Multilingual Multimodal Sentiment Analysis

Amelia Kuang Shuhao Zhang Yuchen Zhou Yuxin Pei Zhe Chen

Abstract

Multimodal sentiment analysis, extracting sentiments from language, acoustic, and visual features, is a widely applicable task around the world. However, very few works have been done in languages other than English. Building a universal model for multilingual sentiment analysis is challenging due to each language's vastly different vocabularies and grammar. To address this issue, we propose to build a multilingual multimodal sentiment detection model in the framework of transfer learning. Specifically, we generate cross-lingual paired texts through translation and adopt Noise Contrastive Estimation (NCE) as the learning objective. We also explore Adaptive MML that filters out non-representative modalities in the flow for efficient and robust sentiment analysis. We empirically show that our methods outperform the baseline models in the multilingual multimodal sentiment analysis task under zero-shot and few-shot settings.

1. Introduction

Sentiment Analysis (SA) is the computational study of people's opinions, attitudes, and emotions (Medhat et al., 2014). People express sentiments in texts, facial expressions, speech utterances, gestures, and much more. SA is thus inherently a multimodal task. It can be applied to fields including consumer information, marketing, social media, and medical services, which heavily depend on the analysis of consumers' sentiments (Hussein, 2018). However, most studies in SA are focused on mainstream languages such as English. The lack of research in low-resource languages (Bagher Zadeh et al., 2020) dramatically limits the applicability of SA.

To address the imbalance across languages, we aim to build a multimodal model to predict the sentiment from visual, acoustic, and language features across different languages. We propose four ideas as follows: first, we explore a transfer learning paradigm in which cross-lingual pairs of transcripts are generated by translation augmentation and the model adopts the Noise-Contrastive Estimation (NCE) as the training objective to learn a multilingual multimodal embedding space. Second, we experiment with Adaptive MML that uses a hard attention mechanism to filter out non-representative modalities in the flow for the sake of efficient and robust sentiment analysis. Third, we perform classifications under a simple few-shot setting based on Adaptive MML. Finally, inspired by the results on sentiment analysis, we further expand the scope of our study to emotion detection with a focus on acoustic-centric methods.

In the following sections, we will first briefly introduce our multilingual multimodal dataset and review prior literature on multimodal sentiment analysis. Then we give a formal description of the addressed research question. Finally, we explore four research ideas by providing a detailed experimental setup and result analysis. We also discuss research directions to explore in the future.

2. Related Work

Unimodal Sentiment Analysis Prior SA works primarily focused on a single modality, namely text or acoustics. For text-based sentiment analysis, early works (Turney, 2002) manually created lexicon patterns while more recent works created models (such as ELMo (Peters et al., 2018) and CoVe (McCann et al., 2017)) for contextualized word representations and demonstrated improvements in sentiment analysis on Stanford Sentiment Treebank (Socher et al., 2013). Recently, state-of-the-art results have been achieved through pre-trained deep language models such as BERT (Devlin et al., 2019). Speech emotion recognition techniques seem to be divided among researchers who use on engineered prosodic features (Heracleous & Yoneyama, 2019) and researchers who apply various deep learning techniques (Abbaschian et al., 2021), including convolutional network and generative networks.

Several studies have extended these monolingual models to multilingual settings. A study (Lausen & Hammerschmidt, 2020) reported good performance by training a language classifier on top of two emotion classifiers for English and German using mel-frequency cepstral coefficients concatenated with shifted delta cepstral coefficients. Another study (Neumann & Vu, 2018) trained an attentive convolutional neural network for bilingual speech emotion recognition. Although the researchers saw promising results from finetuning a monolingual model for cross-lingual recognition, they have noted the differences in the attention mechanism outputs in bilingual recognition, which indicates the presence of language-dependent features to detect emotions in speeches.

Multimodal Sentiment Analysis Sentiment analysis is inherently a multimodal task as human emotions are often expressed by multiple modalities simultaneously. An early comparative research (Rosas et al., 2013) shows that by fusing text, audio, and visual modalities in the most straightforward manner (vector concatenation), the classification accuracy of predicting opinions of Spanish online videos will be significantly improved compared to single modality models. In a similar research (Morency et al., 2011), they proved that the idea of multimodal can also benefit more complicated models like HMM.

Multilingual Representation with Machine Translations Bornea et al., 2020 aims to create a multilingual embedding space that takes semantic meanings instead of the language into account. To enforce the multilingual setting, they augment the original English training data with machine translation-generated data. They also experimented with two frameworks (Adversarial Training and Language Arbitration) so that embeddings of different languages are close to each other. Adversarial Training combines two loss functions; one drives the discriminator to recognize the language label, the other encourages the embeddings to appear uniform to the discriminator. In the Language Arbitration Framework, an additional loss is used to encourage agreement between embeddings of English and translation while also minimizing the cosine-similarity between the answers generated.

Huang et al., 2021 proposes a multilingual, multimodal pretraining strategy, employing a Transformer-based model with contrastive objectives to learn contextual multilingual multimodal representations. The model is trained with multilingual video-text data. The textual and visual modalities are encoded by multilingual BERT (Devlin et al., 2019) and 3D-CNNs, respectively, both followed by a two-layer Transformer pooling head.

3. Dataset

We use the CMU Multimodal Opinion Sentiment, Emotions and Attributes dataset (CMU-MOSEAS) (Bagher Zadeh et al., 2020). It provides recordings of synchronous visual, audio and transcript information in four different European languages (French, Spanish, Portuguese and German). It contains 1000 videos each for Spanish, Portuguese, German, and French. The dataset follows the Likert scale for sentiment label annotations, ranging from $[-3, 3]$: -3 being highly negative, 0 being neutral and 3 being highly positive. French and Spanish are selected as the training and testing languages, with Spanish simulating a low-resource language. For each language, we take 80% of the data for training, 10% for validation and 10% for testing.

As shown in Figure 1, neutral sentiment scores around 0 make up the majority and result in highly imbalanced distribution. There are also more mildly positive sentiments than negative sentiments. Hence, we removed all segments with neutral sentiment (in range $[-0.5, 0.5]$).

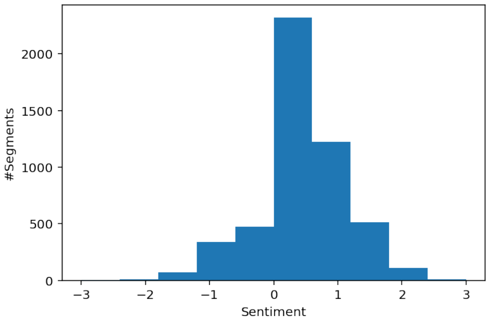

Figure 1. CMU-MOSEAS French subset label distribution

Acoustics We use Librosa (Brian McFee et al., 2015) to perform feature extractions. Based on our experiments, we find that MFCC and melspectrogram features are not as useful in sentiment analysis as chromagram, spectral contrast and tonal centroid features. Therefore, our final models only use the later features totaling 25 dimensions.

Text We use French and Spanish GloVe word embeddings (Pennington et al., 2014) obtained from Ferreira et al. 2016 [1] to get word-level features of 300 dimensions.

Visual We use OpenFace (Baltrusaitis et al., 2018) to extract visual features related to the speakers' facial expressions. Based on past papers using OpenFace features, we only use 35 Facial Action Units to train our models.

Alignment Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) is used to obtain word-level timestamps. The visual and acoustic features are aligned to the words using the average over the utterance intervals.

4. Problem Formulation

We formulate the problem as a regression problem to predict the sentiment score of a speaker in a monologue video on a scale of $[-3, 3]$.

A multilingual multimodal dataset covers L languages. Each

[1] www.cs.cmu.edu/ afm/projects/multilingual_embeddings.html

language subset consists of N_l labeled video segments.

$$\mathbf{X}^l = (\mathbf{X}_0^l, \mathbf{X}_1^l, ..., \mathbf{X}_{N_l}^l) \tag{1}$$

Each segment is defined as

$$\mathbf{X}_i^l = (\mathbf{w}_i^l, \mathbf{v}_i^l, \mathbf{a}_i^l) \tag{2}$$

composed of text ($\mathbf{w}_i^l$), visual ($\mathbf{v}_i^l$), acoustic modalities ($\mathbf{a}_i^l$), respectively, along with the sentiment score.

The language features are given by

$$\mathbf{w}_i^l = (w_i^{(1)}, w_i^{(2)}, ..., w_i^{(T_i)}), w_i^{(t)} \in \mathbb{R}^{D_w} \tag{3}$$

where $w_i^{(j)}$ denotes the jth word and T_i the length of that segment.

Similarly for the visual features,

$$\mathbf{v}_i^l = (v_i^{(1)}, v_i^{(2)}, ..., v_i^{(T_i)}), , v_i^{(t)} \in \mathbb{R}^{D_v} \tag{4}$$

and acoustic features

$$\mathbf{a}_i^l = (a_i^{(1)}, a_i^{(2)}, ..., a_i^{(T_i)}), a_i^{(t)} \in \mathbb{R}^{D_a} \tag{5}$$

The corresponding labels for these N_l segments are denoted as

$$\mathbf{y} = (y_1, y_2, \ldots, y_{N_l}) \tag{6}$$

The objective is to minimize the distance $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ between the predicted labels $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{N_l})$ from the validation set given $\mathbf{X}^l$ from the training set.

$$\min_W \mathcal{L}(\hat{y}, y) \tag{7}$$

Two common choices of the distance function $\mathcal{L}$ are L1 loss $||\hat{\mathbf{y}} - \mathbf{y}||_1$ and L2 loss $||\hat{\mathbf{y}} - \mathbf{y}||_2$.

5. Proposed Approach

5.1. Cross-Lingual Transfer Enabled by NCE Loss and Translation Augmentation

We wish to explore further how to meaningfully represent the different modalities and languages and study how they interact with each other. To this end, we do not simply train a multimodal model in multiple runs with multilingual data without taking into account any interactions between the different languages, but rather we model the interactions between modalities and languages to find a coordinated representation that enables cross-lingual transfer. This setting also allows the possibility of learning multilingual representations with uni-lingual dataset augmented with machine translations. Since the data in different languages in CMU-MOSEAS is independent and we do not have multilingual transcripts for the same video, we utilize translated transcripts (for example, French transcripts and their Spanish translations, along with French videos) during training time to build a cross-lingual multimodal space and study the model's performance with actual data in the translated language (Spanish data in CMU-MOSEAS).

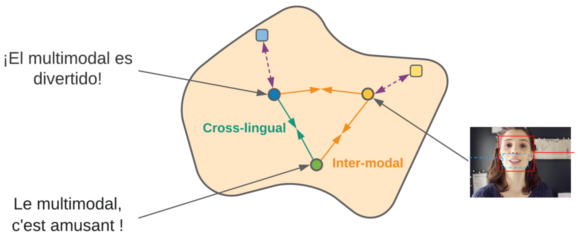

Figure 2. Illustration of NCE Loss

Inspired by the work in Huang et al., 2021, we aim to learn a coordinated, cross-lingual multimodal embedding space that enforces similarities between the different modalities and languages. The learned embeddings will facilitate the downstream task of sentiment analysis. One key intuition is that conditioned on a video, multilingual transcripts should be semantically similar. In addition to the MSE loss for the final regression task, we use two contrastive objectives to model the inter-modal and cross-lingual representations between the modalities and languages.

We use the following Noise-Contrastive Estimation (NCE) objective in our contrastive objectives:

$$\mathcal{L}(X, Y) = \max \sum_{i=1}^{n} \log L_{\text{NCE}}(x_i, y_i)$$

where

$$L_{\text{NCE}}(x_i, y_i) = \frac{e^{x_i^T y_i}}{e^{x_i^T y_i} + \sum_{(x', y') \in \mathcal{N}} e^{x'^T y'}}$$

where the positive pair (x_i, y_i) is defined as the translated pair of transcripts for the same video, and the set of negative pairs $\mathcal{N}$ is defined as unmatched original and translated transcripts.

Pivoted Cross-Lingual Loss Inspired by Huang et al., 2021, we propose an objective that aligns transcripts in different languages that describe the same video. Our intuition is that conditioned in a video, the multilingual transcripts (translated pairs, in our case) should be semantically similar.

Because we have both acoustic and visual modalities for each video, we expand beyond Huang et al., 2021 and explore different ways of conditioning languages based on acoustic or visual modalities.

If we condition on the visual modality only, our cross-lingual loss is defined as

$$\mathcal{L}^{\text{cross}}(T \mid V) = \mathcal{L}(T_1 \mid V_1, T_2 \mid V_1)$$

acoustic modality only:

$$\mathcal{L}^{\text{cross}}(T \mid A) = \mathcal{L}(T_1 \mid A_1, T_2 \mid A_1)$$

both visual and acoustic modalities:

$$\mathcal{L}^{\text{cross}}(T \mid V, A) = \mathcal{L}^{\text{cross}}(T \mid V) + \mathcal{L}^{\text{cross}}(T \mid A)$$

where T_1 and T_2 represent the language features in language L_1 and L_2, respectively, V_1 and A_1 represent the visual and acoustic features in language L_1. Note that T_2 is the corresponding translation from T_1.

Inter-modal Loss Let T and V denotes the subsampled multi-lingual texts and videos. We model the inter-modal loss defined as

$$\mathcal{L}^{\text{inter}} = \mathcal{L}(T, V)$$

MSE Loss We use the MSE loss defined as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

We optimize our model by

$$\max \text{MSE} + \mathcal{L}^{\text{cross}} + \mathcal{L}^{\text{inter}}$$

5.2. Hard Modality-based Attention

After inspecting some of the video segments in the training data, we found that part of the videos contains silence or simply unrelated content such as advertisements. We hypothesize that removing these segments may improve model performance due to the gating of irrelevant information to our task.

Furthermore, Panda et al. suggested that hard modality-based gating at subsegment level, that is, selecting the optimal subset of modalities in each shorter window of frames within the video segment, can be an efficient way of accomplishing multimodal recognition. This modality selection is accomplished through a Policy Network that uses the Gumbel SoftMax trick (Jang et al., 2016), allowing end-to-end training of the Policy Network without the need of secondary loss functions.

In addition, by zeroing out features from specific modalities entirely within each window, the intricate attention mechanism can make the inference process more efficient as all-0 matrix multiplication can be trivially optimized.

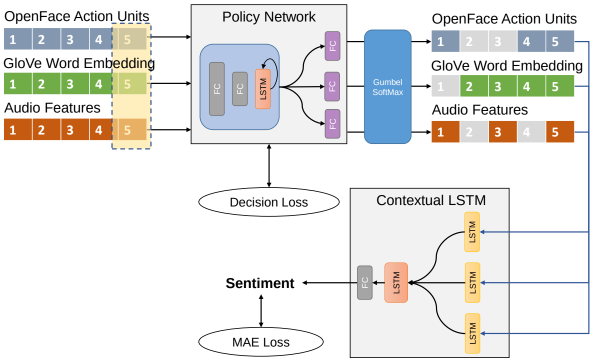

Figure 3. Model Illustration of AdaMML + Contextual LSTM

However, this paper originally use simple late fusion in the model, which may not be ideal for a more complex task like sentiment recognition. As such, we propose to adapt the Contextual LSTM fusion technique introduced by Poria et al. to AdaMML to improve its fusion capability. Figure 3 illustrates our proposed model. We use the standard many-to-one approach (Karpathy, 2015) for predicting sentiment: use the last output of the final contextual LSTM as the sentiment feature vector

Two losses are used to train this model:

Mean Absolute Error (MAE) is used to train the sentiment regression model. It measures the difference between the predicted sentiment intensity $\hat{y}$ and the label y and is defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

The optional decision loss (DL) can be used to control the average number of modalities that the policy network selects at each timestep (window). The target number of modalities can then be varied to study how the policy network learned to select modalities and which modalities are more critical. It is defined as:

$$\text{DL} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i}\sum_{l=1}^{L_i}|\sum_{m}(\text{PL}_{l,m} = 1) - \text{Target}|$$

where $\text{PL}_{l,m}$ is the Policy Network's decision at timestep l regarding modality m.

5.3. Few-shot Transfer Learning

A simple idea in transfer learning is to fine-tune a pre-trained model on the new dataset, hoping that the model would generalize its learned task knowledge on the new dataset. However, neural networks tend to catastrophically forget what was learned in the past (Kemker et al., 2018). This means fine-tuning a pre-trained model on a small new dataset may render it incapable of performing well on the original dataset.

It may even overfit the new dataset instead of generalizing on the new dataset.

To overcome this tendency, we adopt the Rehearsal method mentioned in Kemker et al. that mixes the original dataset and the new dataset during model training.

Specifically, the full French training dataset and a small, adjustable percentage of the Spanish training dataset are mixed together along with their corresponding labels to create the new training dataset used in model training.

5.4. Acoustic-centric Emotion Detection

During our testing on existing state-of-the-art models on sentiment analysis, we found that the text modality is the most dominant and useful modality in predicting sentiment. This is most likely because the task of sentiment analysis is more related to what the speaker spoke, i.e., text (Socher et al., 2013), than how the speaker delivered it, i.e., voice and facial cues.

However, in a multilingual setting with low-resource languages, these low-resource languages likely do not have a good text embedding model for the sentiment analysis model to use.

Hence, we would like to explore the use of predominantly audio and facial cues to predict an attribute that relates more to these features. Given the labels from the CMU-MOSEAS dataset, we want to predict the binary emotion labels.

Formally, given the true labels y_i and the model's predicted labels $\hat{y_i}$ given a set of multimodal features, Binary Cross Entropy loss (equation 8) is used to optimize the model.

$$-\frac{1}{N^l}\left[\sum_{y_i=1}\log\sigma(\hat{y_i})+\sum_{y_i=0}\log(1-\sigma(\hat{y_i}))\right] \tag{8}$$

6. Experimental Setup

Evaluation Metrics To evaluate the model performance, we used Mean Absolute Error (MAE) and Pearson's correlation r. A lower MAE value or higher Pearson's correlation indicates stronger performance.

The Pearson's correlation coefficient is defined as

$$r=\frac{\sum(x_i-\hat{x})(y_i-\hat{y})}{\sqrt{\sum(x_i-\hat{x})^2\sum(y_i-\hat{y})^2}}$$

6.1. Baseline Models

6.1.1. MCTN

Architecture We use a trimodal variation of Multimodal Cyclic Translation Network (Pham et al., 2019) (MCTN) as the baseline model. MCTN uses attention-based Seq2Seq models (Bahdanau et al., 2014) to translate a source modality $\mathbf{X}^l_{i,S}$ into its corresponding target modality $\mathbf{X}^l_{i,T}$. The encoder f_{θ_e} takes in $\mathbf{X}^l_{i,S}$ and encodes it into a joint embedding $\mathcal{E}_{S\to T}$. The decoder f_{θ_d} then decodes $\mathcal{E}_{S\to T}$ into the target modality $\mathbf{X}^l_{i,T}$. The forward translation loss component is defined as

$$\mathcal{L}_t=\frac{1}{N_t}||\mathbf{X}^l_{i,T}-f_{\theta_d}(f_{\theta_e}(\mathbf{X}^l_{i,S}))||_2 \tag{9}$$

This is essentially the Mean Squared Error (MSE) between the decoded information and ground truth.

To further encourage the model in learning a more informative joint embeddings $\mathcal{E}_{S\leftrightarrows T}$ instead of $\mathcal{E}_{S\to T}$, a cycle loss component is added. It forces the model to translate the decoded information $\hat{\mathbf{X}}^l_{i,T}=f_{\theta_d}(f_{\theta_e}(\mathbf{X}^l_{i,S}))$ back to the source modality. The cycle loss is defined as

$$\mathcal{L}_c=\frac{1}{N_t}||\mathbf{X}^l_{i,S}-f_{\theta_d}(f_{\theta_e}(\hat{\mathbf{X}}^l_{i,T}))||_2 \tag{10}$$

This is essentially the Mean Squared Error (MSE) between the cyclic decoded information and ground truth.

To generate sentiment predictions, the encoded joint representation $\mathcal{E}_{S\leftrightarrows T}$ is passed through a prediction RNN $\mathbf{g}$ to get $\mathbf{g}^l_i$, the RNN output of the final time step. The final prediction is generated using a final linear transformation f_{θ_p} of the output from the last time step $\hat{y}_i=f_{\theta_p}(\mathbf{g}^l_i)$

Loss of sentiment prediction is defined as the Mean Absolute Error between the predicted value and ground truth:

$$\mathcal{L}_p=\frac{1}{N_t}||\hat{y}^l_i-y^l_i||_1 \tag{11}$$

The model is trained end-to-end using the coupled objective function defined as

$$\mathcal{L}=\mathcal{L}_p+\lambda_t\mathcal{L}_t+\lambda_c\mathcal{L}_c \tag{12}$$

where λ_t and λ_c are weighing hyperparameters. The model is learned by minimizing $\mathcal{L}$.

At test time, only the source modality $\mathbf{X}^l_S$ is required to generate predictions since the encoder f_{θ_e} has already learned to encode the source modality into a joint embedding $\mathcal{E}_{S\leftrightarrows T}$.

To extend the framework into three modalities, the authors of MCTN tested a few variations of the model and found that the hierarchical architecture worked best. In the hierarchical architecture. A second Seq2Seq model is trained to cyclically translate between $\mathcal{E}_{S\leftrightarrows T}$ and a third modality $\mathbf{X}^l_{T'}$. The encoder $f_{\theta_{e2}}$ of the second Seq2Seq model should produce a joint embedding of all three modalities

$\mathcal{E}_{(S\leftrightarrows T)\leftrightarrows T'}$, which is then fed into an RNN for the final sentiment prediction.

Training details We set $\lambda_t = \lambda_c = 0.5$ for bimodal models and $\lambda_t = \lambda_c = 0.3$ for trimodal models to account for the added translation and cycle losses from the second Seq2Seq model.

A missing training detail from the original MCTN paper is the way to handle feature dimension mismatch between modalities. Our implementation pads the feature dimension of each modality with 0s such that feature vectors of all modalities are of the same length, with features of each modality occupying different subspaces of the padded feature vector. This design allows the model to identify different modalities at the expense of sparser weight matrices.

We also implement a trimodal variant of MCTN that uses one encoder for one source modality $\mathbf{X}_S^l$ and two decoders for two different target modalities using the same embedding $\mathcal{E}_{S\leftrightarrows T,T'}$. This embedding is fed directly to the prediction RNN for the final sentiment prediction. We label this variant as "shared encoder".

6.1.2. Multilogue-Net

Architecture Shenoy & Sardana 2020 proposes the Multilogue-Net, an end-to-end RNN architecture with a bimodal attention mechanism that aims to capture the context of the conversation through all modalities, the dependency between the listener(s) and speaker emotional state, and inter-modal and cross-modal relationships. The model consists of two GRUs (state GRU and emotion GRU) for every modality and participant and a context GRU for each modality common to all participants in the conversation. Because all our videos are monologues that do not involve conversing between more than one party, we set the speaker input to be the same across all video segments.

The context GRU for each modality encodes the utterance representation of that modality ($w_i^{(t)} \in \mathbb{R}^{D_w}, v_i^{(t)} \in \mathbb{R}^{D_v}, a_i^{(t)} \in \mathbb{R}^{D_a}$) and the previous time-stamp speaker state GRU output of that modality.

For a timestamp t, the state GRU takes in input feature representation of that modality and simple attention overall context vectors until that timestamp.

The emotion GRU serves as a decoder for the encoded representation produced by the state GRU: it uses the previous timestamp GRU output and the current timestamp sGRU output to produce a representation that feeds into the pairwise bimodal attention mechanism. The pairwise attention produces the final prediction output.

Loss Function and Evaluation Metrics The Multilogue-Net is able to perform both emotion classification and sentiment regression. We focus on the sentiment score so our choice of loss function is the Mean Squared Error (MSE) along with L2 regularization:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2$$

Training Details The model interleaves dropout with the GRU cells, and the dropout rate is set to 0.25. The Adam optimizer with learning rate 10^{-4} and weight decay 10^{-4} is used with batch size 128 to perform gradient descent. Table 1 shows the performance of Multilogue-Net on the CMU-MOSEAS French subset.

6.1.3. Multimodal Transformer

Model description At present, most multimodal sentiment analysis models require alignments between different modalities before feeding into the model, that is, to separate and reorganize the time-series data based on certain criteria like word or wave diagram of the voice stream. Such a prepossess step seems to be reasonable and would probably lower the implementing difficulty for some models. However, on the other hand, it potentially increases the difficulty of analyzing human language due to the inherent heterogeneity of multimodal data. First, the receptors for different modalities may differ in sample rate. For example, sensors, or even human sense organs, for audio and visual streams have different receiving frequencies, where the existence of an optimal mapping method between them is not guaranteed. Second, the long-lasting relationship on time-series data may not be well captured by a fixed alignment method. For example, a frowning on one face may be related to a pessimistic word occur a long time ago. Therefore, instead of forcing a model to use aligned features, following the unaligned nature of multimodal data may be a reasonable way to further improve the performance of a sentimental analysis.

Multimodal Transformer (MulT) introduced in (Tsai et al., 2019) is an end-to-end model that extends the standard Transformer network(Vaswani et al., 2017) to learn representations directly from unaligned multimodal streams. The key idea of the MulT model is the cross-modal attention module that attends to the cross-modal interactions at the scale of the entire utterances. The overall architecture is as follow:

For all three modalities (here we use text(L), video(V) and audio(A) but not necessary these three), in order to collect sufficient awareness of the neighborhood elements, we first pass the preprocessed modality embeddings to a 1D temporal convolutional layer:

$$\hat{X}_{\{L,V,A\}} = \text{Conv1D}(X_{\{L,V,A\}}, k_{\{L,V,A\}}) \in \mathcal{R}^{T_{\{L,V,A\}}\times d}$$

After the convolution layer, the output is expected to contain the local structure from their source so that the difference

in sampling rates can be eliminated. Besides, the outputs are in the same length d, which makes the cross-modal attention later possible.
Then, in order to attach temporal information, the augment positional embedding is added:

$$Z^{[0]}_{\{L,V,A\}} = \hat{X}_{\{L,V,A\}} + PE(T_{\{L,V,A\}}, d)$$

Where PE is the commonly used positional embedding based on a series of trigonometric functions defined in (Vaswani et al., 2017).
The cross-modal transformers are then defined by:

$$\hat{Z}^{[i]}_{\{\alpha\to\beta\}} = CM^{[i],mul}_{\{\alpha\to\beta\}}(LN(Z^{[i-1]}_{\{\alpha\to\beta\}}), LN(Z^{[0]}_{\{\alpha\}})) + LN(Z^{[i-1]}_{\{\alpha\to\beta\}})$$
$$Z^{[i]}_{\{\alpha\to\beta\}} = f_{\theta^{[i]}_{\{\alpha\to\beta\}}}(LN(\hat{Z}^{[i]}_{\{\alpha\to\beta\}}) + LN(\hat{Z}^{[i-1]}_{\{\alpha\to\beta\}}$$

Where α,β represent two selected modalities, i is the index of layer start from 1, LN means layer normalization, $CM^{[i],mul}$ is the multi-head cross-modal attention defined by:

$$CM_{\beta\to\alpha}(\mathcal{X}_\alpha, \mathcal{X}_\beta) = \text{softmax}(\frac{\mathcal{X}_\alpha \mathcal{W}_{Q\alpha} \mathcal{W}^T_{K\beta} \mathcal{X}^T_\beta}{\sqrt{d_k}})\mathcal{X}_\beta \mathcal{W}_{V\beta}$$

Training and Validating Details All training and validating data are from the CMU-MOSEAS dataset, which each modality is preprocessed as discussed in Chapter 5. Notice that different from the original MulT paper, we use both regression action units and classification action units as visual feature inputs.

In order to be consistent with our other baseline experiments, we still use the MFA-aligned data. We first pad all aligned input blocks to $100 \times d_M$ matrices, where d_M is the length of feature vectors of each modality. Then we still run the model with unaligned mode, which means the transformer is actually working on a relatively short range of stream.

The model is optimized by Adam optimizer with initial learning rate $1e-3$, we use plateau learning rate reduction with a tolerance of 20 steps. The loss function is mean absolute error defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}|$$

7. Results and Discussion

7.1. Cross-Lingual Representations with NCE Loss and Translation Augmentation

We use Multilogue-Net (described in section 6.1.2) as the testbed and experiment with the contrastive objectives. Specifically, we use the evaluation result on Spanish datasets as the metric for language transfer.

We modify the architecture of Multilogue-Net (Shenoy & Sardana, 2020) by feeding the hidden states outputted by the pairwise bi-modal attention to the NCE loss and perform optimization. For example, for cross-lingual NCE loss conditioned on the visual modality, we optimize

$$\mathcal{L}^{\text{cross}}(T \mid V) = \mathcal{L}(T_1 \mid V_1, T_2 \mid V_1)$$

where $T_1 \mid V_1$ is the hidden state of bi-modal attention applied to the textual features T_1 and the visual features V_1, $T_2 \mid V_1$ is the hidden state of bi-modal attention applied to the textual features T_2 (translated from T_1) and the visual features V_1.

As illustrated in Table 1, with TransAug + NCE loss, there is a consistent improvement upon the baseline settings. This proves that our model is able to pick up multilingual cues when trained with translations. However, as expected, the improvements are not as good as directly training on the Spanish CMU-MOSEAS subset. We hypothesize that this is due to the lack of paired video or audio data with the corresponding Spanish texts and the noise introduced by machine translation, resulting in lower-quality data.

In addition, because language is the key component of multilingual multimodal sentiment analysis, we also hypothesize that our performances could be further improved with more robust multilingual language-level features, such as using multilingual BERT (Devlin et al., 2018). In this work, we use GloVe word embeddings (Pennington et al., 2014) to represent textual data, whereas Huang et al., 2021 trains mBERT with gradients from the contrastive objectives. We thus hypothesize that employing mBERT to encode textual features and training mBERT with gradients from the downstream task would improve the performances, but due to time constraints, we leave this as a future direction to explore.

We also perform ablation experiments to study which modalities among visual and acoustic provide the most information to serve as the ”pivot” to align cross-lingual transcripts in the multilingual multimodal embedding space. As shown in Table 1, cross-lingual NCE of the text modality conditioned on the video modality yields the best transfer result.

7.2. AdaMML

Due to time constraints and the limitation of the original fusion methodology, we did not test the original AdaMML and instead tested the AdaMML policy network combined with Contextual LSTM fusion directly. The results are shown in Table 1. We also plotted the average modality utilization rate from the AdaMML's policy network in Figure 4.

We observe that while the original AdaMML with Contextual LSTM fusion performed comparably to the best multimodal baseline models, the introduction of decision loss

Model	Variations	Training Data	Eval Data	r(↑)	MAE(↓)
Baseline Models	*Multilogue-net* T + V + A	FR	FR	0.35	0.50
	Multilogue-net T + V + A	FR	ES	0.28	0.83
	Multilogue-net T + V + A	FR + ES	FR	0.48	0.46
	Multilogue-net T + V + A	FR + ES	ES	0.54	0.55
	MCTN T $\leftrightarrows$ A	FR	FR	0.35	0.53
	MulT T	FR	FR	0.12	0.62
AdaMML + Contextual LSTM	No DL target	FR	FR	0.38	0.54
	DL target=2	FR	FR	0.31	0.53
	DL target=2	ES	ES	0.28	0.57
	DL target=2	FR	ES	0.23	0.60
	DL target=2	FR + 13% ES	ES	0.27	0.53
	DL target=2	FR + 1% ES	ES	0.28	0.56
NCE Loss	$\mathcal{L}^{\text{cross}}(T\|V)$	FR + TransAug	FR	0.40	0.49
	$\mathcal{L}^{\text{cross}}(T\|V)$	FR + TransAug	ES	0.32	0.75
	$\mathcal{L}^{\text{inter}}(T, V)$	FR + TransAug	FR	0.34	0.52
	$\mathcal{L}^{\text{inter}}(T, V)$	FR + TransAug	ES	0.30	0.84
	$\mathcal{L}^{\text{cross}}(T\|V), \mathcal{L}^{\text{cross}}(T\|A)$	FR + TransAug	FR	0.39	0.49
	$\mathcal{L}^{\text{cross}}(T\|V), \mathcal{L}^{\text{cross}}(T\|A)$	FR + TransAug	ES	0.33	0.77

Table 1. Sentiment analysis model performance comparison on CMU-MOSEAS dataset. **r** stands for the Pearson's correlation metric. The arrows associated with the metrics **r** and **MAE** indicate that lower MAE value or higher Pearson's correlation indicates stronger performances. **FR** denotes French training datasets, **ES** denotes Spanish datasets. We used a consistent training and evaluation splits across all experiments. **TransAug** denotes that we augmented the original French training datasets with translated Spanish text data. Note that the results obtained from Multilogue-net and NCE loss are trained with the original dataset without neutral sentiments removed.

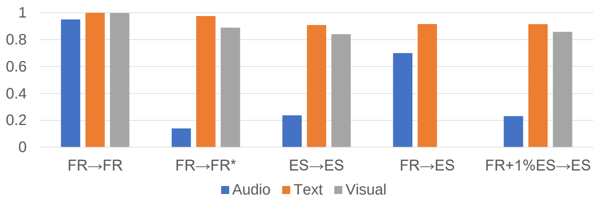

Figure 4. AdaMML Modality Utilization Rate. *: Slight model architecture alteration with Dropout placement

target decreased its performance. This is most likely due to the loss of information needed for the accurate prediction of sentiment when the model is performing comparably. The policy network selects all three modalities most of the time (First FR → FR histogram in Figure 4). This indicates that the policy network may be redundant in feature selection for sentiment prediction, even though our previous experiments with the baseline models showed that the Audio modality is the least useful among the three.

We also test a few architectural variations and test the placement of Dropout to help regularize the model due to the limited amount of training data available and prevent overfitting. We find that when Dropout is applied on both the inputs to policy network and fusion network, the policy network tends to select all three modalities most of the time (First FR → FR histogram in Figure 4). When Dropout is only applied to the fusion network, the policy network tends to select mostly Text and Visual modalities and ignore Audio (Second FR → FR histogram in Figure 4). Since Dropout only works at training time and not test time, this behavior is likely due to the slight architectural difference in training time.

7.3. Few-shot Transfer Learning

We also use the AdaMML + Contextual LSTM model to test out the idea of few-shot transfer learning.

Observing Table 1 and Figure 4, we can see that the model trained on the complete Spanish and French training dataset performs similarly and have similar modality utilization rate distribution from the policy network. However, the model trained only in French does not perform well when tested directly on Spanish. The policy network also outputs a different modality selection distribution as compared to that of the reference monolingual models.

Once we mix in a small subset of the Spanish training dataset, the model immediately performs comparably to the one trained on the full Spanish training set. For reference, French training data contains 1745 labeled segments after filtering, 13% Spanish training set contains 230 labeled segments, and 1% Spanish training set only contains

30 labeled segments.

The main contributing factor for this successful few-shot transfer learning still requires testing, as it might be better feature selection, or better fusion capability, or both. But given that a lot of the state-of-the-art models are attention-based, this result shows that this attention mechanism can be trained in a few-shot manner to generalize to a new language.

7.4. Acoustic-centric Emotion Recognition

CMU-MOSEAS provides a variety of emotion labels, such as anger, confidence, dominance, and happiness. However, most of the emotion labels are highly imbalanced, with positive labels accounting for less than 2% of the total samples. This is tabulated in Table 2. Therefore, we handpicked a few emotions with more positive samples for training, i.e., confidence, happiness, and relaxation. We use the MCTN architecture.

However, despite numerous models trained using various combinations of modalities, our models do not perform well, predicting all test labels as negative without weight adjustment to the positive samples or positive with weight adjustment to the positive samples. This indicates that weight adjustment to the binary cross entropy loss is simply a trade-off parameter between precision and recall. The F-1 score also does not get above 0.5 no matter which model variant that we tried.

Hence, we conclude that the state-of-the-art models currently are not able to predict the emotions of the speaker correctly. There are several possibilities for the failure that are worth further research:

1. The provided labels are not accurate. We observe that most of the videos are presentation-styled. Therefore most of the speakers may not have expressed emotions at all, and the labels can stem from human hallucination.
2. The features we obtained may not be suitable for emotion recognition, which is known as a more challenging task.

8. Conclusion and Future Directions

Main Contribution In this paper, we propose to extend current work on multimodal sentiment analysis to a multilingual setting. We adopt a transfer learning paradigm such that the model trained on one language is then evaluated on another language with zero-shot or few-shot fine-tuning. Towards this goal, we proposed four research ideas, namely NCE objective, Adaptive MML, few-shot learning with AdaMML, and acoustic-centric methods. Empirically, we show that each method makes improvement upon the baseline model, demonstrating success in zero-shot and few-shot transfer learning. The policy network we trained also makes the first attempt to quantitatively compare the contribution of each modality in video recognition across languages.

Emotion	Negative Samples	Positive Samples
anger	9899	201
confident	7799	2301
disgust	10059	41
dominant	9906	194
eloquent	9315	785
entertaining	10057	43
fear	10080	20
happiness	8370	1730
humorous	9993	107
narcissist	10076	24
nervous	9897	203
passionate	8642	1458
persuasive	9314	786
relaxed	4658	5442
reserved	9935	165
sadness	9957	143
sarcastic	10065	35
surprise	10018	82

Table 2. Emotion Label Counts

Future Directions One future direction is to incorporate robust language representation techniques into our work, such as using mBERT (Devlin et al., 2018) and training textual features with gradients from downstream objectives.

Our experiments are primarily based on French and Spanish, both of which have their roots in Europe. In order to better investigate the degree of similarity that sentiment cues share across languages, we would like to test our idea in a broader range of languages, especially those that are from language families other than Indo-European languages. To our knowledge, an extended version of CMU-MOSEAS will be released, which contains Asian languages and African languages. We believe this extended dataset will be a helpful resource for future cross-lingual comparison.

References

Abbaschian, B. J., Sierra-Sosa, D., and Elmaghraby, A. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4), 2021. ISSN 1424-8220. doi: 10.3390/s21041249. URL https://www.mdpi.com/1424-8220/21/4/1249.

Bagher Zadeh, A., Cao, Y., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1801–1812, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.141. URL https://aclanthology.org/2020.emnlp-main.141.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *ArXiv*, 1409, 09 2014.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 59–66. IEEE, 2018.

Bornea, M. A., Pan, L., Rosenthal, S., Florian, R., and Sil, A. Multilingual transfer learning for QA using translation as data augmentation. *CoRR*, abs/2012.05958, 2020. URL https://arxiv.org/abs/2012.05958.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. In Kathryn Huff and James Bergstra (eds.), *Proceedings of the 14th Python in Science Conference*, pp. 18 – 24, 2015. doi: 10.25080/Majora-7b98e3ed-003.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Ferreira, D. C., Martins, A. F. T., and Almeida, M. S. C. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2019–2028, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1190. URL https://aclanthology.org/P16-1190.

Heracleous, P. and Yoneyama, A. A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. 2019.

Huang, P., Patrick, M., Hu, J., Neubig, G., Metze, F., and Hauptmann, A. G. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. *CoRR*, abs/2103.08849, 2021. URL https://arxiv.org/abs/2103.08849.

Hussein, D. M. E.-D. M. A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4):330–338, 2018. ISSN 1018-3639. doi: https://doi.org/10.1016/j.jksues.2016.04.002. URL https://www.sciencedirect.com/science/article/pii/S1018363916300071.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Karpathy, A. The unreasonable effectiveness of recurrent neural networks, 2015. URL http://karpathy.github.io/2015/05/21/rnn-effectiveness/. Accessed on 2021-12-05.

Kemker, R., Abitino, A., McClure, M., and Kanan, C. Measuring catastrophic forgetting in neural networks. In *AAAI*, 2018.

Lausen, A. and Hammerschmidt, K. Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. 2020.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *INTERSPEECH*, 2017.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. Learned in translation: Contextualized word vectors. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6297–6308, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Medhat, W., Hassan, A., and Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014. ISSN 2090-4479. doi: https://doi.org/10.1016/j.asej.2014.04.011. URL https://www.sciencedirect.com/science/article/pii/S2090447914000550.

Morency, L.-P., Mihalcea, R., and Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pp. 169–176, 2011.

Neumann, M. and Vu, N. T. Cross-lingual and multilingual speech emotion recognition on english and french. *CoRR*, abs/1803.00357, 2018. URL http://arxiv.org/abs/1803.00357.

Panda, R., Chen, C.-F., Fan, Q., Sun, X., Saenko, K., Oliva, A., and Feris, R. Adamml: Adaptive multi-modal learning for efficient video recognition. *arXiv preprint arXiv:2105.05165*, 2021.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., and Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6892–6899, Jul. 2019. doi: 10.1609/aaai.v33i01.33016892. URL https://ojs.aaai.org/index.php/AAAI/article/view/4666.

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.

Rosas, V. P., Mihalcea, R., and Morency, L.-P. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45, 2013.

Shenoy, A. and Sardana, A. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pp. 19–28, Seattle, USA, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.challengehml-1.3. URL https://www.aclweb.org/anthology/2020.challengehml-1.3.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

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.

Turney, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 417–424, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073153. URL https://doi.org/10.3115/1073083.1073153.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.