Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis

Hai Pham[1]*, **Thomas Manzini**[1]*, **Paul Pu Liang**[2], **Barnabás Poczós**[2]
{[1]Language Technologies Institute, [2]Machine Learning Department}, CMU, USA
{htpham,tmanzini,pliang,bapoczos}@cs.cmu.edu

Abstract

Multimodal machine learning is a core research area spanning the language, visual and acoustic modalities. The central challenge in multimodal learning involves learning representations that can process and relate information from multiple modalities. In this paper, we propose two methods for unsupervised learning of joint multimodal representations using sequence to sequence (Seq2Seq) methods: a *Seq2Seq Modality Translation Model* and a *Hierarchical Seq2Seq Modality Translation Model*. We also explore multiple different variations on the multimodal inputs and outputs of these seq2seq models. Our experiments on multimodal sentiment analysis using the CMU-MOSI dataset indicate that our methods learn informative multimodal representations that outperform the baselines and achieve improved performance on multimodal sentiment analysis, specifically in the Bimodal case where our model is able to improve F1 Score by twelve points. We also discuss future directions for multimodal Seq2Seq methods.

1 Introduction

Sentiment analysis, which involves identifying a speaker's sentiment, is an open research problem. In this field, the majority of work done focused on unimodal methodologies - primarily textual analysis - where investigating was limited to identifying usage of words in positive and negative scenarios. However, unimodal textual sentiment analysis through usage of words, phrases, and their interdependencies were found to be insufficient for extracting affective content from textual opinions (Rosas et al., 2013). [1] As a result, there has been a recent push towards using statistical methods to extract additional behavioral cues not present in the language modality from the video and audio modalities. This research field is known as multimodal sentiment analysis and it extends the conventional text-based definition of sentiment analysis to a multimodal setup where different modalities contribute to modeling the sentiment of the speaker. For example, (Kaushik et al., 2013) explores modalities such as audio, while (Wöllmer et al., 2013) explores a multimodal approach to predicting 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. As a result, several multimodal datasets, such as CMU-MOSI (Zadeh et al., 2016) and later CMU-MOSEI (Zadeh et al., 2018c), ICT-MMMO (Wöllmer et al., 2013) and YouTube (Morency et al., 2011), take advantage of the abundance of multimodal data on the Internet. At the same time, neural network based multimodal models have been proposed that are highly effective at learning multimodal representations for multimodal sentiment analysis (Chen et al., 2017; Poria et al., 2017; Zadeh et al., 2018a,b).

Recent progress has been limited to supervised learning using labeled data, and does not take advantage of the abundant unlabeled data on the Internet. To address this gap, our work is primarily one of unsupervised representation learning. We attempt to learn a multimodal representation of our data in a structured paradigm and explore whether a joint multimodal representation trained via unsupervised learning can improve the performance for multimodal sentiment analysis. While representation learning has been an area of rapid research

[1] *These authors contributed equally.

in the past years, there has been limited work that explores multimodal setting. To this end, we propose two methods: a *Seq2Seq Modality Translation Model* and a *Hierarchical Seq2Seq Modality Translation Model* for unsupervised learning of multimodal representations. Our results show that using multimodal representations learned from our Seq2Seq modality translation method outperforms the baselines and achieves improved performance on multimodal sentiment analysis.

2 Related Work

In the past, approaches to text-based emotion and sentiment recognition rely mainly on rule-based techniques, bag of words (BoW) modeling or SNoW architecture (Chaumartin, 2007) using a large sentiment or emotion lexicon (Mishne et al., 2005), or statistical approaches that assume the availability of a large dataset annotated with polarity or emotion labels.

Multimodal sentiment analysis has gained a lot of research interests over the last few years (Baltrusaitis et al., 2017). Probably the most challenging task in multimodal sentiment analysis is to find a joint representation of multiple modalities. This problem is has been approached in a number of ways. Earlier works such as (Ngiam et al., 2011; Lazaridou et al., 2015; Kiros et al., 2014) have pushed some progress towards this direction.

Recently, more advanced neural network models were proposed to learn multimodal representations. The Multi-View LSTM (MV-LSTM) (Rajagopalan et al., 2016) was suggested to exploit fusion and temporal relationships. MV-LSTM partitions memory cells and gates into multiple regions corresponding to different views. Tensor Fusion Network (Zadeh et al., 2017) presented an efficient method based on Cartesian-product to take into consideration intramodal and intermodal relations between video, audio and text of the reviews to create a novel feature representation for each utterance. The Gated Multimodal Embedding model (Chen et al., 2017) created an algorithm using reinforcement learning to train an on-off switch that decided what values the video and audio components would have. Noisy modalities are turned off and clean modalities are allowed to pass through. (Zadeh et al., 2018a) utilizes external multimodal memory mechanisms to store multimodal information and create multimodal representations through time. (Zadeh et al., 2018b) proposed using multiple attention coefficient assignments to represent multiple cross-modal interactions. However, all these methods discussed so far are purely supervised approaches to multimodal sentiment analysis and do not leverage the power of unsupervised data and generative approaches towards learning multimodal representations.

Besides supervised approaches, generative methods based on generative adversarial networks (GAN) (Goodfellow et al., 2014) have attracted significant interest in learning joint distribution between two or more modalities (Donahue et al., 2016; Li et al., 2017; Gan et al., 2017). Another method to deal with multimodal problems is to view them as conditional problems which learn to map a modality to the other (Mirza and Osindero, 2014; Kingma et al., 2014; Pandey and Dukkipati, 2017). Our work can be viewed as an extension of the conditional approach, as both utilize unsupervised learning. However, our work differs from those in that it takes into account the sequential dependency within each modality.

Finally, attention based layers have also proved themselves to be effective tools to boost performance of neural network models, such as in neural machine translation (Klein et al.; Bahdanau et al., 2014; Luong et al., 2015), speech recognition (Sriram et al., 2017) and in image captioning (Xu et al., 2015). Our work also employs this mechanism in an attempt to better handle long-term dependencies of variable-length sequences.

3 Problem Formulation

Given a dataset with data $X = (X^{text}, X^{audio}, X^{video})$ where X^{text}, X^{audio}, X^{video} stand for text, audio and video modality inputs, respectively. Typically a dataset is indexed by videos. This means that if we have n videos, then $X = (X_1, X_2, ..., X_n)$ where $X_i = ({X_i}^{text}, {X_i}^{audio}, {X_i}^{video})$, $1 \leq i \leq n$. The corresponding labels for these n videos are $Y = (Y_1, Y_2, ..., Y_n)$, $Y_i \in \mathbb{R}$.

To simplify the problem, we align the input based on words. Typically, researchers often segment each video into a smaller set in which each segmented video will last a couple of seconds, instead of minutes as done in (Chen et al., 2017). After such alignment and segmentation, we have the equal-length inputs of each modality per video. For example, at the i^{th} video, we have ${X_i}^{text} = ({w_i}^{(1)}, {w_i}^{(2)}, ..., {w_i}^{(T_i)})$ where ${w_i}^{(t)}$ stands for the

t^{th} word and T_i is the length of the i^{th} video's text input, *a.k.a* time steps. Note that different videos will have different time steps. Similarly for this video, we have a sequence of audio input ${X_i}^{audio} = ({a_i}^{(1)}, {a_i}^{(2)}, ..., {a_i}^{(T_i)})$ and video input ${X_i}^{video} = ({v_i}^{(1)}, {v_i}^{(2)}, ..., {v_i}^{(T_i)})$.

In this work we are tackling the input learning problem where we want to learn the embedding representation for all text, audio, and video modalities: $\widetilde{X_i} = f(X_i) = f(({X_i}^{text}, {X_i}^{audio}, {X_i}^{video}))$. In our baseline model, the function f is simply the concatenation at time step level: $\widetilde{x_i}^t = [{w_i}^t; {a_i}^t; {v_i}^t]$

In our proposed method, we learn $\widetilde{X_i}$ by using a Seq2Seq model. We do not calculate each embedding representation for each time step, but for the whole sequence. Formally, $\widetilde{X_i} = f(X_i) = Seq2Seq\_Encoder(X_i)$ where $Seq2Seq\_Encoder$ is the encoder part of our Seq2Seq model.

Now, we have the transformed inputs $\widetilde{X} = (\widetilde{X_1}, \widetilde{X_2}, ..., \widetilde{X_n})$ and outputs $Y = (Y_1, Y_2, ..., Y_n)$ for n videos, where $\widetilde{X_i} = (\widetilde{x_i}^1, \widetilde{x_i}^2, ..., \widetilde{x_i}^{T_i})$. For simplicity, in the next formula, we omit the index of video segment i, and so the input becomes $\widetilde{X} = (\widetilde{x}^1, \widetilde{x}^2, ..., \widetilde{x}^T)$, and the labels become $Y = (y^1, y^2, ..., y^T)$.

We will be using a Recurrent Neural Network (RNN) such as LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Chung et al., 2015) to model this sequence. In detail, this RNN has a stack of K hidden layers $h = (h^1, h^2, ..., h^K)$, each contains D hidden neurons: $h^k = (h_1^k, h_2^k, ..., h_D^k)$, $k \in [1, K]$. We denote W and b to be weight and bias, then for the first layer which contacts directly with input:

$$ {h^1}_t = H(W_{xh^1}\widetilde{x_t} + W_{h^1h^1}{h^1}_{t-1} + b_{h^1}) \quad (1) $$

where H is the RNN cell function. For example of LSTM, it contains *input, forget, output* and *cell state*. At hidden layer $k \in [2, K]$:

$$ {h^k}_t = H(W_{h^{t-1}h^t}h_t^{k-1} + W_{h^kh^k}{h^k}_{t-1} + b_{h^k}) \quad (2) $$

Optionally, we apply a soft attention mechanism *on top* of the last hidden layer h^K, with shared weight W_α over T time steps, then we can obtain the attention output α:

$$ \alpha = softmax\left(\begin{bmatrix} W_\alpha h_1^K \\ W_\alpha h_2^K \\ ... \\ W_\alpha h_T^K \end{bmatrix}\right) \quad (3) $$

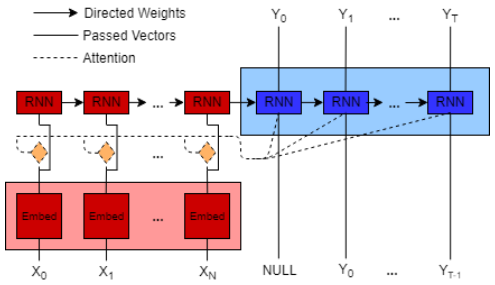

Figure 1: Seq2Seq Modality Translation Model with input $(X_1, ..., X_N)$ and output is $(Y_1, ..., Y_T)$. Seq2Seq makes use of the whole input sequence in the decoding phase for every token Y_i. If attention model (yellow color) is used, for each Y_i, it learns a separate weight vector *w.r.t* each token of input X to see which token should the decoder "attend" more.

The last hidden layer's output now becomes:

$$ A = [h_1^K, h_2^K, ..., h_T^K]\alpha = H^K\alpha \quad (4) $$

And the last output layer with regression score is:

$$ \widetilde{y_t} = W_{Ay}A + b_y \quad (5) $$

Finally, we calculate the loss with respect to the labels. As in (Chen et al., 2017), we choose Mean Absolute Error (MAE) as our loss and later train with stochastic gradient descent:

$$ \mathbb{L}_{MAE}(\widetilde{Y}, Y) = \mathbb{E}[|\widetilde{Y} - Y|] \quad (6) $$

4 Proposed Approach

In this section we describe the different approaches that we plan to take to improve affect recognition through learning multimodal representations.

4.1 Seq2Seq Modality Translation Model

The *Seq2Seq Modality Translation Model* aims to learn multimodal representations that can be used for discriminative tasks. While Seq2Seq models have been predominantly used for machine translation (Bahdanau et al., 2014; Luong et al., 2015), we extend its usage to the realm of multimodal machine learning where we use it to translate one modality to another, or translate a joint representation to another single or joint representation. To do so, we propose a Seq2Seq modality translation model with attention mechanism, as shown in Figure 1. Modality X is translated into modality Y. Our hypothesis is that the intermediate representation of this model, i.e. the output of Seq2Seq's encoder, or the input of its decoder, is close to the joint representation (X, Y) of the two modalities

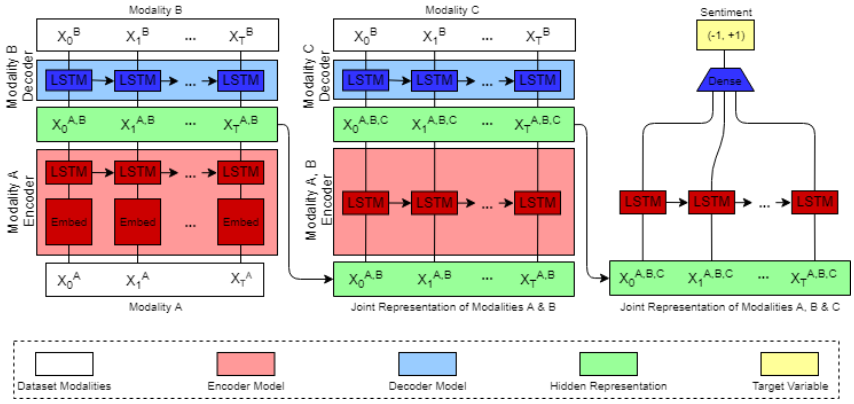

Figure 2: Hierarchical Seq2Seq Modality Translation Model: first we train with 2 modalities, then we add one more on the second phase, from which the results will be fed into RNN for sentiment prediction. The green boxes denote the joint representation learned by Seq2Seq models: the joint representation of modalities A and B will be fed into another Seq2Seq model which in turn learns the joint representation of AB and another modality C. Finally the joint representation of ABC will be fed into a RNN to predict sentiment.

involved. As a result, this representation can be used for tasks that involve learning joint representation across multiple modalities. The detail is in Algorithm 1.

Algorithm 1 Seq2Seq Modality Translation

X, Y, S are 2 modalities and sentiment sequences

Phase 1: Train Seq2Seq
  $\mathcal{E}_{XY} \leftarrow Seq2Seq\_RNN\_Encode(X)$
  $\widetilde{Y} \leftarrow Seq2Seq\_RNN\_Decode(\mathcal{E}_{XY})$
  $loss = cross\_entropy(\widetilde{Y}, Y)$
  Backprop to update params

Phase 2: Sentiment Regression
  $\mathcal{E}_{XY} \leftarrow Seq2Seq\_RNN\_Encode(X)$ ▷ trained encoder in Seq2Seq model
  $R = RNN(\mathcal{E}_{XY})$
  $score \leftarrow Regression(R)$
  $loss \leftarrow MAE(score, S)$
  Backprop to update params

Formally, the Seq2Seq Modality Translation Model consists of 2 separate steps: encoding and decoding, each phase typically consists of a single RNN or a stack of them. This model accepts variable-length inputs of X and Y, and the network should be trained to maximize the translational condition probability $p(Y|X)$. For encoding, it encodes the whole input sequence X into an embedded representation. The hidden state output of each time step is based on the previous hidden state along with the input sequence (refer to Figure 1):

$$h_n = RNN(h_{n-1}, X_n) \tag{7}$$

The encoder's output is the final hidden state's output of the encoding RNN:

$$\mathcal{E} = h_N = RNN(h_{N-1}, X_N) \tag{8}$$

where N is the length of the input sequence X. The decoder tries to decode each token Y_i at a time based on $\mathcal{E}$ and all previous decoded tokens, which is formulated as:

$$p(Y) = \prod_{i=1}^{T} p(Y_i|\mathcal{E}, Y_1, ..., Y_{i-1}) \tag{9}$$

The Seq2Seq training target is to find the best translation sequence which is as close to the ground truth Y as possible, or formally:

$$\widehat{Y} = \arg\max_{Y} p(Y|X) \tag{10}$$

And while there are some other search algorithms such as random sampling or greedy search to decode each token (Neubig, 2017), we use the traditional beam search approach (Sutskever et al., 2014).

4.2 Hierarchical Seq2Seq Modality Translation Model

The Seq2Seq Modality Translation Model only learns joint representation between 2 modalities X and Y. While this might be a strong starting point, we believe an approach that captures the joint interactions between all different modalities X, Y, Z is more effective in modeling the full distribution of the multimodal data and therefore more useful for regression or classification. In response, we propose the *Hierarchical Seq2Seq Modality Translation Model* that learns a joint multimodal representation. Once the Seq2Seq Modality Translation Model is trained for 2 modalities X and Y, we obtain the intermediate representation $\mathcal{E}_{XY}$ which is the joint representation of (X, Y). $\mathcal{E}_{XY}$

is in turn treated as input sequence for the next Seq2Seq Modality Translation Model to decode the third modality Z. The final multimodal representation $\mathcal{E}_{XYZ}$ represents the joint representation of (X, Y, Z). The Hierarchical Seq2Seq Modality Translation Model is described as in Algorithm 2.

Algorithm 2 Hierarchical Seq2Seq Modality Translation: X, Y, Z, S are 3 modalities and sentiment sequences

Phase 1: Train Seq2Seq for 2 modalities
  $\mathcal{E}_{XY} \leftarrow Seq2Seq\_RNN\_Encode(X)$
  $\widetilde{Y} \leftarrow Seq2Seq\_RNN\_Decode(\mathcal{E}_{XY})$
  $loss = cross\_entropy(\widetilde{Y}, Y)$
  Backpropagate to update parameters

Phase 2: Train Seq2Seq for 3 modalities
  $\mathcal{E}_{XYZ} \leftarrow Seq2Seq\_RNN\_Encode(\mathcal{E}_{XY})$
  $\widetilde{Z} \leftarrow Seq2Seq\_RNN\_Decode(\mathcal{E}_{XYZ})$
  $loss = cross\_entropy(\widetilde{Z}, Z)$
  Backpropagate to update parameters

Phase 3: Sentiment Regression
  $\mathcal{E}_{XYZ} \leftarrow Seq2Seq\_RNN\_Encode(\mathcal{E}_{XY})$
  $R = RNN(\mathcal{E}_{XYZ})$
  $score \leftarrow Regression(R)$
  $loss \leftarrow MAE(score, S)$
  Backpropagate to update parameters

This strategy is also illustrated in Figure 2. The output of the second Seq2Seq model is the input of the last RNN model where we will train to predict regression sentiment scores. This last Seq2Seq model will be trained using MAE loss function and it perform subsequent regression process.

5 Experimental Setup

We explored the applications of this model to the CMU-MOSI dataset (Zadeh et al., 2016). We implemented a baseline LSTM model based off the work done in (Chen et al., 2017). Our implementation uses 66.67% of the data for training from which we take a 15.15% held-out set for validation, and the remaining 33.33% is used for testing. Finally, we evaluated our proposed model against the baseline results generated by the implementation of (Chen et al., 2017). Here we compared our results against the various multimodal configurations evaluating our performance using precision, recall, and F1 scores.

5.1 Dataset and Input Modalities

The dataset that we use to explore applications of our model is the CMU Multimodal Opinion-level Sentiment Intensity dataset (CMU-MOSI). The dataset contains video, audio, and transcriptions of 89 different speakers in 93 different videos divided into 2199 separate opinion sentiments. Each video has an associated sentiment label in the range from -3 to 3. The low end of the spectrum (-3) indicates strongly negative sentiment, where as the high end of the spectrum indicates strongly positive sentiment (+3), and ratings of 0 indicate neutral sentiment. The CMU-MOSI dataset is currently subject to much research (Poria et al., 2017; Chen et al., 2017; Zadeh et al., 2018a,b) and the current state of the art is achieved by (Poria et al., 2017) with an F1 score of 80.3 using a context aware model across entire videos. The state of the art using only individual segments is achieved by (Zadeh et al., 2018a) with an F1 score of 77.3.

With respect to raw features that are being given as inputs to our model, we perform feature extraction in the same manner as described in (Chen et al., 2017). In the text domain, pretrained 300 dimensional GLoVe embeddings (Pennington et al., 2014) were used to represent the textual tokens. In the audio domain, low level acoustic features including 12 Mel-frequency cepstral coefficients (MFCCs), 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 et al., 1997, 2002), peak slope parameters and maxima dispersion quotients (Kane and Gobl, 2013) were extracted automatically using COVAREP (Degottex et al., 2014). Finally, in the video domain, Facet (iMotions, 2017) is used to extract per-frame basic and advanced emotions and facial action units as indicators of facial muscle movement (Ekman, 1992; Ekman et al., 1980).

In situations where the same time alignment between different modalities are required, we choose the granularity of the input to be at the level of words. The words are aligned with audio using P2FA (Yuan and Liberman, 2008) to get their exact utterance times. The visual and acoustic modalities are aligned to words using these utterance times.

5.2 Baselines

We use a LSTM model implemented in 3 different ways (one for each different grouping of the modalities). First in the unimodal domain, we run sentiment regression based solely on one modality, second in the bimodal domain we change the input to the concatenation of any pair of modality, and

Method	Feature	BINARY (−1, +1)			7-CLASS (−3, ..., +3)		
		Prec	Recall	F1	Prec	Recall	F1
UniModal-Baseline	Text (T)	**0.77**	**0.76**	**0.76**	**0.32**	**0.35**	**0.33**
	Audio (A)	0.56	0.56	0.56	0.12	0.19	0.14
	Video (V)	0.57	0.47	0.48	0.12	0.19	0.12

Table 1: Unimodal baseline results with 3 metrics: Precision, Recall and F-Score (F1)

finally in the trimodal domain we concatenate all three modalities. This baseline not only serves to act as a benchmark for comparing our results but also acts as a starting point for our code development. As such, any improvements in our metrics are strictly as a result of the representations that we have learned and not structural changes in our model.

5.3 Multimodal Model Variations

Throughout our experimentation, we apply the algorithms in Section 4 with several intuitive variations of how to translate modalities. Below are all approaches that we try to maximize our chances of learning a strong representation.

For bimodal, we translate one modality into another one. For example, A → V stands for translating from Audio to Video, and take the embedding state, which we refer to as `embed`(A+V), to predict sentiment. Here we employ the Seq2Seq Modality Translation Model mentioned in Algorithm 1.

For trimodal, there are a lot more variations as follows. First, since we have 3 different modality and Seq2Seq is only capable of translating one modality to another, we use the Hierarchical Seq2Seq Modality Translation Model which is mentioned in Algorithm 2, e.g. we translate from `T` to `A` to have the joint representation `embed`(T+A), and then continue the translation from `embed`(T+A) to the rest modality which is `V`, which in turn yields the joint representation `embed`(T+A+V) to make sentiment prediction.

Second, we reuse the previous Seq2Seq Modality Translation Model to translate a concatenation of 2 modality to the rest, e.g. `concat`(T+V) to A, and vice versa, e.g. translating from A back to `concat`(T+V).

Finally, we still use the Seq2Seq Modality Translation Model to translate from a concatenation of 2 modality to another concatenation of other 2. With this setting, at least one modality is repeated, and base on many previous works and our experience, we tend to favor text modality (T) over the other two and make it repeated.

6 Results

6.1 Baseline Unimodal Results

We see that with the baseline model, as shown Table 1, the text modality is by far the most discriminative when it comes to detecting emotion. This implies that users rely heavily on their word choice and language to convey meaning and emotion. While this may be true, we know that other works such as (Zadeh et al., 2018a; Poria et al., 2017) have achieved higher scores by combining all these different modalities. This implies that with some careful thinking and pointed model construction, we should be able to improve upon our baseline unimodal results through the integration of additional modalities into our model.

6.2 Baseline Multimodal Results

The results of our different baseline multimodal approaches is shown in Table 2 for bimodal and Table 3 for trimodal. We see that of the multimodal baselines the model which combines the 3 modalities of text, speech, and video performed the best. The baseline model which combined text and audio arrived in second place followed closely by the combined text and video model. The model which combines video and audio arrived in last place by a significant margin. This corroborates our results from our unimodal baselines which implied that the text modality is the most discriminative modality in this dataset.

On the whole we can see that when all three modalities are working in concert we get the best result in a multimodal context, however, it is worth noting that we were not able to match out unimodal baseline with our multimodal models. This implies that there is still more to be drawn from our data when constructing our model and there is generally more work to be done. We believe that incorporating a stronger more robust representation of our data will be beneficial to our later attempts at classification. Though we view this to be out of scope of this work as the focus of this work is on learning informative representations.

Method	Feature	BINARY (−1, +1)			7-CLASS (−3, ..., +3)		
		Prec	Recall	F1	Prec	Recall	F1
BiModal-Baseline	concat(T + V)	**0.78**	**0.67**	0.55	0.01	0.16	0.05
	concat(T + A)	0.44	0.66	0.53	0.02	0.15	0.04
	concat(A + V)	0.55	0.47	0.48	0.13	0.16	0.11
BiModal-Seq2Seq	T → V	0.67	**0.67**	**0.67**	0.26	0.22	**0.22**
	T → A	0.66	0.64	0.65	**0.28**	0.24	0.18
	A → T	0.55	0.60	0.56	0.17	**0.34**	0.11
	A → V	0.55	0.55	0.54	0.16	0.18	0.16
	V → T	0.58	0.58	0.58	0.05	0.16	0.08
	V → A	0.58	0.62	0.58	0.12	0.17	0.01

Table 2: Bimodal results with 3 metrics: Precision, Recall and F-Score (F1)

Method	Feature	BINARY (−1, +1)			7-CLASS (−3, ..., +3)		
		Prec	Recall	F1	Prec	Recall	F1
TriModal-Baseline	concat(T + V + A)	**0.75**	**0.75**	**0.75**	0.24	**0.27**	**0.24**
TriModal-Seq2Seq	embed(T, V) → A	0.56	0.60	0.57	0.10	0.16	0.09
	embed(T, A) → V	0.60	0.55	0.56	0.26	0.15	0.07
	embed(A, V) → T	0.66	0.53	0.44	0.16	0.04	0.09
	embed(A, T) → V	0.59	0.51	0.52	0.13	0.15	0.09
	embed(V, T) → A	0.59	0.60	0.59	0.11	0.17	0.10
	embed(V, A) → T	0.57	0.61	0.58	0.11	0.17	0.09
	concat(T, V) → A	0.67	0.66	0.65	0.22	0.17	0.18
	concat(A, T) → V	0.54	0.55	0.63	0.19	0.15	0.21
	concat(V, A) → T	0.59	0.59	0.58	0.16	0.12	0.12
	T → concat(A, V)	0.70	0.65	0.66	0.23	0.22	0.18
	A → concat(T, V)	0.55	0.53	0.54	0.18	0.20	0.18
	concat(T, A) → concat(T, V)	0.62	0.60	0.61	0.23	0.24	0.22
	concat(T, V) → concat(T, A)	0.68	0.70	0.67	**0.31**	0.24	0.19

Table 3: Trimodal results with 3 metrics: Precision, Recall and F-Score (F1)

6.3 Analysis of Baseline Failure Cases

The common trend that we see among all of those baseline models is the consistent failure to identify extreme cases of either positive or negative emotions. We believe that this phenomenon is due to two possibilities. First we see that there are very few cases of highly positive (+3) and highly negative (−3) examples in the training data. As a result the models that are trained are highly biased towards not selecting +3 or −3 ratings. Secondly, our baseline models are performing categorical classification as opposed to regression or ordinal classification. We plan to solve by training the model to perform this type of prediction as a regression task as opposed to a categorical classification task.

6.4 Bimodal Seq2Seq Results

Our bimodal models require the exploration of two modalities, one for the encoding step and another for the decoding step. We explored several different different encoder/decoder frameworks for these models. The first model that we explored were representations generated from encoding exactly one modality and then decoding exactly one different modality. The results of this approach are included below in Table 2. Here we can see that the Seq2Seq Modality Translation Model outperforms the baseline method in terms of F1 consistently and outperforms in terms of precision and recall in several cases, but not all.

6.5 Trimodal Seq2Seq Results

We try all variations mentioned in Section 5.3 and the full breakdown of these results can be found in the Table 3. According to that, while the Hierarchical Seq2Seq Modality Translation Model is a natural extension to the normal Seq2Seq Modality Translation model, it does not perform well on the CMU-MOSI dataset. Otherwise, using the normal non-hierarchical model with concatenation variations does improve the performance, and particularly beats the baseline (for only F1 score) on the model which translates from concat(T,V) to concat(T+A) for the 7-class case. As mentioned in Section 5.3, we favor the text (T) modality and make it repeated in this setting because it typically contributes more significantly to sentiment prediction. Indeed, we have tried to repeat video or audio

modality but the result shrinks dramatically.

One possible reason for this behavior is the scarcity of training data. Given that at every phase of Seq2Seq translation, we only have 1289 train samples, 230 validation and 269 test samples, Seq2Seq, which typically requires more data for training a good model, does not work efficiently. This affects even more in the hierarchical Seq2Seq cases where we train two phases of Seq2Seq. We project the performance will improve if we work on other dataset which is bigger, or if we pretrain our model on other dataset first before applying it to MOSI.

7 Discussion

The language modality is the most discriminative as well as the most important towards learning multimodal representations. While we outperform the baseline multimodal approach we were unable to outperform the baseline unimodal text approach. Clearly from these results we know that that the text modality is the most discriminative of all of these modalities. However, it appears that these models which we have described are not able to truly separate the importance of the text modality. The fact that we are merging these modalities into a shared representation space is likely decreasing the resolution of the text domain and thus decreasing the modeling power of the domain. This is why we believe that the top performing multimodal model is one that incorporates the text domain so much (see Tables 2 and 3).

It is worth noting that some of the learned representations were quite poor when it came to their use in prediction. For example, representations that were learned using only audio and video generally performed poorly. This is to be expected given the already known information that these modalities are not as discriminative as the language modality. At the same time, some of the worse performing representations were learned in the methodology of learning a representation based on an existing embedding. We believe this to be due to the representation losing the resolution of the original two domains from which the original source embedding was learned and instead being focused on learning the best representation to predict the final modality.

8 Future Directions

This research opens up a promising direction in joint unsupervised learning of multimodal representations and supervised learning of multimodal temporal data. We propose the following extensions that could improve performance:

Firstly, using an Variational Autoencoder (VAE) (Kingma and Welling, 2013) in conjunction with LSTM Encoder/Decoder model (as in the case of VAE Seq2Seq model) would be an interesting avenue to explore. This is because VAEs have been shown to learn better representations as compared to vanilla autoencoders (Kingma and Welling, 2013; Pu et al., 2016).

Secondly, since our method for multimodal representation learning is unsupervised, we could take advantage of larger external datasets to pre-train the multimodal representations before fine-tuning further with CMU-MOSI. We believe this will boost performance because we have limited data in CMU-MOSI for training (CMU-MOSI has 2199 training segments). Some datasets that come to mind include the Persuasion Opinion Multimodal (POM) dataset (Park et al., 2014) with 1000 total videos (longer than segments) and the IEMOCAP dataset with 10000 total segment. Since these datasets also consist of monologue speaker videos, we expect the learnt multimodal representations to generalize.

Thirdly, our method does not train our combined model end to end: the representations that we use to generated during on training run and the sentiment classification model are trained separately. Exploring an end-to-end version of this model end to end could possibly result in better performance where we could additionally fine tune the learned multimodal representation for sentiment analysis.

9 Conclusion

To conclude, this paper investigate the problem of multimodal representation learning to leverage the abundance of unlabeled multimedia data available on the internet. We presente two methods for unsupervised learning of joint multimodal representations using multimodal Seq2Seq models: the *Seq2Seq Modality Translation Model* and the *Hierarchical Seq2Seq Modality Translation Model*. We found that these intermediate multimodal representations can then be used for multimodal downstream tasks. Our experiments indicate that the multimodal representations learned from our Seq2Seq modality translation method are highly informative and achieves improved performance on multimodal sentiment analysis.

10 Acknowledgements

The authors are thankful to the many student peers who commented on and critiqued this work. Specific thanks to Louis-Phillipe Morency and Amir Zadeh for their helpful discussions and thoughtful critiques. We are grateful to our peers who helped us evaluate our methodology, in particular Stephen Tsou and Kshitij Khode. Finally, we also thank the anonymous reviewers for helpful and constructive feedback.

References

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

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

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

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

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

François-Régis Chaumartin. 2007. Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, Stroudsburg, PA, USA, SemEval '07, pages 422–425. http://dl.acm.org/citation.cfm?id=1621474.1621568.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrusaitis, Amir Zadeh, and Morency Louis-Phillippe. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. *ICMI, Glassgow, United Kingdom* .

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

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*. pages 2067–2075.

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

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

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

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

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6(3-4):169–200.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology* 39(6):1125.

Zhe Gan, Liqun Chen, Weiyao Wang, Yunchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. 2017. Triangle generative adversarial networks. *arXiv preprint arXiv:1709.06548* .

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. pages 2672–2680.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

iMotions. 2017. Facial expression analysis. goo.gl/1rh1JN.

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

Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. 2013. Sentiment extraction from natural audio streams. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, pages 8485–8489.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*. pages 3581–3589.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

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

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. ???? OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints* .

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

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

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

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

Gilad Mishne et al. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*. volume 19, pages 321–327.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, pages 169–176.

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

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. pages 689–696.

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

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, New York, NY, USA, ICMI '14, pages 50–57. https://doi.org/10.1145/2663204.2663260.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 873–883.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., pages 2352–2360.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. *Extending Long Short-Term Memory for Multi-View Structured Learning*, Springer International Publishing, Cham, pages 338–353.

Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems* 28(3):38–45.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2017. Cold fusion: Training seq2seq models together with language models. *arXiv preprint arXiv:1708.06426* .

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28(3):46–53.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. pages 2048–2057.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America* 123(5):3878.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 1114–1125.

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

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

Amir Zadeh, Paul Pu Liang, Jon Vanbriesen, Soujanya Poria, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. 2018c. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Association for Computational Linguistics (ACL)*.

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