] to handle data variability.

The specific goal of this work is to design a scheme for automatically using the conventional features as input basis functions for a deep learner to further obtain high level features. Later on these high level features obtain from the SAE based deep learner is feed into the multi kernel SVM and then taken as a multi label problem, solved using a multi label MLP. The proposed algorithm is called a multi label multi kernel sparse autoencoder (MLMK-SAE)

B. *PROPOSED ALGORITHM*

Figure 1 shows the proposed system. Features are extracted and fed into the unsupervised learning SAE to extract features which then goes through the supervised learning in Multi-Kernel SVM and multi label MLP.

1) Data preprocessing and feature extraction

Let music sample $\boldsymbol{F}$ be a $\boldsymbol{k}$ overlapping time series i.e $\boldsymbol{F} = \{\boldsymbol{f^1}, \dots, \boldsymbol{f^k}\}$, where $\boldsymbol{f^i}$ is a sequential $\boldsymbol{i^{th}}$ music feature vectors extracted to form a single instance feature. Here $\boldsymbol{k}$ is the total number of music feature. Let $\boldsymbol{v}$ be the number of total bag of tags of size $|\boldsymbol{v}|$. The content in the bag is represented by a vector $\boldsymbol{c} = [\boldsymbol{c_1}, \dots \boldsymbol{c_{|v|}}]$, where $\boldsymbol{c_k} > 0$ only if there is association between the song and the tag $\boldsymbol{\gamma_k}$, otherwise $\boldsymbol{c_k} = \boldsymbol{0}$. after the feature extraction part we have a dataset D for song feature tag pair as $(\boldsymbol{F}, \boldsymbol{c})$.

We follow a frame based approach to extract the raw features from the waveform as depicted in Figure 1. Te numerical values below refer to CAL-500 [18].

Step 1: The audio signal is sliced into frames of 128 ms window with a hop size of 20ms.

Step2: The bag of frames are then stacked together in a group of 20 to form each instance. Each bag the hop size is 10 frames.

Step 3: For every frame we extract the MFCC's and mel-scale spectrum features. each coefficient is sorted over the frames to extract the features from each bag.

Step 4: To make the single-instance features $\boldsymbol{F}$, we calculate the mean and standard deviation values over the bags.

Step 5: The feature vector $\boldsymbol{F}$ is then fed to a 3-layered deep sparse autoencoder to obtain some high level features

2) Sparse Autoencoder

We use a greedy layer-wise unsupervised pre-training for the SAE [7]. We not only use the smoothness prior but also the sparsity prior is used while designing the algorithm. Since we train the SAE using a self taught learning as in [19], we do not assume the same distribution for the unlabelled and labeled data. By doing so, we design an independent framework for feature extraction which can then be plugged into any classifier. This makes our algorithm robust and universal in nature.

Once we get the k input features $\{\boldsymbol{f^1}, \dots, \boldsymbol{f^k}\}$ with each $f^i \in \mathbb{R}^n$, we minimize the following objective function:

$$min_{w,a} \sum_i \left\| \boldsymbol{f^i} - \sum_j a_j^i w_j \right\|_2^2 + \frac{\lambda}{2} \sum_j \left\| w_j \right\|^2 + \beta \left\| a^i \right\|_1$$
$$s.t. \left\| w_j \right\|_2 \leq 1 \ \forall j \in 1, \dots s. \qquad (1)$$

The optimization variables are the basis vector $\boldsymbol{w} = \{w_1, \dots, w_s\}$with each $w_j \in \mathbb{R}^n$, and the activations $\boldsymbol{a} = \{a^1, \dots, a^k\}$ with each $a^i \in \mathbb{R}^s$. The number of basis s can be much larger than the input dimension n. The L_1norm in above equation takes care of the sparsity. We choose the activation function to be hyperbolic tangent function. λ is the weight decay parameter and β is the weight of the sparsity penalty term. After solving to obtain bases $\boldsymbol{w}$, we computer features for the multi kernel SVM classifier as $\{(\hat{a}(x_l^i), y^i)\}_{i=1}^m$ where

$$\hat{a}(x_l^i) = argmin_{a^i} \left\| \boldsymbol{f_l^i} - \sum_j a_j^i w_j \right\|_2^2 + \beta \left\| a^i \right\|_1 \qquad (2)$$

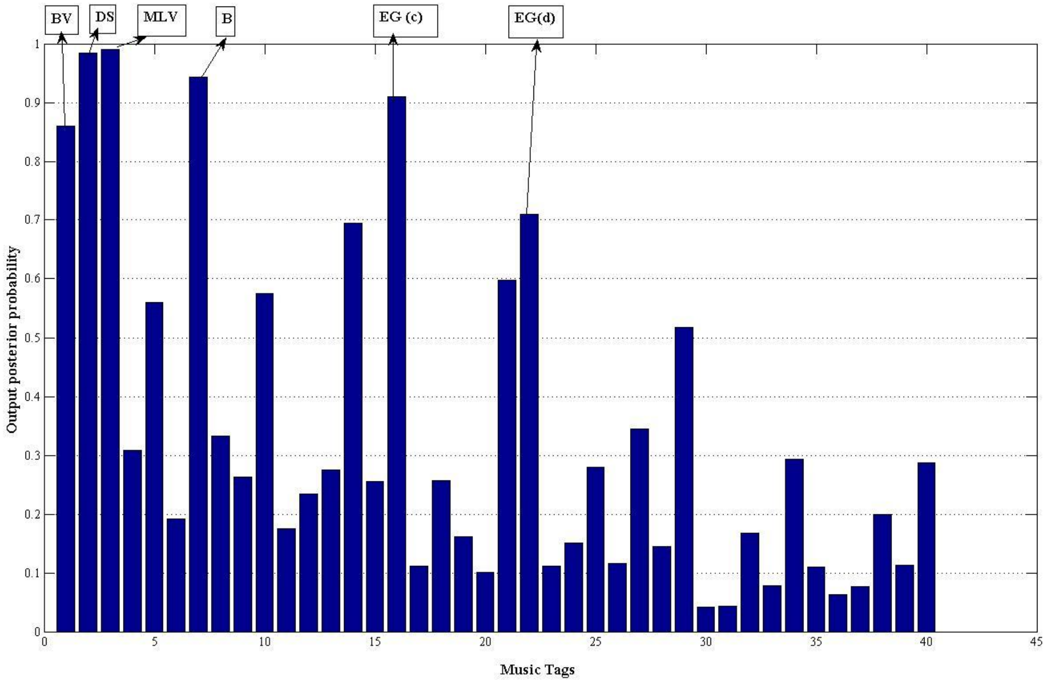

Figure 2: Multinomial distribution over the Genre and Instruments Tags in our dictionary for Stevie Ray Vaughan's "Pride and Joy". the 6 most probable tags have been labeled. Here BV: Backing vocals, DS: Drum set, MLV: Male lead voice, B: Bass, EG(c): Electric guitar (clean) and EG(d): Electrical guitar(distorted)

Table 1 describes various parameters used to train the deep learner. In our settings, we use the Limited memory BFGS algorithm [25],[28] to train the deep learner because of their ability to train easily and it enjoys parallelism by computing the gradient n GPU's [20]. We compute the features after each

layer has been trained and we use unlabeled data to extract the features which then get fed into the multi kernel SVM .

Table 1. Parameter statistics for stacked sparse auto encoder

Number of hidden layers	3
Neurons per layer	300-100-30
Sparsity parameter	0.1
Weight decay parameter	0.003
Sparsity penalty term	3
Number of unsupervised iterations	400
Training time (in hours)	30

3) Multi Kernel SVM

After the unsupervised training part, from a statistical point of view, the algorithm solves a supervised multi label problem. [21],[22]. While designing the kernel machine on top of a deep learner, even though the number of training examples can be large but the resulting basis functions from the deep learner part has low kolmogorov-chaltin complexity, i.e the target function can be modeled with a much fewer number of basis functions. Conventional kernel based algorithms are local in nature and uses smoothness as the only prior. Our approach has lead to a design of algorithms that is not only non –local in nature but also uses general priors apart from the smoothness In other words we increase the prediction bound in a Bayesian sense i.e. we measure the priors over the set of five different kernels [23] i.e. linear, polynomial, quadratic, radial basis function and multilayer perceptron, to expresses the degree to which one classifier will predict well over the other.

4) Multi label MLP

While using the multi-label MLP, our design philosophy is entirely a Bayesian form of probability rather than the frequentist or estimated sense. Therefore while deciding on the "degree of belief" in the output of the multi-kernel, we see this problem from a Bayesian point of view. Hence in order to get the inference from the output of the multi kernel, we use a multi label MLP.

C. EXPERIMENTAL VALIDATION

1) Music dataset

For the evaluation of the algorithm, we use CAL500 [22] dataset which is human-generated musical annotations of 502 popular western musical tracks. Out of 149 tags that are annoted by three humans, we tried to classify 40 tags that describes genres and instruments In our present investigation, we choose these two tags over the others due to their non-perceptual nature. the other tags present in the dataset are highly subjective and hence not a part of present work. Table 3 present the f-score results for the tags and compare the proposed algorithm performance with others. As a comparison algorithm, we used stacked restricted Boltzmann machine (SRBM) [6] and HEM-DTM and HEM-Gaussian mixture models (HEM-GMM) [24].

Table 2: Annotation results for various algorithms on CAL-500 dataset

	Precision	Recall	F-score
MLMK-SAE	0.25	**0.78**	**0.32**
SRBM	0.24	0.75	0.30
HEM-GMM	**0.49**	0.23	0.26
HEM-DTM	0.47	0.25	0.30

The most important features to represent the vocal are timbre, or the spectral features. From Table 2 we see that SAE-MK is better in capturing the timbre or spectral characteristic of the vocals in a song. this is evident from the F-score table where we have high score values for tags that involves vocal. however it requires more training examples to reliably estimate female vocal tags. this is clear from its low F-score values. Hence we can conclude that SAE shows better performance than other algorithms and they are particularly useful for tags that have timbre characteristics.

Table 3. F-score results for some tags

Tag	HEM-DTM	HEM-GMM	MLMK-SAE	SRBM
F-score				
Female lead vocals	0.58	0.42	0.25	0.22
Male lead vocals	0.44	0.08	**0.87**	0.81
Backing vocals	0.30	0.09	**0.48**	0.33
Pop	0.32	0.31	0.12	0.15
Acoustic guitar	0.44	0.31	0.22	0.19
Electrical guitar	0.32	0.14	**0.39**	0.32

Figure 2 shows the multi label MLP output for Stevie Ray Vaughan's "Pride and Joy". If we set a threshold at 0.7, we can observe that the vocal tags are predominant. Table 4 shows the different tags annotation for a song from Stevie Ray Vaugham 'Pride and joy'. We see that HEM-DTM and HEM-GMM are unable to get all the ground truth correctly as compared to SAE-MK, however the genre is identified correctly by all the algorithms.

Table 4. Tag annotations for songs in CAL-500 dataset. ground truth is marked black

Stevie Ray Vaughan	*Pride and Joy*	
Algorithm	Instrument	Genre
HEM-DTM	**Drum set**,	alternative
HEM-GMM	synthesizer	alternative
MLMK-SAE	**Drum set**, **male lead vocal, bass,** backing vocals.	alternative
SRBM	male lead vocals, bass	alternative

D) DISCUSSION AND CONCLUSION

In this paper, multi label tag classification and multi instance features from music signals has been investigated. We see from Table 2 that the F-score value, which gives an estimate about the recall and precision value, is better for the proposed model in vocal tags. The MLMK-SAE is more optimized for the timbre and vocal features. This can be understood as we took the mean over the frames which can be considered as a summation of the timbre features. As an example, we see in Table 4 that the proposed algorithm performs better than HEM-DTM and HEM-GMM when it comes to instruments tagging. We see from Tables 2 and 3 that SRBM performs same as MLMK-SAE, the only difference comes in implementation and ease of training for SAE. In our experience and [28], SRBM and MLMK-SAE have similar performance in most of the music related tasks. Our approach in this work handles the problem of multi labels via high level features from a deep learner and multi kernel SVM's. Finally the result from all the SVM's are fed into a multi-label MLP to get the required outputs

II. PROBLEM 2: AUDIO SCENE CLASSIFICATION :

Acoustic scene classification is a subclass to a wide set of algorithms in machine learning called computational auditory scene analysis (CASA). the goal in this task it to provide a semantic label to the environment in which the audio is recorded [29]. The methodology we used in this task is to treat the scene as a single object and principal component analysis and whitening as pre-processing. then generate features from the deep learner and use them to feed to SVM as classifier. the results are have been tabulated below. The data is provided from the AASP challenge [35]

1) *SCENE CLASSIFICATION DATASET*:

The dataset for the scene classification (SC) challenge consists of 30 sec recordings of various acoustic scenes. The scene classification dataset consists of 2 equally proportioned parts each made up of 10 audio recordings for each scene (class), for a total of 100 recordings per part. The public dataset is published on the C4DM Research Data Repository [30]. The scenes include:

1) Busy street
2) Quiet street
3) Supermarket/store
4) Restaurant
5) Office
6) Park
7) Bus
8) Tube/metro
9) Tubestation
10) Open market

For each scene type, three different recordists visited a wide variety of locations in Greater London over a period of months (Summer and Autumn 2012), and in each scene recorded a few minutes of audio. No systematic variations in the recordings co varied with scene type: all recordings were made in moderate weather conditions, and varying times of day and week, and each recordist recorded each scene type. 30-second segments were selected after careful review of the recordings to ensure they were free of issues such as mobile phone interference or microphone handling noise

Table 5: Confusion matrix for stacked restricted boltzmann machine (SRBM)

Result-----> Data	Bus	Busy street	Office	Open market	Park	Quiet street	Restaurant	Supermarket	tube	tube station
Bus	**1**	**1**	0	0	0	0	0	0	0	0
Busy street	0	**2**	0	0	0	0	0	0	0	0
Office	0	0	**2**	**0**	0	0	0	0	0	0
Open market	**1**	0	0	**1**	0	0	0	0	0	0
Park	0	0	0	0	**2**	0	0	0	0	0
Quiet street	0	0	0	0	**1**	0	**1**	0	0	0
Restaurant	0	0	0	**1**	0	0	0	**1**	0	0
Supermarket	0	0	**1**	**0**	0	0	0	**1**	0	0
Tube	**0**	0	**0**	0	0	0	0	0	**2**	0
Tube station	0	0	0	0	0	0	0	0	0	**2**

Table 6: Confusion matrix for stacked sparse auto encoder (SAE)

Result-----> Data	Bus	Busy street	Office	Open market	Park	Quiet street	Restaurant	Supermarket	tube	tube station
Bus	**1**	**1**	0	0	0	0	0	0	0	0
Busy street	0	**2**	0	0	0	0	0	0	0	0
Office	0	0	**2**	0	0	0	0	0	0	0
Open market	0	0	0	**2**	0	0	0	0	0	0
Park	0	0	0	0	**1**	**1**	0	0	0	0
Quiet street	0	0	0	0	**2**	**0**	**0**	0	0	0
Restaurant	0	0	0	0	0	0	0	**2**	0	0
Supermarket	0	0	0	**1**	0	0	0	**1**	0	0
Tube	**0**	0	0	0	0	0	0	0	**2**	0
Tube station	0	0	0	0	0	0	0	0	0	**2**

2) RESULT AND DISCUSSION

For this train/test task, participating algorithms are evaluated using 5-fold cross validation. Tables 5 and 6 shows the confusion matrix for the performance of the SRBM and SAE features based SVM classifier on the testing data respectively. The system is able to classify busy street , office, park, supermarket and tubestation correctly to distinguish between two classes, some higher level correlation may be needed. the overall classification percentage for SRBM is 65 % and SAE as 67 % . the low classification accuracy is attributed to less training data for feature generation with deep learner.

III. PROBLEM 3 : BIRD CLASSIFICATION PROBLEM

This is a part of MLSP 2013 Bird Classification challenge on kaggle [34]. Identifying bird sounds based on audio data collected in an acoustic monitoring scenario is an open problem in machine learning domain to correctly. Some of the major challenges include multiple simultaneously vocalizing birds, other sources of non-bird sound (e.g. buzzing insects), and background noise like wind, rain, and motor vehicles. in our investigation we observe that by using multiple features from the deep learner and classifying using a multi kernel and multi label MLP these problems can be overcome substantially.

The problem is to classify a set of species present in a audio recording. Apart from being a classic multi instance multi label (MIML) supervised classification problem, we used various techniques including unsupervised deep learner to achieve the desired results. A basic MIML framework is show as in Figure 3.

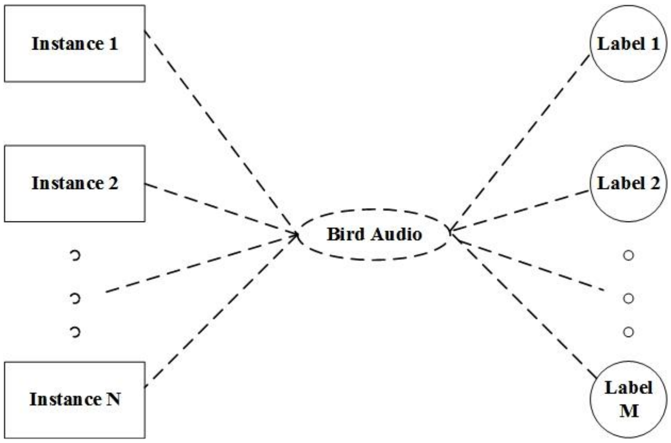

Figure 3 : Basic multi instance multi label framework

The audio dataset for this challenge was collected in the H. J. Andrews (HJA) Long-Term Experimental Research Forest, in the Cascade mountain range of Oregon. Since 2009, members of the OSU Bioacoustics group have collected over 10TB of audio data in HJA using Songmeter audio recording devices. [31] The segmented spectrogram were given as part of data in the competition. Figure 3 shows a segmented spectrogram

The segmented patches in a spectrogram can be viewed as multiple instances in a single audio. Based on these segmented spectrogram, we implemented algorithms as in [33] for extracting single instance form the multi instance audio file. This single instance data file is then used to feed into the stacked sparse auto encoder (SAE) is used to extract the features. The features from the SAE is then used to feed into the multi kernel SVM whose output is then fed into the multi label Multilayer perceptron. The main challenge here was to improve the classifier accuracy by develop heuristic for this specific problem. we used a MLP based MIML algorithms as in [16] Another challenge we faced was to formulate the problem of MIML framework for supervised classification to be aligned features generated from the SAE.

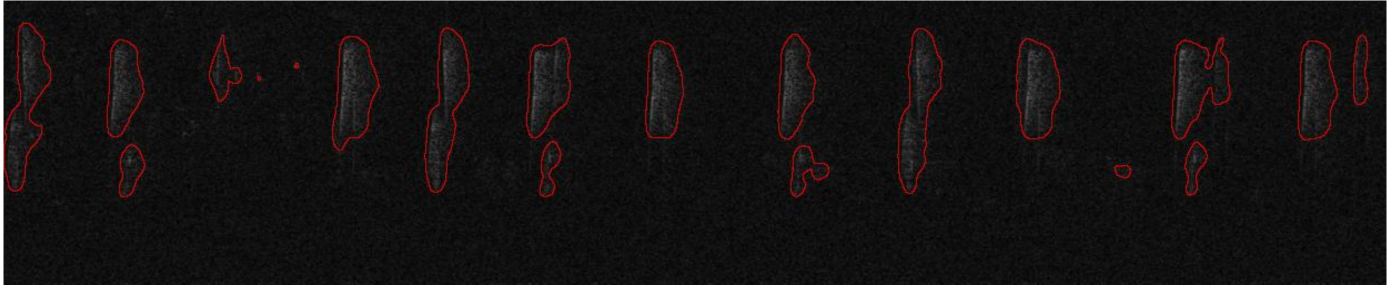

Figure 4: Segmented spectrogram from a bird audio. Each segmentation is an instance that contain different sound including various bird sounds.

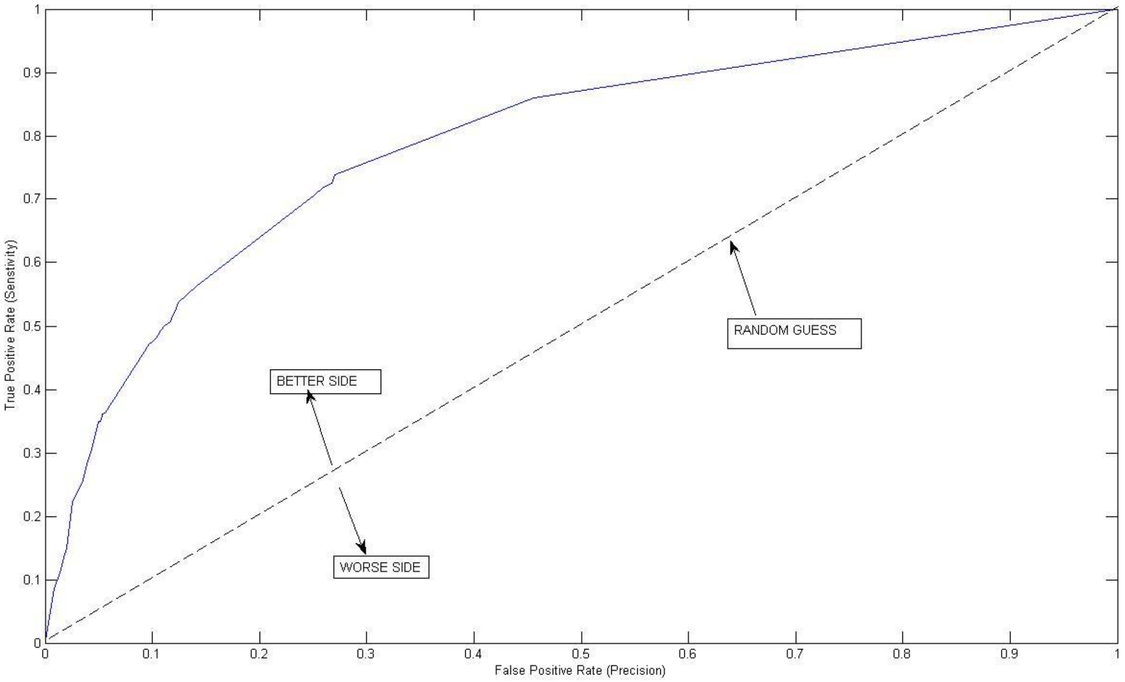

Figure 5: ROC curve for multi kernel SVM

A) RESULTS WITH VARIOUS ALGORITHMS:

As an exhaustive investigation for the bird classification problem, we experimented with various models, the results for which are discussed below.

1) Stacked Restricted Boltzmann Machine (SRBM) Deep Learner :

In order to handle the multi label problem, this technique was tried after giving each the audio file with labels, 19 different times. That means we have to train 19 different RBM's and get it tested on a same testing data. The paradigm used here is same as used for Audio scene classification problem. The performance was 61% for the testing data and it was calculated after using the ROC curve and heuristic.

2) Sparse auto encoder (SAE) Deep Learner:

A similar procedure as the SRBM is followed on this and the results was calculated using the ROC curve to gives an error of 63 %.

3) Multi instance multi label MLP:

This code was designed for multi instance and multi label feature vector given by kaggle dataset. It gives 58 % classification result.

4*) Multi kernel support vector machines:*

This was the most promising among all the algorithms. this worked well and give 68 % classification performance. The location of the bird in the map given is an important heuristic. We design a soft classifier to the kernel output that includes the heuristic (as mentioned above) and draw the ROC curve as shown in Figure5, the performance was increased to 81 % .

The strategy to change a multi instance multi label to single instance multi label was successfully implemented [16] and used in the SVM's mentioned above. Finally as part of submission for the competition, we choose the multi kernel based SVM.

B) WHY DEEP LEARNING PERFORMANCE FAILED ?

One reason is that the two layer RBM or SAE is learning only low level features. To learn the high level features we not only need more deeper architecture but we also need more and large amount of data.[32]

IV CONCLUSION

Music and audio is an interesting and challenging area to work. Most of the algorithms used in music processing are imported from the field of audio. In the current work, we design and develop multi kernel SVMs, stacked RBM's, stacked AE's and multi label MLP's. We show the application of these algorithms on three major tasks namely, music tag classification, audio scene classification and bird audio classification.

V. FUTURE WORK

For music tag classification, we'll develop additional deep learning features and investigate how the performance of the proposed algorithm changes if deep learner features from an entirely different music data In audio scene analysis, we need to improve the feature extraction by training on much larger data in an unsupervised fashion so as to improve the performance of deep learner. In the bird classification, we would like to add better segmentation algorithm that can capture more instances so as to improve the feature extraction performance. For all the three tasks, an open question of developing efficient training algorithm for deep learning apart from gradient descent is an important direction of research.

VI. ACKNOWLEDGMENT

The present research work was partially supported by IEEE Computational Intelligence Society Walter Karplus Summer Research Grant 2013 and Graduate school of Convergence Science, Seoul National University, South Korea. The author would like to thank Il-Young Jeong and Dr.Michael Manry for there feedback that resulted in the successful completion of this research.

REFERENCES

[1] Aucouturier, Jean-Julien, François Pachet, and Mark Sandler. "The way it Sounds": timbre models for analysis and retrieval of music signals." *IEEE Transactions on Multimedia* 7.6 (2005): 1028-1035.

[2] M. Mandel, G. Poliner, and D. Ellis. Support vector machine active learning for music retrieval.

[3] Shawe-Taylor J. and Cristianini N., Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge,2004

[4] Turnbull, Douglas, Luke Barrington, and Gert Lanckriet. "Modelling music and words using a multi-class naıve bayes approach." *Proc. of ISMIR*. 2006.

[5] Coviello, Emanuele, et al. "Automatic Music Tagging With Time Series Models." *ISMIR*. 2010.

[6] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507

[7] Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks." *Advances in neural information processing systems* 19 (2007): 153.

[8] Bengio, Yoshua, Olivier Delalleau, and Nicolas L. Roux. "The curse of highly variable functions for local kernel machines." *Advances in neural information processing systems*. 2005.

[9] Vapnik, Vladimir. The nature of statistical learning theory. springer, 2000.

10] Siddiquie, Behjat, Shiv N. Vitaladevuni, and Larry S. Davis. "Combining multiple kernels for efficient image classification." *Applications of Computer Vision (WACV), 2009 Workshop on*. IEEE, 2009.

[11] Kumar, Abhishek, et al. "A binary classification framework for two-stage multiple kernel learning." *arXiv preprint arXiv:1206.6428* (2012).

[12] T. Bertin-Mahieux, D. Eck and M. Mandel, Automatic tagging of audio: The state-of-the-art., *Machine Audition: Principles, Algorithms and Systems*. IGI Publishing, 2010

[13] Mierswa, Ingo, and Katharina Morik. "Automatic feature extraction for classifying audio data." *Machine learning* 58.2-3 (2005): 127-149.)

[14] Hamel, Philippe, Sean Wood, and Douglas Eck. "Automatic Identification of Instrument Classes in Polyphonic and Poly-Instrument Audio." *ISMIR*. 2009.

[15] Lee, Honglak, et al. "Unsupervised learning of hierarchical representations with convolutional deep belief networks." *Communications of the ACM* 54.10 (2011): 95-103.

[16] Zhou, Zhi-Hua, and Min-Ling Zhang. "Multi-instance multi-label learning with application to scene classification." *Advances in Neural Information Processing Systems*. 2006.

[17] Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." (2013): 1-1.

[18] Turnbull, Douglas, et al. "Semantic annotation and retrieval of music and sound effects." *Audio, Speech, and Language Processing, IEEE Transactions on* 16.2 (2008): 467-476.

[19] R., Battle, A., Lee, H., Packer, B., and Ng, A.Y. Self-taught learning: Transfer learning from unlabelled data. In ICML, 2007.

[20] Raina, Rajat, Anand Madhavan, and Andrew Y. Ng. "Large-scale deep unsupervised learning using graphics processors." *ICML*. Vol. 9. 2009

[21] [Whitman, Brian, and Ryan Rifkin. "Musical query-by-description as a multiclass learning problem." *Multimedia Signal Processing, 2002 IEEE Workshop on*. IEEE, 2002.

[22] Turnbull, Douglas, et al. "Semantic annotation and retrieval of music and sound effects." *Audio, Speech, and Language Processing, IEEE Transactions on* 16.2 (2008): 467-476.

[23] Scholkopf, B., and Smola, A.J., Learning with Kernels, MIT Press, Cambridge, MA. 2002.

[24] Coviello, Emanuele, Antoni B. Chan, and Gert Lanckriet. "Time series models for semantic music annotation." *Audio, Speech, and Language Processing, IEEE Transactions on* 19.5 (2011): 1343-1359.

[25] Lee, H., Battle, A., Raina, R., and Ng, Andrew Y. Efficient sparse coding algorithms. In NIPS, 2007.

[26] Bengio, Yoshua. "Learning deep architectures for AI." *Foundations and trends® in Machine Learning* 2.1 (2009): 1-127.

[27] Bengio, Yoshua, and Yann LeCun. "Scaling learning algorithms towards AI."*Large-Scale Kernel Machines* 34 (2007).

[28] Ngiam, Jiquan, et al. "On optimization methods for deep learning." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.

[29] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *Journal of the Acoustical Society of America*, vol. 122, pp. 881, 2007

[30] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "*Detection and classification of acoustic scenes and events*," an IEEE AASP Challenge, 2013,www.elec.qmul.ac.uk/digitalmusic/sceneseventschallenge).

[31] Briggs, Forrest, et al. "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach." *The Journal of the Acoustical Society of America* 131 (2012): 4640.

[32] Le, Quoc V., et al. "Building high-level features using large scale unsupervised learning." arXiv preprint arXiv:1112.6209 (2011).

[33] Zhang, Min-Ling, and Zhi-Hua Zhou. "Multilabel neural networks with applications to functional genomics and text categorization." *Knowledge and Data Engineering, IEEE Transactions on* 18.10 (2006): 1338-1351.

[34] http://www.kaggle.com/c/mlsp-2013-birds

[35] http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/