Found in Translation:
Learning Robust Joint Representations by Cyclic Translations Between Modalities

Hai Pham[1]*, **Paul Pu Liang**[1]*, **Thomas Manzini**[2], **Louis-Philippe Morency**[1], **Barnabás Póczos**[1]
[1]Carnegie Mellon University, [2]Microsoft AI
{htpham,pliang}@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 by requiring all modalities as input and as a result, the learned representations may be sensitive to noisy or missing modalities at test time. With the recent success of sequence to sequence (Seq2Seq) 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 final sentiment prediction. This ensures that our model remains robust from perturbations or missing information in the other 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 missing or perturbed modalities.

Introduction

Sentiment analysis is an open research problem in machine learning and natural language processing which involves identifying a speaker's opinion (Pang, Lee, and Vaithyanathan 2002). Previously, text-only sentiment analysis through words, phrases, and their compositionality can be found to be insufficient for inferring sentiment content from spoken opinions (Morency, Mihalcea, and Doshi 2011), especially in the presence of rich nonverbal behaviors which can accompany language (Shaffer 2018). As a result, there has been a recent push towards using machine learning methods to learn joint representations from additional information 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 environment. For example, (Kaushik, Sangwan, and Hansen 2013) explore the additional acoustic modality while (Wöllmer et al. 2013) use the language, visual, and acoustic modalities present in monologue videos to predict sentiment. This push has been further bolstered by the advent of multimodal social media platforms, such as YouTube, Facebook, and VideoLectures which are used to express personal opinions on a worldwide scale. The abundance of multimodal data has led to the creation of multimodal datasets, such as CMU-MOSI (Zadeh et al. 2016) and ICT-MMMO (Wöllmer et al. 2013), as well as deep multimodal models that are highly effective at learning discriminative joint multimodal representations (Liang, Zadeh, and Morency 2018; Tsai et al. 2018; Chen et al. 2017). Existing prior work learns joint representations using multiple modalities as input (Liang et al. 2018; Morency, Mihalcea, and Doshi 2011; Zadeh et al. 2016). However, these joint representations also

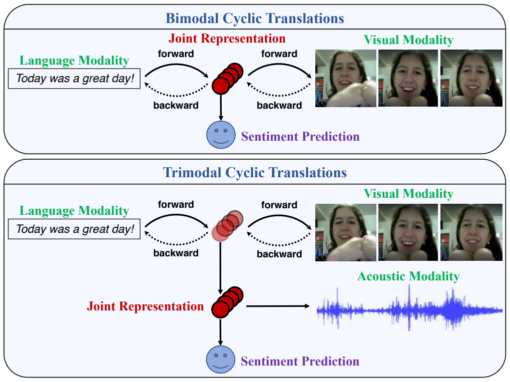


Figure 1: Learning robust joint representations via multimodal cyclic translations. Top: cyclic translations from a source modality (language) to a target modality (visual). Bottom: the representation learned between language and vision are further translated into the acoustic modality, forming the final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

* Equal contributions

regain all modalities at test time, making them sensitive to noisy or missing modalities at test time (Tran et al. 2017; Cai et al. 2018).

To address this problem, we draw inspiration from the recent success of Seq2Seq models for unsupervised representation learning (Sutskever, Vinyals, and Le 2014; Tu et al. 2016). 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 two or three 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 modalities, 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 (*i.e.* multimodal sentiment analysis). Another advantage of MCTN is that once trained with multimodal data, we *only* need data from the source modality at test time to infer the joint representation and label. As a result, MCTN is completely robust to test time perturbations or missing information on other modalities.

Even though translation and generation of videos, audios, and text are difficult (Li et al. 2017b), 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 (Zadeh et al. 2016), ICT-MMMO (Wöllmer et al. 2013), and YouTube (Morency, Mihalcea, and Doshi 2011) public datasets. Additional experiments show that MCTN learns increasingly discriminative joint representations with more input modalities during training.

Related Work

Early work on sentiment analysis focused primarily on written text (Pang, Lee, and Vaithyanathan 2002; Pang, Lee, and others 2008; Socher et al. 2013). Recently, multimodal sentiment analysis has gained more research interest (Baltrusaitis, Ahuja, and Morency 2017). Probably the most challenging task in multimodal sentiment analysis is learning a joint representation of multiple modalities. Earlier work used fusion approaches such as concatenation of input features (Ngiam et al. 2011; Lazaridou, Pham, and Baroni 2015). Several neural network models have also been proposed to learn joint multimodal representations. (Liang et al. 2018) presented a multistage approach to learn hierarchical multimodal representations. The Tensor Fusion Network (Zadeh et al. 2017) and its approximate low-rank model (Liu et al. 2018) presented methods based on Cartesian-products to model unimodal, bimodal and trimodal interactions. The Gated Multimodal Embedding model (Chen et al. 2017) learns an on-off switch to filter noisy or contradictory modalities. Other models have proposed using attention (Cheng et al. 2017) and memory mechanisms (Zadeh et al. 2018) to learn multimodal representations.

In addition to purely supervised approaches, generative methods based on Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have attracted significant interest in learning joint distributions between two or more modalities (Donahue, Krähenbühl, and Darrell 2016; Li et al. 2017a). Another method for multimodal data is to develop conditional generative models (Kingma et al. 2014; Pandey and Dukkipati 2017) and learn to translate one modality to another. Generative-discriminative objectives have been used to learn either joint (Pham et al. 2018; Kiros, Salakhutdinov, and Zemel 2014) or factorized (Tsai et al. 2018) representations. Our work takes into account the sequential dependency of modality translations and 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. One general approach is to infer the missing modalities by modeling the probabilistic relationships among different modalities. Srivastava and Salakhutdinov (2014) proposed using Deep Boltzmann Machines to jointly model the probability distribution over multimodal data. Sampling from the conditional distributions over each modality allows for test-time inference in the presence of missing modalities. Sohn, Shang, and Lee (2014) trained Restricted Boltzmann Machines to minimize the variation of information between modality-specific latent variables. Recently, neural models such as cascaded residual autoencoders (Tran et al. 2017), deep adversarial learning (Cai et al. 2018), or multiple kernel learning (Mario Christoudias et al. 2010) have also been proposed for these tasks. It was also found that training with modalities dropped at random can improve the robustness of joint representations (Ngiam et al. 2011). These methods approximately infer the missing modalities before prediction (Hill, Reichart, and Korhonen 2014; Collell, Zhang, and Moens 2017), leading to possible error compounding. On the other hand, MCTN remains fully robust to missing or perturbed modalities during testing.

Proposed Approach

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

Problem Formulation and Notation

A multimodal dataset consists of N labeled video segments defined as $\mathbf{X} = (\mathbf{X}^l, \mathbf{X}^v, \mathbf{X}^a)$ for the language, visual, and acoustic modalities respectively. The dataset is indexed by N such that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N)$ where $\mathbf{X}_i = (\mathbf{X}_i^l, \mathbf{X}_i^v, \mathbf{X}_i^a)$, $1 \leq i \leq N$. The corresponding labels for these N segments are denoted as $\mathbf{y} = (y_1, y_2, ..., y_N)$, $y_i \in \mathbb{R}$. Following prior work, the multimodal data is synchronized by aligning the input based on the boundaries of each word and zero-padding each example to obtain time-series data of the same length (Liang et al. 2018). The ith sample 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 example. 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 $\mathbf{X}_i^a = (\mathbf{a}_i{}^{(1)}, \mathbf{a}_i{}^{(2)}, ..., \mathbf{a}_i{}^{(L)})$.

Learning Joint Representations

Learning a joint representation between two modalities $\mathbf{X}^S$ and $\mathbf{X}^T$ is defined by 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 (Liang et al. 2018; Liu et al. 2018; Tsai et al. 2018; Zadeh et al. 2018). During training, the parameters θ and w are learned by empirical risk minimization over paired multimodal data and labels in the training set $(\mathbf{X}_{tr}^S, \mathbf{X}_{tr}^T, \mathbf{y}_{tr})$:

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

$$\hat{\mathbf{y}}_{tr} = g_w(\mathcal{E}_{ST}), \tag{2}$$

$$\theta^*, w^* = \arg\min_{\theta, w} \mathbb{E}\left[\ell_{\mathbf{y}}(\hat{\mathbf{y}}_{tr}, \mathbf{y}_{tr})\right]. \tag{3}$$

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):

$$\mathcal{E}_{ST} = f_{\theta^*}(\mathbf{X}_{te}^S, \mathbf{X}_{te}^T), \tag{4}$$

$$\hat{\mathbf{y}}_{te} = g_{w^*}(\mathcal{E}_{ST}). \tag{5}$$

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 an intermediate 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 during test time.

To ensure that our model learns joint representations that retain maximal information from all modalities, we use a cycle consistency loss (Zhu et al. 2017) during modality translation. This method can also be seen as a variant of back-translation which has been recently applied to style transfer (Prabhumoye et al. 2018; Zhu et al. 2017) and unsupervised machine translation (Lample et al. 2018). We use back-translation in a multimodal environment 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 $\mathcal{E}_{S\to T}$:

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

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

$$\mathbf{X}^T = f_{\theta_d}(\mathcal{E}_{S\to T}), \tag{7}$$

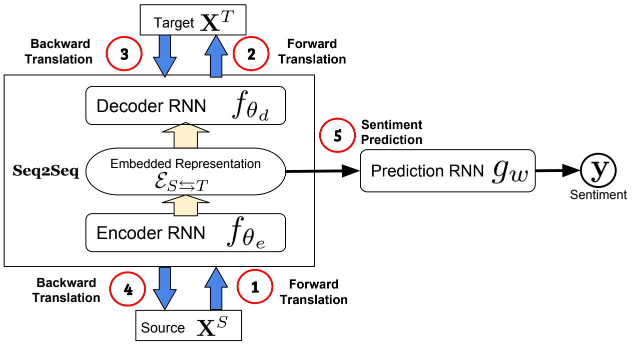


Figure 2: MCTN architecture for two modalities: the source modality $\mathbf{X}^S$ and the target modality $\mathbf{X}^T$. The joint representation $\mathcal{E}_{S\leftrightarrows T}$ is obtained via a cyclic translation between $\mathbf{X}^S$ and $\mathbf{X}^T$. Next, the joint representation $\mathcal{E}_{S\leftrightarrows T}$ 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.

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

$$\mathcal{E}_{T\to S} = f_{\theta_e}(\hat{\mathbf{X}}^T), \ \hat{\mathbf{X}}^S = f_{\theta_d}(\mathcal{E}_{T\to S}). \tag{8}$$

The joint representation is learned by using a Seq2Seq model with attention (Bahdanau, Cho, and Bengio 2014) 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 hidden state output of each time step is based on the previous hidden state along with the input sequence and is constructed using a recurrent network.

$$\mathbf{h}_\ell = \mathtt{RNN}(\mathbf{h}_{\ell-1}, \mathbf{X}_\ell^S) \quad \forall \ell \in [1, L]. \tag{9}$$

The encoder's output is the concatenation of all hidden states of the encoding RNN,

$$\mathcal{E}_{S\to T} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_L], \tag{10}$$

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

$$p(\mathbf{X}^T) = \prod_{\ell=1}^{L} p(\mathbf{X}_\ell^T | \mathcal{E}_{S\to T}, \mathbf{X}_1^T, ..., \mathbf{X}_{\ell-1}^T). \tag{11}$$

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

$$\hat{\mathbf{X}^T} = \arg\max_{\mathbf{X}^T} p(\mathbf{X}^T|\mathbf{X}^S). \tag{12}$$

We use the traditional beam search approach (Sutskever, Vinyals, and Le 2014) for decoding.

To obtain the joint representation for multimodal prediction, we only use the forward translated representation during inference to remove the dependency on the target modality at test time. If cyclic translation is used, we denote the translated representation with the symbol $\leftrightarrows$:

$$\mathcal{E}_{S \leftrightarrows T} = \mathcal{E}_{S \to T}. \tag{13}$$

$\mathcal{E}_{S \leftrightarrows T}$ is then used for sentiment prediction:

$$\hat{\mathbf{y}} = g_w(\mathcal{E}_{S \leftrightarrows T}). \tag{14}$$

Coupled Translation-Prediction Objective

Training is performed with paired multimodal data and labels in the training set $(\mathbf{X}^S_{tr}, \mathbf{X}^T_{tr}, \mathbf{y}_{tr})$ The first two losses are the forward translation loss $\mathcal{L}_t$ defined as

$$\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)], \tag{15}$$

and the cycle consistency loss $\mathcal{L}_c$ defined as

$$\mathcal{L}_c = \mathbb{E}[\ell_{\mathbf{X}^S}(\hat{\mathbf{X}}^S, \mathbf{X}^S)] \tag{16}$$

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

$$\mathcal{L}_p = \mathbb{E}[\ell_{\mathbf{y}}(\hat{\mathbf{y}}, \mathbf{y})] \tag{17}$$

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

Our MCTN model is trained end-to-end with a coupled translation-prediction objective function defined as

$$\mathcal{L} = \lambda_t \mathcal{L}_t + \lambda_c \mathcal{L}_c + \mathcal{L}_p. \tag{18}$$

where λ_t, λ_t are weighting hyperparameters. MCTN parameters are learned by minimizing this objective function

$$\theta_e^*, \theta_d^*, w^* = \underset{\theta_e, \theta_d, w}{\arg\min} \, [\lambda_t \mathcal{L}_t + \lambda_c \mathcal{L}_c + \mathcal{L}_p]. \tag{19}$$

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

$$\mathcal{E}_{S \leftrightarrows T} = f_{\theta_e^*}(\mathbf{X}^S), \tag{20}$$

$$\hat{\mathbf{y}} = g_{w^*}(\mathcal{E}_{S \leftrightarrows T}). \tag{21}$$

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.

Hierarchical MCTN for Three Modalities

We extend the MCTN in a hierarchical manner 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 modality 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.

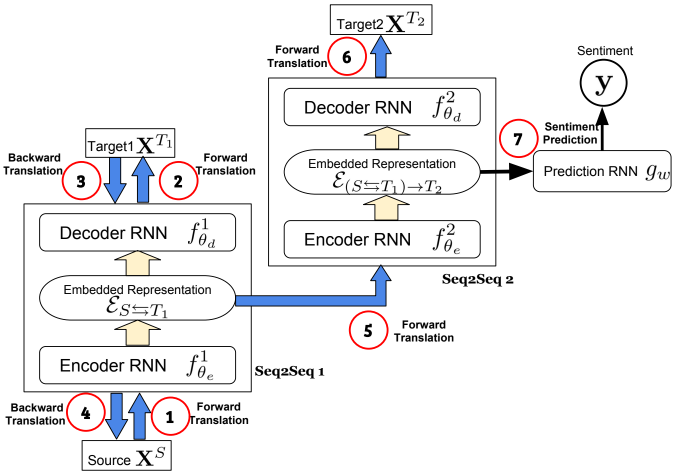

Target2 $\mathbf{X}^{T_2}$
Forward Translation
6
Sentiment
$\mathbf{y}$
Decoder RNN $f^2_{\theta_d}$
Embedded Representation $\mathcal{E}_{(S \leftrightarrows T_1) \to T_2}$
Encoder RNN $f^2_{\theta_e}$
7
Sentiment Prediction
Prediction RNN g_w
Seq2Seq 2
Target1 $\mathbf{X}^{T_1}$
Backward Translation
3
2
Forward Translation
Decoder RNN $f^1_{\theta_d}$
Embedded Representation $\mathcal{E}_{S \leftrightarrows T_1}$
Encoder RNN $f^1_{\theta_e}$
Seq2Seq 1
5
Forward Translation
Backward Translation
4
1
Forward Translation
Source $\mathbf{X}^S$

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 for prediction.

Two Seq2Seq models are used in the hierarchical MCTN for three modalities, denoted 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 forward translation is defined as

$$\mathcal{E}_{S \to T_1} = f^1_{\theta_e}(\mathbf{X}^S_{tr}), \; \hat{\mathbf{X}}^{T_1}_{tr} = f^1_{\theta_d}(\mathcal{E}_{S \to T_1}), \tag{22}$$

and followed by the decoded modality $\mathbf{X}^{T_1}$ being translated back into modality $\mathbf{X}^S$:

$$\mathcal{E}_{T_1 \to S} = f^1_{\theta_e}(\hat{\mathbf{X}}^{T_1}_{tr}), \; \hat{\mathbf{X}}^S_{tr} = f^1_{\theta_d}(\mathcal{E}_{T_1 \to S}). \tag{23}$$

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

$$\mathcal{E}_{S \leftrightarrows T_1} = \mathcal{E}_{S \to T_1}, \tag{24}$$

$$\mathcal{E}_{(S \leftrightarrows T_1) \to T_2} = f^2_{\theta_e}(\mathcal{E}_{S \leftrightarrows T_1}), \; \hat{\mathbf{X}}^{T_2}_{tr} = f^2_{\theta_d}(\mathcal{E}_{(S \leftrightarrows T_1) \to T_2}). \tag{25}$$

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

Training the hierarchical MCTN involves computing a cycle consistent loss for modality T_1, given by the respective forward translation loss $\mathcal{L}_{t_1}$ and the cycle consistency loss $\mathcal{L}_{c_1}$. 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

$$\mathcal{L} = \lambda_{t_1} \mathcal{L}_{t_1} + \lambda_{c_1} \mathcal{L}_{c_1} + \lambda_{t_2} \mathcal{L}_{t_2} + \mathcal{L}_p \tag{26}$$

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 noisy or missing target 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	-	-
MV-HCRF	$\{\ell, v, a\}$	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
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 1: 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.

Experimental Setup

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

Dataset and Input Modalities

We use the CMU Multimodal Opinion-level Sentiment Intensity dataset (CMU-MOSI) which contains 2199 video segments each with a sentiment label in the range $[-3, +3]$. To be consistent with prior work, we use 52 segments for training, 10 for validation and 31 for testing. The same speaker does not appear in both training and testing sets to ensure that our model learns speaker-independent representations. We also run experiments on ICT-MMMO (Wöllmer et al. 2013) and YouTube (Morency, Mihalcea, and Doshi 2011) which consist of online review videos annotated for sentiment.

Multimodal Features and Alignment

Following previous work (Liang et al. 2018), GloVe word embeddings (Pennington, Socher, and Manning 2014), Facet (iMotions 2017), and COVAREP (Degottex et al. 2014) features are extracted for the language, visual and acoustic modalities respectively[2]. Forced alignment is performed using P2FA (Yuan and Liberman 2008) to obtain spoken word utterance times. The visual and acoustic features are aligned by computing their average over the utterance interval of each word.

Evaluation Metrics

For parameter optimization on CMU-MOSI, the prediction loss function is set 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 r. We also perform sentiment classification on CMU-MOSI and report binary accuracy (Acc) and F1 score (F1). On ICT-MMMO and YouTube, we set the prediction loss function as categorical cross-entropy and report sentiment classification and F1 score. For all metrics, higher values indicate stronger performance, except MAE where lower values indicate stronger performance.

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	$\{\ell, v, a\}$	73.8	73.1	45.8	45.0
MV-HCRF	$\{\ell, v, a\}$	68.8	67.1	44.1	44.0
DF	$\{\ell, v, a\}$	65.0	58.7	45.8	32.0
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 2: 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.

Baseline Models

We compare to the following multimodal models: *RMFN* (Liang et al. 2018) uses a multistage approach to learn hierarchical representations (current state-of-the-art on CMU-MOSI). *LMF* (Liu et al. 2018) approximates the expensive tensor products in *TFN* (Zadeh et al. 2017) with efficient low-rank factors. *MFN* (Zadeh et al. 2018) synchronizes sequences using a multimodal gated memory. *EF-LSTM* concatenates multimodal inputs and uses a single LSTM (Hochreiter and Schmidhuber 1997). For a description of other baselines, please refer to the supplementary material.

Results and Discussion

This section presents and discusses 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 previous models[3]. From Table 1, MCTN using language as the source modality achieves new start-of-the-art results on CMU-MOSI for multimodal sentiment analysis. State-of-the-art results are also achieved on ICT-MMMO and YouTube (Table 2). It is important to note that MCTN only uses language during testing, while other baselines 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?

[1] Our source code is released at `https://github.com/hainow/MCTN`.

[2] Details on feature extraction are in supplementary material.

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

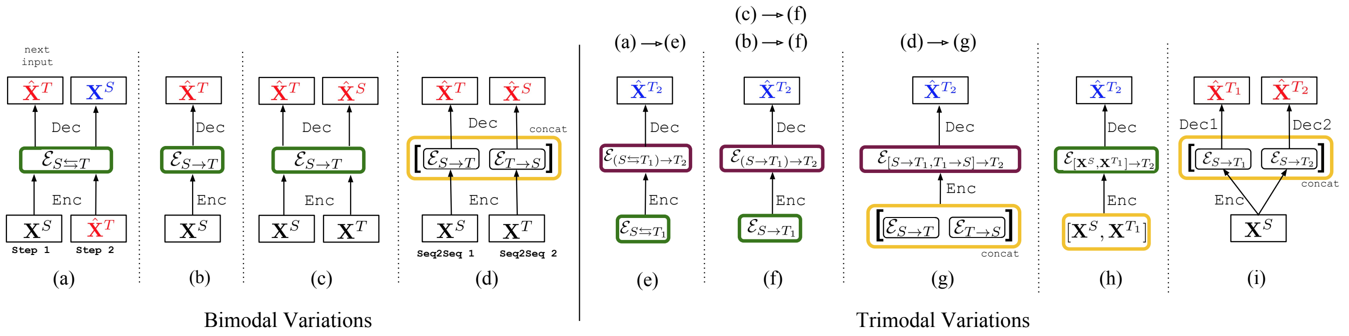
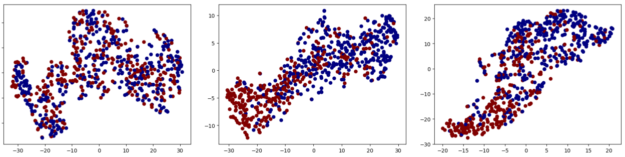

next input
$\hat{\mathbf{X}}^T$
$\hat{\mathbf{X}}^S$
Dec
$\mathcal{E}_{S\leftrightarrows T}$
Enc
$\mathbf{X}^S$
$\hat{\mathbf{X}}^T$
Step 1
Step 2
(a)
$\mathcal{E}_{S\to T}$
(b)
$\mathbf{X}^T$
(c)
$[\mathcal{E}_{S\to T}]$ $[\mathcal{E}_{T\to S}]$
concat
Seq2Seq 1
Seq2Seq 2
(d)
(a) → (e)
(c) → (f)
(b) → (f)
(d) → (g)
$\hat{\mathbf{X}}^{T_2}$
$\mathcal{E}_{(S\leftrightarrows T_1)\to T_2}$
$\mathcal{E}_{S\leftrightarrows T_1}$
(e)
$\mathcal{E}_{(S\to T_1)\to T_2}$
$\mathcal{E}_{S\to T_1}$
(f)
$\mathcal{E}_{[S\to T_1,T_1\to S]\to T_2}$
(g)
$\mathcal{E}_{[\mathbf{X}^S,\mathbf{X}^{T_1}]\to T_2}$
$[\mathbf{X}^S, \mathbf{X}^{T_1}]$
(h)
$\hat{\mathbf{X}}^{T_1}$
Dec1
Dec2
$\mathcal{E}_{S\to T_1}$ $\mathcal{E}_{S\to T_2}$
(i)
Bimodal Variations
Trimodal Variations

Figure 4: Variations of our models: (a) MCTN Bimodal with cyclic translation, (b) Simple Bimodal without cyclic translation, (c) No-Cycle Bimodal with different inputs of the same modality pair, and without cyclic translation, (d) Double Bimodal for two modalities without cyclic translation, with two different inputs (of the same pair), (e) MCTN Trimodal with input from (a), (f) Simple Trimodal for three modalities, with input as a joint representation taken from previous MCTN for two modalities from (b) or (c), (g) Double Trimodal with input from (d), (h) Concat Trimodal which is similar to (b) but with input as the concatenation of 2 modalities, (i) Paired Trimodal using one encoder and 2 separate decoders for modality translations. *Legend*: black modality is ground truth, red ("hat") modality represents translated output, blue ("hat") modality is target output from previous translation outputs, and yellow box denotes concatenation.



(a) MCTN Bimodal *without* cyclic translations

(b) MCTN Bimodal *with* cyclic translations

(c) MCTN Trimodal *with* cyclic translations

Figure 5: t-SNE visualization of the joint representations learned by MCTN. *Legend*: red: videos with negative sentiment, blue: videos with positive sentiment. Adding modalities and using cyclic translations improve discriminative performance and leads to increasingly separable representations.

Dataset		CMU-MOSI			
Model	Translation	Acc	F1	MAE	Corr
MCTN Bimodal (4a)	$V \leftrightarrows A$	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 Trimodal (4e)	$(V \leftrightarrows A) \to T$	56.4	56.3	1.455	0.151
	$(T \leftrightarrows A) \to V$	78.7	78.8	0.960	0.650
	$(T \leftrightarrows V) \to A$	**79.3**	**79.1**	**0.909**	**0.676**

Table 3: MCTN performance improves as more modalities are introduced for cyclic translations during training.

We run experiments with MCTN using combinations of two or three modalities with cyclic translations. From Table 3, 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 discriminative information for sentiment (Zadeh et al. 2017).

Dataset		CMU-MOSI			
Model	Translation	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
MCTN Bimodal (4a)	$V \leftrightarrows A$	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 Bimodal (4b)	$V \to A$	55.4	55.5	1.422	0.119
	$T \to A$	74.2	74.2	0.988	0.616
	$T \to V$	75.7	75.6	1.002	0.617
No-Cycle Bimodal (4c)	$V \to A,\ A \to V$	55.4	55.5	1.422	0.119
	$T \to A,\ A \to T$	75.5	75.6	**0.971**	0.629
	$T \to V,\ V \to T$	75.2	75.3	0.972	0.627
Double Bimodal (4d)	$[V \to A, A \to V]$	57.0	57.1	1.502	0.168
	$[T \to A, A \to T]$	72.3	72.3	1.035	0.578
	$[T \to V, V \to T]$	73.3	73.4	1.020	0.570

Table 4: Bimodal variations results on CMU-MOSI dataset. MCTN Bimodal with cyclic translations performs best.

In addition, we visually inspect the joint representations learned from MCTN as we add more modalities during training (see Table 5). The joint representations for each 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 (van der Maaten and Hinton 2008). Each point is colored red or blue depending on whether the video segment is annotated for positive or negative sentiment. From Figure 5, 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 with more modalities (as seen in Table 3).

Ablation Studies

We use several models to test our design decisions. Specifically, we evaluate the impact of cyclic translations, modality ordering, and hierarchical structure.

For bimodal MCTN, we design the following ablation

Dataset	CMU-MOSI				
Model	Translation	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
MCTN Trimodal (4e)	$(V \leftrightarrows A) \to T$	56.4	56.3	1.455	0.151
	$(T \leftrightarrows A) \to V$	78.7	78.8	0.960	0.650
	$(T \leftrightarrows V) \to A$	**79.3**	**79.1**	**0.909**	**0.676**
Simple Trimodal (4f)	$(V \to T) \to A$	54.1	52.9	1.408	0.040
	$(V \to A) \to T$	52.0	51.9	1.439	0.015
	$(A \to V) \to T$	56.6	56.7	1.593	0.067
	$(A \to T) \to V$	54.1	54.2	1.577	0.028
	$(T \to A) \to V$	74.3	74.4	1.001	0.609
	$(T \to V) \to A$	74.3	74.4	0.997	0.596
Double Trimodal (4g)	$[T \to V, V \to T] \to A$	73.3	73.1	1.058	0.578
Concat Trimodal (4h)	$[V, A] \to T$	55.0	54.6	1.535	0.176
	$[A, T] \to V$	73.3	73.4	1.060	0.561
	$[T, V] \to A$	72.3	72.3	1.068	0.576
	$A \to [T, V]$	55.5	55.6	1.617	0.056
	$T \to [A, V]$	75.7	75.7	0.958	0.634
	$[T, A] \to [T, V]$	73.2	73.2	1.008	0.591
	$[T, V] \to [T, A]$	74.1	74.1	0.999	0.607
Paired Trimodal (4i)	$[T \to A, T \to V]$	73.8	73.8	1.022	0.611

Table 5: Trimodal variations results on CMU-MOSI dataset. MCTN (hierarchical) with cyclic translations performs best.

models shown in the left half of Figure 4: (a) MCTN bimodal between $\mathbf{X}^S$ and $\mathbf{X}^T$, (b) simple bimodal by translating from $\mathbf{X}^S$ to $\mathbf{X}^T$ without cyclic loss, (c) no-cycle bimodal which does not use cyclic translations but rather performs two independent translations between $\mathbf{X}^S$ and $\mathbf{X}^T$, (d) double bimodal: two seq2seq 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 4: (e) MCTN trimodal which uses the proposed hierarchical translations between $\mathbf{X}^S$, $\mathbf{X}^{T_1}$ and $\mathbf{X}^{T_2}$, (f) simple trimodal based on translation from $\mathbf{X}^S$ to $\mathbf{X}^{T_1}$ without cyclic translations, (g) double trimodal extended from (d) which does not use cyclic translations but rather performs two independent translations between $\mathbf{X}^S$ and $\mathbf{X}^{T_1}$, (h) concat trimodal which does not perform a first level of cyclic translation but directly translates the concatenated modality pair $[\mathbf{X}^S, \mathbf{X}^{T_1}]$ into $\mathbf{X}^{T_2}$, and finally, (i) paired trimodal which uses two separate decoders on top of the intermediate representation.

Q3: What is the impact of cyclic translations in MCTN?

The bimodal results are in Table 4. The models that employ cyclic translations (Figure 4(a)) outperform all other models. The trimodal results are in Table 5 and we make a similar observation: Figure 4(e) with cyclic translations outperforms the baselines (f), (g) and (h). The gap for the trimodal case is especially large. This implies that using cyclic translations is crucial for learning discriminative joint representations. Our intuition is that using cyclic translations: (1) encourages the model to enforce symmetry between the representations from source and target modalities thus adding a source of regularization, and (2) ensures that the 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 4(c), which uses one Seq2Seq model for cyclic translations with Figure 4(d), which uses two separate Seq2Seq models: one for forward translation and one for backward translation. We observe from Table 4 that (c) > (d), so using one model with shared parameters is better. This is also true for hierarchical MCTN: (f) > (g) in Table 5. We hypothesize that this is because training two deep Seq2Seq models requires more data and is prone to overfitting. Also, it does not learn only a single joint representation but instead two separate representations.

Q5: What is the impact of varying source and target modalities for cyclic translations?

From Tables 3, 4 and 5, we observe that language contributes most towards the joint representations. For bimodal cases, combining language with visual is generally better than combining the language and acoustic modalities. 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 audio. 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, for the MCTN, language appears to be the most discriminative modality making it crucial to be used as the source modality during translations.

Q6: What is the impact of using two levels of translations instead of one level when learning from three modalities?

Our hierarchical MCTN is shown in Figure 4(e). In Figure 4(h), we concatenate two modalities as input and use only one phase of translation. From Table 5, 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.

Conclusion

This paper investigated learning joint representations via cyclic translations from source to target modalities. During testing, we only need the source modality for prediction which ensures robustness to noisy or missing target modalities. We demonstrate that cyclic translations and seq2seq models are useful for learning joint representations in multimodal environments. In addition to achieving new state-of-the-art results on three datasets, our model learns increasingly discriminative joint representations with more input modalities while maintaining robustness to all target modalities.

Acknowledgements

PPL and LPM are partially supported by the NSF (Award #1833355) and Oculus VR. HP and BP are supported by NSF grant IIS1563887 and the DARPA D3M program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation, DARPA, or Oculus VR, and no official endorsement should be inferred. The authors thank Hieu Pham, Amir Zadeh, and anonymous reviewers for useful discussions and feedback.

References

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

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

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

[Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Baltrusaitis, Ahuja, and Morency 2017] Baltrusaitis, T.; Ahuja, C.; and Morency, L. 2017. Multimodal machine learning: A survey and taxonomy. *CoRR* abs/1705.09406.

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

[Cai et al. 2018] Cai, L.; Wang, Z.; Gao, H.; Shen, D.; and Ji, S. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining*, KDD '18, 1158–1166. New York, NY, USA: ACM.

[Chen et al. 2017] Chen, M.; Wang, S.; Liang, P. P.; Baltrušaitis, T.; Zadeh, A.; and Morency, L.-P. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. *ICMI*.

[Cheng et al. 2017] Cheng, Y.; Huang, F.; Zhou, L.; Jin, C.; Zhang, Y.; and Zhang, T. 2017. A hierarchical multimodal attention-based neural network for image captioning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17. New York, NY, USA: ACM.

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

[Collell, Zhang, and Moens 2017] Collell, G.; Zhang, T.; and Moens, M.-F. 2017. Imagined visual representations as multimodal embeddings. *AAAI*.

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

[Degottex et al. 2014] Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. Covarep - a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE.

[Donahue, Krähenbühl, and Darrell 2016] Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

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

[Drugman et al. 2012] Drugman, T.; Thomas, M.; Gudnason, J.; Naylor, P.; and Dutoit, T. 2012. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing* 20(3):994–1006.

[Goodfellow et al. 2014] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

[Graves, r. Mohamed, and Hinton 2013] Graves, A.; r. Mohamed, A.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE ICASSP*, 6645–6649.

[Hill, Reichart, and Korhonen 2014] Hill, F.; Reichart, R.; and Korhonen, A. 2014. Multi-modal models for concrete and abstract concept meaning. *TACL*.

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

[iMotions 2017] iMotions. 2017. Facial expression analysis.

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

[Kaushik, Sangwan, and Hansen 2013] Kaushik, L.; Sangwan, A.; and Hansen, J. H. 2013. Sentiment extraction from natural audio streams. In *ICASSP*. IEEE.

[Kingma et al. 2014] Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, 3581–3589.

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

[Lample et al. 2018] Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Phrase-based & neural unsupervised machine translation. *CoRR* abs/1804.07755.

[Lazaridou, Pham, and Baroni 2015] Lazaridou, A.; Pham, N. T.; and Baroni, M. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

[Li et al. 2017a] Li, C.; Xu, K.; Zhu, J.; and Zhang, B. 2017a. Triple generative adversarial nets. *arXiv preprint arXiv:1703.02291*.

[Li et al. 2017b] Li, Y.; Min, M. R.; Shen, D.; Carlson, D. E.; and Carin, L. 2017b. Video generation from text. *CoRR* abs/1710.00421.

[Liang et al. 2018] Liang, P. P.; Liu, Z.; Zadeh, A.; and Morency, L.-P. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.

[Liang, Zadeh, and Morency 2018] Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI.

[Liu et al. 2018] Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*.

[Mario Christoudias et al. 2010] Mario Christoudias, C.; Urtasun, R.; Salzmann, M.; and Darrell, T. 2010. Learning to recognize objects from unseen modalities. In Daniilidis, K.; Maragos, P.; and Paragios, N., eds., *Computer Vision – ECCV 2010*.

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

[Morency, Quattoni, and Darrell 2007] Morency, L.-P.; Quattoni, A.; and Darrell, T. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–8. IEEE.

[Ngiam et al. 2011] Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings* *of the 28th international conference on machine learning (ICML-11)*.

[Nojavanasghari et al. 2016] Nojavanasghari, B.; Gopinath, D.; Koushik, J.; Baltrušaitis, T.; and Morency, L.-P. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016.

[Pandey and Dukkipati 2017] Pandey, G., and Dukkipati, A. 2017. Variational methods for conditional multimodal deep learning. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, 308–315. IEEE.

[Pang, Lee, and others 2008] Pang, B.; Lee, L.; et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

[Pang, Lee, and Vaithyanathan 2002] Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: Sentiment classification using machine learning techniques. EMNLP.

[Park et al. 2014] Park, S.; Shim, H. S.; Chatterjee, M.; Sagae, K.; and Morency, L.-P. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *ICMI '14*.

[Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, 1532–1543.

[Pham et al. 2018] Pham, H.; Manzini, T.; Liang, P. P.; and Poczos, B. 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language*. ACL.

[Poria et al. 2017] Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

[Poria, Cambria, and Gelbukh 2015] Poria, S.; Cambria, E.; and Gelbukh, A. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*.

[Prabhumoye et al. 2018] Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style transfer through back-translation. *CoRR* abs/1804.09000.

[Quattoni et al. 2007] Quattoni, A.; Wang, S.; Morency, L.-P.; Collins, M.; and Darrell, T. 2007. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 29(10):1848–1852.

[Rajagopalan et al. 2016] Rajagopalan, S. S.; Morency, L.-P.; Baltrušaitis, T.; and Roland, G. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*.

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

[Shaffer 2018] Shaffer, I. R. 2018. Exploring the performance of facial expression recognition technologies on deaf adults and their children. In *SIGACCESS Conference on Computers and Accessibility*.

[Socher et al. 2013] Socher, R.; Perelygin, A.; Wu, J. Y.; Chuang, J.; Manning, C. D.; Ng, A. Y.; Potts, C.; et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, 1642. Citeseer.

[Sohn, Shang, and Lee 2014] Sohn, K.; Shang, W.; and Lee, H. 2014. Improved multimodal deep learning with variation of information. In *NIPS*.

[Song, Morency, and Davis 2012] Song, Y.; Morency, L.-P.; and Davis, R. 2012. Multi-view latent variable discriminative models for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2120–2127. IEEE.

[Song, Morency, and Davis 2013] Song, Y.; Morency, L.-P.; and Davis, R. 2013. Action recognition by hierarchical sequence summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3562–3569.

[Srivastava and Salakhutdinov 2014] Srivastava, N., and Salakhutdinov, R. 2014. Multimodal learning with deep boltzmann machines. *JMLR* 15.

[Sutskever, Vinyals, and Le 2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

[Tran et al. 2017] Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing modalities imputation via cascaded residual autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 4971–4980.

[Tsai et al. 2018] Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

[Tu et al. 2016] Tu, Z.; Liu, Y.; Shang, L.; Liu, X.; and Li, H. 2016. Neural machine translation with reconstruction. *CoRR* abs/1611.01874.

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

[Wang et al. 2016] Wang, H.; Meghawat, A.; Morency, L.-P.; and Xing, E. P. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *arXiv preprint arXiv:1609.05244*.

[Wöllmer et al. 2013] Wöllmer, M.; Weninger, F.; Knaup, T.; Schuller, B.; Sun, C.; Sagae, K.; and Morency, L.-P. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28(3):46–53.

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

[Zadeh et al. 2016] Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31(6):82–88.

[Zadeh et al. 2017] Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, 1114–1125.

[Zadeh et al. 2018] Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927*.

[Zhu et al. 2006] Zhu, Q.; Yeh, M.-C.; Cheng, K.-T.; and Avidan, S. 2006. Fast human detection using a cascade of histograms of oriented gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, 1491–1498. IEEE.

[Zhu et al. 2017] Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR* abs/1703.10593.

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 (Pennington, Socher, and Manning 2014). 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 (iMotions 2017) is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features (Zhu et al. 2006). 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 (Degottex et al. 2014) is used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features (Drugman and Alwan 2011), glottal source parameters (Childers and Lee 1991; Drugman et al. 2012; Alku 1992; Alku, Strik, and Vilkman 1997; Alku, Bäckström, and Vilkman 2002), peak slope parameters and maxima dispersion quotients (Kane and Gobl 2013). 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.

Multimodal Alignment

We perform forced alignment using P2FA (Yuan and Liberman 2008) 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 (Liang et al. 2018).

Baseline Models

We also implement the Stacked, (*EF-SLSTM*) (Graves, r. Mohamed, and Hinton 2013), Bidirectional (*EF-BLSTM*) (Schuster and Paliwal 1997), and Stacked Bidirectional (*EF-SBLSTM*) LSTMs, as well as the following baselines: *BC-LSTM* (Poria et al. 2017), *EF-HCRF* (Quattoni et al. 2007), *EF/MV-LDHCRF* (Morency, Quattoni, and Darrell 2007), *MV-HCRF* (Song, Morency, and Davis 2012), *EF/MV-HSSHCRF* (Song, Morency, and Davis 2013), *MV-LSTM* (Rajagopalan et al. 2016), *DF* (Nojavanasghari et al. 2016), *SAL-CNN* (Wang et al. 2016), *C-MKL* (Poria, Cambria, and Gelbukh 2015), *THMM* (Morency, Mihalcea, and Doshi 2011), *SVM* (Cortes and Vapnik 1995; Park et al. 2014) and *RF* (Breiman 2001).

Full Results

We present the full results across all baseline models in Table 6 and Table 7. 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 7). 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 6: 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 7: 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.