Learning Robust Joint Representations for Multimodal Sentiment Analysis

Hai Pham*, Paul Pu Liang*, Thomas Manzini, Louis-Philippe Morency, Barnabás Póczos
School of Computer Science, Carnegie Mellon University
{htpham,pliang,tmanzini,morency,bapoczos}@cs.cmu.edu

Abstract

Multimodal sentiment analysis is a core research area that studies speaker sentiment expressed from the language, visual, and acoustic modalities. The central challenge in multimodal learning involves inferring joint representations that can process and relate information from these modalities. However, existing work learns joint representations using multiple modalities as input and may be sensitive to noisy or missing modalities at test time. With the recent success of sequence to sequence models in machine translation, there is an opportunity to explore new ways of learning joint representations that may not require all input modalities at test time. In this paper, we propose a method to learn robust joint representations by translating between modalities. Our method is based on the key insight that translation from a source to a target modality provides a method of learning joint representations using only the source modality as input. We augment modality translations with a cycle consistency loss to ensure that our joint representations retain maximal information from all modalities. Once our translation model is trained with paired multimodal data, we only need data from the source modality at test-time for prediction. This ensures that our model remains robust from perturbations or missing target modalities. We train our model with a coupled translation-prediction objective and it achieves new state-of-the-art results on multimodal sentiment analysis datasets: CMU-MOSI, ICT-MMMO, and YouTube. Additional experiments show that our model learns increasingly discriminative joint representations with more input modalities while maintaining robustness to perturbations of all other modalities.

1 Introduction

Sentiment analysis, which involves identifying a speaker's opinion, is a core research problem in machine learning and natural language processing. However, language-based sentiment analysis through words, phrases, and their compositionality was found to be insufficient for inferring affective content from spoken opinions [34], which contain rich nonverbal behaviors in addition to verbal text. As a result, there has been a recent push towards using machine learning methods to learn joint representations from additional behavioral cues present in the visual and acoustic modalities. This research field has become known as multimodal sentiment analysis and extends the conventional text-based definition of sentiment analysis to a multimodal setup. For example, [22] explore the additional acoustic modality while [62] use the language, visual, and acoustic modalities present in monologue videos to predict sentiment. The abundance of multimodal data has led to the creation of multimodal datasets, such as CMU-MOSI [67] and ICT-MMMO [62], as well as deep multimodal models that are highly effective at learning discriminative joint multimodal representations [30, 58, 8]. Existing work learns joint representations using multiple modalities as input with neural networks [29], graphical models [34] or geometric classifiers [67]. However, this results in joint representations that are sensitive to noisy or missing modalities at test time.

* Equal contributions

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

To address this problem, we draw inspirations from the recent success of sequence to sequence models for unsupervised representation learning [56]. We propose the Multimodal Cyclic Translation Network model (MCTN) to learn robust joint multimodal representations by translating between modalities. Figure 1 illustrates these translations between the language, visual and acoustic modalities. Our method is based on the key insight that translation from a source modality S to a target modality T results in an intermediate representation that captures joint information between modalities S and T. MCTN extends this insight using a cyclic translation loss involving both *forward translations* from source to target, and *backward translations* from the predicted target back to the source modality. Together, we call these *multimodal cyclic translations* to ensure that the learned joint representations capture maximal information from both modalities. We also propose a hierarchical MCTN to learn joint representations between a source modality and multiple target modalities. MCTN is trainable end-to-end with a coupled translation-prediction loss which consists of (1) the cyclic translation loss, and (2) a prediction loss to ensure that the learned joint representations are task-specific. Another advantage of MCTN is that once trained with paired multimodal data (S, T), we *only* need data from the source modality S at test time to infer the joint representation and sentiment prediction. As a result, MCTN is completely robust to test-time perturbations on target modality T and missing modalities.

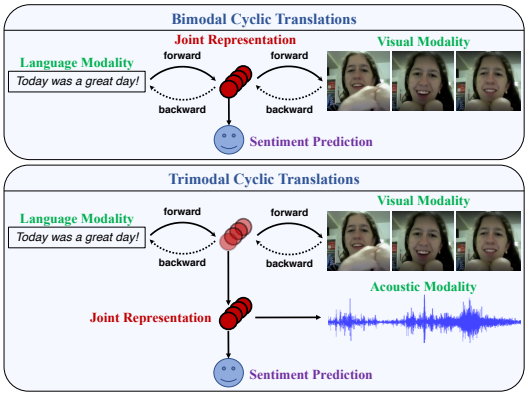


Figure 1: Learning robust joint representations via multimodal cyclic translations. Top: cyclic translations from a source modality (language) to a target modality (visual). Bottom: the representation learned between language and vision are further translated into the acoustic modality, forming the final joint representation. The joint representation is then used for multimodal prediction.

Even though translation and generation of videos, audios, and text are difficult [28], our experiments show that the learned joint representations can help for discriminative tasks: MCTN achieves new state-of-the-art results on multimodal sentiment analysis using the CMU-MOSI, ICT-MMMO, and YouTube public datasets. Additional experiments show that MCTN learns increasingly discriminative joint representations with more input modalities while maintaining robustness to all target modalities.

2 Related Work

Early work on sentiment analysis focused primarily on written text [40, 51]. Recently, multimodal sentiment analysis has gained more research interest [5] since learning joint representation of multiple modalities is a challenging task. Earlier work simply concatenated the input features [37, 26]. Recently, several neural models have also been proposed to learn joint representations [8, 65, 9]. For example, [29] presented a multistage approach to learn hierarchical multimodal representations. The Tensor Fusion Network [64] and the Low-rank Multimodal Fusion model [31] presented methods based on Cartesian-products to model unimodal, bimodal and trimodal interactions.

In addition to purely supervised approaches, generative methods based on Generative Adversarial Networks (GANs) [17] have been used to learn joint distributions between two or more modalities [14, 27]. Another method involves using conditional generative models [33, 23, 39] to translate one modality to another. Generative-discriminative objectives have been used to learn either joint [44, 24] or factorized [58] representations. Our work takes into account the sequential dependency of modality translations and also explores the effect of a cyclic translation loss on modality translations.

Finally, there has been some progress on accounting for noisy or missing modalities at test time. [55] proposed using Deep Boltzmann Machines to model the joint distribution over multimodal data. Sampling from the conditional distributions allow for inference of missing modalities. [52] trained Restricted Boltzmann Machines to minimize the variation of information between modality-specific latent variables. Models based on autoencoders [57], adversarial learning [7], or multiple kernel learning [32] have also been proposed for these tasks. It was also found that training with missing or noisy modalities can improve the robustness of joint representations [37]. These methods approximately infer the missing modalities before prediction, leading to possible error compounding. On the other hand, MCTN remains fully robust to other modalities during testing.

3 Proposed Approach

In this section, we describe our approach for learning joint multimodal representations through modality translations.

Notation: A multimodal dataset consists of data $\mathbf{X} = (\mathbf{X}^l, \mathbf{X}^v, \mathbf{X}^a)$ from the language, visual, and acoustic modalities respectively. It is indexed by n segments $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)$ where $\mathbf{X}_i = (\mathbf{X}_i^l, \mathbf{X}_i^v, \mathbf{X}_i^a)$, $1 \leq i \leq n$. The labels for these n segments are denoted as $\mathbf{y} = (y_1, y_2, ..., y_n)$, $y_i \in \mathbb{R}$. Many datasets are easily synchronized by aligning the input based on the boundaries of each word and zero-padding each segment to obtain time-series data of the same length [29]. The ith segment is given by $\mathbf{X}_i^l = (\mathbf{w}_i{}^{(1)}, \mathbf{w}_i{}^{(2)}, ..., \mathbf{w}_i{}^{(L)})$ where $\mathbf{w}_i{}^{(\ell)}$ stands for the ℓth word and L is the length of each segment. To accompany the language features, we also have a sequence of visual features $\mathbf{X}_i^v = (\mathbf{v}_i{}^{(1)}, \mathbf{v}_i{}^{(2)}, ..., \mathbf{v}_i{}^{(L)})$ and acoustic features $\mathbf{X}_i^a = (\mathbf{a}_i{}^{(1)}, \mathbf{a}_i{}^{(2)}, ..., \mathbf{a}_i{}^{(L)})$.

3.1 Learning Joint Representations

We define learning a joint representation between two modalities $\mathbf{X}^S$ and $\mathbf{X}^T$ as learning a parametrized function f_θ that returns an embedding $\mathcal{E}_{ST} = f_\theta(\mathbf{X}^S, \mathbf{X}^T)$. From there, another function g_w is learned that predicts the label given this joint representation: $\hat{\mathbf{y}} = g_w(\mathcal{E}_{ST})$.

Most work follows this framework during both training and testing [29, 31, 58, 65]. During training, the parameters θ and w are learned by empirical risk minimization over paired multimodal data and labels in the training set $(\mathbf{X}_{tr}^S, \mathbf{X}_{tr}^T, \mathbf{y}_{tr})$:

$$\mathcal{E}_{ST} = f_\theta(\mathbf{X}_{tr}^S, \mathbf{X}_{tr}^T), \ \hat{\mathbf{y}}_{tr} = g_w(\mathcal{E}_{ST}), \tag{1}$$

$$\theta^*, w^* = \arg\min_{\theta, w} \mathbb{E}\left[\ell_{\mathbf{y}}(\hat{\mathbf{y}}_{tr}, \mathbf{y}_{tr})\right]. \tag{2}$$

for a suitable choice of loss function $\ell_{\mathbf{y}}$ over the labels (tr denotes training set). During testing, paired multimodal data in the test set $(\mathbf{X}_{te}^S, \mathbf{X}_{te}^T)$ are used to infer the label (te denotes test set):

$$\mathcal{E}_{ST} = f_{\theta^*}(\mathbf{X}_{te}^S, \mathbf{X}_{te}^T), \ \hat{\mathbf{y}}_{te} = g_{w^*}(\mathcal{E}_{ST}). \tag{3}$$

3.2 Multimodal Cyclic Translation Network

Multimodal Cyclic Translation Network (MCTN) is a neural model that learns robust joint representations by modality translations. Figure 2 shows a detailed description of MCTN for two modalities. Our method is based on the key insight that translation from a source modality $\mathbf{X}^S$ to a target modality $\mathbf{X}^T$ results in a representation that captures joint information between modalities $\mathbf{X}^S$ and $\mathbf{X}^T$, but using only the source modality $\mathbf{X}^S$ as input.

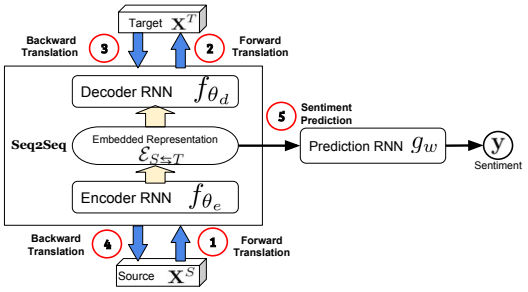


Figure 2: MCTN architecture for two modalities: the source modality $\mathbf{X}^S$ and the target modality $\mathbf{X}^T$. The joint representation $\mathcal{E}_{S \leftrightarrows T}$ is obtained via a cyclic translation between $\mathbf{X}^S$ and $\mathbf{X}^T$. Next, the joint representation $\mathcal{E}_{S \leftrightarrows T}$ is used for sentiment prediction. The model is trained end-to-end with a coupled translation-prediction objective. At test time, only the source modality $\mathbf{X}^S$ is required.

To ensure that our model learns joint representations that retain maximal information from all modalities, we use a cycle consistency loss [68] during modality translation. This method can also be seen as a variant of back-translation which has been recently applied to style transfer [47, 68] and unsupervised machine translation [25]. We use back-translation in a multimodal setup where we encourage our translation model to learn informative joint representations but with only the source modality as input. The cycle consistency loss for modality translation starts by decomposing function f_θ into two parts: an encoder f_{θ_e} and a decoder f_{θ_d}. The encoder takes in $\mathbf{X}^S$ as input and returns a joint embedding $\mathcal{E}_{S \to T}$:

$$\mathcal{E}_{S \to T} = f_{\theta_e}(\mathbf{X}^S), \tag{4}$$

which the decoder then transforms into target modality $\mathbf{X}^T$:

$$\mathbf{X}^T = f_{\theta_d}(\mathcal{E}_{S \to T}), \tag{5}$$

following which the decoded modality T is translated back into modality S:

$$\mathcal{E}_{T\to S} = f_{\theta_e}(\hat{\mathbf{X}}^T),\ \hat{\mathbf{X}}^S = f_{\theta_d}(\mathcal{E}_{T\to S}). \tag{6}$$

The joint representation is learned by using a Sequence to Sequence (Seq2Seq) model with attention [4] that translates source modality $\mathbf{X}^S$ to a target modality $\mathbf{X}^T$. While Seq2Seq models have been predominantly used for machine translation, we extend its usage to the realm of multimodal machine learning. The Seq2Seq model consists of an encoder network and a decoder network, each parametrized as Recurrent Neural Networks (RNNs). The encoder maps the source modality $\mathbf{X}^S$ into an embedded representation $\mathcal{E}_{S\to T}$. Using a recurrent network, the hidden state output of each time step is based on the previous hidden state along with the input sequence

$$\mathbf{h}_\ell = \texttt{RNN}(\mathbf{h}_{\ell-1}, \mathbf{X}^S_\ell) \quad \forall \ell \in [1, L]. \tag{7}$$

The encoder's output is the concatenation of all hidden states of the encoding RNN, $\mathcal{E}_{S\to T} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_L]$, where L is the length of the source modality $\mathbf{X}^S$.

The decoder maps the representation $\mathcal{E}_{S\to T}$ into the target modality $\mathbf{X}^T$. This is performed by decoding each token $\mathbf{X}^T_t$ at a time based on $\mathcal{E}_{S\to T}$ and all previous decoded tokens, which is formulated as

$$p(\mathbf{X}^T) = \prod_{\ell=1}^{L} p(\mathbf{X}^T_\ell | \mathcal{E}_{S\to T}, \mathbf{X}^T_1, ..., \mathbf{X}^T_{\ell-1}). \tag{8}$$

MCTN accepts variable-length inputs of $\mathbf{X}^S$ and $\mathbf{X}^T$, and is trained to maximize the translational condition probability $p(\mathbf{X}^T|\mathbf{X}^S)$. The best translation sequence is then given by

$$\hat{\mathbf{X}^T} = \arg\max_{\mathbf{X}^T} p(\mathbf{X}^T|\mathbf{X}^S). \tag{9}$$

While there are other search algorithms such as random sampling and greedy search that can be used for decoding each token [36], we use the traditional beam search approach [56].

To obtain the joint representation for prediction, we found that simply using one of the translated representations was sufficient for good performance ($\leftrightarrows$ denotes multimodal cyclic translations): $\mathcal{E}_{S\leftrightarrows T} = \mathcal{E}_{S\to T}$. $\mathcal{E}_{S\leftrightarrows T}$ is used for prediction via a recurrent neural network, $\hat{\mathbf{y}} = g_w(\mathcal{E}_{S\leftrightarrows T})$.

3.3 Coupled Translation-Prediction Objective

Training is performed with paired multimodal data and labels in the training set $(\mathbf{X}^S_{tr}, \mathbf{X}^T_{tr}, \mathbf{y}_{tr})$. We evaluate the forward translation loss

$$\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)], \tag{10}$$

and the cycle consistency loss

$$\mathcal{L}_c = \mathbb{E}[\ell_{\mathbf{X}^S}(\hat{\mathbf{X}}^S, \mathbf{X}^S)] \tag{11}$$

for suitable choices of loss functions $\ell_{\mathbf{X}^T}$ and $\ell_{\mathbf{X}^S}$. We use the Mean Squared Error (MSE) between the ground-truth and translated modalities. Finally, the prediction loss $\mathcal{L}_p$ is

$$\mathcal{L}_p = \mathbb{E}[\ell_{\mathbf{y}}(\hat{\mathbf{y}}, \mathbf{y})]. \tag{12}$$

for a loss function $\ell_{\mathbf{y}}$ over the labels.

Equations (10), (11), and (12) are evaluated using the training set and MCTN can be trained end-to-end with a coupled translation-prediction objective function $\mathcal{L} = \lambda_t\mathcal{L}_t + \lambda_c\mathcal{L}_c + \mathcal{L}_p$, where $\mathcal{L}_p$ is the prediction loss, $\mathcal{L}_c$ is the cyclic translation loss, and λ_t, λ_t are weighting hyperparameters. MCTN parameters are learned by minimizing this objective function

$$\theta^*_e, \theta^*_d, w^* = \arg\min_{\theta_e, \theta_d, w} [\lambda_t\mathcal{L}_t + \lambda_c\mathcal{L}_c + \mathcal{L}_p]. \tag{13}$$

Parallel multimodal data is not required at test time. Inference is performed using only the source modality $\mathbf{X}^S_{te}$:

$$\mathcal{E}_{S\leftrightarrows T} = f_{\theta^*_e}(\mathbf{X}^S_{te}),\ \hat{\mathbf{y}}_{te} = g_{w^*}(\mathcal{E}_{S\leftrightarrows T}). \tag{14}$$

This is possible because the encoder $f_{\theta^*_e}$ has been trained to translate the source modality $\mathbf{X}^S$ into a joint representation $\mathcal{E}_{S\leftrightarrows T}$ that captures information from both source and target modalities. Intuitively, the translation model learns to predict target modalities through an informative joint representation.

3.4 Hierarchical MCTN for Three Modalities

We extend the MCTN hierarchically to learn joint representations from more than two modalities. Figure 3 shows the case for three modalities. The hierarchical MCTN starts with a source modality $\mathbf{X}^S$ and two target modalities $\mathbf{X}^{T_1}$ and $\mathbf{X}^{T_2}$. To learn joint representations, two levels of translations are performed. The first level learns a joint representation from $\mathbf{X}^S$ and $\mathbf{X}^{T_1}$ using multimodal cyclic translations as defined previously. At the second level, a joint representation is learned hierarchically by translating the first representation $\mathcal{E}_{S\to T_1}$ into $\mathbf{X}^{T_2}$. For more than three modalities, the modality translation process can be repeated hierarchically.

Two Seq2Seq models are used in the hierarchical MCTN for three modalities. Denote these as encoder-decoder pairs $(f^1_{\theta_e}, f^1_{\theta_d})$ and $(f^2_{\theta_e}, f^2_{\theta_d})$. A multimodal cyclic translation is first performed between source modality $\mathbf{X}^S$ and the first target modality $\mathbf{X}^{T_1}$. The consists of the forward translation:

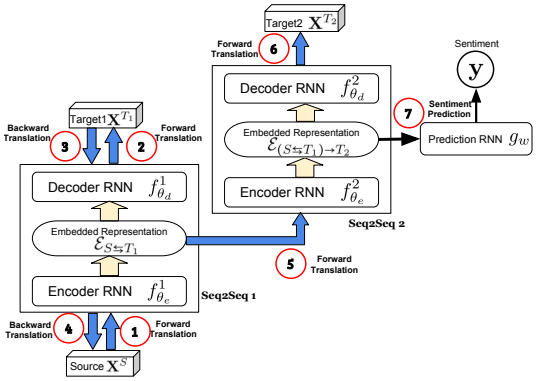


Figure 3: Hierarchical MCTN for three modalities: the source modality $\mathbf{X}^S$ and the target modalities $\mathbf{X}^{T_1}$ and $\mathbf{X}^{T_2}$. The joint representation $\mathcal{E}_{S\leftrightarrows T_1}$ is obtained via a cyclic translation between $\mathbf{X}^S$ and $\mathbf{X}^{T_1}$, then further translated into $\mathbf{X}^{T_2}$. Next, the joint representation of all three modalities, $\mathcal{E}_{(S\leftrightarrows T_1)\to T_2}$, is used for sentiment prediction. The model is trained end-to-end with a coupled translation-prediction objective. At test time, only the source modality $\mathbf{X}^S$ is required.

$$\mathcal{E}_{S\to T_1} = f^1_{\theta_e}(\mathbf{X}^S_{tr}),\ \hat{\mathbf{X}}^{T_1}_{tr} = f^1_{\theta_d}(\mathcal{E}_{S\to T_1}), \tag{15}$$

following which the decoded modality $\mathbf{X}^{T_1}$ is translated back into modality $\mathbf{X}^S$:

$$\mathcal{E}_{T_1\to S} = f^1_{\theta_e}(\hat{\mathbf{X}}^{T_1}_{tr}),\ \hat{\mathbf{X}}^{S}_{tr} = f^1_{\theta_d}(\mathcal{E}_{T_1\to S}). \tag{16}$$

A second hierarchical Seq2Seq model is applied on the time-distributed outputs of the encoder $f^1_{\theta_e}$:

$$\mathcal{E}_{S\leftrightarrows T_1} = \mathcal{E}_{S\to T_1}, \tag{17}$$

$$\mathcal{E}_{(S\leftrightarrows T_1)\to T_2} = f^2_{\theta_e}(\mathcal{E}_{S\leftrightarrows T_1}),\ \hat{\mathbf{X}}^{T_2}_{tr} = f^2_{\theta_d}(\mathcal{E}_{(S\leftrightarrows T_1)\to T_2}). \tag{18}$$

The joint representation between modalities $\mathbf{X}^S$, $\mathbf{X}^{T_1}$ and $\mathbf{X}^{T_2}$ is now $\mathcal{E}_{(S\leftrightarrows T_1)\to T_2}$ and is used for sentiment prediction via a recurrent neural network.

Training the hierarchical MCTN involves computing a cycle consistent loss for modality T_1, given by $\mathcal{L}_{t_1}$ and $\mathcal{L}_{c_1}$. We do not use a cyclic translation loss when translating from $\mathcal{E}_{S\leftrightarrows T_1}$ to $\mathbf{X}^{T_2}$ since the ground truth $\mathcal{E}_{S\leftrightarrows T_1}$ is unknown, and so only the translation loss $\mathcal{L}_{t_2}$ is computed. The final objective for hierarchical MCTN is given by $\mathcal{L} = \lambda_{t_1}\mathcal{L}_{t_1} + \lambda_{c_1}\mathcal{L}_{c_1} + \lambda_{t_2}\mathcal{L}_{t_2} + \mathcal{L}_p$. We emphasize that for MCTN with three modalities, *only* a single source modality $\mathbf{X}^S$ is required at test time. Therefore, MCTN has a significant advantage over existing models since it is robust to noise or missing target modalities.

4 Experimental Setup

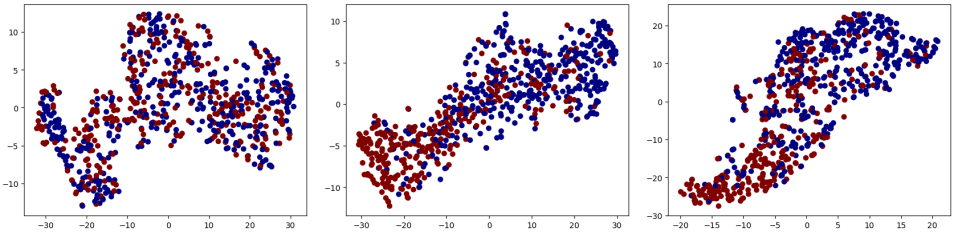
In this section, we describe our experimental methodology to evaluate the joint representations learned by MCTN. [1]

Datasets: We use the CMU-MOSI dataset which contains 2199 video segments each with a sentiment label in the range from −3 to +3. −3 indicates strongly negative sentiment, +3 indicates strongly positive sentiment, and 0 indicates neutral sentiment. CMU-MOSI is subject to much research [58, 65, 8] and the current state of the art is achieved by [29] with a binary accuracy of 78.4%. We additionally perform experiments on the ICT-MMMO [62] and YouTube [34] datasets. These datasets consist of online review videos annotated for sentiment.

Features: Following previous work [29], GloVe word embeddings [42], Facet [20] and COVAREP [12] features are extracted for the language, visual and acoustic modalities respectively. [2] Forced alignment is performed using P2FA [63] to obtain word utterance times and we align the visual and acoustic features by computing their average over each word utterance interval.

[1] Our source code for replicating these experiments will be released at `<anonymous>`.

[2] Details on feature extraction are in supplementary.



(a) Bimodal MCTN *without* cyclic translations (b) Bimodal MCTN *with* cyclic translations (c) Hierarchical MCTN *with* cyclic translations

Figure 4: t-SNE visualization of the joint representations learned by MCTN. Red: videos with negative sentiment, blue: videos with positive sentiment. Adding modalities and using cyclic translations leads to increasingly separable representations and improves discriminative performance.

Metrics: For parameter optimization on the CMU-MOSI dataset, we set the choice of prediction loss function as the Mean Absolute Error (MAE): $\ell_p(\hat{\mathbf{y}}_{train}, \mathbf{y}_{train}) = |\hat{\mathbf{y}}_{train} - \mathbf{y}_{train}|$. We report MAE and Pearson's correlation (Corr). In addition, we also perform sentiment classification on the CMU-MOSI dataset and report binary accuracy (Acc) and F1 score (F1). On the ICT-MMMO and YouTube datasets, we set the choice of prediction loss function as categorical cross-entropy and report classification accuracy (Acc) and F1 score. For all metrics, higher values indicate stronger performance, except MAE where lower values indicate stronger performance.

Baselines: We compare to the following multimodal models: *RMFN* [29] uses a multistage approach to learn hierarchical representations. It is the current state-of-the-art on CMU-MOSI. *LMF* [31] approximates the expensive multimodal tensor products in *TFN* [64] with efficient low-rank factors. *MFN* [65] synchronizes sequences using a multimodal gated memory. *MARN* [66] uses multiple attention coefficients and hybrid LSTM memory components. *GME-LSTM(A)* [8] learns binary gating mechanisms to remove noisy modalities that are contradictory or redundant. *EF-LSTM* concatenates multimodal inputs and uses a single LSTM [19]. For details on all baselines, please refer to the supplementary.

Dataset		CMU-MOSI			
Model	Test Inputs	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
RF	$\{\ell, v, a\}$	56.4	56.3	-	-
EF-LSTM	$\{\ell, v, a\}$	74.3	74.3	1.023	0.622
MV-LSTM	$\{\ell, v, a\}$	73.9	74.0	1.019	0.601
BC-LSTM	$\{\ell, v, a\}$	75.2	75.3	1.079	0.614
TFN	$\{\ell, v, a\}$	74.6	74.5	1.040	0.587
MARN	$\{\ell, v, a\}$	77.1	77.0	0.968	0.625
MFN	$\{\ell, v, a\}$	77.4	77.3	0.965	0.632
LMF	$\{\ell, v, a\}$	76.4	75.7	0.912	0.668
RMFN	$\{\ell, v, a\}$	78.4	78.0	0.922	**0.681**
MCTN	$\{\ell\}$	**79.3**	**79.1**	**0.909**	0.676

Dataset		ICT-MMMO		YouTube	
Model	Test Inputs	Acc(↑)	F1(↑)	Acc(↑)	F1(↑)
RF	$\{\ell, v, a\}$	70.0	69.8	33.3	32.3
EF-LSTM	$\{\ell, v, a\}$	72.5	70.9	44.1	43.6
MV-LSTM	$\{\ell, v, a\}$	72.5	72.3	45.8	43.3
BC-LSTM	$\{\ell, v, a\}$	70.0	70.1	45.0	45.1
TFN	$\{\ell, v, a\}$	72.5	72.6	45.0	41.0
MARN	$\{\ell, v, a\}$	71.3	70.2	48.3	44.9
MFN	$\{\ell, v, a\}$	73.8	73.1	**51.7**	51.6
MCTN	$\{\ell\}$	**81.3**	**80.8**	**51.7**	**52.4**

Table 1: Sentiment prediction results on CMU-MOSI (top), ICT-MMMO and YouTube (bottom). Best results are highlighted in bold. MCTN outperforms the current state-of-the-art across most metrics and uses only language during testing.

5 Results and Discussion

This section discusses several research questions and presents our experimental results.

Comparison with Existing Work: *Q1: How does MCTN compare with existing state-of-the-art approaching for multimodal sentiment analysis?* We compare MCTN with the existing state-of-the-art models [3]. From Table 4, MCTN achieves new start-of-the-art results on binary classification accuracy, F1 score, and MAE on CMU-MOSI. State-of-the-art results are also achieved on the ICT-MMMO and YouTube datasets. These results are even

Dataset		CMU-MOSI			
Model	Translation	Acc	F1	MAE	Corr
MCTN Bi (Fig. 5a)	$V \leftrightarrows A$	53.1	53.2	1.420	0.034
	$T \leftrightarrows A$	76.4	76.4	0.977	0.636
	$T \leftrightarrows V$	76.8	76.8	1.034	0.592
MCTN Tri (Fig. 5e)	$(V \leftrightarrows A) \rightarrow T$	56.4	56.3	1.455	0.151
	$(T \leftrightarrows A) \rightarrow V$	78.7	78.8	0.960	0.650
	$(T \leftrightarrows V) \rightarrow A$	**79.3**	**79.1**	**0.909**	**0.676**

Table 2: MCTN performance improves as more modalities are introduced for cyclic translations during training.

[3] For full results please refer to the supplementary material.

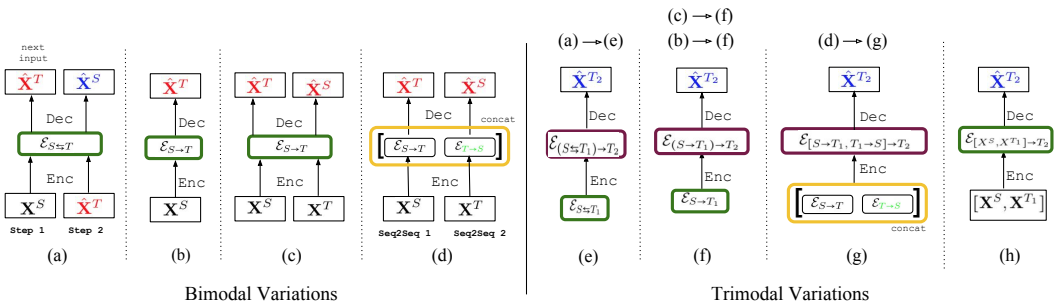


Figure 5: Variations of our models: (a) Bimodal MCTN with cyclic translation, (b) simple bimodal MCTN without cyclic translation, (c) MCTN with different inputs of the same modality pair, and without cyclic translation, (d) two MCTNs for two modalities without cyclic translation, with two different inputs (of the same pair), (e) Hierarchical MCTN with input from (a), (f) Hierarchical MCTN for three modalities, with input as a joint representation taken from previous MCTN for two modalities from (b) or (c), (g) Hierarchical MCTN with input from (d), (h) Concatenation MCTN which is similar to (b) but with input as the concatenation of 2 modalities. *Legend*: black modality is ground truth, red ("hat") modality represents translated output, blue ("hat") modality is target output from previous translation outputs, yellow box denotes concatenation.

more impressive considering that MCTN only uses the language modality during testing, while other baseline models use all three modalities.

Adding More Modalities: *Q2: What is the impact of increasing the number of modalities during training for MCTN with cyclic translations?* We run experiments with MCTN using combinations of two or three modalities with cyclic translations. From Table 2, we observe that adding more modalities improves performance, indicating that the joint representations learned are leveraging the information from more input modalities. This also implies that cyclic translations are a viable method to learn joint representations from multiple modalities since little information is lost from adding more modality translations. Another observation is that using language as the source modality always leads to the best performance, which is intuitive since the language modality contains the most information towards sentiment [64].

In addition, we visually inspect the joint representations learnt from MCTN as we add more modalities during training. The joint representations for each video segment in CMU-MOSI are extracted from the best performing model for each number of modalities and then projected into two dimensions via the t-SNE algorithm [59]. Each point is colored red or blue depending on whether the video segment is annotated for positive or negative sentiment. From Figure 4, we observe that the joint representations become increasingly separable as the more modalities are added when the MCTN is trained. This is consistent with increasing discriminative performance as seen in Table 2.

Ablation Studies: We devise the following ablation models to test each design decision in MCTN: the use of cyclic translations, shared Seq2Seq models, modality ordering, and hierarchical structure.

For bimodal MCTN, we design the following ablation models shown in the left half of Figure 5: (a) is our proposed MCTN between $\mathbf{X}^S$ and $\mathbf{X}^T$, (b) is the MCTN based on translation from $\mathbf{X}^S$ to $\mathbf{X}^T$ without cyclic translations, (c) does not use cyclic translations but rather performs two independent translations between $\mathbf{X}^S$ and $\mathbf{X}^T$, (d) is the pair of MCTN models with different inputs (of the same modality pair) and then using the concatenation of the joint representations $\mathcal{E}_{S\to T}$ and $\mathcal{E}_{T\to S}$ as the final embeddings. For trimodal MCTN, we design the following ablation models shown in the right half of Figure 5: (e) is the proposed hierarchical MCTN between $\mathbf{X}^S$, $\mathbf{X}^{T_1}$ and $\mathbf{X}^{T_2}$, (f) is the MCTN based on translation from $\mathbf{X}^S$ to $\mathbf{X}^{T_1}$ without cyclic translations, (g) is extended from (d) which does not use cyclic translations but rather performs two independent translations between $\mathbf{X}^S$ and $\mathbf{X}^{T_1}$, and finally, (h) does not perform a first level of cyclic translation but directly translates the concatenated modality pair $[\mathbf{X}^S, \mathbf{X}^{T_1}]$ into $\mathbf{X}^{T_2}$.

Q3: What is the impact of cyclic translations in MCTN? The bimodal and trimodal results are shown in Table 3. Only the model in Figure 5(a) employs cyclic translations and they outperform the other baselines. We make a similar observation for hierarchical MCTN: Figure 5(e) with cyclic translations outperforms the trimodal baselines (f), (g) and (h). The gap for the trimodal case is especially large. This implies that using cyclic translations is crucial in learning joint representations. Our intuition

is that using cyclic translations: (1) encourages the model to enforce symmetry between the joint representations from source and target modalities, and (2) ensures that the joint representation retains maximal information from all modalities.

Q4: What is the effect of using two Seq2Seq models instead of one shared Seq2Seq model for cyclic translations? We compare Figure 5(c), which uses one Seq2Seq model for translations with Figure 5(d), which uses two separate Seq2Seq models: one for forward and one for backward translation. We observe from Table 3 that (c) > (d), so using one model with shared parameters is better. This is also true for hierarchical MCTN: (f) > (g). We hypothesize that this is because training two Seq2Seq models requires more data and is prone to overfitting.

Q5: What is the impact when varying source and target modalities for cyclic translations? As shown in Tables 2 and 3, we observe that language contributes most towards the joint representations. For bimodal cases, combining language with visual is generally better than combining language with audio. For hierarchical MCTN, presenting language as the source modality leads to the best performance, and a first level of cyclic translations between language and visual is better than between language and acoustic. On the other hand, only translating between visual and acoustic modalities dramatically decreases performance. Further adding language as a target modality for hierarchical MCTN will not help much as well. Overall, language is still the most important modality for multimodal sentiment analysis and must be used as the source modality during translations.

Dataset	CMU-MOSI				
Model	Translation	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
MCTN Bi (Fig. 5a)	$V \leftrightarrows A$	53.1	53.2	1.420	0.034
	$T \leftrightarrows A$	76.4	76.4	0.977	**0.636**
	$T \leftrightarrows V$	**76.8**	**76.8**	1.034	0.592
Simple Bi (Fig. 5b)	$V \rightarrow A$	55.4	55.5	1.422	0.119
	$T \rightarrow A$	74.2	74.2	0.988	0.616
	$T \rightarrow V$	75.7	75.6	1.002	0.617
No cycle Bi (Fig. 5c)	$V \rightarrow A,\ A \rightarrow V$	55.4	55.5	1.422	0.119
	$T \rightarrow A,\ A \rightarrow T$	75.5	75.6	**0.971**	0.629
	$T \rightarrow V,\ V \rightarrow T$	75.2	75.3	0.972	0.627
Double Bi (Fig. 5d)	$[V \rightarrow A, A \rightarrow V]$	57.0	57.1	1.502	0.168
	$[T \rightarrow A, A \rightarrow T]$	72.3	72.3	1.035	0.578
	$[T \rightarrow V, V \rightarrow T]$	73.3	73.4	1.020	0.570
MCTN Tri (Fig. 5e)	$(V \leftrightarrows A) \rightarrow T$	56.4	56.3	1.455	0.151
	$(T \leftrightarrows A) \rightarrow V$	78.7	78.8	0.960	0.650
	$(T \leftrightarrows V) \rightarrow A$	**79.3**	**79.1**	**0.909**	**0.676**
Simple Tri (Fig. 5f)	$(V \rightarrow T) \rightarrow A$	54.1	52.9	1.408	0.040
	$(V \rightarrow A) \rightarrow T$	52.0	51.9	1.439	0.015
	$(A \rightarrow V) \rightarrow T$	56.6	56.7	1.593	0.067
	$(A \rightarrow T) \rightarrow V$	54.1	54.2	1.577	0.028
	$(T \rightarrow A) \rightarrow V$	74.3	74.4	1.001	0.609
	$(T \rightarrow V) \rightarrow A$	74.3	74.4	0.997	0.596
Double Tri (Fig. 5g)	$[T \rightarrow V,\ V \rightarrow T] \rightarrow A$	73.3	73.1	1.058	0.578
Concat Tri (Fig. 5h)	$[V, A] \rightarrow T$	55.0	54.6	1.535	0.176
	$[A, T] \rightarrow V$	73.3	73.4	1.060	0.561
	$[T, V] \rightarrow A$	72.3	72.3	1.068	0.576
	$A \rightarrow [T, V]$	55.5	55.6	1.617	0.056
	$T \rightarrow [A, V]$	75.7	75.7	0.958	0.634
	$[T, A] \rightarrow [T, V]$	73.2	73.2	1.008	0.591
	$[T, V] \rightarrow [T, A]$	74.1	74.1	0.999	0.607

Table 3: Bimodal (Bi) variations and Trimodal (Tri) ablation results on CMU-MOSI. MCTN and Hierarchical MCTN with cyclic translations performs best.

Q6: What is the impact of using two levels of hierarchical translations instead of one level for three modalities? Our hierarchical MCTN is shown in Figure 5(e). In Figure 5(h), we concatenate two modalities as input and use only one phase of translation. From Table 3, we observe that (e) > (h): both levels of modality translations are important in the hierarchical MCTN. We believe that representation learning is easier when the task is broken down recursively: using two translations each between a single pair of modalities, rather than a single translation between all modalities.

6 Conclusion

To conclude, this paper investigated learning joint representations via cyclic modality translations from source to target modalities. During testing, we only need the source modality for prediction which ensures that our model remains robust from noisy or missing target modalities. We demonstrate that cyclic translations and seq2seq models are especially useful for learning joint multimodal representations. In addition to achieving state-of-the-art results on three datasets, our model learns increasingly discriminative representations with more input modalities while maintaining robustness to all target modalities. Our approach presents several exciting areas for future work, such as: 1) combining our approach with the transformer architecture [60] for modality translations, 2) exploring pretrained deep language models [13, 43] for translations, as well as 3) extending our translation model to work other multimodal tasks involving language and raw speech signals (prosody), videos with multiple speakers (diarization), and combinations of static and temporal data (i.e. image captioning).

References

[1] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118, 1992.

[2] Paavo Alku, Tom Bäckström, and Erkki Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[3] Paavo Alku, Helmer Strik, and Erkki Vilkman. Parabolic spectral parameter - a new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406, 2017.

[6] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[7] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *KDD '18*, pages 1158–1166, 2018.

[8] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. *ICMI*, 2017.

[9] Yong Cheng, Fei Huang, Lian Zhou, Cheng Jin, Yuejie Zhang, and Tao Zhang. A hierarchical multimodal attention-based neural network for image captioning. In *SIGIR '17*, 2017.

[10] Donald G Childers and CK Lee. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[12] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep - a collaborative voice analysis repository for speech technologies. In *ICASSP*. IEEE, 2014.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[14] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[15] Thomas Drugman and Abeer Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, pages 1973–1976, 2011.

[16] Thomas Drugman, Mark Thomas, Jon Gudnason, Patrick Naylor, and Thierry Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006, 2012.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[18] A. Graves, A. r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, May 2013.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[20] iMotions. Facial expression analysis, 2017.

[21] John Kane and Christer Gobl. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179, 2013.

[22] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. Sentiment extraction from natural audio streams. In *ICASSP*. IEEE, 2013.

[23] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014.

[24] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[25] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *CoRR*, abs/1804.07755, 2018.

[26] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

[27] Chongxuan Li, Kun Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. *arXiv preprint arXiv:1703.02291*, 2017.

[28] Yitong Li, Martin Renqiang Min, Dinghan Shen, David E. Carlson, and Lawrence Carin. Video generation from text. *CoRR*, abs/1710.00421, 2017.

[29] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. In *EMNLP*, 2018.

[30] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. In *ICMI*, 2018.

[31] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2018.

[32] C. Mario Christoudias, Raquel Urtasun, Mathieu Salzmann, and Trevor Darrell. Learning to recognize objects from unseen modalities. In *ECCV*, 2010.

[33] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[34] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*. ACM, 2011.

[35] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[36] Graham Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*, 2017.

[37] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011.

[38] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *ICMI*, 2016.

[39] Gaurav Pandey and Ambedkar Dukkipati. Variational methods for conditional multimodal deep learning. In *IJCNN*. IEEE, 2017.

[40] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[41] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *ICMI '14*, 2014.

[42] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[43] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*. 2018.

[44] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Challenge-HML*. ACL, July 2018.

[45] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, 2015.

[46] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *ACL*, 2017.

[47] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style transfer through back-translation. In *ACL*, 2018.

[48] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1848–1852, October 2007.

[49] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Goecke Roland. Extending long short-term memory for multi-view structured learning. In *ECCV*, 2016.

[50] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November 1997.

[51] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[52] Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *NIPS*. 2014.

[53] Yale Song, Louis-Philippe Morency, and Randall Davis. Multi-view latent variable discriminative models for action recognition. In *CVPR*. IEEE, 2012.

[54] Yale Song, Louis-Philippe Morency, and Randall Davis. Action recognition by hierarchical sequence summarization. In *CVPR*. IEEE, 2013.

[55] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *JMLR*, 15, 2014.

[56] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[57] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*, 2017.

[58] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[59] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*. 2017.

[61] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*, 2016.

[62] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 2013.

[63] Jiahong Yuan and Mark Liberman. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 2008.

[64] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1114–1125, 2017.

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

[66] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. *arXiv preprint arXiv:1802.00923*, 2018.

[67] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[68] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.

[69] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*. IEEE, 2006.

A Multimodal Features

Here we present extra details on feature extraction for the language, visual and acoustic modalities.

Language: We used 300 dimensional Glove word embeddings trained on 840 billion tokens from the common crawl dataset [42]. These word embeddings were used to embed a sequence of individual words from video segment transcripts into a sequence of word vectors that represent spoken text.

Visual: The library Facet [20] is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features [69]. These visual features are extracted from the full video segment at 30Hz to form a sequence of facial gesture measures throughout time.

Acoustic: The software COVAREP [12] is used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features [15], glottal source parameters [10, 16, 1, 3, 2], peak slope parameters and maxima dispersion quotients [21]. These visual features are extracted from the full audio clip of each segment at 100Hz to form a sequence that represent variations in tone of voice over an audio segment.

B Multimodal Alignment

We perform forced alignment using P2FA [63] to obtain the exact utterance time-stamp of each word. This allows us to align the three modalities together. Since words are considered the basic units of language we use the interval duration of each word utterance as one time-step. We acquire the aligned video and audio features by computing the expectation of their modality feature values over the word utterance time interval [29].

B.1 Baseline Models

We also implement the Stacked, (*EF-SLSTM*) [18], Bidirectional (*EF-BLSTM*) [50], and Stacked Bidirectional (*EF-SBLSTM*) LSTMs, as well as the following baselines: *BC-LSTM* [46], *EF-HCRF* [48], *EF/MV-LDHCRF* [35], *MV-HCRF* [53], *EF/MV-HSSHCRF* [54], *MV-LSTM* [49], *DF* [38], *SAL-CNN* [61], *C-MKL* [45], *THMM* [34], *SVM* [11, 41] and *RF* [6].

C Full Results

We present the full results across all baseline models in Table 3 and Table 4. MCTN using all modalities achieves new start-of-the-art results on binary classification accuracy, F1 score, and MAE on the CMU-MOSI dataset for multimodal sentiment analysis. State-of-the-art results are also achieved on the ICT-MMMO and YouTube datasets (Table 4). These results are even more impressive considering that MCTN only uses the language modality during testing, while other baseline models use all three modalities.

Dataset		CMU-MOSI			
Model	Test Inputs	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
RF	$\{\ell, v, a\}$	56.4	56.3	-	-
SVM	$\{\ell, v, a\}$	71.6	72.3	1.100	0.559
THMM	$\{\ell, v, a\}$	50.7	45.4	-	-
EF-HCRF	$\{\ell, v, a\}$	65.3	65.4	-	-
EF-LDHCRF	$\{\ell, v, a\}$	64.0	64.0	-	-
MV-HCRF	$\{\ell, v, a\}$	44.8	27.7	-	-
MV-LDHCRF	$\{\ell, v, a\}$	64.0	64.0	-	-
CMV-HCRF	$\{\ell, v, a\}$	44.8	27.7	-	-
CMV-LDHCRF	$\{\ell, v, a\}$	63.6	63.6	-	-
EF-HSSHCRF	$\{\ell, v, a\}$	63.3	63.4	-	-
MV-HSSHCRF	$\{\ell, v, a\}$	65.6	65.7	-	-
DF	$\{\ell, v, a\}$	74.2	74.2	1.143	0.518
EF-LSTM	$\{\ell, v, a\}$	74.3	74.3	1.023	0.622
EF-SLSTM	$\{\ell, v, a\}$	72.7	72.8	1.081	0.600
EF-BLSTM	$\{\ell, v, a\}$	72.0	72.0	1.080	0.577
EF-SBLSTM	$\{\ell, v, a\}$	73.3	73.2	1.037	0.619
MV-LSTM	$\{\ell, v, a\}$	73.9	74.0	1.019	0.601
BC-LSTM	$\{\ell, v, a\}$	75.2	75.3	1.079	0.614
TFN	$\{\ell, v, a\}$	74.6	74.5	1.040	0.587
GME-LSTM(A)	$\{\ell, v, a\}$	76.5	73.4	0.955	-
MARN	$\{\ell, v, a\}$	77.1	77.0	0.968	0.625
MFN	$\{\ell, v, a\}$	77.4	77.3	0.965	0.632
LMF	$\{\ell, v, a\}$	76.4	75.7	0.912	0.668
RMFN	$\{\ell, v, a\}$	78.4	78.0	0.922	**0.681**
MCTN	$\{\ell\}$	**79.3**	**79.1**	**0.909**	0.676

Table 3: Sentiment prediction results on CMU-MOSI. Best results are highlighted in bold. MCTN outperforms the current state-of-the-art across most evaluation metrics and uses only the language modality during testing.

Dataset		ICT-MMMO		YouTube	
Model	Test Inputs	Acc(↑)	F1(↑)	Acc(↑)	F1(↑)
RF	$\{\ell, v, a\}$	70.0	69.8	33.3	32.3
SVM	$\{\ell, v, a\}$	68.8	68.7	42.4	37.9
THMM	$\{\ell, v, a\}$	53.8	53.0	42.4	27.9
EF-HCRF	$\{\ell, v, a\}$	50.0	50.3	44.1	43.8
EF-LDHCRF	$\{\ell, v, a\}$	73.8	73.1	45.8	45.0
MV-HCRF	$\{\ell, v, a\}$	36.3	19.3	27.1	19.7
MV-LDHCRF	$\{\ell, v, a\}$	68.8	67.1	44.1	44.0
CMV-HCRF	$\{\ell, v, a\}$	36.3	19.3	30.5	14.3
CMV-LDHCRF	$\{\ell, v, a\}$	51.3	51.4	42.4	42.0
EF-HSSHCRF	$\{\ell, v, a\}$	50.0	51.3	37.3	35.6
MV-HSSHCRF	$\{\ell, v, a\}$	62.5	63.1	44.1	44.0
DF	$\{\ell, v, a\}$	65.0	58.7	45.8	32.0
EF-LSTM	$\{\ell, v, a\}$	66.3	65.0	44.1	43.6
EF-SLSTM	$\{\ell, v, a\}$	72.5	70.9	40.7	41.2
EF-BLSTM	$\{\ell, v, a\}$	63.8	49.6	42.4	38.1
EF-SBLSTM	$\{\ell, v, a\}$	62.5	49.0	37.3	33.2
MV-LSTM	$\{\ell, v, a\}$	72.5	72.3	45.8	43.3
BC-LSTM	$\{\ell, v, a\}$	70.0	70.1	45.0	45.1
TFN	$\{\ell, v, a\}$	72.5	72.6	45.0	41.0
MARN	$\{\ell, v, a\}$	71.3	70.2	48.3	44.9
MFN	$\{\ell, v, a\}$	73.8	73.1	**51.7**	51.6
MCTN	$\{\ell\}$	**81.3**	**80.8**	**51.7**	**52.4**

Table 4: Sentiment prediction results on ICT-MMMO and YouTube. Best results are highlighted in bold. MCTN outperforms the current state-of-the-art across most evaluation metrics and uses only the language modality during testing.