MM-SPCL for Multimodal Emotion Recognition in Conversation

Athiya Deviyani [*1] **Abuzar Khan** [*1] **Neil Skarphedinsson** [*1] **Prasoon Varshney** [*1]

Abstract

Multimodal Emotion Recognition in Conversation (mERC) is an important stepping stone towards intelligent systems interacting with humans where emotional awareness is crucial to the system. As a result of advancements in Multimodal Machine Learning as well as availability of datasets for the mERC task has proliferated with novel ways for computer systems to classify emotions and sentiments based on multi-sensory input. In this report we introduce two novel multimodal research contributions for this task, which we apply on the Multimodal EmotionLines Dataset (MELD) containing scenes from the Friends TV Series. Our first contribution is fusing modalities by performing cross-modal pruning of attention heads. The second contribution is a multimodal stochastic voting ensemble which randomly drops out modality combinations during training and achieves a new multimodal state of the art on both F1-score and accuracy. We also delve into ablation studies on two other recent fusion approaches: pairwise cross-modal attention fusion, and attention bottlenecks for multimodal fusion. Our code is available on github[1].

1. Introduction

Emotion Recognition in Conversations (ERC) is an important area of study in the pursuit of building machines that interact with humans in an empathetic and understanding manner. It is an area of research rooted in natural language processing (NLP) largely due to the ability to retrieve conversational data from social media platforms (Poria et al., 2018). As a result of this ease of access to textual corpora, most methods have taken a text-only unimodal approach.

However, in human dialogue, we decipher emotions by leveraging information beyond just text such as intonations in speech, facial expressions, gestures, and context. Intonations in voice can change the emotion conveyed by a sentence (Levis, 1999), facial expressions can convey sarcasm, and speaker context can help a listener recognize emotion based on the who the speaker is.

If we want to model things such as intonations, facial expressions, gestures, and context then we need multiple modalities beyond just text. In the last few years, multimodal ERC (mERC) has become a popular research topic with frequent state of the art advancements on popular multimodal datasets (Poria et al., 2018; Zadeh et al., 2018) for mERC.

The mERC task is quite difficult due to the multiple ways we can perform reasoning to incorporate these different types of information which are often obvious to humans. This includes (1) modelling speaker context, (2) extracting information such as facial expressions from the visual modality, (3) understanding relationships between speakers in a dialogue, (4) interpreting an utterance by taking previous parts of the dialogue into context.

In this paper we present two novel methods for multimodal machine learning and apply them to the task of mERC, one of which achieves a multimodal state-of-the-art results on the Multimodal EmotionLines Dataset (MELD). Our main contributions are as follows:

1. **Cross-Modal Pruning:** We present a novel cross-modal pruning approach to reduce redundancy of information between modalities by splitting representations into multiple heads and shutting off some of them based on similarity metrics.

2. **Multimodal Stochasic Voting Ensembles:** We present an algorithm where modality combinations vote for labels. During training we perform random dropouts dropout on the majority of modality combinations. This method achieves a new multimodal state of the art F1-score and accuracy on MELD.

3. **Evaluating various fusion techniques:** We present ablation studies for three different types of representation fusion techniques: namely, linear projection, pairwise cross-modal attention fusion, and bottlenecked attention fusion.

The rest of the paper is organized as follows: in section 2 we look at related work. In section 3 we formally describe the problem of mERC. In section 4 we describe a baseline we build on top of in our paper. In section 5, we will go

[1] Github Link

over Multimodal Supervised Prototypical Contrastive Learning in detail. In section 6, we will present our results and ablation studies. In section 7, we will perform a thorough analysis of our results. In section 8, we will discuss potential future research directions given our findings. Finally, we will conclude our paper in section 9.

2. Related Work

The task of ERC extends further back to 1974, where Ekman (1974) conducted a psychology study demonstrating that emotions could be classified given enough data. These emotions, namely joy, fear, anger, sadness, disgust, and surprise have been referred to as Ekman's universal emotions. The EmotionLines dataset (Hsu et al., 2018) is a supervised ERC dataset based on text utterances where the goal is to label each utterance as one of these six emotions (in addition to neutral, signifying the lack thereof). The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018) is a multimodal extension of the EmotionLines dataset that adds an acoustic and visual modality to these utterances *fa, v, tg*.

It should be noted that MELD is not the first multimodal ERC (mERC) dataset. Earlier datasets include IEMOCAP (Busso et al., 2008) and SEMAINE (McKeown et al., 2011). However, the conversations are dyadic in nature, which precludes the difficulty in tracking individual speaker states and handling co-reference. There are other, more recent mERC datasets such as CMU-MOSEI (Zadeh et al., 2018), MOSI, (Zadeh et al., 2016), and MOUD (Pérez-Rosas et al., 2013). However, unlike MELD they are not conversational. For this reason it is not possible to develop methods which leverage useful information such as speaker context, dialogue context, etc.

The authors of MELD implemented a number of baseline models for their dataset. These include text-CNN (Kim, 2014), bcLSTM (Poria et al., 2017), and DialogueRNN (Majumder et al., 2018). They applied these models with different combinations of the three modalities. However, they did not include the video modality in any of these experiments. The best performing model was the DialogueRNN (text + audio) with the weighted average F1 score of 60.25.

Since MELD was released, there has been a considerable amount of different, novel methods proposed. A commonality amongst twelve highest scoring methods on MELD[2] is the use of RoBERTa (Liu et al., 2019). What differentiates these approaches is how they leverage the additional information present in MELD to outperform methods simply classifying text utterances based on mere sentence-level embeddings. EmoBERTa (Kim & Vossen, 2021) does this by making the model speaker-aware by prepending speaker names to the utterances. Just as DialogueRNN achieved state of the by introducing speaker-context, RoBERTa at the time achieved state-of-the-art on MELD. As these two examples have shown it can be highly beneficial to apply reasoning by building the structured nature of a dialogue into the inference. Saxena et al. (2022) used graph neural networks to model both the dialogue participants and speaker personality. M2F-Net (Chudasama et al., 2022), the current multimodal state-of-the-art does not leverage context information to the same extent as suggested by EmoBERTa. Rather, it uses novel feature extractors that leverage features such as facial expressions in the video. The authors of M2F-Net do not specify whether or not they tried prepending the speaker names in front of each sentence. This does demonstrate, however, the advantage of extracting relevant features such as facial expressions. The current state-of-the-art, SPCL (Song et al., 2022), is unimodal in nature. SPCL takes advantage of prototypical networks (Snell et al., 2017b) to address the class imbalance problem in MELD. To the best of our knowledge there exists no published method in the research literature which attempts to utilize multiple modalities to extend SPCL.

3. Problem Statement

3.1. Dataset

For this project, we focus on the mERC task on MELD (Poria et al., 2018). MELD is a multimodal extension of the EmotionLines Dataset introduced by Hsu et al. (2018), which contains data in the acoustic, visual, and text modalities *fa, v, tg*. MELD contains a total of 1432 dialogues where each dialogue contains a sequence of utterances. In total there are 13708 utterances. Each utterance consists of a video clip *fa, vg* and a textual transcript *ftg*, along with two sets of labels that describe the emotion and sentiment, respectively. There are 7 emotion classes in the dataset, `joy, sadness, surprise, fear, disgust, anger, neutral`, and three sentiment classes, `positive, negative, neutral`. From a preliminary exploratory data analysis on the dataset, we observe a class imbalance, where the `neutral` label presents itself as a large majority in both the emotion and sentiment classes.

The dataset has already been split into train, validation, and test sets. The train set contains 1038 dialogues (9989 utterances). The validation set contains 114 dialogues (1109 utterances). Finally, the test set contains 280 dialogues (2610 utterances).

[2] https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld

3.2. Task description

In the MELD dataset, the goal is to classify a video clip utterance u_i as having emotion k_j where $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, K\}$. N is the number of utterances and $K = 3$ when classifying sentiments and $K = 7$ when classifying emotions.

In the dataset, there are M speakers, where $M < N$. This follows from the fact that each utterance u_i has only one speaker. Further, we use u_t to denote the utterance at a particular time-step. Given a dialogue n of length T, we can denote the sequence of all utterances in that dialogue in the following manner $u_0^{(n)}, u_1^{(n)}, \ldots, u_T^{(n)}$. Since we never reason about more than one dialogue at any given time, we will drop the n and write the sequence of utterances instead as $u_0, u_1, \ldots, u_T$.

The problem can then either be framed as a sequence tagging problem for every dialogue or simplified to a multiclass classification problem for every utterance. If framed as a sequence tagging problem (akin to parts-of-speech tagging or named entity recognition), each utterance can be treated as a token in a dialogue, and one would jointly decode the most likely sequence of emotions for every utterance within the dialogue using algorithms like Viterbi decoding. However, in this domain, given each video clip utterance depends only on past context, not lookahead, and also contains multiple words, intonations, or facial expressions, a simplification to a multiclass classification problem is feasible provided the context of previous utterances within a dialogue is encoded.

4. Baseline Methology

We implement Supervised Prototypical Contrastive Learning (SPCL) for Emotion Recognition in Conversation (Song et al., 2022) as a baseline for this task. The paper addresses the issue of class imbalance by leveraging prototypical networks (Snell et al., 2017a). Interestingly, the authors also address a problem that arises with MELD having being collected in a multimodal fashion which resulted in some utterances' text being misleading about the emotion that the utterance carries. To counter this, the authors leveraged curriculum learning (Bengio et al., 2009) to counter such extreme samples.

4.1. Encoding the context

We will extend the notation introduced in section 3 to effectively describe SPCL. Here, we use D to denote a dialogue. The textual transcript for an utterance u_t is denoted as u_t (e.g. "Oh, honey") and the name of its respective speaker is denoted as $s(u_t)$ (e.g. "Rachel"). Mathematically, we can describe a particular dialogue n with length T^n as:

$$D^{(n)} = [s(u_1); u_1; s(u_2); u_2; \ldots; s(u_{T_n}); u_{T_n}] \quad (1)$$

In our problem setting, we consider the context of k utterances for a dialogue and denote it as $D_t^{(n)}{}_{k:t}$. This context is then used as input to predict the label for u_t. We call this the context for D^n at timestep t.

$$D_t^{(n)}{}_{k:t} = [s(u_{t-k}); u_{t-k}; \ldots; s(u_t); u_t] \quad (2)$$

During training and inference, $D_t^{(n)}{}_{k:t}$ is converted into a string by concatenating all items in the context. We denote it as $S_t^{(n)}{}_{k:t}$. Here is an example of such as string where $k = 3$.

> "Rachel: What are you doing here? Ross: hey, you know, this building is on my paper route. Rachel: Oh, honey."

Natural language can be encoded using BERT-like models by using the encoding of the CLS token as a representation of the input string. Instead, Song et al. (2022) use prompt-based learning (Liu et al., 2021) where they construct a prompt which extends the string $S_t^{(n)}{}_{k:t}$ by appending "s_t feels <mask>" at the end of it.

We denote the embedding for the "<mask>" token as z_i, which is then used for the downstream task of classifying the utterance u_t.

4.2. Prototypical Contrastive Learning

An important contribution from Song et al. (2022) is that of combining prototypical learning with contrastive learning to alleviate the class imbalance problem. To lay down terminology, we have a batch I which is a set of "<mask>" token encodings $I = \{z_1, \ldots, z_N\}$, a score function G, and the vanilla supervised contrastive loss for $i \in I$ as:

$$F(z_i; z_j) = \exp(G(z_i; z_j)/\tau) \quad (3)$$

$$P_{sup}(i) = \sum_{z_p \in P(i)} F(z_i; z_p) \quad (4)$$

$$N_{sup}(i) = \sum_{z_j \in I_i} F(z_i; z_p) \quad (5)$$

$$L_i^{sup} = \log \frac{1}{|P(i)|} \frac{P_{sup}(i)}{N_{sup}(i)} \quad (6)$$

Where $P(i)$ is the set of positive samples in I, and $I_i = \{z_j \in I \mid i \neq j\}$. This setup, however, is impacted by class imbalance. To counter this, prototypical learning is employed.

A fixed-length queue for each emotion is maintained as $Q_j = [z_1^j ... z_L^j]$ for the j^{th} emotion, a support set of size K is uniformly sampled from this. A prototype T_j is derived as the mean of the support set. With this, we augment the negative scores of the i^{th} sample as $N_{spcl}(i) = N_{sup}(i) + \sum_{k \in E_{y_i}} F(z_i, T_k)$ and the positive scores as $P_{spcl}(i) = P_{sup}(i) + F(z_i, T_{y_i})$. E is the set of all emotion labels and $E_{y_i} = \{y_j \in E \mid i \neq j\}$. Finally, this gives the SPCL loss:

$$L_i^{spcl} = \log \frac{1}{|P(i)| + 1} \frac{P_{spcl}(i)}{N_{spcl}(i)} \quad (7)$$

4.3. Towards Multimodal SPCL

MELD as a dataset was designed for multimodal tasks, with human annotations performed by watching video clips and all three modalities $\{t, a, v\}$ in mind. The current state-of-the-art (SPCL) however, utilizes just the $text$ modality. This causes two issues; (1) The model fails to incorporate information from other modalities, and (2) using this dataset for unimodal tasks, as in SPCL, requires some way to mitigate effects of extreme samples wherein the information present in the modality in question (for SPCL, $text$) is almost misleading, because the right information needed to predict the emotion correctly lies in other modalities. Thus, we propose multimodal extensions of SPCL (MM-SPCL) and the rest of this work focuses on various methods for these multimodal extensions to the SPCL pipeline. Section 5.1 first discusses feature extraction and attempted fusion approaches, and then dives into our novel contributions of cross-modal pruning and multimodal stochastic voting ensembling technique for this method.

5. Multimodal Supervised Prototypical Contrastive Learning

5.1. Feature Extraction

5.1.1. Text Embeddings

We found extending SPCL to MM-SPCL to be non-trivial primarily since most of the SPCL pipeline is devoted to curriculum learning and clustering on the unimodal text embeddings. In fact, the curriculum learning was used as a workaround to the fact that the SPCL framework tackles a multimodal dataset with a unimodal approach. Consequently, as mentioned in subsection 4.3 we chose the extracted embeddings for the $mask$ token, z_t, as the point of fusion. However, we did so in two ways. One, for mid-fusion within the SPCL pipeline, that is, fuse representations from video (z_v) and / or audio (z_a) modalities with the text representation, hereafter z_t, and feed the fused representations z_f back into the SPCL pipeline for training and inference. And two, we explored whether the fine-tuning of the BERT model whose training is influenced by the prompting (which produces the prompt's mask embedding) and the novel SPCL pipeline (which happens downstream to the prompting) serve to generate better text embeddings. Thus, we attempted training SPCL end-to-end, extracting z_t for each utterance in the data via inference after convergence, and training a set of Multi-Layer Perceptrons (MLP) on various unimodal, bimodal, and trimodal combinations of utterance representations towards late fusion. Figure 1 demonstrates where the feature extraction takes place in this setting.

Figure 1: We train SPCL in its entirety and then use the mask embedding from its prompt as an embedding for an utterance

5.1.2. Video Embeddings

For the video embeddings we utilized Clip-ViT (Radford et al.). Upon analyzing the video data in MELD it is obvious that there is a considerable cutting back-and-forth between different camera angles in each scene. We estimated that the most reliable way to capture the speaker of an utterance would be to select the middle frame. The reason being that the start of a clip might still show the speaker of the previous utterance. Thus we try to avoid this by selecting the middle frame. Figure 2 demonstrates the extraction.

5.1.3. Audio Embeddings

For the audio embeddings we make use of the audio features provided by the authors of MELD (Poria et al., 2018). These features were extracted with openSMILE (Eyben et al., 2010). We believe a new method for extraction of contextualized audio representations is important for this task, however, we leave it to future work due to resource constraints in this work.

5.2. Fusion Approaches

5.2.1. Linear Projection

In this approach, we simply concatenate embeddings for different modalities and pass them through a single linear

Figure 2: Video embedding extraction

layer to project the concatenated vector back to original dimension of 1024.

5.2.2. Cross-Modal Attention

We follow the approach in (Tsai et al., 2019) to perform pairwise contextualization of modalities. For the three modalities, text , audio , and video , this translates to 6 different transformer encoder blocks. Each encoder block has 5 layers of cross-modal attention, followed by 2 layers of self-atttention.

For brevity, let us denote the triplet of modalities, text , audio , and video , by $\{t, a, v\}$, then, the six transformers would perform pairwise attention with pairs $\{t, a\}$, $\{t, v\}$, $\{a, t\}$, $\{a, v\}$, $\{v, t\}$, and $\{v, a\}$ respectively. That is, for each pair of modalities, the first modality's embedding denoted as z_1^l forms the query, and the second modality's vector embedding z_2^l forms the keys and values for the attention mechanism (Vaswani et al., 2017).

Figure 3: Architecture for pairwise cross-modal fusion (Tsai et al., 2019).

5.2.3. Attention Bottlenecks for Multimodal Fusion

Nagrani et al. (2021) use special-purpose bottleneck nodes for multimodal fusion. The bottleneck nodes are hypothesized to encourage sharing of only the salient information betweeen modalities, by restricting information to flow between modalities through these nodes and not allowing direct interactions.

Inspired by this approach, we augment the cross-modal attention in section 5.2.2 by adding bottleneck nodes z_{fsn}^l between each pair of modality vectors z_1^l and z_2^l. We use the following equations as presented in (Nagrani et al., 2021).

$$[z_2^l \,||\, z_{fsn}^l] = \text{CrossModal}([z_2^l \,||\, z_{fsn}^l\ ;\ ]) \tag{8}$$

$$[z_1^{l+1} \,||\, z_{fsn}^{l+1}] = \text{CrossModal}([z_1^l \,||\, z_{fsn}^l\ ;\ ]) \tag{9}$$

5.3. Cross-modal pruning

Based on the assumptions that all modalities have some amount of redundancy in the information they carry, and that the representations for each modality are not aligned, we introduced a novel cross-modal pruning module prior to the fusion module in the MM-SPCL pipeline. The idea is to remove redundant information by explicitly zeroing out sections of a pruning modality that are similar to sections in the base modality.

Figure 4: Cross-modal pruning in four steps

Figure 4 depicts the process in four steps. First, split both the base and the pruning modality into H_B and H_P heads of size K (we tested with $K \in \{16, 32, 64\}$, and H_B and H_P varied accordingly). Second, measure similarity between

each of the base heads and pruning heads. Third, for each base head, find the pruning head with the maximum similarity and shut it off by multiplying it by 0. And fourth, concatenate the heads and then apply the fusion module of choice.

5.4. Multimodal Stochastic Voting Ensemble

We have a set of seven different modality combinations: $M = \{t, a, v, ta, tv, av, tav\}$. Three of them are unimodal, three are bimodal and one trimodal. Here, we present an algorithm that we call a Multimodal Stochastic Voting Ensemble (MSVE). The ensemble voting happens over different modality combinations instead of over the modalities separately. It is stochastic since at training time we randomly sample modality combinations for each batch.

Note that the dropout for the ensemble is performed only during training, at inference time, all modalities are always provided. We hypothesize that this combats strong modality combinations overpowering weak ones since a weak modality combination can not rely on a strong modality combination being present.

Algorithm 1 MSVE training for a single batch

Inputs: Input feature set $X = \{x_m \mid m \in M\}$, labels y, set of MLPs $H = \{h_m \mid m \in M\}$, modality combination sample size k.

```
M_k ← SAMPLE(M, k)
z ← Σ_{m ∈ M_k} h_m(x_m)
ŷ ← SoftMax(z)
L ← CrossEntropy(ŷ, y)
backpropagate L
update weights for all h_m ∈ H
```

Algorithm 1 describes how training is performed for a single batch. We have a set of input features and a set of Multi Layer Perceptrons (MLPs) for each modality combination. As the input features goes x_t, x_a, and x_v have been extracted prior to training as described in Section 5.1. For the bimodal and trimodal feature combinations we perform simple concatenation of the unimodal features. As an example, for audio and video (ta) we denote that $x_{ta} = x_t \frown x_a$.

In each batch we sample k modalities where $k \in N$ and $k \leq |M|$. For the modalities we sample we feed the input features into their respective MLPs and sum up their predictions for the seven emotion classes, we empirically chose $k = 2$. Figure 5 explains this process pictorially.

With this simple algorithm we achieve state of the art results on MELD on both F1-score and accuracy as presented in Section 6.4. Interestingly, all of the MLPs we employ are very simple. All but two have a single linear layer, one has two layers and one has three layers. The method does not need to perform any fine-tuning on large pre-trained models and as a result has only 350K trainable parameters.

Figure 5: A graphical representation of MSVE

6. Results and Ablation Studies

In this section we present the results of our various multimodal approaches and contrast them with SPCL (Song et al., 2022) and M2FNet (Chudasama et al., 2022) which currently top the leaderboard for MELD[3].

Before we delve into the results we want to mention that we have standardized our evaluation to be based on peak validation F1-Score instead of peak test F1-Score (as performed in the codebase provided by Song et al.). By doing so we make sure that the test F1-score and accuracy are unbiased metrics of the models performance and ensure that SPCL is evaluated in a consistent manner with M2FNet and our proposed methods. By ensuring that SPCL is validated in a similar way to other methods, its performance drops significantly. The reported F1-score was 67.25 %. Our reproduction using their code achieved 66.53%. This drop, we hypothesize comes partially as a result of us using a batch size of 8 instead of 64 due to resource constraints since a smaller batch size will negatively affect contrastive learning (Song et al., 2022). However, upon standardization of evaluation the model's F1-score drops further to 65.58%.

Table 1 compares our methods to both M2FNet and SPCL.

6.1. Hyperparameters

A learning rate of 1e-4 was chosen empirically by performing a grid search over the hyperparameter values {1e-2, 1e-3, 5e-4}. For cross-modal and bottleneck fusion, the number

[3] Leaderboard can be found here

Method	F1 Score (%)	Accuracy (%)
M2FNet (multimodal SOTA)	66.71	67.85
Unimodal SPCL (reported)	67.25	NR
Unimodal SPCL (reproduction)	65.58	
MM-SPCL-CMPr (ours)	66.12	66.96
MSVE (ours)	66.98	68.54

Table 1: MM-SPCL-CMPr refers to multimodal SPCL pipeline with Cross-Modal Pruning. NR = Not Reported

of transformer encoder blocks consistently comprise of 5 cross-modal and 2 self-attention blocks.

For bottleneck fusion, results in Table 2 are reported with a bottleneck size of 8 and mid fusion starting after layer 2 (of 5). These two were chosen based on ablation studies presented in Table 3.

6.2. Fusion Approaches

Table 2 shows the results of various fusion approaches on MELD. With most fusion methods $\{t, a\}$ seems to perform worse than $\{t, v\}$. We hypothesize that this is due to the openSMILE audio representations not being speaker or dialog context-aware audio representations, and building such contextualized features might be necessary for the task.

6.3. Cross-modal pruning

We performed ablations with varying sizes of pruning heads with $K \in \{16, 32, 64\}$. Table 4 shows the weighted F1-scores achieved with various modality combinations as well as head sizes. Throughout our ablations, we maintained *text* as the base modality, and *audio* or *video* to be the pruning modality. We then used a single linear layer as our fusion module of choice when running ablations for cross-modal pruning. Lastly, we used cosine similarity as our similarity metric.

In all combinations, except for $\{t, a\}$, increasing the size of the heads led to an improvement in performance and in all cases the best performance was achieved by using all three modalities, followed by $\{t, a\}$, and lastly $\{t, v\}$. It should be noted, however, that 5 out of 9 of our ablations resulted in higher F1 scores than the reproduction of SPCL.

6.4. Multimodal Stochastic Voting Ensemble

We performed an ablation study over different modality combinations and notice that adding modalities and more interaction monotonically increases both the weighted F1-score and the accuracy. We found that including $\{a, v\}$ degraded the performance and thus we did not include this combination in our final model.

We further experimented with randomly dropping out modality combinations during training and observe that this doing so increases the performance even further. MSVE, both with and with-out dropout achieves a new multimodal sate-of-the-art results on both test F1-score and test accuracy (Table 5). We perform early stopping based on the dev split, unlike SPCL which performs early stopping on the test split.

7. Analysis and Discussion

7.1. Cross-modal pruning

We hypothesize that in general, the worse performance when pruning the video modality results out of the fact that the text and the video modality have little information in common, and shutting off heads essentially takes away valuable information from the modality. Further, it is important to note here that there are several ways to tune cross-modal pruning to achieve better results, but due to resource constraints, we push these to future work as highlighted in section 8.3.

Given that the pruning head most similar to each base head was shut off, a larger head size could lead to a larger part of the pruning modality being shut off, but considering that we deliberately did not stop multiple base-heads from shutting off the same pruning head, it could also lead to a smaller part of the pruning modality being shut off.

7.2. Multimodal stochastic voting ensemble

We observe interesting characteristics of MSVE such as monotonically increasing F1-score and accuracy whenever we increase the number of different modality interactions. This is in line with what one would expect when adding more information where the multiple modalities and modality combinations don't overpower or confound each other, which was one of the motivations for the modality combination drop-out. We further hypothesize that since this is essentially a type of (very) late fusion, the issue of alignment doesn't cause the performance to decrease.

Further, we observe that by performing sampling (dropping out modality combinations during training) improves metrics at test time. During inference we do not perform any sampling and add up outputs from all modality combinations. We hypothesize that by sampling during training, we are essentially making the model robust to presence or absence of information from modalities, effectively regularizing the model.

7.3. Qualitative Analysis

In table 6, we present some examples we manually verified by watching the original video clip. MM-SPCL with text and audio inputs gets audio context right in (dialogue 111, utterance 6), where the information about sadness was not present in text, and therefore, SPCL doesn't get it right. However, in another example (dialogue 1, utterance 2),

Modalities	F1-Score (%)			
	SPCL	Linear Projection (LP)	Cross-Modal Fusion (CM)	Bottleneck Fusion (BT)
f t g	65.58	-	-	-
f t,a g	-	65.66	64.92	64.45
f t,v g	-	65.43	66.15	65.91
f t,a,v g	-	66.02	65.98	65.36

Table 2: Weighted F1-score for various fusion approaches. Number of bottleneck nodes is consistently kept at 8. Mid fusion starts at layer 2 of 5.

Bottleneck Fusion	F1 Score		
# of bottleneck nodes	2	5	8
(Mid fusion start layer xed at 2)			
f ta g	63:83	64:45	65.09
f tv g	65:55	65:10	65.68
f tav g	64:58	65.36	65:02
Mid fusion start layer	1	2	3
(# of bottleneck nodes xed at 8)			
f ta g	65.18	65:09	65:17
f tv g	65:26	65.68	65:05
f tav g	63:90	65.02	64:10

Table 3: Ablation studies for bottleneck fusion

Modality combinations	F1-Score (%)		
	K=16	K=32	K=64
f ta g	66:03	66:03	65:39
f tv g	63:74	63:74	65:58
f tav g	66.09	66.05	66.12

Table 4: Ablation studies for cross-modal pruning

Method	F1 Score
MSVE (uni)	65:52
MSVE (bi)	66:39
MSVE (uni,bi)	66:46
MSVE (tri)	66:54
MSVE (uni,bi,tri)	66:93
MSVE (uni,bi,tri) + dropout	66.98

Table 5: Weighted F1-scores for MSVE

while audio tone sounds angry, the ground truth label is joy instead of anger , based on the situation in the scene. The situational information is not present in either modality, and thus, the prediction is wrong. More example dialogues are available in the Appendix 7.

8. Future Work

8.1. Sequence Alignment

By compressing each modality into a 1-D vector per utterance as described in Section 5.1, we effectively average or sample over the temporal dimension for each utterance and effectively loose out on valuable information. However, the transformer-based fusion approaches like pairwise cross-modal attention that we are using are more adept at handling temporal sequences in each modality.

Therefore, in future work, to tackle this problem, our specific proposals for the three modalities are:

1. For BERT-based text embeddings optimized using prototypical contrastive learning, use the whole last hidden state layer instead of the mask embedding.
2. For audio representations, build a new feature extractor. We will cover this in Section 8.2 in depth.
3. For video representations, instead of sampling the middle frame of a video, retain representations of all frames with a higher frame rate.

8.2. Generating better audio representations

For the analysis in this paper, we directly fused audio features present in MELD generated through openSMILE (Eyben et al., 2010) with the textual representation of emotions in SPCL. It is important to note that these audio features are not learned for emotion prediction or to have explicit character awareness. The audio features are also devoid of any temporal context within the dialogues. Further, we know from our previous analysis (scatter plot in gure 6 in Appendix B) that audio features are not separable by emotions using t-SNE unlike glove-based text features. This points to a need for contrastive-learning based audio representations that help hard examples belonging to different emotions to be pushed farther and similar ones to come closer.

Therefore, we propose to learn audio features via multi-task contrastive learning. The framework will follow a feature extractor F_a followed by two classi ers for the tasks in

Dia,Utt ID	Utterance	True Label	SPCL f tg	MM-SPCL f tag	Remarks
1, 2	Joey: Push 'em out, push 'em out, harder, harder.	joy	joy	anger	Assertive, almost angry tone
1, 3	Joey: Push 'em out, push 'em out, way out!	joy	anger	anger	Assertive, almost angry tone
1, 4	Joey: Let's get that ball and really move, hey, hey, ho, ho.	joy	joy	joy	Switches to playful joyous tone
111, 3	Chandler: You kissed my best Ross!	anger	joy	anger	
111, 4	Mrs. Bing: O-kay. Look, it, it was stupid.	sadness	sadness	sadness	
111, 5	Chandler: Really stupid.	anger	disgust	disgust	
111, 6	Mrs. Bing: Really stupid.	sadness	disgust	sadness	MM-SPCL gets audio context right

Table 6: Disagreements between MM-SPCL tag and SPCL on the test set, and our remarks after listening to the raw audio files for the corresponding dialogue and utterance files. The values under SPCL and MM-SPCL denote the predictions of the corresponding model.

question, i.e. emotion recognition and character prediction, to allow for the features learned by F_a to have information about emotion in the utterance as well as the speaker. The features from F_a will also be used for contrastive learning to allow for these features to be separable in the emotion space. Next, considering each dialogue as a sequence of audios, we will use a recurrent network (Cho et al., 2014; Hochreiter & Schmidhuber, 1997) that takes in the learned features from F_a for the audio corresponding to each turn, and train it for ERC (in conversation now, since we have context). The final encodings will be residually added to the input representations to preserve information.

These new audio features will also help with the sequence alignment issue in Section 8.1.

8.3. Better pruning techniques

As highlighted in section 6.3, the cross-modal pruning module has several points of tuning available which were not exploited here owing to resource constraints. Therefore, we propose three avenues for future work that we believe can give large boosts to performance. This comprises of 1) using more sophisticated fusion techniques than linear projection, 2) using a similarity threshold instead of looking for the most similar pruning head for each base head, and 3) projecting the base and the pruning modalities in a common space before splitting into multiple heads.

9. Conclusion

In this work, we experiment with various approaches to build a multimodal supervised prototypical learning framework, including three representation (early) fusion approaches, and one late fusion ensembling technique with residual learning. In similar training settings (primarily constrained batch size), our MSVE algorithm achieves a weighted average F1 score of 66.98, which beats the previous multimodal SOTA by +0.27 points, and beats our reproduction of the current unimodal SOTA by +1.40 points.

References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., and Onoe, N. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4652–4661, 2022.

Ekman, Friesen, O. C. D.-T. H. K. L. P. R.-B. Universals and cultural differences in the judgments of facial expressions of emotion, 1974. URL https://pubmed.ncbi.nlm.nih.gov/3681648/ .

Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H., and Ku, L.-W. EmotionLines: An emotion corpus of multi-party

conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1252.

Kim, T. and Vossen, P. Emoberta: Speaker-aware emotion recognition in conversation with roberta, 2021. URL https://arxiv.org/abs/2108.12009.

Kim, Y. Convolutional neural networks for sentence classification, 2014. URL https://arxiv.org/abs/1408.5882.

Levis, J. M. Intonation in theory and practice, revised. TESOL quarterly, 33(1):37–63, 1999.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. URL https://arxiv.org/abs/2107.13586.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations, 2018. URL https://arxiv.org/abs/1811.00405.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE transactions on affective computing, 3(1):5–17, 2011.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems, 34: 14200–14213, 2021. URL https://proceedings.neurips.cc/paper/2021/file/76ba9f564ebbc35b1014ac498fafadd0-Paper.pdf.

Pérez-Rosas, V., Mihalcea, R., and Morency, L.-P. Utterance-level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 973–982, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1096.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1081. URL https://aclanthology.org/P17-1081.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations, 2018. URL https://arxiv.org/abs/1810.02508.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. URL https://arxiv.org/abs/2103.00020.

Saxena, P., Huang, Y. J., and Kurohashi, S. Static and dynamic speaker modeling based on graph neural network for emotion recognition in conversation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pp. 247–253, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-srw.31. URL https://aclanthology.org/2022.naacl-srw.31.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. Advances in neural information processing systems, 30, 2017a.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning, 2017b. URL https://arxiv.org/abs/1703.05175.

Song, X., Huang, L., Xue, H., and Hu, S. Supervised prototypical contrastive learning for emotion recognition in conversation. arXiv preprint arXiv:2210.08713, 2022.

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 conference. Association for Computational Linguistics. Meeting, volume 2019, pp. 6558. NIH Public Access, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259, 2016.

Zadeh, A. B., 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), pp. 2236–2246, 2018.

A. Appendix: Detailed Dialogue Examples

Dia,Utt ID	Utterance	True Label	SPCL *ftg*	MM-SPCL *ftag*	Remarks
1, 0	Joey: Come on, Lydia, you can do it.	neutral	neutral	neutral	
1, 1	Joey: Push!	joy	joy	anger	
1, 2	Joey: Push 'em out, push 'em out, harder, harder.	joy	joy	anger	Assertive, almost angry tone
1, 3	Joey: Push 'em out, push 'em out, way out!	joy	anger	anger	Assertive, almost angry tone
1, 4	Joey: Let's get that ball and really move, hey, hey, ho, ho.	joy	joy	joy	Switches to playful joyous tone
85, 0	Joey: But um, I don't think it's anything serious.	neutral	neutral	neutral	Chandler and Joey are scared
85, 1	Chandler: This sounds like a hernia. You have to—you-you—Go to the doctor!	surprise	anger	fear	since Joey is in pain, and
85, 2	Joey: No way!	anger	anger	fear	fear explains the emotion better than anger.
85, 3	Joey: 'Kay look, if I have to go to the doctor for anything it's gonna be for this thing sticking out of my stomach!	anger	anger	fear	Therefore, the ground truth is questionable
111, 3	Chandler: You kissed my best Ross!	anger	joy	anger	
111, 4	Mrs. Bing: O-kay. Look, it, it was stupid.	sadness	sadness	sadness	
111, 5	Chandler: Really stupid.	anger	disgust	disgust	
111, 6	Mrs. Bing: Really stupid.	sadness	disgust	sadness	MM-SPCL gets audio context right
279, 11	Rachel: Yeah, I mean, come on Ross, no one will even notice...	neutral	anger	neutral	
279, 12	Ross: They're not listening too me?	surprise	anger	surprise	MM-SPCL gets audio context right
279, 13	Rachel: Of course they're listening to you! Everybody listens to you.	neutral	anger	neutral	

Table 7: (Full) Disagreements between MM-SPCL and SPCL on the test set, and our remarks after listening to the raw audio files for the corresponding dialogue and utterance files. The values under SPCL and MM-SPCL denote the predictions of the corresponding model.

Dia,Utt ID	Utterance	True Label	SPCL *ftg*	MM-SPCL *ftag*	Remarks
17, 3	Ewww! Oh! It's the Mattress King!	disgust	surprise	disgust	
17, 4	Don't look honey. Change the channel! Change the channel!	disgust	anger	anger	Disgust portrayed in audio and video, not text
17, 5	Wait! Wait! I wanna see this. After I divorce him, half of that kingdom is gonna be mine.	joy	surprise	surprise	
17, 6	What a wank!	anger	disgust	disgust	Angry voice and facial expressions
88, 8	Joey: Can we please turn this off?	sadness	anger	anger	Sadness clearly conveyed in audio and video
88, 9	Rachel: Noo way, Kevin.	joy	disgust	disgust	Teasing, sarcastic, but joyous not disgust

Table 8: Examples of erroneous agreements between SPCL and MM-SPCL, and our remarks after listening to the raw video files for the corresponding dialogue and utterance files. The values under SPCL and MM-SPCL denote the predictions of the corresponding model.