Interspeech 2018
2-6 September 2018, Hyderabad

Speech Emotion Recognition from Variable-Length Inputs with Triplet Loss Function

Jian Huang[1,2], *Ya Li*,[1] *Jianhua Tao*[1, 2,3], *Zheng Lian*[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{jian.huang,yli,jhtao,zheng.lian}@nlpr.ia.ac.cn

Abstract

Automatic emotion recognition is a crucial element on understanding human behavior and interaction. Prior works on speech emotion recognition focus on exploring various feature sets and models. Compared with these methods, we propose a triplet framework based on Long Short-Term Memory Neural Network (LSTM) for speech emotion recognition. The system learns a mapping from acoustic features to discriminative embedding features, which are regarded as basis of testing with SVM. The proposed model is trained with triplet loss and supervised loss simultaneously. The triplet loss makes intra-class distance shorter and inter-class distance longer, and supervised loss incorporates class label information. In view of variable-length inputs, we explore three different strategies to handle this problem, and meanwhile make better use of temporal dynamic process information. Our experimental results on the Interactive Emotional Motion Capture (IEMOCAP) database reveal that the proposed methods are beneficial to performance improvement. We demonstrate promise of triplet framework for speech emotion recognition and present our analysis.

Index Terms: speech emotion recognition, triplet loss, variable-length inputs

1. Introduction

As one of the main communication media between humans, speech not only contains basic language information, but also a wealth of emotional information. Emotion recognition has been attracting increasing attention recently, owing to its essential role in human behavioral signal processing [1].

Emotion can be quantified with discrete categories (*e.g.* neutral, happy, sad, angry, *etc.*) statically over utterances [2]. Several studies have focused on prediction of the emotional state from the speech. Prior researches on speech emotion recognition usually follow conventional paradigm of pattern recognition. They focus either on constructing robust and discriminative emotional features [3], or find an effective recognition model [4], or a combination of both.

However, emotions are naturally ambiguous [5]. Specifically, different types of emotions can be confused with each other, increasing the difficulty of emotion recognition [6]. Therefore, some researchers propose new thoughts. Lee *et al.* [7] design a hierarchical computational structure to solve the easiest classification tasks first. The structure maps an utterance into one of the multiple emotion classes through subsequent layers of binary classifications, which achieves effective results in multiple database contexts. Based on this work, Xi *et al.* [8] propose an emotion-pair based framework to generate more precise emotion bi-classification results. The framework uses a Naive Bayes classifier based decision fusion strategy to capture emotion distribution information, which achieves better performance than [7].

Due to strong ability of Siamese network in face verification [9][10], Lian *et al.* [11] utilize a Siamese network structure to release the ambiguity of emotion. The system attempts to learn the similarity and distinction between two audios. They essentially utilize a binary classification loss to train the model, which can represent the probability that two utterances in the pair are of the same category or not. However, it can't obtain a largest separation between positive and negative pairs due to lack of the margin threshold.

Different from their work, we utilize the triplet loss to train recognition models and generate more discriminative emotional features. The triplet loss is proposed in FaceNet [12] to achieve superior performance for face recognition. The input of model is the triplets including two utterances of same category and one utterance of other categories. The function of triplet loss is to separate the positive pair from the negative by a distance margin, making emotional categories more discriminative. Choosing the optimal triplets turns out to be very important for achieving good performance [13]. To improve the efficiency of training, we utilize hard-positive mining techniques to ensure consistently increasing difficulty of triplets as the network trains [12].

Another difficulty in emotional speech processing is that the utterances have variable length, which bothers the researchers all the time. Conventional static models compute global information to get fixed dimensional feature vector for every sample. They extract frame-level low-level descriptors (LLD) followed by utterance-level information extraction with different functionals, such as mean, maximum *etc.* [14][15]. But most of neural network models need full frame-level information and equal-length inputs for all samples. Like image processing, some researchers obtain equal-length inputs by clipping the utterances, meaning that longer turns are cut and shorter ones are padded with zeros [16]. Moreover, Yusuf *et al.* [17] use a global pooling strategy to down-sample variable-length inputs to a fixed dimensional vector by using a fully convolutional network. In view of this problem, we explore three different methods to get equal-length inputs, and meanwhile make better use of temporal dynamic process information.

In this paper, we propose a triplet framework for speech emotion recognition from variable-length inputs. The system handles the problem of variable-length inputs and model emotional dynamic information. The triplet loss makes intra-class distance shorter and inter-class distance longer. The goal

10.21437/Interspeech.2018-1432

of this paper is to investigate the application of triplet loss to enhance the state of research in speech emotion recognition.

The rest of the paper is as follows: in section 2, we introduce the proposed method. Section 3 presents the database and acoustic features. In section 4, we describe experimental results and analysis. Finally, we conclude the paper in section 5.

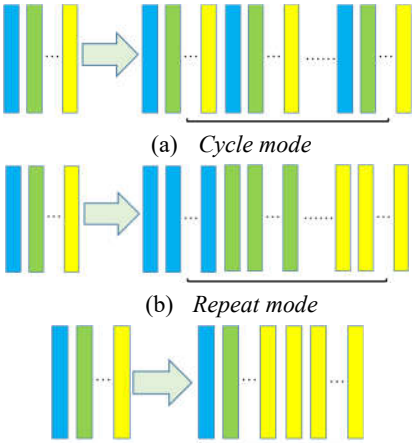
2. Model

In this study, we propose the triplet framework based on LSTM for speech emotion recognition, as shown in Figure 1. The input of system is a triplet consisting of two matching utterances (*wav1 and wav2*) and a non-matching utterance (*wav3*). Considering that the utterances have variable length after framing, the variable-length processing module is added to get equal-length inputs. Our proposed triplet framework gets discriminative embedding features with LSTM. We explore hard-positive mining techniques to select hard triplets for training the network more effectively. The system computes the triplet loss to separate the positive pair from the negative by a distance margin, and supervised loss to incorporate class label information simultaneously. Once this embedding has been produced, we utilize SVM to get the final predicted results during testing.

2.1. Variable-length processing

In view of variable-length inputs, most of researches only cut or pad the utterances to obtain equal-length inputs. However, it would lose and disturb the original emotional information. Some researchers also explore the influence of different length inputs on speech emotion recognition, and the results indicate that the performance increases as the length inputs are longer until they are too long [16][18].

In this study, we explore three strategies to handle this problem. Firstly, one suitable value F is set to decide the final length. We consider pad mode shown in Figure 2(c) for comparison, which pads the utterance with last frame until the length inputs are equal to F. The other two modes repeats the utterances a few times such that the total length is greater than F. Then, we obtain F consecutive frames from the middle part randomly. The difference is that cycle mode repeats the whole utterance at every turn, while repeat model repeats the single frame, as shown in Figure 2(a)(b). The key idea of the cycle mode is to make emotional dynamic cyclic and longer, whereas the repeat mode is to make emotional dynamic repeating and longer. The motivation is to display emotional dynamic information well. Besides, these modes are applied to the utterances whose length is less than F. For the utterances longer than F, their valid emotional dynamic information usually locates in the middle part and too long length inputs would have negative effect on emotion prediction. Therefore, we only cut the head and tail of the utterance to get F frames.



(a) *Cycle mode*

(b) *Repeat mode*

(c) *Pad mode*

Figure 2. *Three strategies of variable-length processing.*

2.2. Loss function

The network encodes the utterance x into a d-dimensional space and the embedding feature is represented by $f(x) \in \mathbb{R}^d$. Here, we want to ensure that the utterance x_i^a (anchor) is closer to all other utterances x_i^p (positive) of same category than it is to any utterance x_i^n (negative) of other categories, visualized in Figure 3. Thus, we want,

$$\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2,$$
$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \Gamma \quad (1)$$

where α is a margin that is enforced between positive and negative pairs. Γ is the set of all possible triplets in the training set and has cardinality N.

Then the triplet loss is minimized:

$$L = \sum_i^N [\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha]_+ \quad (2)$$

where $[z]_+ = \max(z, 0)$.

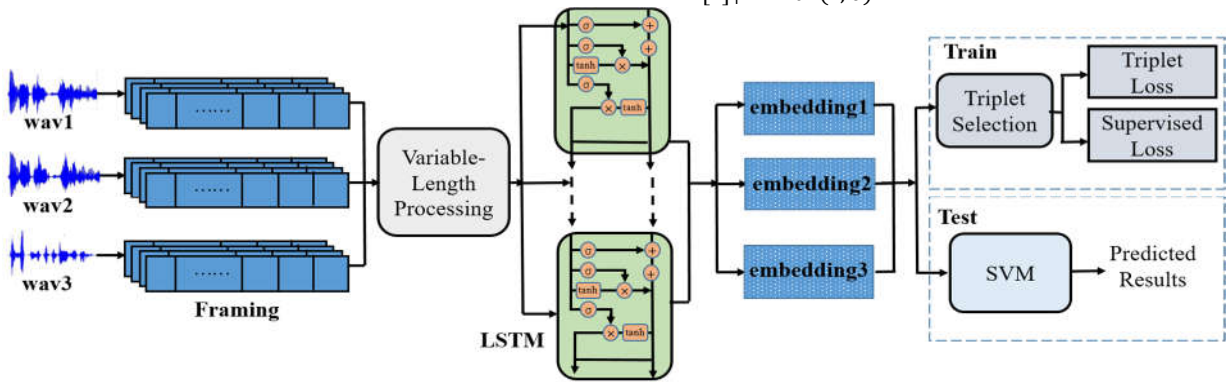


Figure 1. *Overview of the proposed triplet speech emotion recognition system. The input is a triplet consisting two matching utterances (wav1 and wav2) and a non-matching utterance (wav3). The variable-length processing module is added to get equal-length inputs, followed by LSTM layer to get embedding features. We utilize hard-positive mining techniques to select hard triplets. The network is trained with triplet loss and supervised loss simultaneously. Finally, we utilize SVM to get the final predicted results based on embedding features during testing.*

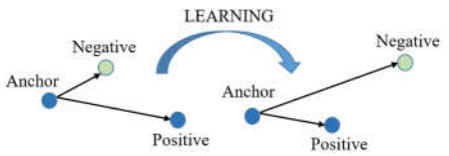


Figure 3. *The Triplet Loss minimizes the distance between an anchor and a positive, and maximizes the distance between the anchor and a negative [12].*

The triplet loss can obtain large separation between positive and negative pairs due to the margin threshold, which reduces intra-class distance and enlarges inter-class distance. This allows the utterances from one category to live on a manifold, still enforcing the distance and discriminability to other categories. Besides triplet loss, we also utilize the supervised cross entropy to incorporate class label information, which provides the guidance for emotional clustering.

2.3. Triplet selection

Even for small dataset, it can produce overwhelming number of triplet samples. But most of triplets would not contribute to the training and result in slower convergence. Therefore, it is crucial to select hard triplets that violate the triplet constraint in Equation (1). This means that, given x_i^a, we want to select an x_i^p such that $argmax_{x_i^p}\|f(x_i^a) - f(x_i^p)\|_2^2$, and similarly an x_i^n such that $argmin_{x_i^n}\|f(x_i^a) - f(x_i^n)\|_2^2$. It is infeasible to compute the argmin and argmax across the whole training set. In practice, we only compute the argmin and argmax within a mini-batch.

3. Dataset and feature set

3.1. Dataset

We use Interactive Emotional Dyadic Motion Capture (IEMOCAP) [19] to evaluate our proposed method. This corpus was designed for studying multimodal expressive dyadic interactions, recording (approximately a total of 12 hours) over 5 dyadic sessions with 10 subjects. Each interaction is around 5 minutes in length, and is segmented into sentence levels. At least three evaluators annotated each utterance in the database with the categorical emotion labels such as happy, sad and angry. We consider only the utterances with majority agreement (at least two out of three evaluators gave the same emotion label). Similar to prior studies [16][18], the utterances that bore the following four emotions are included: "angry", "happy", "sad", and "neutral", with "excited" considered as "happy". In total we use 5,531 utterances: 20.0% "angry", 19.6% "sad", 29.6% "happy", and 30.8% "neutral". The experiment protocol is leave-one-speaker-out which means there is no speaker overlap between training and testing set.

3.2. Feature set

The input of LSTM layer is short-time acoustic features extracted from every frame. The feature set is based on the INTERSPEECH 2014 Computational Paralinguistics Challenge [20], including energy, spectral and voicing related LLDs as well as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity. We also add the first dimension of the MFCC, the first order derivatives of all the LLDs, as well as the second order derivatives of MFCC 0-14. The resulting features 147 LLDs are extracted by openSMILE [21].

4. Experiments and Analysis

4.1. Experiment setup

In the experiments, the acoustic features are mapped to fixed dimensional embedding using LSTM. There is one LSTM layer with 64 memory cells. We use dropout after LSTM with the rate 0.5. The maximum training epochs are 100. The dimension of embedding is 128. We use a batch size of around 1000 triplets. Adadelta [22] optimization algorithm is utilized. In addition, we use the Gaussian noise with standard deviation 0.01 to obtain robust performance. The value α and F will be explored in the following sections.

4.2. Effect of variable-length processing

In IEMOCAP, the mean frame number of all utterances is 731 (max.:3409, min.:54, median:654) as the frame length is set 0.06s. As mentioned above, we explore three strategies to get equal-length inputs. On the one hand, the short utterances are enriched to make emotional processes longer; on the other hand, the long utterances are clipped to retain valid emotional dynamic information. In this section, we utilize basic LSTM network with cross-entropy objective to compare their performance.

Firstly, we explore the effect of frame number (F) on recognition accuracy, as shown in Figure 4. The performance of cycle mode is better than repeat mode in general, and the pad mode is worst. The cycle mode achieves best performance when F is 800, while best performance of repeat model lies in shortest frame number and pad mode lies in longest frame number. It is unexpected that these modes get worse accuracy when F is near median value.

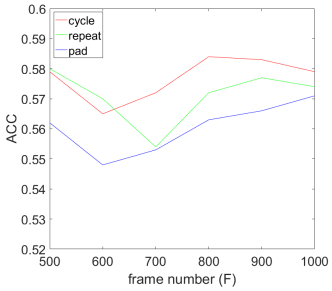


Figure 4: *System performance with different frame number.*

Then, the experimental results of three modes optimizing other hyper-parameters under optimal F are listed in Table 1. We can observe that the cycle mode and repeat mode achieve better performance than the pad mode, and improve by 3.3% and 1.5% respectively. Besides, the cycle mode is superior to repeat mode. The effect of cycle mode generates longer emotional process which can be better inferred by LSTM network structure. However, the repeat mode duplicates single frame several times, and the effect is to dilute emotional expression which is detrimental to emotion recognition.

Table 1: *Performance comparison under three modes based on basic LSTM network.*

Models	Accuracy
Cycle mode	**0.596**
Repeat mode	0.581
Pad mode	0.563

4.3. Speech emotion recognition with triplet loss

To improve the performance of speech emotion recognition, we utilize the triplet loss to reinforce emotional clustering. The triplet loss tries to enforce a margin between each pair of utterances from one category to other categories. As mentioned before, correct triplet selection is crucial for fast convergence. We conduct the experiments about different α value based on the network of Figure 1. The experimental results are shown in Figure 5. We can observe that the performance of cycle mode and repeat mode increase as the value of α is larger, while the best α of pad mode is in the middle. The cycle mode and repeat mode achieve best performance when α is about 0.05. Then the performance decreases as α is larger, which indicates that the system chooses the triplets between hard positive pair and hard negative pair, but not too harder triplets.

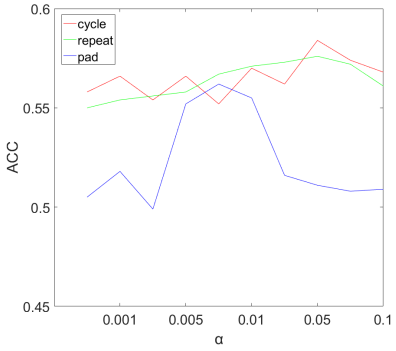


Figure 5: *System performance with* α value

The final experimental results of three modes based on the triplet framework, optimizing other hyper-parameters under optimal *F* and α are listed in Table 2. The cycle mode achieves best performance 0.604. Compared with Table 1, the performance of three modes improve by 0.8%, 1.1% and 1.6% respectively. Therefore, the triplet loss can increase the performance of speech emotion recognition, which verifies the effectiveness of proposed method.

Figure 6 shows embedding features of some samples learnt on the 2-D encoding space during training. The yellow represents "sad", the red is "neutral", the green is "happy" and the blue is "angry". The initial sight represents the visualization of acoustic features, showing disorder state in Figure 6(a). The following figures show that the samples of similar category cluster and the samples of different categories far away gradually. Particularly, Figure 6(c) show the manifold structure influenced by triplet loss. The final state of Figure 6(d) shows that "neutral" and "sad" can cluster well, but "angry" is too scattered and "happy" is confused with other categories. The method needs further research to obtain better performance.

We also compare the proposed method with two other methods of the literature. Our proposed method achieves better performance than 0.585 of Lee's work [7], which introduces a hierarchical binary decision tree method to recognize emotions. Haytham et al. [18] introduce a frame-based formulation to model intra-utterance dynamics with end-to-end deep learning , whose accuracy is 0.609. Compared with [18], our method also achieves comparable result. Furthermore, this paper provides a new thought which concentrates on increasing the emotional discriminability to enhance the performance of speech emotion recognition.

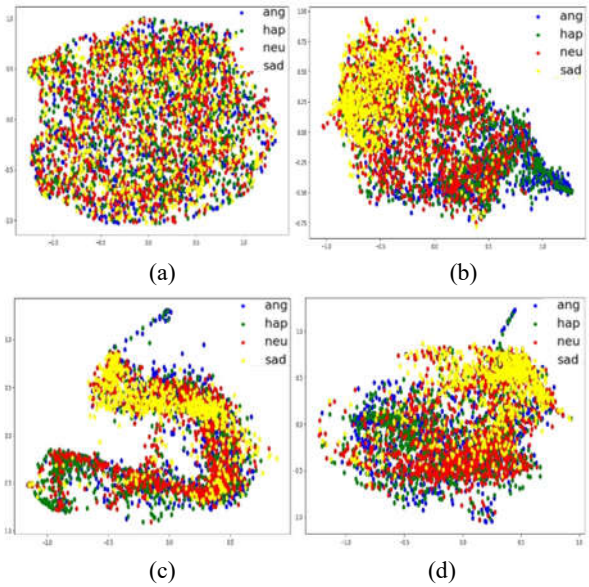


Figure 6: *Embedding features of some samples learnt on the 2-D encoding space during training*

Table 2: *Performance comparison based on the triplet framework between three modes and other methods.*

Model	Accuracy
Cycle mode	**0.604**
Repeat mode	0.592
Pad mode	0.579
Hierarchical binary decision tree [7]	0.585
Frame-based formulation [18]	0.609

5. Conclusion

In this paper, we propose the triplet framework to reinforce emotional clustering for speech emotion recognition. The system learns a mapping from acoustic features to fixed discriminative embedding features based on LSTM. The proposed model is trained with triplet loss and supervised loss simultaneously. The triplet loss reduces intra-class distance and enlarges inter-class distance, and the supervised loss captures class label information. The experimental results reveal that the triplet loss can highlight the performance, also make the samples of similar category cluster and the samples of different categories far away. Compared with other methods, our method also achieves comparable result. In view of variable-length inputs, we propose the cycle mode to generate longer emotional processes. The method can make better use of temporal dynamic information and is beneficial to performance improvement. This paper provides a new thought to enhance the state of research in speech emotion recognition. In the future, we will explore more effective measures to improve the performance.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379, No. 61332017, No. 61603390, No. 61771472) and the National Key Research & Development Plan of China (No. 2017YFB1002802).

7. References

[1] J. Tao, T. Tan, “Affective computing: A review,” International Conference on Affective Computing and Intelligent Interaction, pp. 981-995, 2005.

[2] H. Gunes, B. Schuller, “Categorical and dimensional affect analysis in continuous input: Current trends and future directions,” Image & Vision Computing, vol. 31, no.2, pp. 120-136, 2013.

[3] F. Eyben, K. R. Scherer, B. W. Schuller, et al, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” IEEE Transactions on Affective Computing, vol. 7, pp. 190-202, 2016.

[4] K. Han, D. Yu, I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[5] E. Mower, A. Metallinou, C. C. Lee, et al, “Interpreting ambiguous emotional expressions,” Affective Computing and Intelligent Interaction and Workshops, ACII 2009, 3rd International Conference on. IEEE, pp. 1-8, 2009.

[6] L. Chao, J. Tao, M. Yang, et al, “Long short term memory recurrent neural network based encoding method for emotion recognition in video,” Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on IEEE, pp. 2752-2756, 2016.

[7] C. C. Lee, E. Mower, C. Busso, et al, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Communication, vol. 53, no. (9-10), pp.1162-1171, 2011.

[8] X. Ma, Z. Wu, J. Jia, et al, “Speech Emotion Recognition with Emotion-Pair based Framework Considering Emotion Distribution Information in Dimensional Emotion Space,” Proc. Interspeech 2017, pp.1238-1242, 2017.

[9] Y. Sun, Y. Chen, X. Wang, et al, “Deep learning face representation by joint identification-verification,” Advances in neural information processing systems, pp.1988-1996, 2014.

[10] R. R. Varior, B. Shuai, J. Lu, et al, “A Siamese Long Short-Term Memory Architecture for Human Re-identification,” European Conference on Computer Vision, Springer, Cham, pp.135-153, 2016.

[11] Z. Lian, Y. Li, J. Tao, et al, "A pairwise discriminative task for speech emotion recognition," arXiv preprint arXiv:1801.01237 , 2018.

[12] F. Schroff, D. Kalenichenko, J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp.815-823, 2015.

[13] Y. Bengio, J. Louradour, R. Collobert, et al, “Curriculum learning,” Proceedings of the 26th annual international conference on machine learning, ACM, pp. 41-48, 2009.

[14] B. Schuller, S. Steidl, A. Batliner, et al, “The INTERSPEECH 2010 paralinguistic challenge,” INTERSPEECH, pp.2794-2797, 2010.

[15] J. Huang, Y. Li, J. Tao, “Effect of Dimensional Emotion in Discrete Speech Emotion Classification,” ASMMC, 2017.

[16] M. Neumann, N. T. Vu, “Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech,” INTERSPEECH. 2017, pp. 1263-1267, 2017.

[17] Y. Aytar, C. Vondrick, A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” Advances in Neural Information Processing Systems, pp.892-900, 2016.

[18] H. M. Fayek, M. Lech, L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” Neural Networks, pp. 60-68, 2017.

[19] C. Busso, M. Bulut, C. C. Lee, et al, “IEMOCAP: Interactive emotional dyadic motion capture database,” Language resources and evaluation, vol. 42, no.4, pp. 335, 2008.

[20] B. Schuller, S. Steidl et al, “The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load”, Interspeech, pp. 427-431, 2014.

[21] F. Eyben, F. Weninger, F. Gross et al, “Recent developments in opensmile, the munich open-source multimedia feature extractor”, Proceedings of the 21st ACM international conference on Multimedia, pp.835-838, 2013.

[22] M. D. Zeiler, “ADADELTA: an adaptive learning rate method,” arXiv preprint arXiv:1212.5701, 2012.