Multimodal Sentiment Analysis with Common-sense Modulation

Anonymous ACL submission

Abstract

Our world is inherently multimodal and recent work highlights the importance of machine learning models leveraging multiple streams of information in making decisions. Multimodal sentiment analysis has been an active area of research that requires models to take advantage of the linguistic, acoustic, and visual signals available in an utterance. However, most current models do not take into account any social common-sense knowledge which is crucial in how we perceive sentiment in a conversation. To address that, in this paper, we aim to influence or modulate modality representations with common-sense knowledge obtained from a generative social common-sense knowledge base. We provide a novel way to modulate the linguistic, acoustic, and visual features corresponding to an utterance by scaling and shifting these representations. We use the knowledge base to obtain knowledge latent representations for an utterance corresponding to different states of the speaker such as the intent and the reaction and use it to shift and scale the three modalities. Our experiments on popular multimodal sentiment analysis benchmark datasets show that our proposed method is on par and often surpasses the current state-of-the art models.

1 Introduction

Human communication often employs different modalities of communication – language, audio, and video being some of the most common ones. The meaning of utterances and the perception of mood and sentiment during conversations depend not only on its content but also on its gestures and intonations. With the large volumes of video available on the internet, multimodal sentiment analysis has increasingly become a more important and active area of research. Advances in the field of multimodal machine learning allow us to mine sentiment from videos, for example, which have multiple applications such as social media monitoring for business intelligence. In addition to the importance of leveraging multiple modalities of information, recent work in neuroscience sheds light on the importance of background knowledge or common-sense reasoning in the way our brains perceive emotion or sentiment of utterances during a conversation.

In this work, we explore methods to prime our models with common-sense reasoning and inject background knowledge to better interpret the three modality signals when determining sentiment. However, it is neither obvious how to obtain this common-sense knowledge nor how to influence the modality representations to take advantage of the background information. Recent work by Bosselut et al., 2019 proposed a generative model called COMET for automatic knowledge graph construction, which can be fine-tuned on an atlas of everyday common-sense reasoning to generate social common-sense knowledge inferences. However, COMET only generates phrases of knowledge inferences. In this work, we instead propose leveraging COMET to obtain representations for the different attributes of the speaker given an utterance. To take advantage of this background social knowledge, we propose using feature-wise transformations to shift and scale representations of the three modalities with the knowledge representations. We also introduce an adaptive fusion of the three modalities which allows the model to dynamically weigh the importance of the individual modalities in determining the sentiment.

We evaluate our approach on two benchmark datasets for multimodal sentiment analysis - CMU MOSI and CMU MOSEI, and our results show the effectiveness of our approach which surpasses current state-of-the-art models. To further evaluate the adaptability of our approach, we test its performance on a similar benchmark for multimodal humour detection – UR_FUNNY. We additionally include ablation experiments which prove the importance of the common-sense modulation.

2 Related work

Human multimodal sentiment analysis requires inferring the sentiment of an utterance from three modalities – language, audio, and video, and requires fusing information from three modality signals while also accounting for temporality. Early works adopted simple fusion strategies, such as early concatenation of the three modalities (Ngiam et al., 2011; Lazaridou et al. 2015) or a late-fusion approach, where higher level modality representations learnt independently from the three modality signals are combined (Nguyen et al., 2018; Ranganathan et al., 2016). Later works placed greater emphasis on sophisticated fusion mechanisms that not only captured the individual modality-specific information but also captured correlations among the signals. Such approaches include models that synchronize multimodal sequences using a multi-view gated memory, recurrent models that capture intra-modal and multiple cross-modal interactions by assigning multiple attention coefficients, tensor-based fusion mechanisms that capture unimodal, bimodal and trimodal interactions across time (Liu et al., 2018; Mai et al., 2019; Zadeh et al., 2017) and models that capture modality invariant and specific information using independent subspaces (Hazarika et al., 2020). Other recent works either influenced word representations with non-verbal cues (Wang et al., 2018) or used cyclic translation between modalities to effectively model correlations (Pham et al., 2018).

Transformer architectures have also been extended to the multimodal sentiment analysis tasks where it is crucial to model cross-modality interactions in the temporal domain. Tsai et al., 2019 used stacks of pairwise and bidirectional cross-modal attention blocks that attend to low-level modality signals to model cross modal interactions. Lv et al., 2021 extended this idea by introducing a separate message hub so that higher level modality interactions could be captured via self-attention to higher-level representations. Rahman et al., 2019 introduced a mechanism to integrate multimodal information into large pre-trained transformers by shifting pre-trained weights using audio and visual modalities. Contrary to these approaches, in this work, we shift three modality representations by incorporating background common-sense knowledge.

Using common-sense knowledge to infer emotion in conversations was explored in prior work (Ghosal et al., 2020). However, concatenating knowledge from a generative knowledge-base to the input of the model, by Ghosal et al., 2020, has the implicit assumption that the common-sense reasoning is only used by the initial layers of the model. In our approach, we do not impose any such restrictions on the model and allow it to modulate representations deeper within a network based on common-sense reasoning.

3 Proposed Method

In this section, we provide an overview of our proposed approach to address the task of multimodal sentiment analysis using common-sense knowledge modulation. Our task is to predict the sentiment of an utterance given a sequence of word-aligned feature vectors for language ($\mathcal{R}^{T\times d_l}$), video ($\mathcal{R}^{T\times d_v}$), and audio ($\mathcal{R}^{T\times d_a}$), where T is the length of the sequence and d_l, d_v, and d_a are the dimensions of the language, visual, and acoustic features, respectively. At a high level, our approach includes modality-specific transformers for each of the language, vision, and audio signals followed by an attention-based adaptive fusion mechanism of the three modalities. Our main contribution is to introduce a novel way to influence latent variables for the different modalities using a generative common-sense knowledge base. Unlike prior works (Tsai et al., 2019; Lv et al., 2021) that used either three pairs of transformers or an additional transformer to capture cross-modal interactions, we only use a single modality-specific transformer and inject background knowledge through feature-wise transformations. The overall architecture is shown in Figure 1. The following sections elaborate on our approach.

3.1 Modality Specific Transformers

We use the three modalities – language, vision, and audio. Each modality-specific backbone takes as input $X_m = \mathcal{R}^{T\times d_m}$, where $m \in \{l, a, v\}$ represents the three modalities. Our modality-specific transformers consist of layers of transformer blocks that have scaled dot product multihead self-attention and feed forward sub-layers together with residual connections and layer norms. The scaled dot product attention is given by

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right),$$

where Q, K, and V are queries, keys, and values which are linear projections of the input to a block

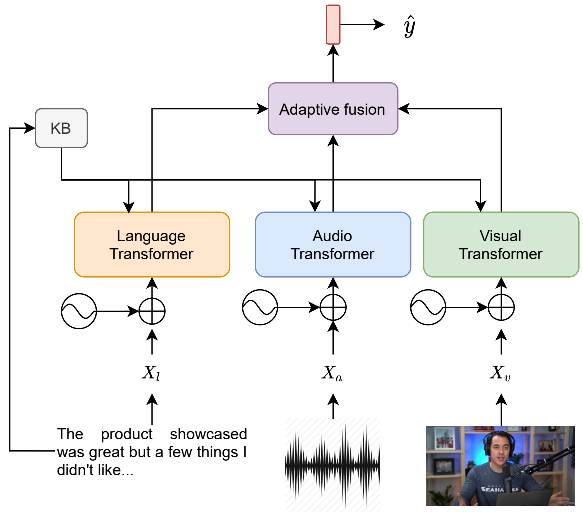

Figure 1: Overall architecture of our proposed model with modality-specific transformer backbones and adaptive modality fusion. Common-sense representations from the knowledge base (KB) are used to influence the modality transformers. Audio and visual transformers additionally cross-attend to raw language features X_l (arrows left out for brevity).

for modality m, I_m. In our model, we use multihead self-attention, which allows the model to attend to different representational subspaces and is given by

$$\text{MAttn}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

$$\text{where head}_i = \text{Attn}(QW_i^q, KW_i^k, VW_i^v).$$

After the multihead self-attention, we use position-wise feed forward layers. Additionally, we add sinusoidal position embeddings to X_m before it is input into the modality transformers. For the audio and visual transformers, we also add a cross-attention layer to the raw language features after the self-attention sub-layer. Similar to the work by Tsai et al., 2019, we empirically noticed that attending to low-level language features works better than attending to higher-level representations.

3.2 Generative Commonsense Knowledge

To inject common-sense knowledge into our model, we leverage large pre-trained language models fine-tuned on a social common-sense knowledge base. Prior work by Bosselut et al., 2019 introduced a COMET model that fine-tuned a GPT-2 (medium) model (Radford et al., 2019) on an atlas of 'everyday' common-sense reasoning, called ATOMIC (Sap et al., 2018). ATOMIC contains 877k textual descriptions of inferential knowledge for if-then reasoning. Fine-tuning a pre-trained GPT-2 model on ATOMIC and using it to generate novel commonsense reasoning to input into our multimodal pipeline gives our model the ability to make social common-sense inferences based on an utterance. The COMET model proposed in Bosselut et al. (2019) can generate novel common-sense inferences with respect to 9 different attributes (*eg* xAttr). However, the model is only able to generate discrete textual inferences. To leverage the model in our approach, we ignore the discrete inferences and use the hidden state of the decoder corresponding to the attribute token. More specifically, we use the utterance as a input to the COMET model and use the final hidden state from the decoder corresponding to a specific attribute as a representation for that attribute. We then concatenate the representations for multiple attributes to get the aggregated common-sense representation, CSK. The attributes used in our approach are the relevant social reasoning knowledge to inferring sentiment:

- **xReact**: speaker's reaction to an utterance
- **oReact**: listeners' reaction to an utterance
- **xEffect**: utterance's effect on the speaker
- **oEffect**: utterance's effect on listeners
- **xIntent**: the intent of the speaker

3.3 Sequence summarizer

We use attention to summarize each of the modality sequences. After l layers of modality-specific transformer blocks, we obtain $\{x_l^1, ..., x_l^T\}$ T-length sequences corresponding to the particular modality. The sequence is passed through a linear projection and softmax to get a score for each timestep, γ^t and the summarized sequence, $\tilde{x}_l$, is a weighted combination of the timesteps, *i.e.*,

$$\gamma^t = \frac{\exp(W_l x_l^t)}{\sum_{t=1}^T \exp\left(W_l x_l^t\right)}$$

$$SeqSum(l) = \tilde{x}_l = \sum_{t=1}^{T} \gamma^t x_l^t$$

3.4 Latent modulation

Using the common-sense features extracted from the COMET model, our main contribution is to introduce a CSKMod module before the self attention sub-layer to modulate the intermediate representations of the transformer blocks for the three modalities as illustrated in Figure 2. Our approach

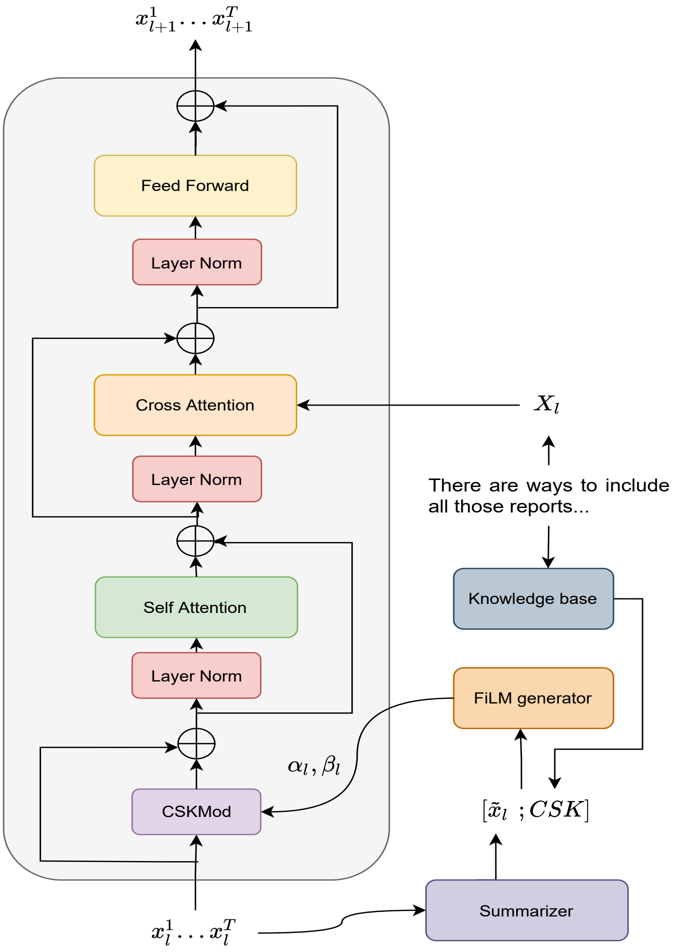

Figure 2: Transformer block including the CSKMod sub-layer. Audio and visual transformers additionally cross-attend to raw language features, X_l. The language transformer does not have the cross-attention sub-layer. The FiLM generator is the auxiliary network used to obtain α_l and β_l.

to modulate representations within a network is motivated by feature-wise transformations that have been followed in prior works where representations within a network are shifted and scaled based on some conditioning input (Perez et al., 2017a; Perez et al., 2017b; Strub et al., 2018; Dumoulin et al., 2016). In our approach, for the language modality, we first summarize the input sequence to the transformer block at layer l to get $\tilde{x}_l$. This, together with the CSK representation obtained from COMET, is used to get shifting and scaling parameters α_l and β_l respectively from a separate network (FiLM generator (Perez et al., 2017a)). Finally, the CSKMod module modulates representations using dynamic layer normalization in the following steps:

$$\alpha_l, \beta_l = \text{MLP}([\tilde{x}_l; CSK])$$

$$x_l^t = \alpha_l \odot \left(\frac{x_l^t - \mu_l}{\sigma_l} \right) + \beta_l,$$

where $\mu_l = \frac{1}{T} \sum_{t=1}^{T} x_l^t$ and $\sigma_l = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(x_l^t - \mu_l \right)^2}$. The other modalities are also modulated using the same steps. Finally, identical to the other sub-layers within the transformer block, we add a residual connection around the CSKMod sub-layer.

Feature-wise transformations based on common-sense background information allows the model to leverage the information at different layers. Additionally, in Section 5, we empirically show that common-sense knowledge derived from the COMET model is crucial to the boost in performance. The increased performance can be attributed to the fact that social knowledge derived from the COMET, which is fine-tuned on a social knowledge base, is more specific, tailored, and relevant to interpreting sentiment compared to the generic knowledge derived from a large pre-trained language model. This is analogous to how one's perception of sentiment in a conversation is influenced by his prior knowledge and exposure to different social settings from the past.

3.5 Adaptive modality fusion

After L layers of the modality-specific transformers, we use the sequence summarizer to get modality summaries $\tilde{x}^l_L$, $\tilde{x}^a_L$, and $\tilde{x}^v_L$ corresponding to the language, acoustic, and visual modalities. The concatenated $[\tilde{x}^l_L; \tilde{x}^a_L; \tilde{x}^v_L]$ is then passed through a linear projection and softmax to get scores s^l, s^a and s^v which are the mixture weights for the linguistic, acoustic, and visual modalities, respectively. We reason that not all modalities are equally informative for every utterance to determine the sentiment and giving the model the ability to assign contribution weights to the different modalities gives it the ability to emphasize contributions of some modalities over others. The final fused representation is given by

$$\tilde{x} = s^l \tilde{x}^l_L + s^a \tilde{x}^a_L + s^v \tilde{x}^v_L,$$

which is then used to obtain the prediction, $\hat{y}$, through a linear transformation.

3.6 Modality correlation loss

To further capture correlations among the modality representations $\tilde{x}^l_L$, $\tilde{x}^a_L$, and $\tilde{x}^v_L$ of the same utterance, we first project these representations to a shared embedding space with a linear projection head and L2 normalized. The correlation loss between two modalities, $\text{Corr}(m, m')$, is then computed as

$$-\frac{1}{N}\sum_{k=1}^{N}\log\left(\frac{\exp(\hat{x}_k^m\cdot\hat{x}_k^{m'}/\tau)}{\sum_{k'=1}^{N}\exp(\hat{x}_{k'}^m\cdot\hat{x}_{k'}^{m'}/\tau)}\right),$$

where $\hat{x}_k^m$ is the normalized projection of the m^{th} modality for the k^{th} sample in the batch. To account for correlations among the pairs of modalities, the final correlation loss is computed as

$$\text{CL}(l, a, v) = \frac{1}{3}\left[\text{Corr}(l, a) + \text{Corr}(l, v) + \text{Corr}(a, v)\right].$$

4 Experiments

4.1 Datasets

To empirically evaluate our approach of modulating hidden states with common-sense knowledge, and to compare to prior works (Tsai et al., 2019; Hazarika et al., 2020), we consider two benchmark datasets for multimodal sentiment analysis – CMU MOSI and CMU MOSEI.

CMU MOSI (Zadeh et al., 2016) is a small multimodal dataset often used as a benchmark for mutimodal sentiment analysis in prior work (Tsai et al., 2019; Hazarika et al., 2020). The dataset is a collection of 2199 opinion video clips obtained from YouTube. CMU MOSEI (Bagher Zadeh et al., 2018) is similar to CMU-MOSI but of a much larger scale; indeed, it is the largest dataset for multimodal sentiment analysis with a collection of 22,840 sentence utterance videos from more than 1000 YouTube speakers. Both CMU MOSI and CMU MOSEI are collections of monologues of speakers expressing their opinions and sentiment is annotated in the $[-3, 3]$ range for sentiment intensity with -3 and 3 representing strongly negative and strongly positive sentiments, respectively. MOSI and MOSEI are accessed through the SDK[1] available online.

Similar to Hazarika et al., 2020, we also evaluate the adaptability of our model on a similar task of multimodal humour detection on the UR_FUNNY dataset. This dataset is a multimodal balanced collection of punchlines from 1866 TED talk videos (Hasan et al., 2019) chosen from 1741 different speakers across 417 topics. Each punchline is annotated with a binary label that is used for humour/non-humour classification. Each punchline is also accompanied by its preceding context in the form of the linguistic, acoustic, and visual features for the utterances leading up to the punchline.

All of the datasets contain multimodal information for a single utterance corresponding to the linguistic, acoustic, and visual modalities. All of the datasets were word aligned.

4.2 Features

For all the datasets, the language features used are fixed pre-trained 300-dimensional GloVe embeddings (Pennington et al., 2014) for each of the words in the utterance. The acoustic features of the utterances are extracted using the COVAREP (Degottex et al., 2014) toolkit. These are various low-level acoustic features of the audio signal such as 12 Mel-frequency cepstral coefficients, pitch, tracking, voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters, and maxima dispersion quotients. For the MOSI and MOSEI datasets, these acoustic features are 74 dimensional. For UR_FUNNY, these are 84 dimensional.

Visual features for the MOSI and MOSEI datasets are extracted using Facet [2] to indicate facial action units based on Facial Action Coding System (Ekman and Rosenberg, 1997), which record facial muscle movements that correspond to displayed emotions. These are 47 and 35 dimensional for MOSI and MOSEI respectively. For UR_FUNNY, we use OpenFace (Baltrušaitis et al., 2016) to extract the facial action units at a rate of 30 frame/sec. The visual features for UR_FUNNY are 75 dimensional. The linguistic, acoustic, and visual features have different temporal resolution. These features are word-aligned using a forced aligner called P2FA (Yuan and Liberman, 2008) to obtain aligned timesteps segmented on a word-level for the audio and visual signals. The acoustic and visual features are then averaged over the duration of a word to obtain a word-level representation.

4.2.1 BERT language features

In addition to GloVe embeddings, we consider using BERT embeddings for the language features to fairly compare with some prior works. In this setting, the language transformer in Figure 1 is replaced with BERT and the last 4 layers of the model is fine-tuned. We do not modulate BERT hidden

[1] https://github.com/A2Zadeh/CMU-MultimodalSDK

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

Model	MAE (L)	Corr (H)	Acc-2 (H)	F1-score (H)	Acc-7 (H)
MFN*	-	-	76.0 / -	76.0 / -	-
RAVEN*	0.614	0.662	79.1 / -	79.5 / -	50.0
MCTN*	0.609	0.670	79.8 / -	80.6 / -	49.6
MulT*	0.580	0.703	- / 82.5	- / 82.3	51.8
PMR†	-	-	- / 83.3	- / 82.6	52.5
CSKMod	**0.550**	**0.732**	**79.9 / 83.3**	**80.7 / 83.3**	**53.8**
ICCN (B)*	0.565	0.713	- / 84.18	- / 84.2	51.6
MISA (B)*	0.555	0.756	83.6 / 85.5	83.8 / 85.3	52.2
CSKMod (B)	**0.536**	**0.764**	**83.4 / 85.8**	**83.7 / 85.6**	**54.0**

Table 1: Results on CMU MOSEI. Acc-2 has two values represented as '-/-' with left values for 'neg/non-neg' classification accuracy and right values for 'neg/pos' classification accuracy of sentiment values. Models that take BERT embeddings for language input are marked with 'B'. MAE **lower** is better. For all other metrics, **higher** is better. * from Hazarika et al., 2020; † from Lv et al., 2021.

Model	MAE (L)	Corr (H)	Acc-2 (H)	F1-score (H)	Acc-7 (H)
MFN*	0.965	0.632	77.4 / -	77.3 / -	34.1
RAVEN*	0.915	0.691	78.0 / -	76.6 / -	33.2
MCTN*	0.909	0.676	79.3 / -	79.1 / -	35.6
MulT*	0.871	0.698	- / 83.0	- / 82.8	40.0
PMR†	-	-	**- / 83.6**	**-/ 83.4**	40.6
CSKMod	**0.862**	**0.712**	80.9 / 81.9	79.8 / 81.9	**41.5**
ICCN (B)*	0.860	0.710	- / 83.0	- / 83.0	39.0
MISA (B)*	0.783	0.761	81.8 / 83.4	81.7 / 83.6	42.3
CSKMod (B)	**0.770**	**0.769**	**81.8 / 83.8**	**81.7 / 83.8**	**42.9**

Table 2: Results on CMU MOSI. Acc-2 has two values represented as '-/-' with left values for 'neg/non-neg' classification accuracy and right values for 'neg/pos' classification accuracy of sentiment values. Models that take BERT embeddings for language input are marked with a 'B'. MAE **lower** is better. For all other metrics, **higher** is better. * from Hazarika et al., 2020; † from Lv et al., 2021.

states because modulating pre-trained weights performed worse empirically. The remainder of the architecture is similar to Figure 1.

4.3 Evaluation Criteria

We use the same evaluation criteria as in prior works (Hazarika et al., 2020; Tsai et al., 2019). The sentiment intensity prediction for the MOSI and MOSEI datasets are regression tasks. Standard metrics include the mean absolute error (MAE) and Pearson correlation (Corr). Additionally, the classification score for the seven-class accuracy corresponding to seven integer sentiment labels from -3 to 3 is also used to empirically evaluate performance. Finally, binary classification scores – binary accuracy and F1-score – are also used. Following the approach in (Hazarika et al., 2020), for binary classification, we report the two metrics using a segmentation marker -/-. The left number corresponds to the score on binary classification of neg/non-neg, where sentiment values for the instances are divided into classes of < 0 and ≥ 0. The right number corresponds to the score on binary classification of neg/pos where sentiment values for the instances are divided into classes of < 0 and > 0.

4.4 Baselines

We choose several competitive baselines to evaluate our approach. MFN (Zadeh et al., 2018) performs temporal modeling and modality fusion. A Recurrent Attended Variation Network (RAVEN) (Wang et al., 2018) models the fine-grained structure of non-verbal subword sequences and shifts word representations based on non-verbal cues. MCTN (Pham et al., 2018) learns robust representations via cycle consistency loss to maximize the information captured from all modalities. MulT (Tsai et al., 2019) and the more recently proposed PMR (Lv et al., 2021) use transformer architectures to capture cross-modal interactions. ICCN (Sun et al., 2019) learns representations via explicitly capturing correlation among modalities of the same utterance. Finally, MISA (Hazarika et al., 2020) learns modality representations by projecting each modality into modality-invariant and modality-specific subspaces.

4.5 Results

We evaluate our model on word-aligned sequences for multimodal sentiment analysis. Results obtained on MOSEI and MOSI are reported on Tables 1 and 2, respectively. Models marked with 'B' take BERT embeddings for the language features. Most notably, our approach performs better across most metrics than competitive baselines on both datasets. We noticed a significant improvement of over 3% in MAE on the CMU MOSEI dataset compared to MISA for our experiments using BERT embeddings. Our multiclass accuracy (Acc-7) was also significantly higher on this dataset, with approximately 3.5% relative improvement (we improved to 54.0 from MISA's 52.2). Even with GloVe embeddings, our approach provided superior results across most metrics relative to competitive reported baseline (PMR) on both datasets.

Additionally, to evaluate the adaptability of our model, we compare the binary classification accuracy of our approach multimodal humor detection on the UR_FUNNY dataset with baselines as reported in Table 3. The results point to the effectiveness of modulating modality representations using commonsense knowledge representations not just for sentiment analysis but also for the task of humour detection.

Model	Context	Acc-2
C-MFN*		64.47
TFN*		64.71
LMF*		65.16
C-MFN*	✓	65.23
MISA*		68.60
CSKMod-COMET		69.20

Table 3: Binary accuracy results for models trained on UR_FUNNY dataset. Results marked * are from Hazarika et al., 2020.

5 Analysis

5.1 Subsample Experiments

To further evaluate the efficacy of our approach, we run two additional experiments on subsets of the MOSEI dataset. Because there is greater variance in the scores with smaller datasets, we consider separate experiments where we randomly sample either 5%, 10%, or 100% of the MOSEI training data to train the models. We evaluate the models on the entirety of the dev and test sets. For these experiments, we compute the mean and standard deviation of the metrics over runs using 15 different but fixed seeds. 5% of MOSEI is approximately 800 training instances which is half the size of the MOSI dataset. All experiments use GloVe embeddings for the language features.

Additionally, to quantify the influence of the CSKMod layer, we also evaluate the performance of our base model without any modulation for these experiments (NoMod). Finally, to ablate the effects of information injected from just the GPT-2 pretrained model, we evaluate a variant of our original model that just modulates using GPT-2 hidden states (CSKMod - GPT2). In this model, each utterance is input into a pre-trained GPT-2 model and hidden state corresponding to the last time-step from the encoder is used to modulate the modality representations instead.

The results of our experiments, reported in Tables 4, 5, and 6, show that CSKMod-COMET outperforms the baseline models on all three experiments. Binary classification scores are only reported for the neg/pos classification. Secondly, in the experiments, CSKMod-GPT2 performs approximately similar to our base model, NoMod, which lacks any form of modulation or external knowledge. These two observations point to the efficacy of modulating modality representations using a external generative social common-sense knowledge base. The difference in performance between modulation using pre-trained GPT-2 and the model fine-tuned on ATOMIC can be attributed to the fact that the knowledge obtained from the generative knowledge base is more specific and relevant to social contextual understanding.

5.2 Layer of modulation

In our experiments, we also noticed the importance of the layer at which we start modulating the modality representations. We noticed that early modulation degrades model performance. For our 8-layer

Model	MAE (L)	Corr (H)	Acc-2 (H)	F1-score (H)	Acc-7 (H)
MulT[†]	0.826 ± 0.015	0.355 ± 0.098	65.7 ± 3.8	62.5 ± 5.6	40.0 ± 1.0
MISA[†]	0.763 ± 0.016	0.461 ± 0.026	73.1 ± 0.9	72.8 ± 1.0	40.4 ± 1.3
NoMod	0.711 ± 0.019	0.557 ± 0.027	77.1 ± 0.7	77.1 ± 0.7	42.7 ± 1.4
CSKMod - GPT2	0.697 ± 0.020	0.578 ± 0.026	77.4 ± 1.2	77.4 ± 1.1	43.8 ± 1.8
CSKMod - COMET	**0.637 ± 0.010**	**0.650 ± 0.010**	**80.8 ± 1.1**	**80.8 ± 1.0**	**47.3 ± 1.0**

Table 4: Results for models trained on 5% of the CMU MOSEI dataset over 15 runs. Acc-2 and F1-score are the binary classification metrics for neg/pos classes. † indicates results obtained using publicly available code and applicable hyper-parameters.

Model	MAE (L)	Corr (H)	Acc-2 (H)	F1-score (H)	Acc-7 (H)
MulT[†]	0.753 ± 0.013	0.521 ± 0.023	75.6 ± 1.0	75.5 ± 0.9	40.8 ± 1.1
MISA[†]	0.722 ± 0.013	0.535 ± 0.017	75.4 ± 0.8	75.2 ± 0.9	42.5 ± 1.2
NoMod	0.664 ± 0.009	0.614 ± 0.011	78.7 ± 0.4	78.5 ± 0.6	46.3 ± 1.0
CSKMod - GPT2	0.661 ± 0.016	0.624 ± 0.012	78.8 ± 0.8	78.5 ± 0.8	46.5 ± 1.7
CSKMod - COMET	**0.615 ± 0.008**	**0.672 ± 0.009**	**81.5 ± 0.8**	**81.5 ± 0.7**	**49.3 ± 0.8**

Table 5: Results for models trained on 10% of CMU MOSEI over 15 runs. † indicates results obtained using publicly available code and applicable hyper-parameters.

Model	MAE (L)	Corr (H)	Acc-2 (H)	F1-score (H)	Acc-7 (H)
MulT[†]	0.582 ± 0.007	0.692 ± 0.010	81.7 ± 0.4	81.4 ± 0.5	49.2 ± 0.7
MISA[†]	0.578 ± 0.010	0.701 ± 0.011	81.6 ± 0.9	82.2 ± 0.9	50.2 ± 0.8
NoMod	0.595 ± 0.007	0.684 ± 0.007	81.1 ± 0.5	81.0 ± 0.5	49.3 ± 0.5
CSKMod - GPT2	0.569 ± 0.005	0.699 ± 0.006	81.6 ± 0.8	81.3 ± 0.9	52.0 ± 0.4
CSKMod - COMET	**0.556 ± 0.004**	**0.730 ± 0.004**	**83.2 ± 0.5**	**83.2 ± 0.5**	**53.4 ± 0.3**

Table 6: Results for models trained on entirety of CMU MOSEI over 15 runs. † indicates results obtained using publicly available code and applicable hyper-parameters.

architectures, we achieved the best results when we modulated representations for layers 6, 7, and 8. A possible explanation for the degraded performance for early modulation might stem from the fact that the model does not sufficiently capture discriminative features from the individual modalities before incorporating commonsense knowledge to modulate these representations.

5.3 Segmentation by emotions

For additional fine-grained analysis, we segmented the test set based on the emotion annotation for the examples. For each of the emotions, we computed the F1-score for the sentiments (multi-class), and we noticed a significant improvement (over 0.05 F1-score) in the F1-score for the neutral sentiment when using common-sense modulation with COMET compared to the absence of any background information. This could result from the effectiveness of using common-sense inferences and background knowledge to discriminate subtleties in the sentiment expressed when the utterance has positive/negative words, which could confuse the model without any background information otherwise.

6 Conclusion

Background knowledge and common-sense reasoning play crucial roles in the way we perceive sentiment and mood in a conversation. While most prior works emphasize fusion mechanisms of the multiple streams of signals – linguistics, acoustic, and visual – in this work, we propose a way to modulate modality representations using a common-sense knowledge base. This is done by shifting and scaling higher-level representations with a transformer architecture and by introducing a CSKMod sublayer within a transformer block. Empirical results prove the effectiveness of our approach. Additionally, ablation studies highlight the importance of the CSKMod module in the overall architecture.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

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

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. *CoRR*, abs/1906.05317.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP — A collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A learned representation for artistic style. *CoRR*, abs/1610.07629.

P. Ekman and E. Rosenberg. 1997. What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (FACS). Oxford University Press, USA.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. *CoRR*, abs/2010.02795.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. *CoRR*, abs/2005.03545.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *CoRR*, abs/1501.02598.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *CoRR*, abs/1806.00064.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2554–2562.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492, Florence, Italy. Association for Computational Linguistics.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. ICML'11, page 689–696, Madison, WI, USA. Omnipress.

Dung Nguyen, Kien Nguyen, Sridha Sridharan, David Dean, and Clinton Fookes. 2018. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding*, 174:33–42.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2017a. Film: Visual reasoning with a general conditioning layer. *CoRR*, abs/1709.07871.

Ethan Perez, Harm de Vries, Florian Strub, Vincent Dumoulin, and Aaron C. Courville. 2017b. Learning visual reasoning without strong priors. *CoRR*, abs/1707.03017.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2018. Found in translation: Learning robust joint representations by cyclic translations between modalities. *CoRR*, abs/1812.07809.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Wasifur Rahman, Md. Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. M-BERT: injecting multimodal information in the BERT structure. *CoRR*, abs/1908.05787.

Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146.

Florian Strub, Mathieu Seurin, Ethan Perez, Harm de Vries, Jérémie Mary, Philippe Preux, Aaron C. Courville, and Olivier Pietquin. 2018. Visual reasoning with multi-hop feature modulation. *CoRR*, abs/1808.04446.

Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. 2019. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *CoRR*, abs/1911.05544.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *CoRR*, abs/1906.00295.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *CoRR*, abs/1811.09362.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *CoRR*, abs/1707.07250.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. *CoRR*, abs/1802.00927.

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

A t-SNE embeddings

t-SNE embeddings of hidden states prior to the last layer for the CMU-MOSEI dataset with and without COMET common-sense modulation.

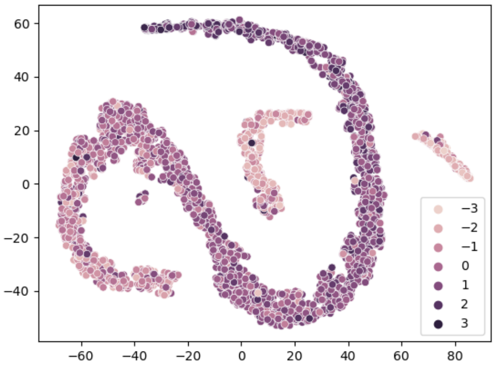

Figure 3: t-SNE embeddings without knowledge modulation

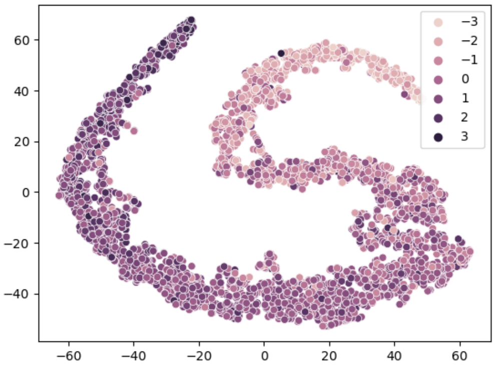

Figure 4: t-SNE embeddings with COMET common-sense modulation.

B Datasets

We use three benchmark datasets to evaluate our model - CMU-MOSEI, CMU-MOSI and UR_FUNNY. Table below lists the dataset sizes.

Dataset	train	dev	test
CMU-MOSI	1283	229	686
CMU-MOSEI	16315	1871	4654
UR_FUNNY	10598	2828	3290

Table 7: Number of utterances for the datasets in experiments.

C Hyper-parameters

Hyper-parameters for our results

Hyper-parameter	MOSI	MOSEI	UR_FUNNY
learning-rate	1e-4	1e-4	1e-4
batch-size	32	32	32
dropout	0.1	0.1	0
d_model	40	40	40
n_heads	4	4	8
τ	1.0	1.0	1.0
gradient clip	1.0	0.8	1.0
activation	ReLU	ReLU	ReLU
correlation loss weight	0.1	1.0	0.1