2018 Third International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)
14-15, December 2018

Music Instrument Identification Based on a 2-D Representation

Alekhya Ghosh,
Electronics and Communication Engineering
IRPEL, University of Calcutta
Kolkata, India
alekhyaghosh.cu@gmail.com

Dibakar Sil
Electronics and Communication Engineering
National Institute of Technology,
Durgapur, India
dibakar.sil2@gmail.com

Arghadeep Pal
Electronics and Communication Engineering
IRPEL, University of Calcutta
Kolkata, India
parghadeep@gmail.com

Sarbani Palit
Computer Vision and Pattern Recognition Unit
Indian Statistical Institute,
Kolkata, India
sarbanip@isical.ac.in

***Abstract*— Automatic recognition of musical instruments has been a great problem in the field of digital signal processing. In this work a novel, compact set of features obtained by transforming a music signal to the spatial domain is introduced with promising results. Preliminary results on a popular music database, presented in terms of precision, recall and identification accuracy, are highly encouraging. The best result comprises an accuracy of 84.02% by Decision Tree (DT) for a set of 9 instruments belonging to different families. The accuracy for predicting the instrument family 96.07% for string family and for wind instrument the overall prediction accuracy is 90.78%. Further inclusion of some other features from some available works are also checked leading to slight increase in the accuracy. The obtained results are also compared with the result predicted by the VGG – 16 network on the same dataset. The feature set containing 8 features that introduces less computational complexity compared to available works, may be considered as a major contribution. The comparison shows, we can replicate nearly accurate classification through our proposed method in less amount of time than deep neural architecture.**

Keywords*— *music instrument recognition; spatial features; random forest; decision tree.

I. INTRODUCTION

Automatic identification of instruments from a music excerpt is one of the biggest challenges in the field of Music Information Retrieval (MIR). The timbre differences of various instruments that are easily detected by a trained musician are quite difficult to identify through signal processing and mechanical methods. A general approach is to extract some efficient features from the duly preprocessed music signal and then employ some classification algorithm.

Over the decades different kinds of features have been constructed for this task. Music signals have been projected on different domains for this purpose. Time and frequency domain properties such as rise time, slope of line fitted into rms-energy curve after attack, mean of spectral centroid, average cepstral coefficients during onset were investigated for instrument recognition [1,2]. MFCC, linear prediction and delta cepstral coefficients resulted in 35% accuracy for recognition of individual instruments and 77% for identification of instrument families. Gaussian Mixture Models (GMM) and k-Nearest Neighbor (k-NN) model classifier were employed for instrument classification using line spectral frequencies (LSF) as features [3] produced instrument family recognition of 95% and 90% at the individual instrument level. For polyphonic audio signals, instrument recognition is done with the help of source filter mode modelled on Mel frequency scale [4] yielded 59% accuracy.

Timbre classification based on spectral centroid and spectral fine structure has been used for perpetual relevance of the acoustic [5]. Fractional Fourier Transform based Mel-frequency cepstral coefficient (MFCC) features forming the input to a Counter Propagation Neural Network helped to maximize the discrimination between interclass instruments and minimize that of intra-class instruments [6]. A Convolutional Neural Network framework is very useful in instrument recognition in polyphonic music [7] yielding an improvement of 23.1% and 18.8% in performance. The problems related to automatic music transcription and audio source separation is dealt in [8]. Autocorrelation based method can be used to tackle with monophonic audio transcription (a single note is present at a time). This paper highlights the fact that a blackboard model or a multiple-cause/sparse coding method which are related to independent component analysis (ICA) can be used to deal with polyphonic music transcription (several notes are present at any time). For efficient and compact analysis, wavelet coefficient histogram features were used for recognition of 18 musical instruments [9]. It resulted in an accuracy of 76.83% which exceeds that of MFCC and other features (73.82%). Frequency and wavelet domain analysis was used to create the feature set for distinguishing instruments within same family [10].

The available works deal with a large set of features of different domains while achieving low accuracy level within the same instrument class. In this paper, we aim to develop a compact and efficient spatial feature set employing a novel two-dimensional representation of a

signal. A considerable reduction of computational complexity is achieved thereby. The method is described in Section II with a detailed description of the proposed spatial feature set in Section II-B. Section III presents the results and comparison with a recent approach. Section IV concludes the paper.

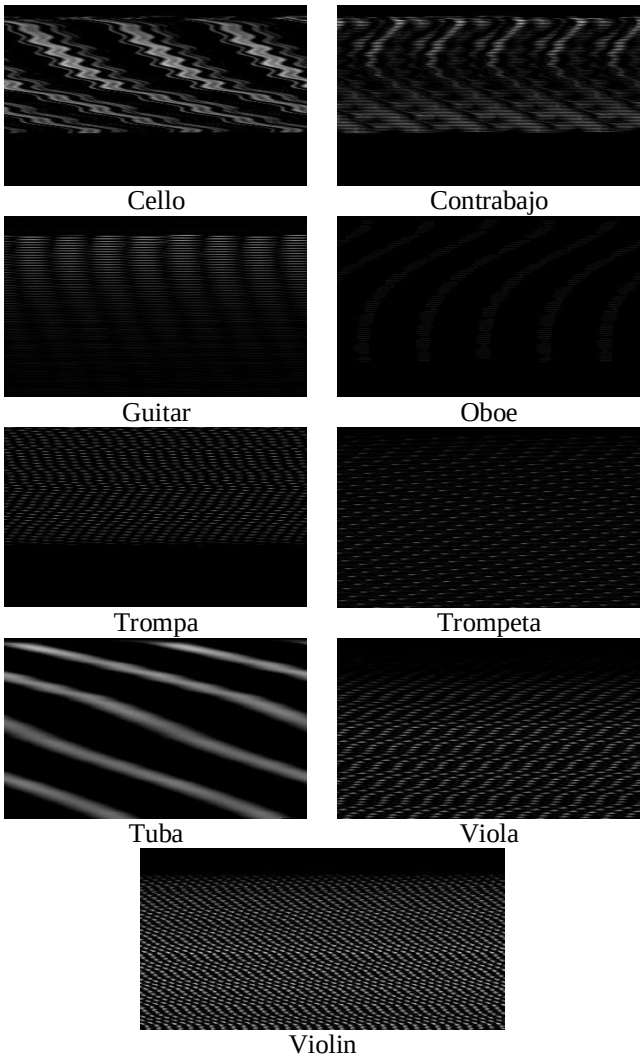

Fig. 1. Pictorial representation of different notes of different instruments.

II. The Proposed Approach

Notes form the basic unit of music which can combine to form different musical pieces. The main characteristics of notes are pitch, loudness and timbre. Out of these, it is timbre which causes the sounds generated from different instruments to sound different. The conventional Attack-Decay-Sustain-Release envelop of the notes varies from instruments to instruments reflecting that the timber differences among them may be studied from their time domain as well as frequency domain representations. As the ADSR frames show significant differences for different instruments so it is apparent that the distribution of the values will not be identical in the newly formed matrices for various instruments as shown in the pictorial representation of the matrices in Fig. 1. Signals corresponding to various notes of musical instruments are used as the inputs for the proposed approach.

A. *Configuring the Signal into an Image Matrix:*

The one-dimensional music signal (corresponding to a note) is converted into a matrix. From each of the signals a square matrix is formed in such a way that, there is no loss of signal. It is done in the following manner:

Step1: Z_1 number of zeros is appended to the signal where

$Z_1 = (\lceil(\sqrt{N})\rceil + 1)$ and N is the length of the original signal. So size of the new signal, $N_1 = (Z_1 + N)$.

Step 2: An $(s \times s)$ matrix I is formed from the signal $A(n)$, where $n = 0, \dots, s^2 - 1$, $and \quad s = \sqrt{N_1}$ such that $I(i,j) = A\big((i \times s) + j\big)$, $0 < i,\ j < (s-1)$.

Hence,

$$I = \begin{bmatrix} A(0) & A(1) & \cdots & A(s-1) \\ A(s) & A(s+1) & \cdots & A(2s-1) \\ \vdots & \vdots & \cdots & \vdots \\ A(s^2-s) & A(s^2-s+1) & \cdots & A(s^2-1) \end{bmatrix} \quad (1)$$

Matrices obtained from notes of various instruments, have been displayed as intensity images in Fig.1. maintaining the Integrity of the Specifications.

B. *Feature Extractions:*

For the classification of the images thus obtained, we considered the statistical features defined in [11], [12] for textural analysis, applied on gray level co-occurrence matrices. In this paper, however, we extracted these features directly from the images.

Denoting $I(i,j)$ as the $(i,j)^{th}$ entry of the matrix, let $I_x(i) = \sum_{j=1}^{s} I(i,j)$, $s \times s$ is the dimension of the matrix.

$I_y(j) = \sum_{i=1}^{s} I(i,j)$

$I_{x+y}(l) = \sum^{s}_{\substack{i=1 \\ i+j=l}} \sum_{j=1}^{s} I(i,j) \quad where, l = 2,3, \dots\dots, 2s$

$I_{x-y}(l) = \sum^{s}_{\substack{i=1 \\ |i-j|=k}} \sum_{j=1}^{s} I(i,j) \quad where, l = 2,3, \dots\dots, 2s$

The following features are now computed as,

1) *Auto-correlation:* $\boldsymbol{A} = \sum_i \sum_j (ij) I(i,j)$

2) *Correlation:*

$$\text{corr} = \left\{\frac{\sum_i \sum_j (ij) I(i,j) - \mu_x \mu_y}{(\sigma_x \sigma_y)}\right\} \quad (2)$$

where $\mu_x, \mu_y, \sigma_x, \sigma_y$ are the mean and standard deviation of I_x, I_y respectively.

3) *Sum Average:* $S_{av} = \sum_{i=2}^{2s} i I_{x+y}(i)$ (3)

4) *Sum Variance:* $S_{var} = \sum_{i=2}^{2s} (i - \varphi)^2 I_{x+y}(i)$ (4)

where, $\varphi = -\sum_{i=2}^{2s} I_{x+y}(i) \log I_{x+y}(i)$ is the entropy and, $I_{x+y}(l) = \sum_{i=1}^{s} \sum_{j=1}^{s} I(i,j)$.

5) *Difference Average: D = variance of* I_{x-y}.

6) *Normalised Inverse Difference:*

$$\boldsymbol{I} = \sum_{i,j=1}^{s} \left[\frac{I(i,j)}{1 + \{\frac{(i-j)^2}{s^2}\}}\right], \quad (5)$$

7) *Information Measure of Correlation:*

$$IMC = (1 - \exp[-2.0(E_{xy_2} - E_{xy})])^{1/2} , \quad (6)$$

where, $E_{xy} = -\sum_i \sum_j I(i,j) \log\big(I(i,j)\big)$ and

$$E_{xy_2} = -\sum_i \sum_j I_x(i)\, I_y(j) \log\{I_x(i) I_y(j)\}$$

8) Entropy: : $E = -\sum_i \sum_i I(i,j) \log_2 I(i,j)$. (7)

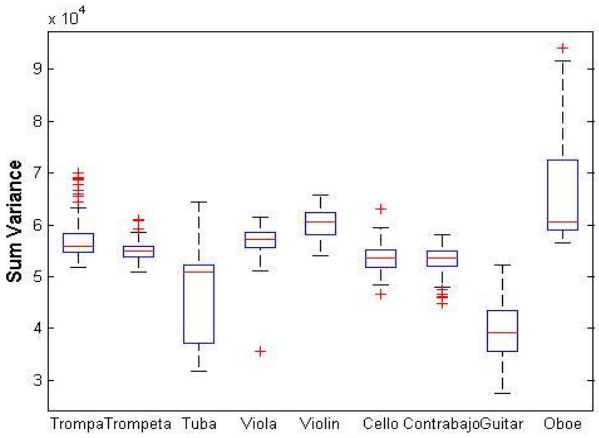

Fig. 2. Box plot of Sum Variance of the spatial representation of the solo notes.

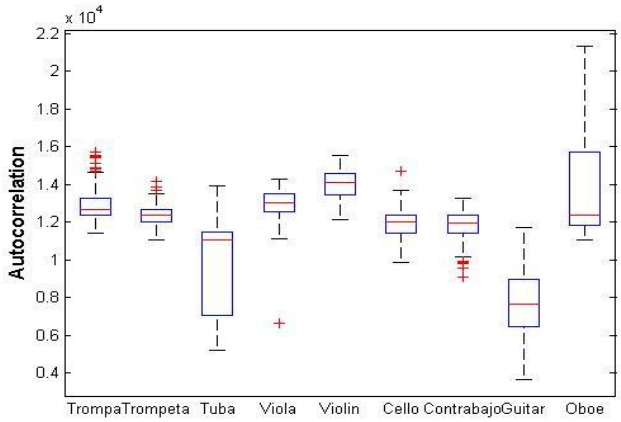

Fig. 3. Box plot of Autocorrelation of the spatial representation of the solo notes.

From Fig.2 and Fig.3, it can be seen that some of the instruments namely, Viola, Violin, Guitar, Oboe and others can be separated using the features, but other instruments have some overlapping. So considering the difference of each of the instruments, in this paper, a set of 8 spatial features are proposed which proved to be efficient in classification.

III. RESULTS

The dataset used for the present study contains individual notes obtained from 9 different instruments namely Cello, Contrabajo, Guitar, Oboe, Trompa, Trompeta, Tuba, Viola, and Violin [13], each of about 2 second duration and sampling frequency 22050Hz. The instruments belong to either of the string and wind instrument groups. Fig.2 shows a boxplot of the feature 'Sum variance' for all the instruments. It may be observed that though this feature is sufficient for distinguishing between Violin and Contrabajo but it is not sufficient when all are considered. So a combination of the features described in Section II-B is used for the classification task.

The efficiency of the proposed feature set containing the features as proposed in section II-B is compared to that of a VGG – 16 network. Each images (I) are resized to a (224 × 224) image before they are applied to the VGG – 16 network.

Two different classifiers, namely Decision Tree [14] and Random Forest [15] have been employed for the classification task. 50% of the dataset have been used for training and 50% have been used for testing. The outcome of the classifiers have been analysed in terms of Precision and Recall, defined as:

$$\text{Precision, } P = \frac{(True\ positive)}{(True\ positive + False\ positive)} \quad (8)$$

$$\text{Recall, } R = \frac{(True\ positive)}{(True\ positive + False\ Negative)} \quad (9)$$

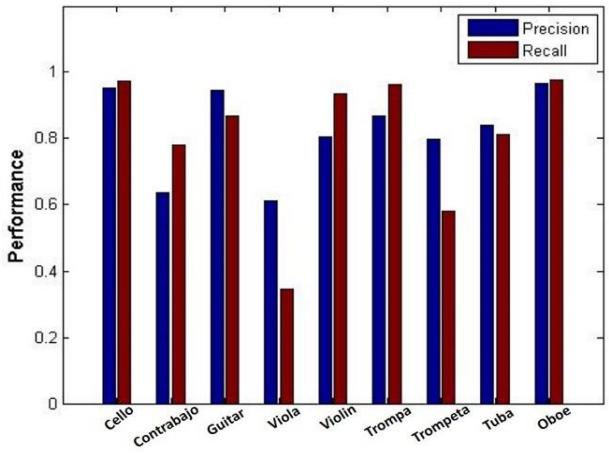

Fig. 4. Performance using Decision Tree in terms of Precision and Recall.

A bar chart of these values computed for the notes of all the instruments using Decision Tree, is shown in Fig.4. It may be observed that the proposed feature set yields the best precision and recall for Oboe. The precision value is higher than the recall for Guitar, Viola, Trompeta and Tuba. On the other hand, Contrabajo, Cello, Violin, Oboe and Trompa show the reverse trend. It is interesting to note that, though recall is low for Viola, its precision is fairly high.

Table I shows the accuracy (rounded off values) of identification, using Decision Tree. The highest accuracy among the string instruments is obtained for Cello (97.53%) while Oboe yields best results (97.78%) among the wind instruments. The result is poor for Viola as it suffers from inter class confusion. The proposed method shows a high accuracy while predicting the instrument family. For string instrument the accuracy is 96.07%, whereas, for wind instrument the accuracy is 90.78%. So, it may be assumed that due to overlapping of the ranges of feature values within the same instrument family, the individual precision or recall values are poor for some instruments.

TABLE I. CONFUSION MATRIX FOR IDENTIFICATION OF NOTES USING DECISION TREE IN TERMS OF PERCENTAGE OF ACCURACY.

	Cello	Contrabajo	Guitar	Viola	Violin	Trompa	Trompeta	Tuba	Oboe
Cello	97	0	0	1.2	1.2	0	0	0	0
Cont.	1.3	78	0	5.2	5.2	0	3.9	6.5	0
Guitar	0	4.8	87	0	0.9	1.9	0	3.8	1.9
Viola	3.6	27	0	34	34.5	0	0	0	0
Violin	0.7	2.1	0	2.2	93.6	0	1.4	0	0
Trom--pa	0	0	0	0	0	96	3.7	0	0
Trom--peta	0	9.7	1.6	6.4	11.3	13	58	0	0
Tuba	0	8.5	6.8	0	0	0	1.7	81	1.7
Oboe	0	0	0	0	0	2.2	0	0	98

TABLE II. COMPARISON OF ACCURACY (PERCENTAGE) MEASURE FOR THE TWO METHODS

Instrument	Proposed feature set and Random Forest	Proposed feature set and Decision Tree	Proposed Feature set + features of [10] and Decision Tree
Cello	**97.53**	**97.53**	91.35
Contrabajo	**77.92**	**77.92**	73.08
Guitar	90.48	86.67	**94.28**
Viola	**41.82**	34.55	40.74
Violin	92.14	**93.57**	84.28
Trompa	95.12	**96.34**	95.2
Trompeta	45.16	58.06	**80.00**
Tuba	79.66	81.36	**85.00**
Oboe	**100**	97.78	98.89
Overall	83.75	84.02	**84.59**

Table II presents a comparison of the performances of Decision Tree and Random Forest, in the first 2 columns. It may be noted that for string instruments, Random Forest shows better result, whereas, for wind instruments Decision Tree generates better accuracy. The overall performance improves when a combination of the proposed feature set along with the wavelet based and harmonic features of [10] is considered.

Table III shows a comparison with results of [10] which had an overall accuracy of 89.85% among 4 different types of string instrument. It can be seen that our proposed feature set yields an overall accuracy of 96.00% while the combination of the proposed feature set and the features proposed in [10] generates 97.57% accuracy on the same set of string instruments, using Decision Tree. For Cello and Violin the accuracy for proposed image classification method is more than the aforementioned paper but in case of Guitar and Contrabajo, harmonic and wavelet domain features overcome the proposed feature set in terms of accuracy.

TABLE III. COMPARISON OF ACCURACY (PERCENTAGE) FOR THE DIFFERENT FEATURE SET AMONG STRING INSTRUMENTS.

Instrument	Features of [9]	Proposed feature set	Proposed Feature set + features of [10]
Cello	85.36	**100**	**100**
Guitar	**97.14**	92.38	92.38
Violin	83.57	97.14	**100**
Cont.	**96.10**	90.91	**97.40**
Total	89.85	**96.00**	**97.51**

Among the two classification methodologies, Decision Tree yielded more overall accuracy. Thus, only the precision and recall values for Decision Tree is presented here.

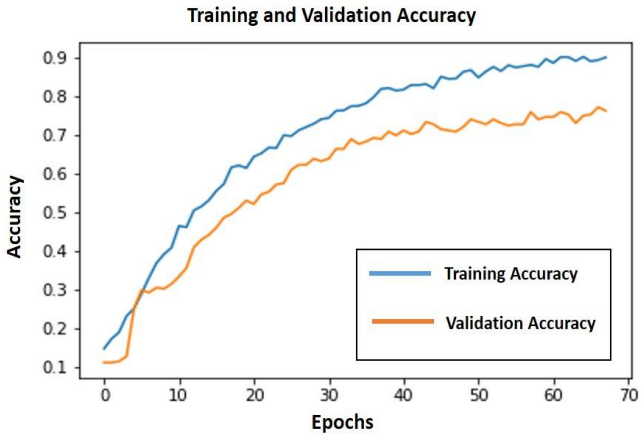

Fig. 5. Plot of Accuracy versus Epochs for VGG-16.

The result is compared to the outcome predicted by VGG-16 [16]. The overall accuracy achieved is 76.19%. Fig. 5 shows the train accuracy and test accuracy for the VGG – 16 network. The relatively poor result for the deep neural architecture may be attributed to the fact that the dataset used here is not sufficiently large for the proper training of the network.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

The conversion of audio signals to 2D matrices and deploying spatial features effectively increases the performance of instrument recognition process. The feature set proposed here may be applied for classification within string instruments along with Random Forest classifier. On the other hand, Decision Tree classifier proved to be better for the recognition task within wind instrument family. The family identification of the instruments also provides promising results (96.07% for string instruments). The overall performance improved when the training and testing was done with 70% and 30% data respectively. The accuracy of Viola significantly increases to 48% in the second case. Moreover, the proposed feature set yields a satisfactorily better result than the deep neural network architecture (VGG - 16) for this case.

The compact and efficient feature set are encouraging enough to motivate the application of other 2D features for music instrument identification. The availability of large data bases is being explored in order to test the efficiency of other deep learning algorithms.

REFERENCES

[1] A. Eronen And A. Klapuri, "Musical Instrument Recognition Using Cepstral Coefficients And Temporal Features*," Proc.Of Icassp 2000*, (Cat. No.00ch37100).

[2] A. Eronen, "Comparison Of Features For Musical Instrument Recognition." 2001 Ieee Workshop On Appl. Of Signal Pro.To Audio And Acoustics, 2001, Pp. 19-22.

[3] A.G. Krishna, T.V. Sreenivas, "Music Instrument Recognition: From Isolated Notes To Solo Phrases." *Proc. Ieee Int. Conf. On Assp*, 2004, Vol. 4, Pp. Iv-265--Iv-268.

[4] T. Heittola, A. Klapuri, And T. Virtanen, "Musical Instrument Recognition In Polyphonic Audio Using Source-filter Model For Sound Separation", *Proc. Int. Soc. Music Information Retriev* , 2009.

[5] A. Caclin, S. Mcadams, B. K. Smith, And S. Winsberg, "Acoustic Correlates Of Timbre Space Dimensions: A Confirmatory Study Using Synthetic Tones," *The Journal Of The Acoustical Society Of America*, July 2005, Vol. 118, No. 1, Pp. 471–482.

[6] D. G. Bhalke, C. B. R. Rao, And D. S. Bormane, "Automatic Musical Instrument Classification Using Fractional Fourier Transform Based- Mfcc Features And Counter Propagation Neural Network," *Journal Of Intelligent Information Systems*, Vol. 46, No. 3, May 2015, Pp. 425–446.

[7] Y. Han, J. Kim, And K. Lee, "Deep Convolutional Neural Networks For Predominant Instrument Recognition In Polyphonic Music," *Ieee/Acm Transactions On Audio, Speech, And Language Processing*, Vol. 25, No. 1, Jan. 2017, Pp. 208–221.

[8] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti & M. B. Sandler (2002) Automatic Music Transcription and Audio Source Separation, Cybernetics and Systems, 33:6, Pp. 603-627.

[9] R.S. Kothe And D.G. Bhalke, "Musical Instrument Recognition Using Wavelet Coefficient Histograms." Proceedings Of Emerging Trends In Electronics And Telecommunication Engineering (Ncet 2013). Pp. 37-41, 2013.

[10] A. Banerjee, A. Ghosh, S. Palit And M. A. F. Ballester, "A Novel Approach To String Instrument Recognition", International Conference On Image And Signal Processing, 2018, Cherbourg, France.

[11] R. M. Haralick, K. Shanmugam, And I. Dinstein, "Textural Features Of Image Classification", *Ieee Transactions On Systems, Man And Cybernetics*, Vol. Smc-3, No. 6, Nov. 1973, Pp. 610-621.

[12] D. A. Clausi, "An Analysis Of Co-Occurrence Texture Statistics As A Function Of Grey Level Quantization," *Canadian Jour. Of Remote Sensing*, Vol. 28,No. 1, Jan. 2002, Pp. 45–62.

[13] Institute For Research And Coord. In Acoustics/Music (Ircam), Pompidou, France.

[14] L Breiman, Friedman Jh, Olshen Ra And Stone Cj., *Classification And Regression Trees*, Belmont California: Wadsworth, Inc.; 1984.

[15] L. Breiman, "Random Forests", *Machine Learning*, Vol. 45, Issue 1, 2001, Pp.5-32.

[16] K. Simonyan And Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv preprint arXiv:1409.1556, 2014.