Audio-Visual Fusion for Sentiment Classification using Cross-Modal Autoencoder

Sri Harsha Dumpala, Imran Sheikh, Rupayan Chakraborty, Sunil Kumar Kopparapu
TCS Research and Innovation - Mumbai, INDIA
{d.harsha, imran.as, rupayan.chakraborty, sunilkumar.kopparapu}@tcs.com

Abstract

Humans often correlate information from multiple modalities, particularly audio and visual modalities, while learning as well as interacting with the external environment. In this learning mechanism, they also acquire the capability to interpret the information about missing modalities from the available modalities. Imparting these capabilities to machines might help build better Human-Machine interfaces and interactions. In this paper, we consider the task of multi-modal sentiment classification, using the audio and visual modalities, under the scenario where both the modalities are available during training but only one modality is available during test. We propose a novel model combining deep canonical correlation analysis (DCCA) with cross-modal autoencoders. These autoencoders try to reconstruct the representations corresponding to the missing modality, using the DCCA transformed representations of the available input modalities. Experiments on the CMU-MOSI and CMU-MOSEI datasets for sentiment classification on Youtube videos show the effectiveness of our proposed model.

1 Introduction

Speech and vision are two powerful modalities which are often correlated by humans to learn, understand and interact with the external environment. Human auditory perception may be grounded by vision and in turn, the visual understanding may be grounded by auditory cues. This grounding not only helps humans correlate information across modalities but also helps them to interpolate the missing modality from the available modality [1, 2, 3]. We hypothesize that this observation can help classification systems to perform better when one of the modalities is missing during test conditions.

In this paper, we consider the task of multi-modal sentiment classification from videos which have both visual and audio modalities[1] [4, 5, 6, 7]. We consider the problem where both modalities are available during training but only one modality is available during testing. We specifically examine the performance of deep canonical correlation analysis (DCCA) [8] on sentiment classification task. DCCA uses deep neural networks (DNN) to encode representations of the two modalities into a space where they are highly correlated. As a result, the encoded representation of a single modality carry information from the other modality as well. We propose a novel DCCA model which consists of two additional decoder DNNs. These decoders try to reconstruct the representations corresponding to a single modality, using the encoded representation of the available input modalities. The additional decoder in our model resembles a cross-modal autoencoder (CAE), so we refer to the proposed model as deep canonically correlated cross-modal autoencoder (DCC-CAE).

[1] We will use modality and view interchangeably in this paper.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

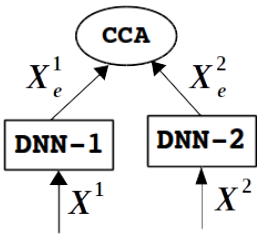

Figure 1: DCCA

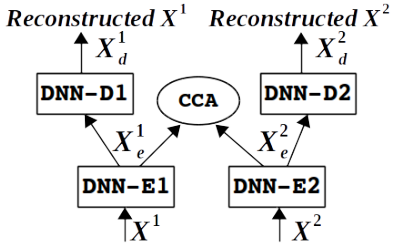

Figure 2: DCCAE

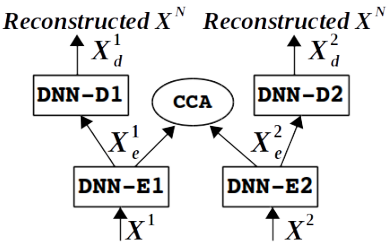

Figure 3: DCC-CAE

2 Related Work

Modeling correspondence between multiple modalities such as vision, language and speech has attained huge attention in recent times [9, 10, 3]. As a result, interesting and new problems such as visual question answering [11, 9, 12, 13], object discovery by multi-modal dialogue [9, 14] and text-to-image generation [15, 16, 17] have emerged. Most of these approaches are at the intersection of vision and text. Only a few studies have explored speech modality in conjunction with image and text [10, 18, 19, 20, 21]. In literature, the speech and visual modalities are jointly mapped onto a common representation space, where the two modalities are not maximally correlated. Moreover, these approaches require all the modalities considered during training to be available during testing. In this paper, we will consider models based on DCCA [8, 22] which can perform even in the absence of an input modality. To the best of our knowledge, there is no prior work which explores DCCA-based approaches combining speech and vision. While DCCA-based models proposed have shown improvements in the representation of ASR transcripts [23, 24], we show that the proposed cross-modal decoder based DCCA architecture gives better audio-visual representations and sentiment classification accuracies. The main contribution of this paper is (a) coupling audio and video in DCCA framework and (b) improved performance of DCC-CAE for sentiment classification.

3 DCCA-based Approaches

DCCA: Canonical Correlation Analysis (CCA) [25] is a statistical technique for finding a linear projection for two views into a common space where they are maximally correlated. DCCA (see Figure 1) is a non-linear version of CCA, in which both input views are passed through DNNs, and are then correlated with a CCA loss [8]. In Figure 1, (X^1, X^2) denote the representations of the two modalities corresponding to the same input data. DNN-1 and DNN-2 are used to obtain the non-linear transformation (X_e^1, X_e^2) of each view (X^1, X^2). During training, (X^1 and X^2) are extracted from the training data, and then used to learn (DNN-1, DNN-2) such that canonical correlation between the transformed representations is maximized. Thus, the objective to be optimized for DCCA is:

$$\max_{\theta^1,\theta^2} \; CCA(X_e^1, X_e^2); \;\; X_e^1 = g^1(X^1;\theta^1), X_e^2 = g^2(X^2;\theta^2)$$

where, g^1, g^2 denote the nonlinear transformations of DNN-1 and DNN-2, respectively. θ^1, θ^2 refer to the weight matrices of DNN-1 and DNN-2, respectively. CCA refers to the CCA cost function. The trained DNNs can be used to obtain the transformed representations carrying information from both the views, using any one of the input views.

DCCAE: DCCAE [22] includes two additional components in a DCCA model. Apart from maximizing the correlation between the input views, DCCAE has the additional objective of reconstructing the input representations from the encoded vectors (X_e^1 and X_e^2). As shown in Figure 2, DCCAE consist of two autoencoders, and the objective to be optimized is:

$$\min_{\theta_e^1,\theta_e^2,\theta_d^1,\theta_d^2} \; -CCA(X_e^1, X_e^2) + \left\|(X^1 - X_d^1)\right\|^2 + \left\|(X^2 - X_d^2)\right\|^2$$

$$X_e^1 = g_e^1(X^1;\theta_e^1), \; X_e^2 = g_e^2(X^2;\theta_e^2), \; X_d^1 = g_d^1(X_e^1;\theta_d^1), \; X_d^2 = g_d^2(X_e^2;\theta_d^2),$$

where, g_e^1, g_e^2 denote the nonlinear transformations of the encoders DNN-encoder-1 (**DNN-E1**), DNN-encoder-2 (**DNN-E2**), respectively. g_d^1, g_d^2 denote the transformation of the decoders DNN-decoder1

(**DNN-D1**) and DNN-decoder2 (**DNN-D2**), respectively. $\theta_e^1, \theta_e^2, \theta_d^1$ and θ_d^2 refer to the weight matrices of **DNN-E1**, **DNN-E2**, **DNN-D1** and **DNN-D2**, respectively. Here, **DNN-D1** is used to reconstruct view-1 which is encoded by **DNN-E1**. Similarly, **DNN-D2** is used to reconstruct view-2 which is encoded by **DNN-E2**. X_d^1 and X_d^2 refer to the reconstructed versions of X^1 and X^2, respectively.

3.1 Proposed Model: DCC-CAE

The proposed DCC-CAE model is an extension to DCCAE. As shown in Figure 3, DCC-CAE is similar in structure to DCCAE except that a single input view is reconstructed by both the decoders. Let us consider the case when both view-1 and view-2 are available during training, but only view-2 is available during testing. While training, the networks are trained to reconstruct view-1 by both the decoders i.e., **DNN-D1** and **DNN-D2**. Thus the objective to be optimized for DCC-CAE gets modified as:

$$\min_{\theta_e^1,\theta_e^2,\theta_d^1,\theta_d^2} -CCA(X_e^1, X_e^2) + \left\|(X^N - X_d^1)\right\|^2 + \left\|(X^N - X_d^2)\right\|^2$$
$$X_e^1 = g_e^1(X^1;\theta_e^1),\ X_e^2 = g_e^2(X^2;\theta_e^2),\ X_d^1 = g_d^1(X_e^1;\theta_d^1),\ X_d^2 = g_d^2(X_e^2;\theta_d^2),\ N = 1,$$

when we test for view-2. Here, X^N denotes the view not available during testing. g_e^1 and g_e^2, as earlier, denote the nonlinear transformations of the encoders **DNN-E1** and **DNN-E2**, respectively. g_d^1 and g_d^2 denote the transformation function of the decoders **DNN-D1** and **DNN-D2**, respectively. Here, it is to be noted that **DNN-D**, which takes as input the encoded version of view-2 (i.e., X_e^2), has to reconstruct view-1 while **DNN-D1**, as in DCCAE, is used to reconstruct view-1 from the encoded version of view-1 (X_e^1). X_d^1 and X_d^2 refer to the reconstructed versions of X^1 obtained from **DNN-D1** and **DNN-D2**, respectively.

This approach of training not only ensures maximum correlation between the two views but also encourages the network to project both views into a space which is more inclined towards the missing view. During testing, this helps DCC-CAE to better estimate the missing view, compared to DCCAE.

4 Experiments and Results

We evaluate the performance of proposed DCC-CAE, in comparison to DCCA and DCCAE, by considering two datasets, namely, CMU-MOSI [26] and CMU-MOSEI [27]. CMU-MOSI consists of 93 movie review videos segmented into 2199 clips/utterances. CMU-MOSEI consists of 2550 multi-domain monologue videos segmented into 18350 clips/utterances. Both, CMU-MOSI and CMU-MOSEI datasets are annotated with utterance-level sentiment labels in the range $[-3, 3]$. In this work, we perform binary sentiment classification in which labels $[-3, 0]$ are considered as negative and $(0, 3]$ are considered as positive sentiments. The train, validation and test split for CMU-MOSI and CMU-MOSEI datasets are utterances from 52, 10 and 31 videos, and utterances from 2100, 250 and 200 videos, respectively.

The visual modality (**V**) is represented with features (100-dimensional) extracted, at clip-level, using FACET framework and OpenFace toolkit [28, 29]. The audio modality (**A**) is represented with features (384-dimensional) extracted, at utterance-level, using openSMILE toolkit [30].

4.1 Results

For experimental validation of the sentiment classification task, we will consider both, audio (**A**) and video (**V**) modalities while training, but consider either **A** or **V** during testing. The modality available during testing is passed through the corresponding trained DNN-encoder, and the output of the DNN-encoder is considered as the projected view. The projected representations of audio and video are represented as $\mathbf{A}_p$ and $\mathbf{V}_p$, respectively. The projected views ($\mathbf{A}_p$ and $\mathbf{V}_p$) are concatenated with their original views to obtain the concatenated views ([**A**,$\mathbf{A}_p$] and [**V**,$\mathbf{V}_p$]). It is to be noted that the concatenated views are obtained by considering only a single modality (either **A** or **V**) during testing. We use bidirectional LSTM-RNN, as explained in [7] for sentiment classification, to label the clip/utterance level sentiment.

Table 1 shows the sentiment classification performances in terms of % accuracy (Acc.) and F-Score (F_1) obtained for CMU-MOSI and CMU-MOSEI datasets. It can be observed that the projected

Table 1: Multi-modal sentiment classification performance for a single test view.

		Audio					Video			
		MOSI		MOSEI			MOSI		MOSEI	
		Acc.	F_1	Acc.	F_1		Acc.	F_1	Acc.	F_1
–	**A**	50.6	50.0	59.4	58.0	**V**	53.8	53.5	59.6	59.5
DCCA	$\mathbf{A}_p$	50.7	50.2	59.6	58.2	$\mathbf{V}_p$	53.9	53.7	59.7	59.8
DCCA	[**A**,$\mathbf{A}_p$]	52.8	53.0	63.5	63.2	[**V**,$\mathbf{V}_p$]	55.7	55.9	64.1	64.2
DCCAE	[**A**,$\mathbf{A}_p$]	54.3	54.1	64.7	64.5	[**V**,$\mathbf{V}_p$]	57.2	57.1	65.3	65.4
DCC-CAE	[**A**,$\mathbf{A}_p$]	**56.2**	**56.5**	**66.4**	**66.1**	[**V**,$\mathbf{V}_p$]	**59.4**	**59.3**	**67.3**	**67.5**

Table 2: Multi-modal sentiment classification performance on **CMU-MOSI** dataset, when both, audio and visual, modalities are available during test.

	[**A**,**V**]	DCCA [**A**,$\mathbf{A}_p$,**V**,$\mathbf{V}_p$]	DCCAE [**A**,$\mathbf{A}_p$,**V**,$\mathbf{V}_p$]	**DCC-CAE** [**A**,$\mathbf{A}_p$,**V**,$\mathbf{V}_p$]
Acc.	56.1	56.4	58.3	**60.8**
F_1	56.3	56.6	58.5	**61.0**

views $\mathbf{A}_p$ and $\mathbf{V}_p$ obtained using DCCA did not result in significant improvements in performance over the original representations **A** and **V**, respectively. Similar observations were also made for the projections obtained from DCCAE and DCC-CAE. However, the concatenated views ([**A**,$\mathbf{A}_p$] or [**V**,$\mathbf{V}_p$]) obtained using each of the DCCA-based models give significant improvements over the original representations **A** and **V**, while DCC-CAE giving the best performance. For the audio only view, DCC-CAE when compared with DCCA and DCCAE, achieved an absolute improvement in Acc. (F_1) of 3.4% (3.5) and 1.9% (2.4) on CMU-MOSI, and 2.9% (2.9) and 1.7% (1.6) on CMU-MOSEI datasets, respectively. Similarly, for the video only view, DCC-CAE achieved absolute improvements in Acc. (F_1) of 3.7% (3.4) and 2.2% (2.2) on CMU-MOSI, and 3.2% (3.3) and 2.0% (2.1) on CMU-MOSEI datasets, when compared with DCCA and DCCAE, respectively, .

We also observed that representations obtained from our DCC-CAE model give a better performance compared to the audio-video cross-modal sequence-2-sequence models [31]. In [31] the entire video context was used for training the cross-modal sequence-2-sequence models for extracting representations for sentiment classification on the CMU-MOSI and CMU-MOSEI datasets. Our DCC-CAE model does not use the context from adjacent utterances and still achieves a better sentiment classification performance. In addition to the results obtained from a single view, Table 2 provides the sentiment classification performance on CMU-MOSI dataset, when both audio and visual modalities are available during test. It can be observed that using representations from the DCCA-based models improve the performance over the model considering the original representations. Proposed DCC-CAE model outperforms other DCCA models even when both audio and visual modalities are available.

5 Conclusion

In this paper, we proposed a novel model, called DCC-CAE to address the task of multi-modal sentiment classification under the scenario of missing modalities during test. Our model combines deep canonical correlation analysis (DCCA) with cross-modal autoencoder. The training objective of the model not only ensures maximum correlation between the two modalities but also encourages the model to project both modalities into a space which is more inclined towards the missing modality. Experimental results on two datasets show the effectiveness of the proposed model. DCC-CAE achieved an absolute improvement in accuracy of around 3% over DCCA, and around 2% over DCCAE when the video modality is missing. Similarly, an improvement in accuracy of around 4% over DCCA and around 2% over DCCAE is achieved by DCC-CAE, when the audio modality is missing. DCC-CAE model performs better than DCCA and DCCAE even when both the modalities are used in test.

References

[1] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[2] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[3] Ashwin K Vijayakumar, Ramakrishna Vedantam, and Devi Parikh. Sound-word2vec: Learning word representations grounded in sounds. *arXiv preprint arXiv:1703.01720*, 2017.

[4] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramaanyam. Benchmarking multimodal sentiment analysis. *arXiv preprint arXiv:1707.09538*, 2017.

[5] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927*, 2018.

[6] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[7] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 873–883, 2017.

[8] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[9] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, volume 1, page 3, 2017.

[10] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

[11] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[12] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017.

[13] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[14] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.

[15] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1060–1069. JMLR. org, 2016.

[16] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2989–2998. IEEE, 2017.

[17] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. *arXiv preprint arXiv:1803.08495*, 2018.

[18] David Harwath and James Glass. Learning word-like units from joint audio-visual analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 506–517, 2017.

[19] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. In *Interspeech*, 2017.

[20] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. *arXiv preprint arXiv:1804.01452*, 2018.

[21] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 613–622, 2017.

[22] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[23] Imran Sheikh, Sri Harsha Dumpala, Rupayan Chakraborty, and Sunil Kumar Kopparapu. Sentiment analysis using imperfect views from spoken language and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), ACL*, pages 35–39, 2018.

[24] Sri Harsha Dumpala, Imran Sheikh, Rupayan Chakraborty, and Sunil Kumar Kopparapu. Sentiment classification on erroneous asr transcripts: A multi view learning approach. In *Spoken Language Technology Workshop (SLT)*. IEEE, 2018.

[25] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[26] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016.

[27] Amir Zadeh. CMU-MOSEI dataset. `http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/`, 2018. Accessed: 2018.

[28] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[29] Amir Zadeh, Yao Chong Lim, Tadas Baltrušaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, volume 7, 2017.

[30] Florian Eyben, Felix Weninger, Martin Woellmer, and Bjoern Schuller. openSMILE. `http://www.audeering.com/research/opensmile`, 2009. Accessed: 2017.

[31] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. *arXiv preprint arXiv:1807.03915*, 2018.

Appendix

Datasets

CMU-MOSI: CMU Multi-modal Corpus of Sentiment Intensity and Subjectivity Analysis (CMU-MOSI) dataset consists of 93 movie opinion related videos collected from Youtube. These 93 videos are segmented into 2199 clips/utterances. In this work, train, validation and test set details are as follows:

- Train set: 1284 clips from 52 videos
- Validation set: 229 clips from 10 videos
- test set: 686 clips from 31 videos

CMU-MOSEI:CMU Multi-modal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset consists of 2550 multi-domain monologue videos collected from Youtube. These videos are segmented into 18350 clips/utterances. Train, validation and test sets considered in this work are as follows:

- Train set: 15529 clips from 2100 videos
- Validation set: 1257 clips from 200 videos
- test set: 1564 clips from 250 videos

Representation of Audio and Visual Modality

Audio Modality: It is represented by a set of 12 high level descriptors (HLDs) extracted from 16 low level audio descriptors (LLDs). LLDs include acoustic features such as voice probability, Mel-frequency cepstral coefficients (MFCCs), pitch, RMS energies and their corresponding delta regression coefficients. The HLDs carry more relevant information about the sentiments and emotions expressed in the utterance as compared to the LLDs. The HLDs being the higher order statistics of the LLDs, the dimension of the audio representation remains same (i.e. 384) for all utterances. We used openSMILE toolkit to extract the above mentioned audio features.

Video Modality: Most of the videos in the considered datasets are focused on a single person speaking to the audience through a close-up camera. Therefore, the features extracted from the person's face are used to build representations for the visual modality. The pre-computed features comprise of two sets. Set1 captures indicators of emotions, viz. anger, contempt, disgust, fear, joy, sadness, surprise, frustration and confusion, indicators of facial muscle movements and 20 facial action units. Set2 contains estimates of head position and rotation, and other facial landmarks. These features are extracted by the FACET framework and OpenFace toolkit. These features were originally extracted at a frame level, each video having 30 frames/second. To obtain utterance level representations, we averaged these features over an utterance and normalized them using statistics obtained over the training set.

DCCA-based Models Configuration

DCCA model: Each encoder (i.e., DNN-1 and DNN-2) consists of 3 hidden layers. Each hidden layer consists of 500 hidden units. ReLU activation function is used for the hidden layers. The output layer of each encoder has 10 units with linear activation function.

DCCAE model: Each encoder (i.e., DNN-E1 and DNN-E2) and decoder (DNN-D1 and DNN-D2) consists of 3 hidden layers. Each hidden layer consists of 500 hidden units. ReLU activation function is used for the hidden layers. The output layer of each encoder has 10 units with linear activation function. The output layer of the decoder also have linear activation.

DCC-CAE model: Each encoder (i.e., DNN-E1 and DNN-E2) and decoder (DNN-D1 and DNN-D2) consists of 3 hidden layers. Each hidden layer consists of 500 hidden units. ReLU activation function is used for the hidden layers. The output layer of each encoder has 10 units with linear activation function. The output layer of the decoder also have linear activation.

Training: All the DCCA-based models are trained using Adam optimizer with a batch size of 10. All models were trained for 100 epochs with a dropout of 0.4 and a patience of 4. The hyperparameters such as learning rate, momentum, learning rate decay parameter etc are tuned using the validation data for each model on each dataset.