Sensitivity of Multimodal Representation Learning Frameworks for Different Input Representations

Team Members

- Eray Erturk, 8747550178
- Yusuf Umut Ciftci, 6238630165
- Ada Toydemir, 8016967027

Problem Definition

While benchmarking multiple multimodal learning frameworks for sentiment analysis, training paradigms or multimodal fusion algorithms, the importance of input features such as different word representations obtained with different vector embedding techniques, or different audio signal processing techniques, have been overlooked when algorithmic superiority conclusions are made. In this project, we will investigate how sensitive these comparisons are with respect to different input feature representations.

Literature Review

Research on multimodal sentiment analysis has been accelerated due to advances in deep learning and several approaches have been proposed which utilize audio, visual and text modalities to extract multimodal representations. In our work, we focus on four multimodal learning frameworks, which are Modality-Invariant and -Specific Representations (MISA) for multimodal sentiment analysis [6], Multimodal Transformer (MulT) [12], MultiModal InfoMax (MMIM) [4] and Gradient Blending (GB) [13].

MISA projects each modality into two distinct subspaces, which are modality-invariant modality-specific subspaces. The representations learnt in modality-invariant subspace cover the shared information and reduce the information gap across modalities, where the representations learnt in modality-specific subspace are private to each modality and capture modality-specific characteristics. These representations provide a holistic view of multimodal data, which is used for fusion that leads to task predictions [6]. In this work, text modality is encoded with a pre-trained BERT model. For visual and acoustic features, they used Facet[1] to extract facial expression features and COVAREP [2] to extract various low-level statistical acoustic features. Authors used word-aligned visual and acoustic features.

MulT processes each modality first with temporal convolutional layers to extract local temporal information and project each modality to the same dimensional subspace to enable cross-modal attention computation. The main backbone of MulT is the cross-modal attention mechanism and with 3 modalities, MulT has overall 6 cross-model attention transformers, which is a stack of cross-model attention modules and MLP networks. Cross-modal attention mechanism allows the model to capture complex interactions between modalities [12]. In this work, they use GloVe embeddings as text features and for visual and acoustic features, they used Facet[1] and COVAREP [2] features, respectively. MulT supports the usage of both aligned and non-aligned features but in our work, we considered aligned features only.

[1] https://imotions.com/platform/

MultiModal InfoMax (MMIM) hierarchically maximizes the mutual information in unimodal input pairs (inter-modality) and between multimodal fusion result and unimodal input in order to maintain task related information through multimodal fusion [4]. The framework is jointly trained with the main task to improve the performance of the downstream task. The authors are using the head embedding from the last layer's output of pretrained BERT model to encode an input sentence. For visual modality, they are using COVAREP [2] and for acoustic P2FA [17] to extract features from unaligned raw data. These features are then fed to modality specific unidirectional LSTMs [8].

Gradient blending is an extension of supervised learning through dynamic weighting of modalities. Multimodal models receive more information compared to unimodal models, however, they are prone to overfitting due to their increased capacities. Moreover, different modalities overfit at different rates, therefore training the multimodal model with a single optimization objective doesn't leverage the full potential of all modalities. Gradient blending computes an optimal blending of modalities based on their overfitting behaviors by minimizing overfitting to generalization ratio [13]. The authors are using ResNet3D [11] as their visual backbone for RGB and optical flow modalities, and ResNet [7] for the audio backbone. For visual input they are using raw video clips and for audio input they are using log-Mel features with 40 Mel filters. The audio and visual features are temporally alligned.

When algorithmic comparisons are made to show the superiority of the model of interest with respect to other models in the literature, the differences in the data processing such as modality alignment and feature extraction, i.e., different encoder architectures, raw signals or hand-crafted features, etc., are mostly overlooked. Also, empirical studies in previous works have shown that text modality is the dominant modality in sentiment analysis [3, 12, 14, 19]. Therefore, the robustness of performance comparisons reported in the literature should be investigated carefully, especially with respect to different text representations. Such an investigation has been performed by Yin and Soleymani, where they have used several report sentiment analysis performance comparisons between MISA, MulT, MMIM, early fusion, late fusion and unimodal models, across three text encoders, BERT, RoBERTa and DeBERTaV3 [16]. Their results replicate the text modality dominance phenomenon in sentiment analysis, and show that straightforward fusion algorithms can achieve comparable performance to state-of-the-art models and none of these state-of-the-art models report these straightforward fusion performances in their reports. Also, they show that previously reported performance comparisons across these algorithms are sensitive to different text representations.

Hazarika et al. [5] also studied modality robustness in multimodal sentiment analysis with a different perspective. They investigated how several multimodal learning frameworks perform under missing modality and modality perturbation scenarios. First, their results show that even under no missing modality and no perturbation scenario, MulT outperforms MISA in regression correlation for CMU-MOSI dataset, which is not in line with the original MISA paper. Second, they show that the performance comparisons are prone to change under missing modality and modality perturbation scenarios, where MulT happens to be more robust than MISA, where MISA was reported to outperform MulT in its original paper [5, 6].

Data

For our preliminary analysis, we used CMU-MOSI [18] and CMU-MOSEI [1] datasets. CMU-MOSI dataset is a collection of YouTube monologues and collected from 93 videos from 89 distinct speakers. In total, it consists of 2199 opinion video segments. CMU-MOSEI dataset is the improved version of CMU-MOSI dataset with 22856 opinion video segments collected from approximately 5000 videos, 1000 distinct speakers and 250 different topics. For both datasets, utterances are manually labeled with a sentiment score continuously ranging from -3 (strongly negative) to 3 (strongly positive).

Method

As our preliminary analysis, our first step is to reproduce the results of Yin and Soleymani. For this purpose, we trained MISA, MulT, MMIM, early fusion, late fusion and unimodal models with BERT, RoBERTa and DeBERTaV3 text encoders. We also trained GB models and used GloVe embeddings in addition to text encoders stated above. For MISA, MulT and MMIM, we used AdamW optimizer [10] with learning rate 1e-5 and batch size of 32 for CMU-MOSI and 16 for CMU-MOSEI. For MulT with GloVe embeddings, we used learning rate of 1e-3. We used the default settings for the remaining hyperparameters. For visual and acoustic encoders, we used 2 layer LSTM network (each one is bidirectional, the hidden states of the first LSTM layer are inputs to the second LSTM layer) and the final representation is obtained by concatenation of the last hiddens states of the first and second LSTM layers. The regression heads used for early and late fusion are MLP with 3 hidden layers where each layer has 384, 256 and 128 neurons, respectively, and with a dropout rate of 0.1. For GB, we used a single layer bidirectional LSTM network for both visual and acoustic features and the fusion head is constructed by MLP with 2 hidden layers where each layer has 128 neurons, and with a dropout rate of 0.1.

Let h_a, h_v and h_t denote the hidden representations of acoustic, visual and text features obtained after their corresponding encoders, respectively. The early fusion model is:

$$\begin{aligned} h &= \text{Concat}(h_a, h_v, h_t) \\ \hat{y} &= \text{MLP}(h) \end{aligned} \tag{1}$$

The late fusion model is:

$$\begin{aligned} \hat{y}_m &= \text{MLP}(h_m) \qquad\qquad (m \in \{a, v, t\}) \\ \hat{y} &= \frac{1}{3}(\hat{y}_a + \hat{y}_v + \hat{y}_t) \end{aligned} \tag{2}$$

It should be noted that the feature encoders and regression heads used in late fusion are same as their unimodal counterparts in terms of the architecture, but distinct in terms of how they are trained since they are trained jointly. The loss function for training early fusion, late fusion, unimodal models and GB is:

$$L = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{3}$$

where n is the number of samples, y_i and $\hat{y}_i$ are the ground truth and predicted sentiments, respectively.

To extend the robustness study done by Hazarika et al. [5] we applied noise peturbation to analyze the sensitivity of different embeddings. After calculating the encoded hidden layers for text using GloVe, BERT, RoBERTa, and DoBERTa, we sampled 5%, 15%, and 30% of the representations from the testing set and modify them by adding white Gaussian noise with a kernel size of 1. We did not modify the models or pipeline for robustness analysis.

Results

Here we demonstrate our preliminary results on CMU-MOSI and CMU-MOSEI datasets, across all methods and text-encoders. For metrics, we used mean absolute error (MAE), regression Pearson correlation (Corr), seven-class classification accuracy (Acc$_7$), positive-negative binary classification accuracy (Acc$_7$) and its weighted F1 score.

Models	MAE ↓	Corr ↑	Acc_7 ↑	Acc_2 ↑	F1 ↑
Visual-only	1.476	0.035	0.172	0.518	0.509
Acoustic-only	1.416	0.057	0.155	0.537	0.539
GloVe text encoder					
Text-only	1.061	0.578	0.290	0.750	0.750
MulT	1.040	0.598	0.290	0.761	0.758
MISA	1.164	0.518	0.241	0.695	0.697
MMIM	1.231	0.439	0.255	0.626	0.626
GB	1.205	0.430	0.231	0.651	0.650
Early-fusion	1.079	0.603	0.279	0.748	0.749
Late-fusion	1.068	0.597	0.278	0.728	0.730
BERT text encoder					
Text-only	0.815	0.754	0.427	0.811	0.812
MulT	0.836	0.757	0.394	0.820	0.820
MISA	0.779	0.761	0.421	0.814	0.814
MMIM	0.806	0.749	0.434	0.816	0.817
GB	0.754	0.763	0.445	0.809	0.809
Early-fusion	0.805	0.754	0.443	0.816	0.816
Late-fusion	0.801	0.750	0.423	0.816	0.816
RoBERTa text encoder					
Text-only	0.679	0.814	0.479	0.854	0.854
MulT	0.720	0.821	0.433	0.856	0.857
MISA	0.724	0.787	0.445	0.843	0.843
MMIM	1.474	-0.035	0.155	0.448	0.277
GB	1.478	-0.006	0.155	0.448	0.277
Early-fusion	0.668	0.819	0.455	0.867	0.868
Late-fusion	0.670	0.820	0.461	0.857	0.857
DeBERTa text encoder					
Text-only	0.649	0.833	0.469	0.870	0.871
MulT	0.746	0.826	0.415	0.852	0.852
MISA	0.726	0.783	0.449	0.848	0.848
MMIM	0.688	0.802	0.465	0.839	0.839
GB	0.721	0.824	0.422	0.849	0.849
Early-fusion	0.632	0.845	0.464	0.875	0.875
Late-fusion	0.622	0.853	0.481	0.881	0.881
Averaged Performance					
Text-only	0.810	0.745	**0.416**	0.821	0.822
MulT	0.835	0.750	0.383	0.822	0.821
MISA	0.848	0.712	0.389	0.800	0.801
MMIM	1.049	0.506	0.327	0.682	0.639
GB	1.039	0.505	0.313	0.689	0.646
Early-fusion	0.796	**0.755**	0.410	**0.827**	**0.827**
Late-fusion	**0.790**	**0.755**	0.411	0.821	0.821

Table 1: CMU-MOSI results. Averaged performance is obtained by averaging over the four text encoders.

In line with the literature, for both datasets, visual-only and acoustic-only unimodal models have poor performance, near random-chance performance for CMU-MOSI and slightly better performance for CMU-MOSEI. It should be noted that our unimodal models seem to perform slightly worse compared to [16]. The unimodal acoustic-only model with HuBERT Embeddings showed improvements for CMU-MOSEI in two class accuracy and f1 score with slightly better than random-chance for CMU-MOSI for both 2 and 7 class accuracy.

MMIM has been perceived as the state-of-the-art architecture for multimodal sentiment analysis in the literature [4], but even if we used the same hyperparameters used in [16], MMIM is outperformed by fusion methods. This supports our claim that algorithmic superiority comparisons reported in the literature may not always hold. We suspect that this performance gap between our MMIM runs and MMIM runs in [16] may be caused by aligned features. For all our comparisons, we are using word-aligned visual and acoustic features to have a fairer comparison across different models by using the exact same data. However, in [16], for MulT, MMIM, GB, visual-only and acoustic-only unimodal models, visual and acoustic features are not aligned to

Models	MAE ↓	Corr ↑	Acc_7 ↑	Acc_2 ↑	F1 ↑
Visual-only	0.825	0.179	0.417	0.613	0.589
Acoustic-only	0.829	0.141	0.413	0.627	0.549
GloVe text encoder					
Text-only	0.638	0.668	0.464	0.806	0.806
MulT	0.677	0.633	0.442	0.791	0.793
MISA	0.657	0.653	0.459	0.797	0.796
MMIM	0.682	0.606	0.451	0.762	0.763
GB	0.688	0.593	0.434	0.756	0.758
Early-fusion	0.636	0.659	0.476	0.807	0.808
Late-fusion	0.637	0.665	0.472	0.799	0.799
BERT text encoder					
Text-only	0.569	0.761	0.509	0.848	0.849
MulT	0.576	0.754	0.499	0.848	0.848
MISA	0.548	0.764	0.515	0.861	0.861
MMIM	0.553	0.745	0.516	0.822	0.824
GB	0.554	0.758	0.518	0.789	0.797
Early-fusion	0.546	0.766	0.522	0.853	0.853
Late-fusion	0.561	0.745	0.515	0.847	0.848
RoBERTa text encoder					
Text-only	0.506	0.807	0.542	0.881	0.879
MulT	0.539	0.798	0.519	0.875	0.876
MISA	0.499	0.806	0.556	0.882	0.882
MMIM	0.821	0.228	0.416	0.639	0.640
GB	0.841	0.213	0.413	0.710	0.590
Early-fusion	0.502	0.805	0.544	0.881	0.880
Late-fusion	0.511	0.794	0.545	0.868	0.868
DeBERTa text encoder					
Text-only	0.496	0.807	0.563	0.867	0.867
MulT	0.534	0.808	0.524	0.874	0.874
MISA	0.521	0.803	0.525	0.869	0.870
MMIM	0.601	0.730	0.490	0.756	0.767
GB	0.554	0.751	0.521	0.807	0.813
Early-fusion	0.492	0.815	0.554	0.872	0.872
Late-fusion	0.499	0.809	0.548	0.878	0.878
Averaged Performance					
Text-only	0.552	**0.761**	0.520	0.851	0.850
MulT	0.582	0.748	0.496	0.847	0.848
MISA	0.556	0.757	0.514	0.852	0.852
MMIM	0.664	0.577	0.468	0.744	0.748
GB	0.659	0.578	0.471	0.765	0.739
Early-fusion	**0.544**	**0.761**	**0.524**	**0.853**	**0.853**
Late-fusion	0.552	0.753	0.520	0.848	0.848

Table 2: CMU-MOSEI results. Averaged performance is obtained by averaging over the four text encoders.

text features. This shows that the performance improvement over existing multimodal learning frameworks may not be caused by architectural novelty but different data modality representations.

Based on average performance over multiple text encoders, text-only unimodal models outperform MulT and MISA (seems to outperform MMIM as well but we cannot make a conclusion due to incomplete runs), and for both datasets, fusion methods appear to achieve the best performance with early fusion being slightly better than late fusion. Thus, our results suggest that even if there is an intensive research on developing complex architectures for multimodal sentiment analysis, simple and straightforward multimodal fusion should not be overlooked and taken into account for performance comparisons.

Next, to investigate the sensitivity of multimodal fusions, MISA, MMIM and GB with respect to different acoustic feature representations, we performed the same analysis by using acoustic features extracted by HuBERT [9]. We used a HuBERT base model pretrained on IEMOCAP dataset for SUPERB emotion recognition task [15] to extract the acoustic features. In this setting, visual features are aligned to text features but not the

Models	MAE ↓	Corr ↑	Acc_7 ↑	Acc_2 ↑	F1 ↑
Acoustic-only	1.284	0.321	0.210	0.625	0.627
GloVe text encoder					
MISA	1.144	0.524	0.273	0.709	0.711
MMIM	1.219	0.457	0.266	0.661	0.662
GB	1.205	0.430	0.231	0.651	0.650
Early-fusion	1.087	0.574	0.277	0.752	0.752
Late-fusion	1.072	0.571	0.281	0.742	0.743
BERT text encoder					
MISA	0.797	0.764	0.420	0.835	0.836
MMIM	0.824	0.722	0.428	0.795	0.795
GB	0.887	0.721	0.365	0.791	0.791
Early-fusion	0.792	0.762	0.431	0.817	0.818
Late-fusion	0.782	0.763	0.430	0.826	0.827
RoBERTa text encoder					
MISA	0.644	0.832	0.471	0.864	0.864
MMIM	1.405	0.096	0.161	0.542	0.540
GB	1.470	-0.086	0.154	0.447	0.276
Early-fusion	0.672	0.825	0.488	0.866	0.867
Late-fusion	0.666	0.826	0.470	0.848	0.848
DeBERTa text encoder					
MISA	0.659	0.831	0.478	0.866	0.865
MMIM	0.890	0.705	0.316	0.810	0.805
GB	0.721	0.824	0.422	0.849	0.849
Early-fusion	0.635	0.840	0.474	0.867	0.867
Late-fusion	0.614	0.852	0.481	0.860	0.860
Averaged Performance					
MISA	0.811	0.738	0.411	0.819	0.819
MMIM	1.084	0.495	0.292	0.702	0.700
GB	1.070	0.515	0.293	0.684	0.641
Early-fusion	0.797	0.750	**0.418**	**0.826**	**0.826**
Late-fusion	**0.784**	**0.753**	0.416	0.819	0.820

Table 3: CMU-MOSI results for HuBERT Embeddings. Averaged performance is obtained by averaging over the four text encoders.

acoustic features, since HuBERT architecture accepts raw audio signals sampled at 16 kHz as input features. Results in Table 3 and Table 4 shows the performances obtained by using HuBERT acoustic features. For both datasets, using HuBERT features improves unimodal acoustic classifiers' performance. Specifically, for MOSI, the acoustic classifier's correlation performance improves from 0.057 to 0.321 and for MOSEI, it improves from 0.141 to 443. Thus, we can conclude that acoustic features extracted by HuBERT are more informative for sentiment analysis task. However, we did not observe significant performance improvements across multimodal fusions and multimodal learning frameworks with HuBERT features. For MOSI, when the results in Table 1 and Table 3 are compared, the performances for MISA and multimodal fusions have slight improvements but the top performing frameworks across metrics are similar. The same observations applies for MOSEI as well, as Table 2 and Table 4 indicates. We could not obtain HuBERT results for MulT due to memory issues since MulT requires embeddings at each time step, unlike the other models which require a single embedding per utterance, but we believe that the same observations would be valid for MulT as well.

Our results for noise perturbation are in line with the original paper [5] showing that MISA outperforms MulT in model robustness even across all of the different text encoders we tested for. MulT consistently saw greater negative impacts from noise perturbation over MMIM and MISA. This sensitivity analysis suggests that noise perturbation shows greatest improvements for MulT. GloVe embeddings consistently showed improvement in model robustness over any of the other text encoders, while RoBERTa and DeBERTa always had the lowest scores for model robustness. While DeBERTa and RoBERTa embeddings improve results over GloVe embeddings, this directly trades-off with model robustness.

Models	MAE ↓	Corr ↑	Acc$_7$ ↑	Acc$_2$ ↑	F1 ↑
Acoustic-only	0.773	0.443	0.396	0.729	0.717
GloVe text encoder					
MISA	0.669	0.651	0.441	0.804	0.797
MMIM	0.676	0.609	0.448	0.750	0.754
GB	0.688	0.593	0.434	0.756	0.758
Early-fusion	0.627	0.676	0.475	0.813	0.813
Late-fusion	0.626	0.680	0.476	0.810	0.809
BERT text encoder					
MISA	0.552	0.765	0.518	0.857	0.857
MMIM	0.566	0.745	0.508	0.816	0.819
GB	0.554	0.758	0.518	0.789	0.797
Early-fusion	0.547	0.766	0.519	0.860	0.860
Late-fusion	0.544	0.764	0.526	0.853	0.854
RoBERTa text encoder					
MISA	0.503	0.806	0.547	0.879	0.877
MMIM	0.764	0.437	0.418	0.718	0.714
GB	0.841	0.212	0.413	0.710	0.590
Early-fusion	0.504	0.804	0.545	0.875	0.873
Late-fusion	0.509	0.792	0.554	0.869	0.870
DeBERTa text encoder					
MISA	0.504	0.818	0.545	0.871	0.868
MMIM	0.841	0.044	0.413	0.710	0.590
GB	0.560	0.752	0.522	0.826	0.829
Early-fusion	0.495	0.816	0.552	0.878	0.878
Late-fusion	0.516	0.800	0.543	0.867	0.866
Averaged Performance					
MISA	0.557	0.760	0.513	0.853	0.850
MMIM	0.711	0.458	0.446	0.748	0.719
GB	0.660	0.578	0.471	0.770	0.743
Early-fusion	**0.543**	**0.766**	0.522	**0.857**	**0.856**
Late-fusion	0.549	0.759	**0.525**	0.850	0.850

Table 4: CMU-MOSEI results for HuBERT Embeddings. Averaged performance is obtained by averaging over the four text encoders.

Models	5% Noise Perturbation		15% Noise Perturbation		30% Noise Perturbation	
	Corr	F1	Corr	F1	Corr	F1
GloVe text encoder						
MulT	↓ 0.032	↓ 1.560	↓ 0.095	↓ 3.675	↓ 0.186	↓ 7.126
MISA	↓ 0.022	↓ 1.254	↓ 0.067	↓ 3.081	↓ 0.158	↓ 7.021
MMIM	-	-	-	-	-	-
BERT text encoder						
MulT	↓ 0.042	↓ 1.826	↓ 0.116	↓ 4.086	↓ 0.241	↓ 8.112
MISA	↓ 0.037	↓ 1.610	↓ 0.100	↓ 3.694	↓ 0.214	↓ 7.499
MMIM	↓ 0.037	↓ 1.583	↓ 0.098	↓ 3.757	↓ 0.221	↓ 7.461
RoBERTa text encoder						
MulT	↓ 0.049	↓ 2.040	↓ 0.129	↓ 5.184	↓ 0.272	↓ 9.152
MISA	↓ 0.044	↓ 1.805	↓ 0.116	↓ 4.508	↓ 0.244	↓ 8.591
MMIM	↓ 0.054	↓ 1.924	↓ 0.125	↓ 4.828	↓ 0.261	↓ 8.762
DeBERTa text encoder						
MulT	↓ 0.049	↓ 2.089	↓ 0.138	↓ 5.102	↓ 0.280	↓ 9.128
MISA	↓ 0.044	↓ 1.801	↓ 0.123	↓ 4.512	↓ 0.257	↓ 8.584
MMIM	-	-	-	-	-	-

Table 5: Change in CMU-MOSI results across text encoders with noise perturbation.

Conclusions and Lessons Learned

Before we started the project, we wanted to analyze the sensitivity of multimodal frameworks used for sentiment analysis with respect to different feature representations. Thus, for multimodal fusions, unimodal classifiers, MISA, MMIM, MulT and GB, we compared the average performance across several metrics by using different text feature representations, which include GloVe embeddings and embeddings obtained from pretrained BERT, RoBERTa and DeBERTa. As Table 1 and Table 2 indicates, the algorithmic performance comparisons across different frameworks are prune to change when different text feature representations are used. Also, averaged performance across all text encoders indicate that multimodal fusions, which are perceived to be straight-forward information aggregation techniques compared to multimodal frameworks which try more complex approaches, have the top performance, thus, this result suggests that multimodal fusions should not be overlooked for multimodal sentiment analysis as they serve as a powerful tool. Another observation is that, in line with the previous work, unimodal text classifiers have a strong performance on its own, compared to multimodal frameworks and it shows that text features are the dominant features in sentiment analysis. However, unimodal classifiers of acoustic and visual features have poor performance when COVAREP and Facet features are used (Table 1 and Table 2) and we thought that dominance of text features may be caused by the poor quality of COVAREP and Facet features and may not be an inherent property of text features. To further investigate this, we used acoustic features extracted by a pretrained HuBERT model and performed the same analysis for unimodal acoustic classifier and multimodal frameworks across both datasets. We found out that even if HuBERT extracted embeddings are more informative than COVAREP features in this task, the performance across multimodal frameworks does not have huge benefit with HuBERT features, thus, it supports the hypothesis that text features are the dominant features in sentiment analysis. Finally, we extended our analysis by measuring and comparing model robustness with respect to input feature encodings. In real-world use, outside of academic superiority comparisons, model robustness is important to consider in making generalizable and applicable systems. We found that for text representations, GloVe embeddings trained more robust networks and that MulT depended more heavily on text modality over the other models. This gives us insight on the differences and use cases for models and embeddings, but also helps us create a larger framework for making algorithmic superiority comparisons and conclusions.

Contributions

Everyone in the group were responsible for running the following for both datasets and all text encoders:

- Eray Erturk: MISA, unimodal models and multimodal fusions, HuBERT embeddings extraction, MISA, MulT and unimodal models with HuBERT embeddings
- Yusuf Umut Ciftci: MMIM, GB and MMIM, GB and multimodal fusions with HuBERT embeddings
- Ada Toydemir: MulT, Noise Perturbation for MISA, MMIM, and MulT

The write-up and presentation was prepared with equal contribution of all group members.

References

[1] Bagher Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 2236–2246.

[2] Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. COVAREP — A collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2014), pp. 960–964. ISSN: 2379-190X.

[3] Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L.-p., and Poria, S. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal QC Canada, Oct. 2021), ACM, pp. 6–15.

[4] Han, W., Chen, H., and Poria, S. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis, Sept. 2021. arXiv:2109.00412 [cs].

[5] Hazarika, D., Li, Y., Cheng, B., Zhao, S., Zimmermann, R., and Poria, S. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, United States, July 2022), Association for Computational Linguistics, pp. 685–696.

[6] Hazarika, D., Zimmermann, R., and Poria, S. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis, Oct. 2020. arXiv:2005.03545 [cs].

[7] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR abs/1512.03385* (2015).

[8] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[9] Hsu, W., Bolte, B., Tsai, Y. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *CoRR abs/2106.07447* (2021).

[10] Loshchilov, I., and Hutter, F. Decoupled Weight Decay Regularization, Jan. 2019. arXiv:1711.05101 [cs, math].

[11] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. A closer look at spatiotemporal convolutions for action recognition. *CoRR abs/1711.11248* (2017).

[12] Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 6558–6569.

[13] Wang, W., Tran, D., and Feiszli, M. What Makes Training Multi-Modal Classification Networks Hard?, Apr. 2020. arXiv:1905.12681 [cs].

[14] Wu, Y., Lin, Z., Zhao, Y., Qin, B., and Zhu, L.-N. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (Online, Aug. 2021), Association for Computational Linguistics, pp. 4730–4738.

[15] Yang, S., Chi, P., Chuang, Y., Lai, C. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G., Huang, T., Tseng, W., Lee, K., Liu, D., Huang, Z., Dong, S., Li, S., Watanabe, S., Mohamed, A., and Lee, H. SUPERB: speech processing universal performance benchmark. *CoRR abs/2105.01051* (2021).

[16] Yin, Y., and Soleymani, M. Does Multimodal Learning Benefit Sentiment Analysis? Some Empirical Studies.

[17] Yuan, J., and Liberman, M. Speaker identification on the scotus corpus. *The Journal of the Acoustical Society of America 123*, 5 (2008), 3878–3878.

[18] Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems 31*, 6 (Nov. 2016), 82–88. Conference Name: IEEE Intelligent Systems.

[19] Zeng, Y., Mai, S., and Hu, H. Which is Making the Contribution: Modulating Unimodal and Cross-modal Dynamics for Multimodal Sentiment Analysis, Nov. 2021. arXiv:2111.08451 [cs].