Adversarial Auto-encoders for Speech Emotion Recognition

Saurabh Sahu*, Rahul Gupta+, Ganesh Sivaraman*, Carol Espy-Wilson*

INSTITUTE FOR SYSTEMS RESEARCH
A. JAMES CLARK SCHOOL OF ENGINEERING

Introduction

- Emotion recognition has applications in psychiatry, psychology, medicine and designing human-computer interaction systems.
- Speech is a non-invasive way of collecting data which makes speech emotion recognition a widely researched problem.
- The standard approach is feature extraction followed by classification.
- This approach has two drawbacks:
  - High dimensionality of features used makes it difficult to analyze.
  - Unavailability of a large dataset to train the classifier models.
- We address these problems by employing an adversarial auto-encoder framework [1]. We conduct two specific experiments : (1) Using low dimensional code vectors as features for classification purposes and (2) Classification using synthetically generated samples from the adversarial auto-encoder to address the problem of small datasets.

Experimental Set-up

- We are doing a 4-way classification with the classes being angry, sad, neutral and happy. We do a batch-wise training.
- We used 4490 utterances from IEMOCAP dataset [2]. It consists of scripted and spontaneous dyadic interaction sessions performed by actors. There are 5 such sessions, two actors participating in each session. Each session has different actors involved. We did a cross-validation experiment with utterances from four sessions used for training and one session being used as a test set. So, it was a speaker-independent evaluation. We have 1708 neutral, 1103 angry, 1084 sad and 595 happy utterances.
- The feed forward model is shown in Fig 1. 1582 dimensional opensmile [3] features were input to the model (denoted as *x* in Fig 1, *x'* is its reconstruction). The layer generating the code vectors has 2 neurons. Thus, 1582-D features were compressed onto a 2-D space. The code vectors are mapped onto a mapping space distribution (MSD). MSD is a Gaussian mixture comprising of four 2-D Gaussians because we are performing a 4-way classification.
- We perform the following steps on each batch of 128 training samples:
  - Reconstruction error between x and x' is minimized
  - Input is transformed by encoder and an equal number of samples are sampled from the MSD. Weights of encoder and discriminator are updated so that the discriminator gets better at distinguishing coded samples from MSD samples.
  - We then freeze the discriminator weights. The weights of encoder are updated so that the discriminator is fooled into thinking that the code vectors have been sampled from the MSD.

Authors' affiliations : *Speech Communication Laboratory, ISR, UMD; +Amazon.com, USA

Adversarial Auto-encoder

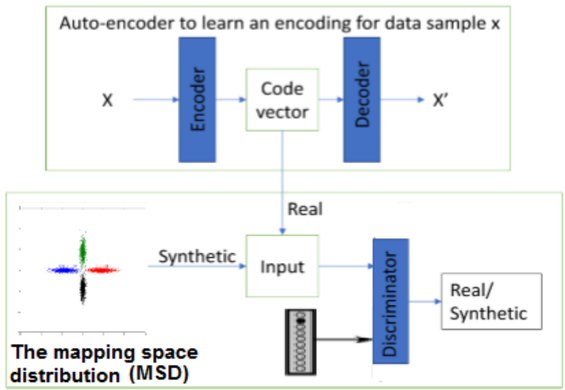
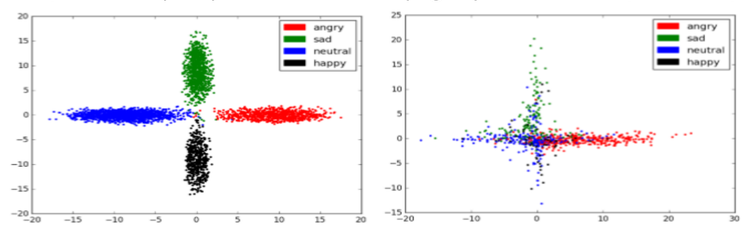


Fig 1: Code vectors generated from training data belonging to a particular class gets mapped to a specific mixture component of MSD

Plots, Results and Conclusion

- Mapping the code vectors : Code vectors generated from training samples belonging to a particular class are perfectly encoded onto a specific MSD mixture component (left). Test cases (right) are also quite separable.



- Classification : We used SVM to do 4-way classification of the utterances. We compare the unweighted average recall (UAR) obtained using 1582-D raw opensmile features, the 2-D code vectors as features and performing other compression techniques on raw opensmile features. The numbers for raw opensmile and code vectors are fairly close.

	Opensmile features (1582-D)	Code Vectors (2-D)	Auto-encoder (100-D)	LDA (2-D)	PCA (2-D)
UAR (%)	57.88	56.38	53.92	48.7	43.12

- Synthetic sample generation : We randomly sample 2-D points from each of the mixture component in MSD and use the decoder part of the auto-encoder to generate 1582-D samples. These were used as synthetic data for training a SVM. The results show that the synthetic samples do carry some discriminative information.

Dataset	Chance	Synthetic only	Real only	Real + Synthetic
UAR (%)	25	33.75	57.88	58.38

References

[1] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," arXiv preprint arXiv:1511.05644, 2015.
[2] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008.
[3] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013, pp. 835–838.