USC-SIPI Report #447

Human Behavior Understanding from Language through Unsupervised Modeling

By
Shao-Yen Tseng

December 2020

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
USC Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

Human Behavior Understanding from Language through Unsupervised Modeling

by

Shao-Yen Tseng

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

December 2020

To my wife and my parents

Acknowledgements

The completion of this dissertation would not have been possible without the guidance and support of my advisers, colleagues, friends, and family.

Firstly, I would like to thank my advisers Dr. Shrikanth Narayanan and Dr. Panayiotis Georgiou for their tremendous support throughout my journey. They have both offered great guidance in academic research but more importantly built a mirthful environment for my colleagues and I to work under. I thank the members of my qualifying and defense committee, Dr. Jonathan Gratch, Dr. C.-C. Kuo, and Dr. Gayla Margolin, for their valuable insights and suggestions. I would also like to thank Dr. Brian Baucom for his contributions of indispensable knowledge in the field of Psychology as well as overall enthusiastic support.

I would like to thank my colleagues at the Signal Analysis and Interpretation Laboratory (SAIL) and the Signal Processing for Communication Understanding and Behavior Analysis Laboratory (SCUBA) for not just their support in research but also for the friendships and fun times we had together.

Finally, I give thanks to my family. To my parents for their love and support throughout my life. And to my beloved wife, Juiyu Lin, for supporting me through the thick and thin of this journey.

Table of Contents

List of Tables

List of Figures

Abstract

The recognition of behavioral cues in human interactions is an integral task towards the mapping of psychological processes in the speaker. The understanding of these processes form a basis for provision of corrective counseling in many psychotherapy applications. It may also serve as a point of research in many observational studies in psychology, with objectives such as improvement of patient care-giving or therapist proficiency. An assistive method for the above applications in the form of automated behavior annotation is an attractive goal, however identifying behavioral cues in organic conversational interactions is a challenging task even for humans and is the culmination of decades of social experience and personal inflection. While many tasks can be adequately handled by applying deep learning on "big data", in the domain of behavior understanding data is often scarce and application-specific, leading to poor performance in automated systems. In this thesis I aim to overcome some challenges in development of human behavior recognition systems. I address issues commonly seen in the behavioral domain, such as noisy weak labels and data scarcity, by proposing various unsupervised representation learning techniques and neural networks for human behavior recognition.

In the first part of this thesis I propose novel architectures for identifying and rating human behaviors from lexical transcripts of extended conversations. I tackle the issue of training deep models with limited in-domain data and weak labels by leveraging out-of-domain word embeddings in a recurrent neural network framework. In the second part of this thesis I investigate how concepts of behavior can be transferred between different behavioral applications through methods such as contextual learning, online mutlitask learning, and multimodal approaches.

I show that through the proposed learning methods, models can become increasingly adept at rating behaviors in multiple scenarios. Experiments in this thesis are conducted on annotating

behaviors in couples counseling sessions. I also demonstrate the applicability of these techniques to the task of emotion recognition, which is a subset of human behaviors requiring shorter time-frames.

Chapter 1

Introduction

1.1 Background

1.1.1 Understanding human behavior

The evaluation of mental health is heavily based on the understanding of behaviors exhibited in human communication. In couples therapy, for example, behaviors in couple interactions are identified and annotated across numerous dimensions, such as *negativity*, *blame*, or *acceptance*. Counseling therapists may then identify reasons of conflict based on these assessments and encourage mediatory action to improve couple relationships [1]. The understanding of human behavior is also prominently applied in many observational studies in psychology. For example in studies of spousal communication and its effect on care-giving in cancer patients [2]. Or studies in identifying new cognitive markers of risk for suicide among veterans [3].

Behavior understanding is the complex task of recognizing behavioral cues in human interactions and encodes many layers of complexity: the dynamics of the interlocutors, their perception, appraisal, and expression of emotion, their thinking and problem solving intents, skills and creativity, the context and knowledge of interlocutors, and their abilities towards *emotion* regulation [4]. Behavior is not the same as emotions, but it is encoded in part through the modulation of emotional expression and affected by the perception and actions of those emotions, and thus shares a tight relationship with emotional expression.

1.1.2 Challenges

Annotation of human behaviors is usually performed by trained human coders and is an expensive and time-consuming process. As such it only takes place in limited cases for research purposes. Human annotators first have to be trained in accordance with detailed coding manuals [5], [6] to provide accurate and consistent ratings without the influence of personal bias. Trained annotators may then be evaluated to select those with the highest agreements for the final annotation task. The overall process is lengthy and strainful, but even so, agreement in human annotations can still be quite low [6]. This is possibly caused by inherent cognitive biases of annotators which affect the degree to which behavioral intensities are scaled.

Human behavior also manifests over longer time frames than emotions and requires a larger context window to be identified correctly. In addition, many different dimensions of behavior have only subtle differences between them which makes tracking the presence of a specific behavior over a long time range more difficult.

One final critical challenge is that behavioral expressions vary between individuals in general. This difference can stem from the uniqueness of each person in terms of physical and mental states, health status, current mood, and personality traits, *etc*. Additional variability might arise from associated cognitive loads from the task scenario, specifics in the study environment, as well as behaviors of other individuals. Nonetheless for humans, while tedious, it is possible to identify common behavioral expressions given enough contextual observations.

1.2 Dissertation Overview

The main theme of this thesis is the exploration of computational models suited for behavior understanding from spoken conversations and methods in unsupervised representation learning that can be used to improve their effectiveness.

1.2.1 Behavior annotation using recurrent neural networks

Recurrent neural networks (RNN) have demonstrated incredible abilities in handling long range dependencies in data with longer timeframes [7]. In Natural Language Processing (NLP), LSTM units have produced state-of-the-art results in many tasks [8], [9]. However, the use of RNNs in behavior estimation has seen limited success due to limitations in available data. Firstly, due to privacy restrictions, data with rich information of behavior in psychotherapy sessions is often severely limited in quantity. This greatly restricts the ability to train deep models with good convergence and generalizable performance. Secondly, due to the additional effort required for fine-grained annotation, labels are restricted to covering entire sessions with no indication of ground truth for shorter duration segments. This introduces further challenges in training deep models due to the large observational window required for a single prediction.

I propose a neural network framework for identifying and rating human behaviors from lexical transcripts of extended conversations. The neural network composes of recurrent layers to handle sequences of words obtained from a sliding window. I address the issue of training deep models with limited in-domain data and weak labels by leveraging out-of-domain word embeddings in the recurrent neural network framework. The network is trained with weak labels by directly assigning the session-level rating as targets for all frames contained in respective sessions. A fusion layer is then applied over frame-level scores in a session to map back to a predicted behavior rating. The proposed framework is evaluated by annotating behaviors observed in couples therapy sessions [10].

1.2.2 Unsupervised domain transfer in sentence embeddings

Representation learning is a crucial method for obtaining superior results in many machine learning tasks [11]. In the scope of natural language processing notable examples of transforming input into highly informative abstractions are *word embeddings* such as *word2vec* [12] or *GloVe* [13]. The use of neural networks in our proposed behavior annotation framework allows us to leverage

these embeddings even further through joint optimization. However, as current models are extended to include more context for better annotation, better embeddings for longer windows such as sentences are required.

In the second part of this thesis I investigate how concepts such as conversational content or behavior information can be encoded into meaningful representations from sentences. I present an unsupervised deep learning method for deriving vector representations from sentences that are well-suited for behavior annotation. To train the sentence encoder I leverage out-of-domain conversational data by sourcing from online databases of movie dialogues. I show that conversational data is a rich source to learn meaningful embeddings and is suited for use in behavior annotation due to its relatedness of structure to in-domain behavioral datasets [14].

Following this I then propose the use of an online multitask objective to expand sentence embeddings with domain knowledge. I connect sentiment to behavioral expression in text as a fundamental basis and aim to predict sentiment labels as an additional training objective. I show through experiments that such embeddings are a viable and potent feature for multiple tasks of behavior and emotion recognition [15].

1.2.3 Multimodal and multitask approaches

Word embeddings such as ELMo and BERT have recently been shown to model word usage in language with greater efficacy through contextualized learning on large-scale language corpora, resulting in significant performance improvement across many natural language processing tasks. In last part of this thesis I integrate paralinguistic information into contextualized lexical embeddings through the addition of acoustic features to a bidirectional language model followed by a more recent Transformer architecture. The multimodal models are trained on spoken language data that includes both text and audio modalities. I then show that such models pretrained on unsupervised language modeling tasks can provide embeddings which combine paralinguistic cues with lexical content which improve performance in emotion recognition.

Chapter 2

Behavior Annotation Using Recurrent Neural Networks

2.1 Introduction

In this chapter we address the issues relating to training recurrent neural networks on long speech conversations with limited data and weak labels. We propose an LSTM-RNN system for capturing behavior trajectories in couples interactions in a such an environment. To allow for training of the RNN with limited data we use pretrained word representations learned from out-of-domain corpora and joint optimization. We also show the viability of using session-level labels for learning frame-level behavior. Using a fusion of the frame-level behavior trajectories we show that the ratings predicted by our proposed system achieve inter-annotator agreement comparable to those of trained human annotators.

2.2 Behavior Modeling

2.2.1 Maximum likelihood model

In previous works [16], [17], a *Maximum Likelihood* (ML) model with n-gram statistical language models of the interlocutor's language was implemented for behavior recognition. This model assumed that all the utterances observed in a particular session have been generated from the same behavioral state. While *n*-gram language models provide a compact approximation of the joint

probability of *n*-length word sequences, they have limitations. First, the framework suffers when presented with *Out-Of-Vocabulary* (OOV) test data. Secondly ML models are inflexible to variable length *n*-grams based on data availability (backoff helps but doesn't solve the problem) and this reduces robustness when longer context is introduced. Finally and very importantly, ML models are applicable for classification tasks but not estimation of continuous rating values.

2.2.2 Behavior modeling with LSTM

Recurrent neural networks have become increasingly popular for sequence learning tasks as they are adept at integrating temporal information from the entire sequence history as opposed to a fixed window of data in feed-forward neural networks. This dynamic context is especially valuable in natural language processing where semantic meaning may have long-term dependencies across any number of words. RNNs have been shown to perform better than statistical language models in such data-sparse situations by learning distributed representations for words [18], [19]. However the training of RNNs generally requires large amounts of data with accurate labels; something generally not available in our domain. Therefore, we propose the use of pretrained distributed representations of words from out-of-domain large corpora to alleviate the problem of data sparsity. In addition we train the RNN using a weakly supervised method to account for the missing frame-level labels. The details of our proposed RNN system are described in Section 2.4.

2.3 Data and Associated Challenges

For our experiments, we use the corpus of 134 couples from the UCLA/UW Couple Therapy Research Project [20]. The dataset contains audio and video recordings, along with transcripts, of real couples with marital issues interacting. In each session, the couples discuss a specific topic (e.g. "why can't you leave my stuff alone ?") chosen in turn for around 10 minutes. The behaviors of each speaker are rated by multiple annotators based on the Couples Interaction [21] and Social Support [22] Rating Systems. This results in 33 behavioral codes such as "Acceptance", "Blame",

and “Positivity”. Each annotator provides session-level subjective ratings for these codes on a Likert scale of 1-9, where 1 indicates absence of the behavior and 9 implies a strong presence. The sessions are rated by 2-12 annotators with majority of the sessions ($\sim$ 90%) rated by 3-4 annotators. Finally, these ratings are averaged to obtain a 33-dimensional vector of session level behavior ratings per interlocutor per interaction.

In this thesis, we focus primarily on the “Negativity” behavioral code. As was also done in our earlier work [16], [17], [23] we only consider sessions with mean annotator ratings in the top 20% (‘High Negativity’) and bottom 20% (‘Low Negativity’) of the code range for the sessions with good audio quality. This is less than 25% of the whole data set. A more comprehensive description of this corpus is reported in [21], [22], [24].

2.3.1 Associated challenges

Since human raters do not provide behavioral ratings for each utterance in the session we instead use the global rating as training labels for the individual sequences. In other words, all word sequences within a session are trained with the same label as the global rating. This method assumes that sequences of words from a session are related to global rating in a non-linear, complex manner. This is depicted in Figure 2.1 where the session-level label ρ is assumed to be a proxy for the label of the i-th frame ρ_i'. This also infers that the longer our sliding context-window the less the mismatch between the global rating ρ and ρ_i'. Ideally one would like the whole session to be passed as a training sample, however this would drastically decrease our training set and make training difficult. Nevertheless, a larger window can help identify lexical combinations that contribute towards the expression, and consequently estimation, of specific behaviors.

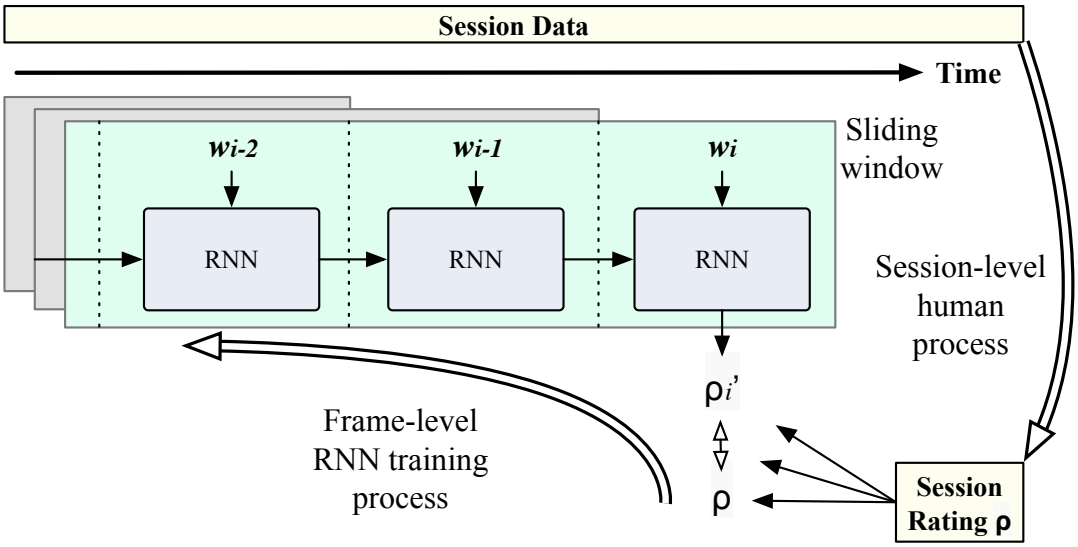


Figure 2.1: Training frame-level RNN using global rating values

2.4 Methodology

2.4.1 Proposed architecture

We encode the input as a one-hot vector **w**, where the n-th word in our dictionary is represented by setting the n-th element in **w** to 1 and all other elements to 0. We assume a vocabulary of N unique words and $0 \leq n \leq N$. The first layer in our RNN maps the one-hot vectors **w** into intermediate continuous vectors using an embedding layer [12].

The next hidden layer consists of the LSTM blocks that, employing memory cells, will store non-linear representations of the current sequence history and be better able to encode context. To prevent overfitting a dropout layer is added after the LSTM.

Finally the last layer is a feed-forward layer that performs non-linear mappings to better approximate the human scale of behavior. The RNN is then trained for a fixed number of epochs using an adaptive learning rate optimizer [25].

For evaluation purposes, and to better approximate the human annotation process we also require a fusion layer after the RNN to combine the behavior metrics over all the time-steps and

obtain a prediction of the global rating. The proposed 3-layer recursive neural network architecture is shown in Figure 2.2.

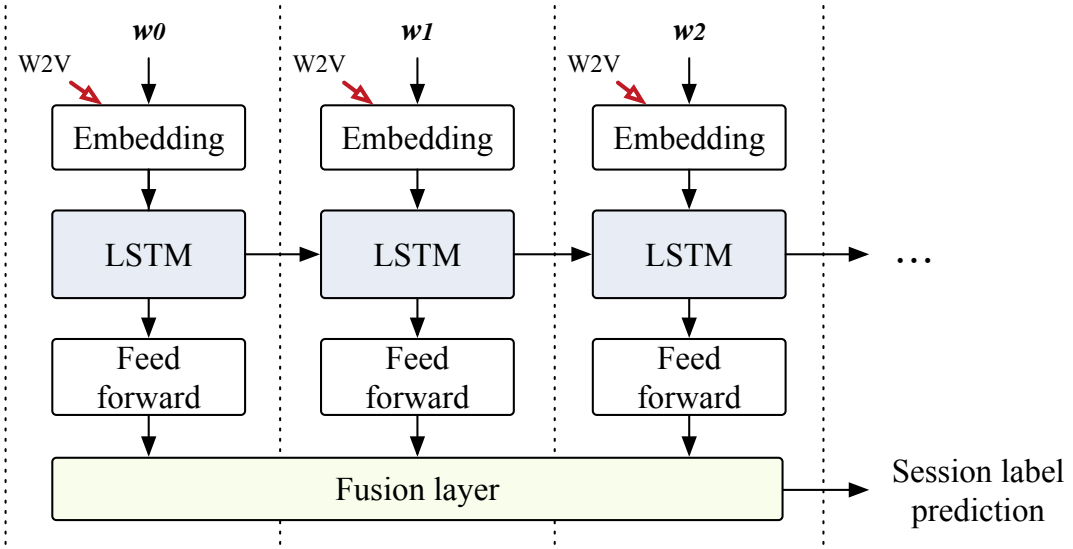


Figure 2.2: Recurrent Neural Network system for predicting behavior

2.4.2 Incorporating out-of-domain word representations

Past work has shown that distributed representations of words in a vector space can be trained to capture syntactic and semantic relationships between words [12], [26]. Such learned representations of words allow learning algorithms to combine semantic knowledge of words and achieve better performance in natural language processing tasks.

In our work, we investigate two options for generating such representations. One is to directly train this on our limited, but domain-specific training data. We will denote this as `1Hot`. Another option that also addresses the problem of data sparsity and allows for a more generalized model, is to incorporate out-of-domain knowledge by pretraining word representations on larger corpora, and we will denote this as `w2v`.

We expect that employing this second method will have advantages: First, by using pre-trained word representations we can mitigate the issue of data sparsity in our training data. High-quality word representations will map similar words to closely spaced points in the vector representation

space. This allows us to use a smaller number of parameters and hyper-parameters in constructing and training our RNN. Second, by training on the word representations the system will generalize well in regards to out-of-vocabulary words. Words that were not seen during training will still be mapped to a continuous vector that preserves its semantic relationships to words that were seen during training. The RNN will therefore be able to produce reasonable if not accurate predictions when encountering out-of-vocabulary words in a sequence.

To learn high-quality word representations we use the Google News corpus [27] which contains more than 4 billion words. We also introduce 1 million words from the General Psychotherapy corpus transcripts from [28] to allow the word representations to be more representative of our target domain. The word representations are learned through the methods described in [12] using the Google *word2vec* toolkit [29]. Since our final objective is to estimate the behavior metrics for word sequences we reduce the vector dimensionality from the commonly-used size of 300. In our experiments we tried vector dimensionality configurations of 300, 50, and 10.

The continuous word representations are incorporated into the RNN system by fixing the weights in the embedding layer with the learned word to vector mappings. These weights are then maintained during training to preserve the learned word representations.

2.4.3 Joint optimization

Using pretrained word representations the RNN learns to predict the behavior ratings from continuous vectors that capture the semantic relationships between words. However, although these word vectors encode a lot of semantic information they are not optimized for predicting behavior. By jointly training these word vectors with the behavior ratings the word representations become more indicative of behavior where appropriate while still maintaining semantic relationship. In training our RNN with pretrained word representations we initialize with the above learned word vectors and allow the weights in the embedding layer to be updated to allow for this joint optimization. We will denote this by `w2v-joint`.

2.4.4 Fusion layer

Our RNN system is trained to predict behavioral ratings for different sequences of words. Since we do not have local-level annotations to compare these predictions with, we evaluate the system at the global session-score level. We do this by fusing the local predictions to arrive at a global predicted score, similar to the human process of integrating behavioral information over time to arrive at a gestalt opinion of the session.

We observed that, in general, the median predicted rating exhibited lesser bias as an estimator of the true rating than the mean rating, possibly due to the former's robustness to outliers. Therefore, we used an RBF-Kernel Support Vector Regressor to learn a mapping from the median predicted rating to the true rating on our training data. At test time, we applied this map on the median predicted test rating to obtain the predicted session-level rating, which we then compared against the true session-level rating that had been used to train our RNN system.

2.5 Experimental Results

In our experiments we used a leave-one-couple-out cross-validation scheme to separate train and test data. In each fold one couple is held out while the rest are used for training. We applied a sliding window with a 1-word shift across each session to generate multiple training sequences and trained each RNN architecture for 25 epochs. We also tried different dimension sizes for the pretrained word vectors and found that the best results can be obtained from a dimension size of 10.

2.5.1 Binary classification of behavior

We first focused on binary classification of "Negativity" at the session level which is easier to compare with human annotations. A threshold was applied to the average of behavior metrics in a session to classify that session into High or Low Negativity. For each configuration an Equal-Error

Table 2.1: Classification accuracy (%) on negativity for different input sequence lengths

RNN Configuration	Input sequence length (words)		
	unigram	*bigram*	*trigram*
`1Hot`	87.86	85.71	86.43
`w2v`	87.5	87.1	86.8
`w2v-joint`	**88.93**	**88.21**	**87.86**

Rate threshold for the binarization task was obtained from the training data. We trained using different context length for each of the proposed RNN configurations.

The classification accuracy for the different RNN configurations with varying input sequence lengths is shown in Table 2.1. We observe, as expected due to limited data, a slightly decreasing accuracy as context is increased, but we also see that the accuracy drop is minimal. We also observe that the pretrained word representations (`w2v`) are more robust than embeddings that only employ only domain data (`1hot`) but can become even more robust by joint training (`w2v-joint`).

Note that while the relative improvement is significant it is also limited by the upper limit – even humans do not agree 100% – so the binary evaluation task is limiting our evaluation abilities. For instance, if the upper limit was 100% then we have about 15% relative improvement but if the upper limit is 92% then this jumps to a relative 40% improvement.

2.5.2 Predicting true behavior ratings

2.5.2.1 Behavioral distribution

Observing the individual markers of negativity throughout a couples interaction per session we see that the `w2v-joint` system provides a more reasonable distribution of behavior metrics: the behavioral histogram is more skewed towards the true rating value, while the `1-hot` system has very few discriminating data points. For example, Figure 2.3 shows the distribution of the sequence scores for one session.

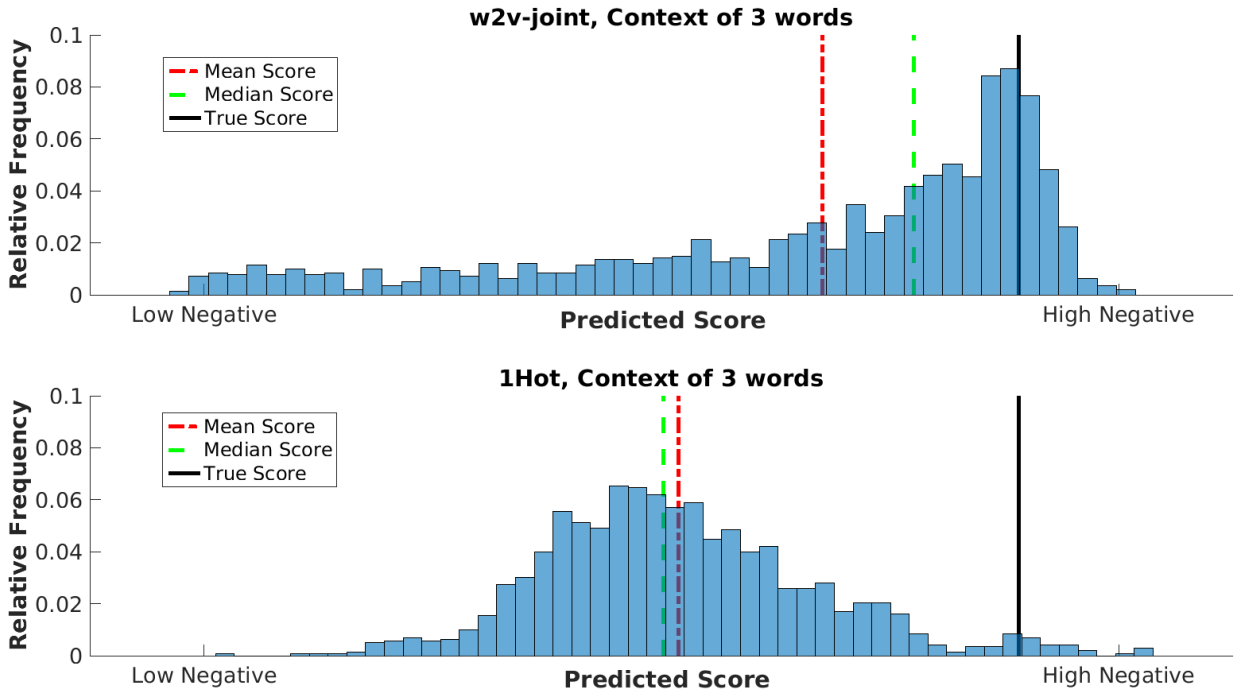


Figure 2.3: Comparison of distribution of scores for words in a session for 1-hot system (bottome) versus w2v-joint (top)

2.5.2.2 Handling out-of-vocabulary words

We also analyzed the performance of our RNN system on unseen data: words that were out-of-vocabulary during training. The pretrained system (`w2v-joint`) is able to exploit information from domain-OOV words through their similarity in the general pretraining corpus to seen domain words. Table 2.2 shows some examples of domain-OOV words and their estimated behavior metrics for negativity, where 0 and 1 indicate absence and presence respectively.

Table 2.2: Examples of out-of-vocabulary words and their behavior metrics

OOV Word	Behavior Metric for Negativity
denies	0.91
kill	0.87
dissatisfaction	0.75
funner	0.26
doggie	0.22
coordination	0.09

2.5.2.3 Agreement with human annotators

To better evaluate our system performance we estimated the behavior ratings which are obtained through the fusion layer. We compared the estimated behavior ratings to those from human annotators using Krippendorff's alpha. In the first comparison we randomly replaced a human annotation with our predicted rating for all sessions. We found that the jointly optimized word representations gave ratings that had better agreement with human ratings than conventional one-hot vectors. Next, we replaced human annotations that deviated most from the mean with our predicted ratings. In this setting we found that our predicted ratings had higher inter-annotator agreement than human-only annotations. This shows that with jointly optimized word representations our RNN system can achieve better inter-annotator agreement than outlier human annotators. The inter-annotator agreement of our predicted ratings for the different comparisons is shown in Table 2.3.

Table 2.3: Comparison of agreement using Krippendorff's alpha

Annotator Configuration	**Krippendorff's alpha**	
	`1Hot`	`w2v-joint`
All human annotators	0.821	
Random replacement with random predictions (average)	0.492	
Random replacement with machine predictions (average)	0.7611	0.7739
Outlier replaced with machine prediction	0.7997	**0.8249**

2.6 Conclusions

In psychological evaluations of therapy sessions, ratings for behaviors are very often annotated at the global session-level. This coarse resolution drastically increases the difficulty of learning frame-level or utterance-level behaviors. In this chapter we have developed a RNN system for

estimating behavior in variable-length context windows at the frame level. This enables us to observe continuous metrics of behavior in a sliding window and allows for fusion of behavior from different modalities. The RNN was trained in a data limited environment and only global ratings. We showed that by pretraining word representations on out-of-domain large vocabulary corpora and performing joint optimization we can solve the issue of data sparsity in our data and achieve increased robustness to out-of-vocabulary words. Finally we applied top level fusion on the frame-level behavior metrics to evaluate the behavior trajectories and estimate the true session rating. The estimated behavior rating from our system achieves high agreement with trained human annotators and even outperforms outlier human annotations.

This chapter proposed a RNN system that can be trained in a data limited environment to obtain meaningful behavior trajectories in a couples interaction session. This is the first step in allowing for detailed online analysis by psychologists of the interplay of behaviors in couples interactions at a finer resolution. Current observational studies in psychology often involve the time-consuming and expensive process of annotating specific behaviors in lengthy sessions. In the future this model can be deployed for a more automated method of evaluating behavior in human interactions.

Chapter 3

Unsupervised Learning of Deep Sentence Embeddings

Previous NLP methods have mainly utilized N-grams for behavior annotation using count-based or RNN language models [10], [30], [31]. While N-gram models are suited for modeling language structure they are unable to capture semantic information of entire speech segments. To conceptualize linguistic information from larger context Tanana *et al.* [32] proposed two methods for deriving sentence features. One was a discrete sentence feature model using N-grams and dependency relations in the parse tree. The other was a *recursive* neural network based on word embeddings in addition to the parse tree. These methods sought to encode contextual meaning of sentences into vector representations based on functional relationships of words from the dependency tree.

In this chapter we present an unsupervised deep learning method for deriving distributed vector representations of sentences that are well-suited for behavior annotation. We explore different methods of training such vectors and demonstrate the benefits of unsupervised training on closely-matched out-of-domain data. We also propose a comprehensive framework for modeling human behavior using recurrent neural networks (RNN) and the learned deep sentence embeddings. The RNN framework estimates the trajectories of behavior in conversational interactions using sequences of sentence embeddings. Finally, we evaluate our system on behavior ratings from the Couples Therapy Corpus using a regression model on top of the behavior trajectories.

3.1 Deep Sentence Embeddings

Recurrent neural networks have shown great ability in capturing temporal information in sequences by embedding past history information within hidden states in intermediate layers. Using these hidden states the network then makes the best output decision conditioned on this representation of history [33].

Later, it was shown in [34] that significant improvements in NLP tasks such as machine translation could be obtained by embedding the whole history before generating the output. These networks were referred to as sequence-to-sequence models and incorporate an encoder-decoder architecture that encodes the entire input before generating the output at the decoder stage. The power of sequence-to-sequence models in NLP tasks stems from the fact that the structure of language is non-deterministic and highly dependent upon context [34]. By encoding the entire input as an embedding the network learns how to extract relevant information from the whole input before generating the output. In a way, we can say that the hidden states of the encoder represent the contextual concept that is conveyed by the input sentence. These hidden states are sometimes referred to as deep sentence embeddings and have been shown to be more adept in many NLP tasks than knowledge-based or handmade features [35], [36].

3.2 Methodology

3.2.1 Deep conversational sentence embeddings

Deep sentence embeddings represent input sentences at a higher or "deeper" level of abstraction. However, the quality of embeddings depends greatly on the training methodology and learning criteria. Our goal is to estimate behavior in human interactions, therefore it follows that the sentence embeddings should represent expressed behaviors and conveyed concepts within the conversations. To this end we employ neural conversation models (CM) [37]. These are encoder-decoder

networks that have been trained to give responses from queries and are capable of basic conversations. Embeddings from these networks represent real conversations and encode relevant content of the conversation. While they are not explicitly trained to identiy behavior, given the short-term stationarity of behavior we hypothesize that behavioral information is also represented in these embeddings. Therefore, we extract sentence embeddings from conversational encoder-decoder networks to use as input features for behavior annotation. The encoder-decoder architecture is shown in Figure 3.1 and is described in further detail in Section 4.

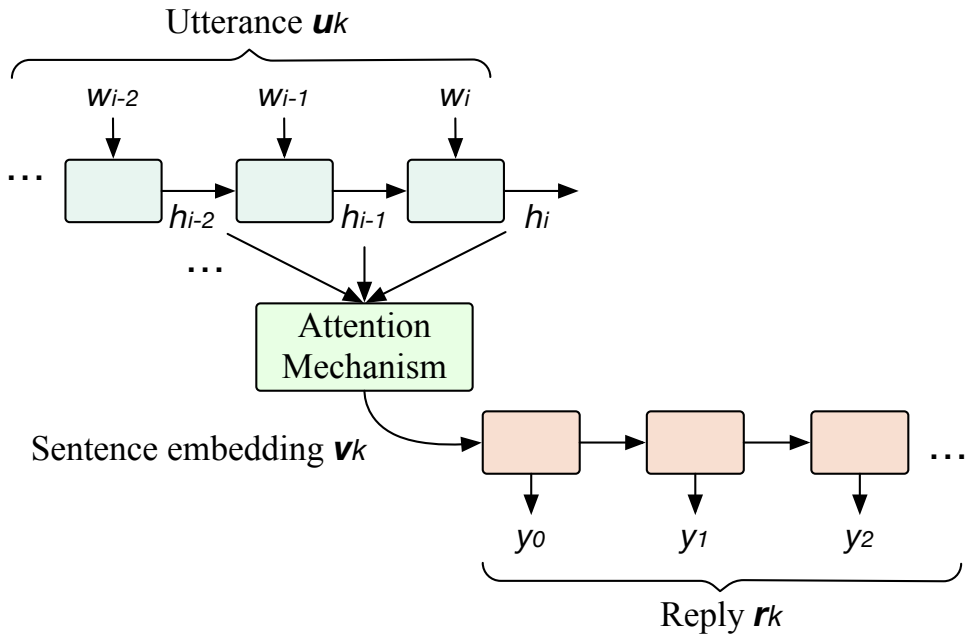


Figure 3.1: The encoder-decoder conversation model for generating deep sentence embeddings

3.2.2 Behavior annotation

The method described in the previous section is used to generate deep sentence embeddings for each utterance in our dataset. We then combine multiple utterances into sequences of sentence embeddings. We view these sequences as a representation of conversational information within the interaction over time. Our assumption is that these sentence embeddings generalize information from the speaker in a much richer form than those obtained from the word-level while maintaining

temporal information. These sequences of sentence embeddings are therefore ideal for use as features in identifying behavior throughout the interaction.

We apply a sliding window to generate frames of sentence embedding sequences and train an RNN to estimate behavior ratings for each frame. The RNN consists of an LSTM and a feedforward layer in the hidden layers with dropout added after each layer. Since we are estimating a normalized behavior rating between 0 and 1 we use a sigmoid function at the output layer. This architecture is based on the LSTM-RNN described in Chapter 2.4.1 which was proven to be effective in modeling behavior with limited training data. Figure 3.2 shows the architecture of the LSTM-RNN.

3.3 Corpora and Learning Methods

3.3.1 Training deep sentence embeddings

In our experiments we used the OpenSubtitles dataset [38] to train the encoder-decoder model for generating deep sentence embeddings. This dataset contains dialogue from movies which is similar to the back-and-forth structure of the interaction in our target behavior dataset. We processed the data by segmenting paragraphs into sentences. We also removed various generic expressions such as "*I don't know*" to prevent overtraining on responses that have little relation to queries. Similar to [37] we treat any two consecutive sentences as an utterance-reply pair without considering who uttered the sentence. We then trained the encoder-decoder model to predict the reply given the utterance. However, our work differs in that our final goal is not a working conversation model, but rather a rich semantic representation of the utterance to be used as input feature for behavior annotation. Therefore we want the sentence embeddings to be as compact as possible while still capturing rich contextual information. In our experiments we tried embedding sizes of 100, 500, and 1024 with 3 LSTM layers in the encoder and decoders. We also add an attention mechanism [39] after the encoder to allow the network to focus on more salient portions of the input. The final training set consists of 35 million utterance-reply pairs.

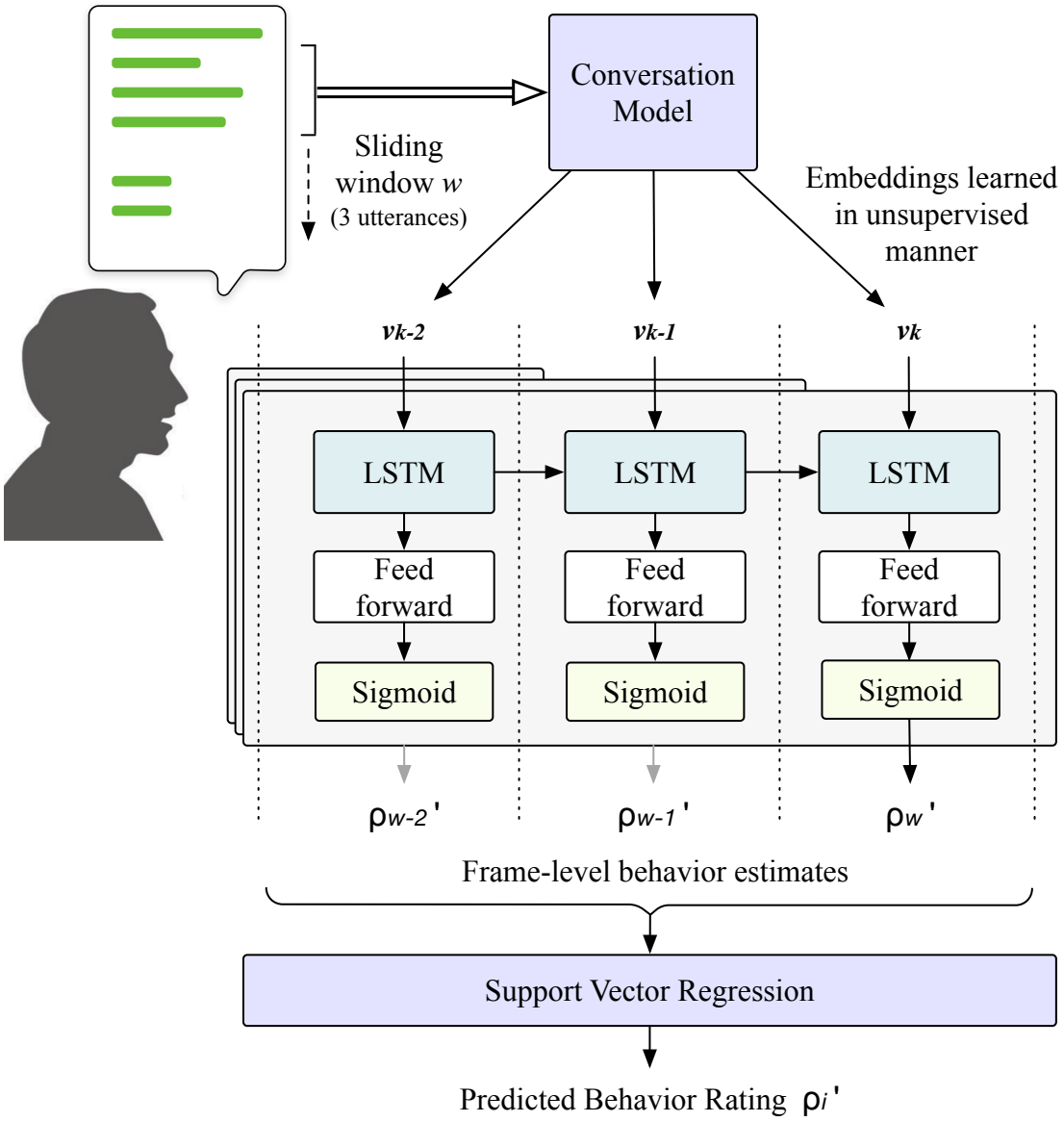


Figure 3.2: Architecture of the RNN for estimating frame-level behavior from sequences of sentence embeddings

3.3.2 Behavior Annotation in Couples Therapy

3.3.2.1 Couples Therapy Corpus

To train behavior models we use data from the UCLA/UW Couple Therapy Research Project [1] which contains audio and video recordings of 134 couples with real marital issues interacting over multiple sessions. In each session, couples discuss a specific topic chosen in turn for around 10 minutes. The behaviors of each speaker are then rated by multiple annotators based on the Couples

Interaction [5] and Social Support [40] Rating Systems. The rating system contains 33 behavioral codes such as “Acceptance”, “Blame” and “Negativity”. Each annotator provides ratings for these codes on a Likert scale of 1 to 9 for every session, where 1 indicates strong absence and 9 indicates strong presence of the behavior. There are 2 to 12 annotators per session with the majority of sessions ($\sim$ 90%) having 3 to 4 annotators. Finally, these ratings are averaged to obtain a 33 dimensional vector of behavior ratings per interlocutor for every session.

In this Chapter, we focus primarily on the behavior code “Negativity”. For our experiments we use manual transcriptions of sessions with mean annotator ratings in the top and bottom 20% of the code range for sessions with good audio quality.

3.3.2.2 Frame-level behavior metrics

Behavior ratings in the Couples Therapy Corpus are annotated for entire sessions and no labels are provided for individual utterances. However, it is infeasible to treat entire sessions as a single sequence of embeddings due to the high complexity of such a model and data scarcity issues. We also want fine annotations of utterances in addition to session-level annotations. Therefore we employ weakly supervised learning and assign session-level ratings as target values for all short embedding sequences in a session. This assumes that all utterances in a session relate to the overall rating in a non-linear and complex manner and by only considering shorter sequences we can still map back to the session rating.

In our experiments we generate frames of embedding sequences using a sliding window of 3 utterances with a shift of 1 utterance. The RNN takes these embedding sequences as input and is trained to predict the session-level rating using an SGD optimizer. The result is an overlapping trajectory of frame-level behavior metrics over the session. We refer to these values as metrics since they convey information of behavior and indirectly relate to the session rating in some form.

3.3.2.3 Annotating behavior

Since we do not have annotations for individual utterances to compare with, we validate our system with session-level ratings. We do this by fusing, using the techniques described below, the frame-level behavior metrics to derive an estimate for the session-level score. In a sense this method is similar to the human process of integrating behavioral information over time to arrive at a gestalt opinion of the session.

To compare results we apply the fusion method used in our previous work [10]. Specifically, we used an RBF-Kernel Support Vector Regressor to learn a mapping from the median of the frame-level behavior metrics in a session to the true rating. At test time, we apply this map on the median of the behavior metrics to obtain an estimated rating for the entire session. Although there are many different fusion techniques we implement this method for consistency with prior work. Session-level fusion is not the focus of this study. An overview of our proposed behavior annotation framework is shown in Figure 3.3.

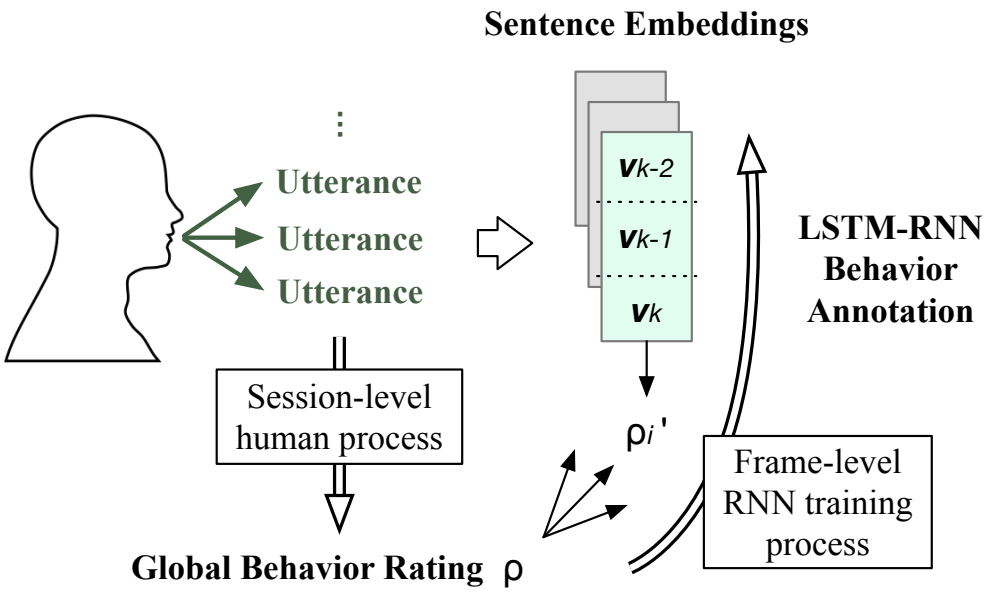


Figure 3.3: The behavior annotation framework

3.4 Experimental Evaluation

In all our experiments we used a leave-one-couple-out cross-validation scheme to separate the Couples Therapy data into train and test sets. For each fold data from one couple is held out from training and used for evaluation. This resulted in a total of 134 folds. The approach of our experiments are as follows:

- Train a conversation model using an encoder-decoder architecture by predicting replies to utterances in the OpenSubtitles dataset. This first step is domain-data independent and is only done once. The following steps are run on the per-fold split.
- Extract deep sentence embeddings for all utterances in the Couples Therapy Corpus. Use the attention layer of the conversation model as an embedding.
- Use a sequence of sentence embeddings as features and train an RNN to estimate the session-level ratings from each embedding sequence. This is the first supervised step.
- To obtain session-level behavior ratings we train an RBF-Kernel SVR to map the median of frame-level behavior metrics to a final score for each fold.

3.4.1 Predicting true behavior ratings

For validation we compared the estimated behavior rating from the fusion outputs to scores given by human annotators. We trained different RNNs for behavior annotation with various types of embedding sequences for comparison. These included our previous work of word-level sequences [10], the sum of *word2vec* [12] embeddings in a sentence, and deep sentence embeddings extracted from an English-to-French neural machine translation (NMT) model [34]. For fair comparison the dimensions of the embeddings were fixed to the same size.

It is important to note that there is no absolute truth in the reference annotations. These are subjective, and human annotators have disagreements. As such we can at best achieve to reach

agreement with the mean comparable to the inter-annotator agreement. We thus have two validation metrics: *(1)* The Mean Absolute Error (MAE) with the average rating, that we know is not necessarily the gold standard; and *(2)* Treat our system as another annotator and see how it compares to existing human expert annotators in terms of inter-annotator agreement.

Table 3.1 shows the Mean Absolute Error (MAE) between estimated ratings of different models and the average score of human annotators. Our proposed model using deep sentence embeddings performs significantly better than prior work [10] (*Mann-Whitney U-test,* $p < 0.05$).

Table 3.1: MAE of estimated ratings using different models

Model	**MAE**
Word-level sequences [10]	1.53
Sum of *word2vec*	1.43
NMT embeddings	1.68
CM embeddings	**1.37**

To evaluate inter-annotator agreement we mixed our estimated ratings with human annotations and calculated Krippendorff's alpha coefficient for two different configurations. In the first configuration we randomly replaced one human annotation with estimated ratings. We found that while all models achieved lower inter-annotator agreement than human-only annotations, the system trained on embedding sequences from the conversation model gave the best results. Next, we selected human annotations that deviated most from the mean and replaced them with estimated ratings. Again we found that our proposed method gave the highest agreement and even outperformed outlier human annotators in terms of agreement with other annotators. Table 3.2 shows the inter-annotator agreement between estimated ratings and human annotators under the different configurations. The Krippendorff's alpha for random replacement with random values is also shown as reference.

Table 3.2: Comparison of inter-annotator agreement using Krippendorff's alpha

Annotator Configuration	**Krippendorff's Alpha**			
All human annotators	0.821			
Random replacement with random predictions	0.492			
	Word seq. [10]	Sum W2V	NMT embd.	CM embd.
Random replacement with machine predictions	0.7739	0.7776	0.7511	0.7832
Outlier replaced with machine prediction	0.8249	0.8368	0.8010	**0.8403**

3.4.2 Rating behaviors in text

To see how our system performs on out-of-domain data we rated negativity on dialogue from the television series "Friends" as an example. We used transcripts from the show and tokenized each speaker turn into one or more utterances. To track behavior in the overall interaction we assumed that all utterances originated from a single person and applied the sliding window on all the data. Even though "Friends" is a comedy and is expected to be mostly positive, our system was able to identify many utterances that seem to exhibit negative behavior. Some examples of negativity in the dialogues are shown in Table 3.3. These results are encouraging in that they show how our behavior annotation framework is able to learn from weak labels and be transferable to other domains.

3.5 Conclusions and Future Work

In this chapter we proposed a behavior annotation framework based on deep sentence embeddings trained using neural conversation models. We theorize that sentence embeddings from conversation

Table 3.3: Examples of negativity in utterances from Friends

Less Negative Sentences
Alright, this barbecue is gonna be very fun
I'm not saying he has to spend the whole evening with me, but at least check in .
More Negative Sentences
I'm the girl in the veil who stomped on your heart in front of your entire family.
Joey, this is sick, it's disgusting, it's not really true, is it?

models are more adept at capturing conversational concepts which relate better to behavior. We then modeled interactions using sequences of these embeddings and trained an LSTM-RNN to estimate trajectories of behavior in Couples Therapy Sessions. Finally, we evaluated our system by fusing local behavior metrics into a session-level rating and compared with human annotations. The results of our experiments showed that using embedding sequences from conversation models as input features for behavior modeling achieves higher inter-annotator agreement with human annotators over other types of sentence embeddings. Such an approach gives session-level behavior ratings close to human annotators and even outperforms outlier humans.

Our system seeks to alleviate the expensive and time-consuming process of manual behavior annotation required for observational studies in psychotherapy. In addition, through weakly supervised learning, we provide objective behavior ratings at a finer resolution of per utterance. The utterance-level behavior ratings are more capable than previous works at capturing behavior trajectories in a couples interaction session and allow for more detailed analysis by psychologists.

Chapter 4

Online Multitask Learning

4.1 Introduction

Word embeddings exploit the use of language by learning semantic regularities based on a context of neighboring words. This form of contextual learning is unsupervised, which allows learning from large-scale corpora and is the main reason for its effectiveness in improved performance on many tasks such as constituency parsing [41], sentiment analysis [42], [43], natural language inference [44], and video/image captioning [45], [46].

With the introduction of sequence-to-sequence models (*seq2seq*) [47], embeddings were extended to encode entire sentences and allowed representation of higher level concepts through longer context. For example, [48] obtained *sentence embeddings*, which they referred to as *skip-thought* vectors, by training models to generate the surrounding sentences of extracts from contiguous pieces of text from novels. The authors showed that the embeddings were adept at representing the semantic and syntactic properties of sentences through evaluation on various semantic related tasks. [49] extracted sentence embeddings from an LSTM-RNN which was trained using user click-through data logged from a web search engine. They then showed that embeddings generated by their models were especially useful for web document retrieval tasks. Later, [14] extracted sentence embeddings from a conversation model and showed the richness of semantic content by applying an additional weakly-supervised architecture to estimate the behavioral ratings of couples

therapy sessions. More recently, [50] learned unsupervised sentence embeddings using an extension of the training objective used in *word2vec* [12]. The authors proposed an unsupervised model which composes sentence embeddings from word vectors and n-gram embeddings through joint optimization. They then showed the generalizability of their sentence embeddings by evaluating on a wide range of downstream NLP tasks.

Sentence representations that are not task-specific but rather *general-purpose* and can be applied directly to multiple NLP tasks have also been proposed. [51] achieved this by training for various tasks such as machine translation, constituency parsing, and image caption generation, to produce embeddings which improved the translation quality between English and German. Subsequently in [52] it was hypothesized that a single Natural Language Inference (NLI) task [53] was sufficient in learning general purpose embeddings due to it being a high-level understanding task. The authors then showed the effectiveness of the sentence embeddings in 12 transfer tasks, examples of which include semantic relatedness, sentiment analysis, and caption-image retrieval. Later, [54] presented a large-scale multitask framework for learning general purpose sentence embeddings by training with a multitude of NLP tasks, including skip-thought training, machine translation, entailment classification, and constituent parsing. Similarly, [55] proposed a transformer based sentence encoding model trained on multiple tasks which also include skip-thought training, conversational response generation, and NLI.

The benefit of many of the methods in the aforementioned work is that the embedding transformation is learned on large amounts of data. Since the generation of natural language is an extremely complex process, it is crucial to leverage large corpora when training embeddings so as to capture *true* semantic concepts instead of regularities of the data, *e.g.* domain-specific topics [56]. Previously this was achieved through the use of abundant unlabeled datasets and unsupervised learning techniques [48], [50], [57]. However, as recent work [54], [55] has shown, learning sentence representations from multiple labeled datasets can produce significant improvements over prior unsupervised methods.

A common issue with unsupervised training of word or sentence embeddings is the unpredictability of the resulting embedding transformation. In other words, the information carried by embeddings is highly uninterpretable and may often contain redundant or irrelevant information [58]. In addition, depending on training conditions such as architecture or dataset, the representations might fail to capture informational concepts or even semantics of the input data [52].

It has also been noted that the quality of sentence embeddings is often highly dependent on the training dataset [14], [49]. In fact, the benefit of using matched datasets may be so prominent that embeddings trained on small *domain-relevant* datasets could yield results better than those trained on larger generic unlabeled datasets [48]. And while many general purpose sentence embeddings have been trained with large amounts of labeled data through multitasking, applications by others to their respective domains might not guarantee the same significant improvement of results. This problem is inherent in the fact that a domain adaptation step is generally still required over the embeddings.

One way that unsupervised representations can better gain domain-specificity is through *multi-task learning* (MTL). For example prior work has shown the benefits of leveraging MTL to enhance the informational content of word embeddings in many NLP applications [59]–[61]. In recent years, through the advancement of computational methods, MTL has been applied to the learning of sentence embeddings that allow for a larger context window. For example, [62] jointly learned sentence embeddings with an additional pivot prediction task in conjunction with sentiment classification. [63] predicted neighboring words as a secondary objective to improve accuracy of various sequence labeling tasks.

One of the challenges in MTL is that the labels required by the secondary task are not often available for the vast amounts of unlabeled datasets employed in representation learning. One of the contributions towards this direction is that we do not require the existence of such labeling. Our work differs in that we build on *unsupervised contextual learning* to learn the sentence representation and attempt to guide the sentence embeddings to become domain relevant through a related multitask objective. The underlying assumption of our work is that the behavior expressed

in two adjacent sentences will be the same due to short term stationarity. However, the resulting representation encodes a vast amount of information, which we hope to further attune towards domain-relevance. We achieve this through the related task of emotion-related labels. Unlike prior works however, our second emotion-related guiding task does not require prior labeling. We target unsupervised scenarios and use a naive scheme based on limited human-knowledge to *automatically generate multitask labels from unlabeled data in an online manner*. We hypothesize that by adopting an extremely simple form of sentiment analysis [64] as the multitask objective the unsupervised sentence embeddings will become more adept in behavior understanding.

Specifically in this work we aspire to combine the advantages of unsupervised learning with multitask learning to derive representations that are better suited for affect and behavior recognition tasks. We propose an online MTL framework which aims to *guide* unsupervised sentence embeddings into a space that is more discriminative in the targeted application scenario even under the use of mismatched and limited data. In our framework, transfer of domain-knowledge is achieved through an additional task in parallel with contextual learning. The labels for the multitask are generated online from the data to maintain an unsupervised scenario. We show that embeddings trained through this framework offer improved deftness in multiple supervised affective tasks.

4.2 Unsupervised Multitask Embeddings

In this section we describe the methods used to learn domain-adapted unsupervised sentence embeddings. We introduce the learning of sentence embeddings using sequence-to-sequence models followed by the formulation of our online multitask training objective and its architecture.

4.2.1 Sequence-to-sequence sentence embeddings

The sequence-to-sequence model maps input sequences to output sequences using an encoder-decoder architecture. Given an input sentence $\mathbf{x} = (x_0, x_2, ..., x_T)$ and output sentence $\mathbf{y} = (y_0, y_2, ..., y_{T'})$,

where x_t and y_t represent individual words, the standard sequence model can be expressed as computing the conditional probability

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{t=0}^{T'} P(y_t \mid y_{i<t}, \mathbf{s}, h) \tag{4.1}$$

where $\mathbf{s}$ is the sequence of outputs s_t from the encoder and h is the internal representation of the input given by the last hidden state of the encoder. For a given dataset $\mathscr{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$, we denote the learned internal representation as

$$h_{\theta} \equiv f(\mathbf{x} \mid \mathscr{D}) = f(\mathbf{x} \mid \boldsymbol{\theta})$$

where $f(\cdot)$ is the encoder function and $\boldsymbol{\theta}$ is the set of parameters resulting from $\mathscr{D}$.

The internal representation $h_{\boldsymbol{\theta}}$ encodes the input $\mathbf{x}$ into a vector space that allows the decoder to generate a good estimate of $\mathbf{y}$. In cases where $\mathscr{D}$ contains semantically-related data pairs, $h_{\boldsymbol{\theta}}$ can be viewed as a semantic vector representation of the input, or sentence embedding, which can be useful for subsequent NLP tasks. In our case we apply contextual learning and designate consecutive sentences in continuous corpora as $\mathbf{x}$ and $\mathbf{y}$.

While this model allows us to obtain semantically rich embeddings through training on unlabeled data, the quality of the embeddings is highly influenced by biases in the data and prevents the embeddings from becoming specialized in any target task [52]. Therefore we propose to enhance the quality of unsupervised sentence embeddings through multitask learning.

4.2.2 Multitask embedding training

The addition of a multitask objective can guide embeddings into a space that is more discriminative in a target application. We hypothesize that this holds true even when the multitask labels are generated online from unlabeled data with no assumption of label reliability, as long as there is some relation between the multitask and target application.

Assuming an online system which generates multitask labels $\mathbf{b}$ for each input $\mathbf{x}$ we can augment the dataset to yield $\mathscr{D}_{\text{aug}} = \{(\mathbf{x}_n, \mathbf{y}_n, \mathbf{b}_n)\}_{n=1}^{N}$. We then aim to predict this new label $\mathbf{b}$ in conjunction with the original output sequence $\mathbf{y}$. This is implemented in our *seq2seq* model by adding another *head* to the internal representation h, shown in Figure 4.1, which we will refer to as the *multitask network*. In addition to Eq. 4.1, the model now also estimates the conditional probability

$$P(\mathbf{b} \mid \mathbf{x}) = g(h \mid \mathscr{D}_{\text{aug}}) = g(h_{\theta_{\text{aug}}})$$

where $g(\cdot)$ is the network function for online transfer learning using the multitask network and $h_{\theta_{\text{aug}}}$ is the new internal representation given by $\mathscr{D}_{\text{aug}}$. In this work, $g(\cdot)$ is implemented using a multilayer perceptron. The overall architecture is shown in Figure 4.1.

The training loss is then the weighted sum of losses from the multiple tasks, defined as

$$J = \lambda \cdot L_1(\mathbf{y}, \mathbf{x}) + (1 - \lambda) \cdot L_2(\mathbf{b}, \mathbf{x})$$

where L_1 and L_2 are the cross entropy losses for contextual learning and the additional task, respectively.

With most multitask setups there is an issue on how to control the training ratio λ to account for different data sources. For example, if there is no overlap in inputs of the multiple tasks then λ can only alternate between 0 and 1 during training to switch between the different tasks. However, since we propose a multitask objective whose labels are generated from incoming data we are able to freely adjust λ. It is possible to adjust the multitask ratio as training progresses to put emphasis on different tasks but we do not make any assumptions on the optimal weighting scheme and give equal importance to both tasks by setting λ to 0.5.

4.2.3 Online multitask label generation

To guide the embeddings in becoming more suitable for affective tasks, we select a multitask objective that classifies the polarity in sentiment (positive or negative) of input sentences. Tasks

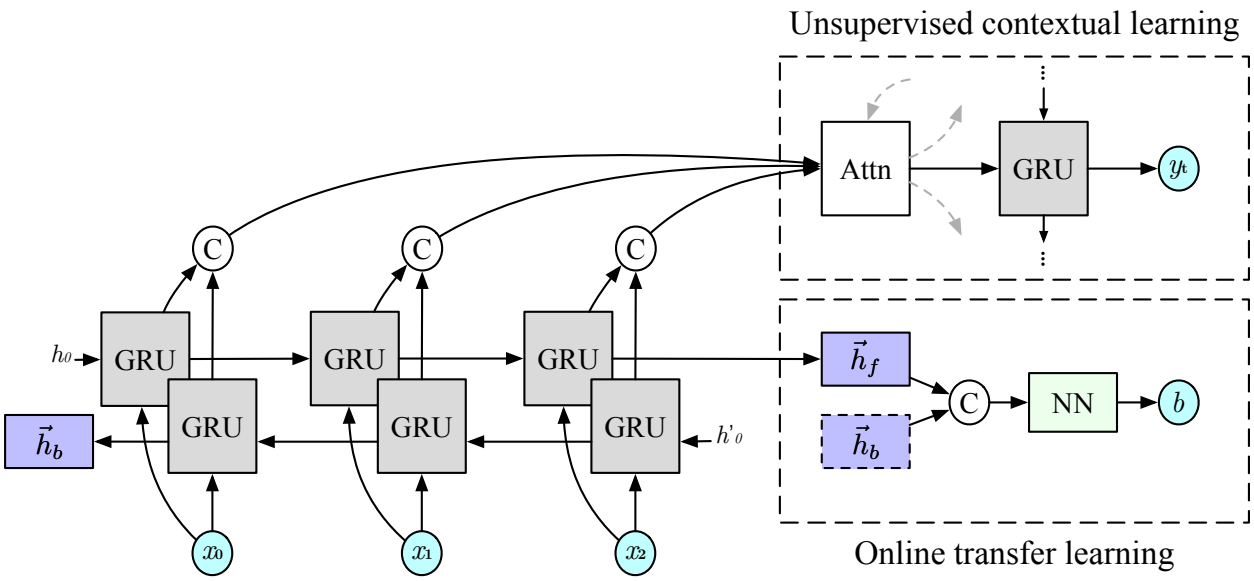


Figure 4.1: Bidirectional sequence-to-sequence conversation model with multitask objective The `GRU` blocks represent multi-layered RNNs using GRU units, `C` is the concatenation function, and `Attn` is an attention mechanism [65] with dotted arrows representing connections to and from other timesteps. For simplicity, only one timestep (y_t) of the decoder is shown.

such as emotion recognition or human behavior analysis [66] are more complicated than these two affective states, however we hypothesize this is a related task allowing for domain knowledge transfer into the sentence embeddings.

We generate the affective labels for each input during training using an online mechanism. In our online approach we apply the simplest method by automatically labeling inputs using a simple, knowledge-driven, look-up table of likely affect of single words [67]. Specifically, we use words categorized in the two top-level affective states: negative and positive emotion. An input sentence is assigned a *Negative* or *Positive* label based on the majority number of words corresponding to each affective state. Some examples of affective words in the affective look-up table are shown in Table 4.1.

Evidently, this labeling approach differs slightly from sentiment analysis [64], which mostly focuses on classifying the polarity of subjective opinions. In our case we label all the inputs naively based on the count of affective words and do not consider semantic context or even simple word negation. We expect this approach to deviate greatly from the ground truth, and that truth may

be contextual, subjective, and fluid, however we hypothesize the inclusion of affective knowledge in embeddings will still be beneficial in identifying more complex behaviors or emotions later. Specifically, we do not want to constrain the system through methods such as [64] but rather place emphasis and focus on domain relevant terms.

Table 4.1: Examples of positive and negative affect words

Affective State			
Positive		**Negative**	
cute	*love*	*ugly*	*hate*
rich	*nice*	*hurt*	*nasty*
special	*sweet*	*wicked*	*distraught*
forgive	*handsome*	*shame*	*overwhelm*

4.3 Evaluating on Behavior Identification using Embeddings

After MTL training, the encoder in the *seq2seq* model is used to extract embeddings for use as features in behavior identification in long pieces of text (which we refer to as sessions). Each session has a behavior label and contains multiple sentence embeddings. We define sentence embeddings to be the concatenation of the final output states of both the forward and backward RNNs in the encoder. We also concatenated the output states from all the intermediate layers of the encoder. This is an extension of history-of-word embeddings [68] and is motivated by the intuition that intermediate layers represent different levels of concept. By utilizing intermediate representations of the sentence, we expect that more information related to human behavior can be captured.

To evaluate the ability of the proposed system in creating behavior-tuned embeddings we apply the embeddings to task of behavior and emotion analysis. We do this in multiple ways: from minimal information about the domain, to training supervised neural networks over the unsupervised sentence embeddings. These methods are described below.

4.3.1 Unsupervised clustering of embeddings

As an initial evaluation step we analyzed the performance of the embeddings on a binary behavior classification task using minimal training on the Couples Therapy Corpus which will be described below. We applied a simple k-means clustering method on sentence embeddings from training sessions to obtain two clusters. We then labeled the clusters by randomly selecting a single session from the training set as seed and assigning the session label to the cluster which the majority of embeddings in that session belonged to. The other cluster was subsequently labeled as the opposite class label. Final test session labels were predicted based on which cluster the majority of embeddings from a session were in. Although this method of behavior classification is very rudimentary with the possibility the randomly selected session being an outlier, it nonetheless gives valuable insight on the discriminative power of the sentence embeddings. It should be noted that we do not make any assumptions on the meaning behind the clusters other than their adeptness in classifying behavior.

4.3.2 Embeddings as features in supervised learning

We also evaluated two supervised techniques on both the IEMOCAP and Couples Therapy Corpus. The two methods are k-nearest neighbor and a more advanced neural network-based method, both of which utilize the unsupervised embeddings as features in supervised learning.

4.3.2.1 k-Nearest neighbors

In this evaluation scenario we used the labels in the training data towards constructing a very simple classifier using the k-nearest neighbors (k-NN). All embeddings in the training set were assigned the label of the session they belonged to. A test embedding was then labeled according to its k-nearest neighbors in the training set. The final session label was obtained by a majority vote over all embeddings in the session.

4.3.2.2 Neural networks

Finally, we employed neural networks to estimate behavior ratings as well as recognize emotions. For behavior annotation we applied the framework proposed in [14]. Sessions were segmented into sentences and represented as a sequence of embeddings. A sliding window of size 3 was applied over the embeddings followed by an RNN using LSTM units. LSTM units were used instead of GRUs, which were used in the *seq2seq* model, to allow direct comparison with results from [14]. However we do not expect significant differences in performance between the two types of units, as was shown by [69] in their own applications.

The network was trained to predict the session rating from each window of multiple sentences representations. The final rating was obtained by training a Support Vector Regressor to map from the median value of all window predictions in a session to the session rating.

4.4 Experimental Setup

4.4.1 Datasets

In this section we describe the datasets that were used in the experiments. We used the OpenSubtitles2016 corpus [70] to pre-train sentence embeddings in the online multitask framework. To evaluate the embeddings in domain-specific tasks, we used the Couples Therapy Corpus [1] and IEMOCAP [71].

4.4.1.1 OpenSubtitles

Since our final task is emotion and behavior analysis of human interactions, we applied a dataset that contains conversational speech to pre-train our embeddings. A natural choice for a source rich in dialogue is subtitles from movies and TV shows. To this end we used the OpenSubtitles2016 corpus [70] to train the unsupervised sentence embeddings.

The OpenSubtitles2016 corpus was compiled from a database dump of the `opensubtitles.org` repository and comprises of subtitles from 152,939 movies and TV episodes spanning a time period of over 20 years. Out of more than 60 languages in the corpus we selected only subtitles in the English language for use in our training. The original corpus applied basic pre-processing through text standardization and segmentation of the subtitles into sentences [72]. We then used further techniques to clean up the text by applying auto-correction of commonly misspelled words, contraction removal, and replacement of proper nouns through parts-of-speech tagging.

To generate back-and-forth conversations we assigned consecutive sentences in the subtitles as turns in an interaction. Since there is no speaker information in the corpus, distinguishing between dialogues and monologues without the use of more advanced content analysis methods is nontrivial. However, we assume that this difference in conversational continuity will be dampened by the large amount of data available. We also reason that monologues also represent some form of internal dialogue which also ties the concepts between sentences. More importantly, since our final task is to represent behavior, we desire that sentence pairs carry information related to behavior. This can be achieved through the concept of *short-term behavior stationarity* in which two nearby sentences are likely to represent the same behavior, irrespective of turn-taking. This property was also shown by [73] wherein correlations in behavior were observed across interlocutors.

After forming all utterance/reply pairs from the corpus we randomly sampled 30 million sentence pairs as the final training data.

4.4.1.2 Couples Therapy Corpus

We evaluated our sentence embeddings in the task of annotating behaviors in human interactions using data from the UCLA/UW Couple Therapy Research Project [1]. This corpus pertains to the training of unsupervised, k-NN, and neural network learning methods described in the previous section.

The Couples Therapy Corpus contains recordings of 134 real couples with marital issues interacting over multiple sessions. In each session the couples each discussed a self-selected topic for

around 10 minutes. The recordings of the session were then rated by multiple annotators based on the Couples Interaction [5] and Social Support [40] rating systems. The combined rating system describes 31 behavioral codes rated on a Likert scale of 1 to 9, where 1 indicates strong absence and 9 indicates strong presence of the given behavior. The number of annotators per session ranged from 2 to 12, however the majority of sessions (~90%) had 3 to 4 annotators. Annotator ratings were then averaged to obtain a 31 dimensional vector of behavior ratings per interlocutor for every session. The ratings were binarized to produce labels for the classification task and the Likert scale values were used for behavior rating estimation.

In this chapter we focused on the behaviors *Acceptance*, *Blame*, *Humor*, *Sadness*, *Negativity*, and *Positivity*. While the behaviors *Negativity* and *Positivity* are more certain to benefit from the affect labels in MTL, which may be loosely similar, the remaining behaviors have more specific definitions which may be more challenging in identifying. We formulated two tasks for each of the behaviors: (1) binary classification on the presence of a behavior and (2) regression on the rating of a behavior in the whole session.

Similar to prior works [14], [74] we used only those sessions that had averaged ratings in the top and bottom 20% of the dataset. In total, 85 individual couples were included in our evaluation dataset. Evaluation of the models was performed using a leave-one-*couple*-out cross-validation scheme. That is, for each fold, sessions from one couple were used as the test set while the remaining sessions were used as the training and validation set. We report evaluation metrics averaged across these 85 folds.

4.4.1.3 IEMOCAP

We also evaluated the effectiveness of our sentence embeddings in emotion recognition using the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [71]. We use this corpus for domain-supervised learning using the embeddings as features. This dataset contains recordings from five male-female pairs of actors performing both scripted and improvised dyadic interactions.

Utterances from the interactions were then rated by multiple annotators for dimensional and categorical emotions. Similar to other works [75], [76], we focused on four categorical labels where there was majority agreement between annotators: *happiness*, *sadness*, *anger*, and *neutral*, with *excitement* considered as *happiness*. We used the transcripts from the dataset and removed any acoustic annotations such as "laughter" or "breathing". After discarding empty sentences our final dataset consisted of 5,500 utterances (1103 for *anger*, 1078 for *sadness*, 1615 for *happiness*, and 1704 for *neutral*). To evaluate the domain-supervised layers we used a leave-one-*pair*-out cross-validation testing scheme and report the evaluation metrics averaged across 5 folds.

4.4.2 Model architectures and training details

4.4.2.1 Sentence embeddings

The sequence-to-sequence model with multitask objective comprises three sections: encoder, decoder, and the multitask network. The encoder was implemented using a multi-layered bidirectional RNN using GRU units. We performed a grid search using hyper-parameter settings of 2 and 3 layers, and, 100 and 300 dimensions in each direction per layer. For the decoder a unidirectional RNN using GRU units was used instead of bidirectional. The number of layers in the decoder were the same as the encoder while the dimension size was doubled to account for the concatenation of states and outputs from both directions.

The multitask network was implemented using a neural network with four hidden layers of sizes 512, 512, 256, and 128. The final output had a dimension size of 2 to represent *Positive* and *Negative* affect class labels. We used the rectified linear unit (ReLU) function as activation functions in the hidden layers and a softmax activation function in the final output layer. No other network hyper-parameters were tried for the multitask network.

The sentence embedding models were trained with the OpenSubtitles dataset for 5 epochs using stochastic gradient descent with an added momentum term. The learning rate was set to 0.05 and momentum set to 0.9. We also reduced the learning rate by a factor of 10 every epoch.

4.4.2.2 Supervised behavior annotation

Similar to [14] we used a recurrent neural network to estimate behavior ratings in the Couples Therapy Corpus. The network had a single recurrent layer implemented using LSTM units with dimension size 50. A sigmoid function was applied before the output to estimate the normalized rating value. In each fold one couple was randomly selected as validation to select the best model.

4.4.2.3 Supervised emotion recognition

A neural network with four hidden layers was used to classify emotions using embeddings of sentences from the IEMOCAP dataset. The hidden layers were of size 256 and used ReLU as the activation function. The model was trained for 20 epochs using Adagrad [77] as the optimization method. No other network hyper-parameters were tried for the emotion recognition network. A subset of the training data (~10%) was used as validation in selecting the best model.

4.5 Experimental Results

We evaluated the performance of our unsupervised multitask sentences embeddings on the task of behavior annotation in the Couples Therapy Corpus, as well as emotion recognition on the IEMOCAP dataset. We also compared to multiple state-of-the-art general purpose embeddings such as InferSent [52], GenSen [54], and Universal Sentence Encoder [55].

4.5.1 Results on Couples Therapy Corpus

We defined two sub-tasks in behavior annotation on the Couples Therapy Corpus: (1) binary classification of the presences of behaviors and (2) regression for real-valued session ratings of the behaviors.

For the classification sub-task we used the accuracy averaged across all test folds as the evaluation metric. Table 4.2 shows the accuracy results on different behaviors in the Couples Therapy Corpus. The addition of the multitask objective improved the classification accuracy of sentence

embeddings from the conversation model across all behaviors except *Positivity* in unsupervised classification with k-Means. Under supervised learning using k-NN, our multitask embeddings improved accuracy on all behaviors except *Humor*. In terms of mean accuracy, our multitask embeddings performed better than other sentence embeddings with an absolute improvement over no multitasking of 1.07% and 3.24% for unsupervised and supervised methods respectively. Our multitask embeddings also achieved the highest mean accuracy over all the behaviors. The improvement over the second best results obtained from GenSen was statistically significant with p-value < 0.006 using McNemar's test.

For the regression sub-task we evaluated performance using Krippendorff's alpha coefficient [78]. Krippendorff's alpha is a reliability measure of the agreement between independent observers in regards to their annotation of data, commonly known as the inter-annotator agreement. We used this metric to evaluate how well trained models would function as a replacement for human annotators. Similar to [14] we evaluate the agreement with various ways of incorporating machine-generated ratings. In the first method, human annotations were randomly replaced by the estimated ratings in each session. This was performed 10 times to obtain the average Krippendorff's alpha of random injection. In the second method, the outlier annotation (rating farthest from the mean) in each session was replaced by the estimated ratings.

Table 4.3 shows the inter-annotator agreement of the different injection methods. While no system was consistently optimal, we observed that our online MTL embeddings were comparable with state-of-the-art general purpose embeddings. In fact, statistical tests using Mann–Whitney U test on the annotation errors showed no significant differences between the best model and ours.

Table 4.2: Accuracy (%) of behavior identification using sentence embeddings

Method	Embedding Model	Acceptance	Blame	Humor	Negativity	Positivity	Sadness	Mean Accuracy
k-Means	InferSent [52]	58.9	63.6	60.7	61.4	62.1	58.9	60.93
	GenSen [54]	53.9	66.4	58.9	61.4	61.4	59.6	60.27
	Universal Sentence Encoder [55]	59.3	65.7	59.6	61.8	64.3	59.6	61.72
	Conversation Model [14]	61.9	65.4	59.1	64.6	**65.7**	57.9	62.43
	+ Online MTL (proposed)	**64.0**	**66.4**	**62.1**	**65.0**	62.1	**61.4**	**63.50**
k-NN	InferSent [52]	83.2	81.1	57.1	85.4	78.6	65.7	75.27
	GenSen [54]	**85.0**	85.0	56.1	85.7	81.1	63.2	76.02
	Universal Sentence Encoder [55]	80.0	82.5	**60.4**	83.9	79.6	66.8	75.53
	Conversation Model [14]	79.6	80.0	59.6	85.7	82.5	64.6	75.53
	+ Online MTL (proposed)	**85.0**	**85.4**	60.0	**87.9**	**86.8**	**67.9**	**78.77**

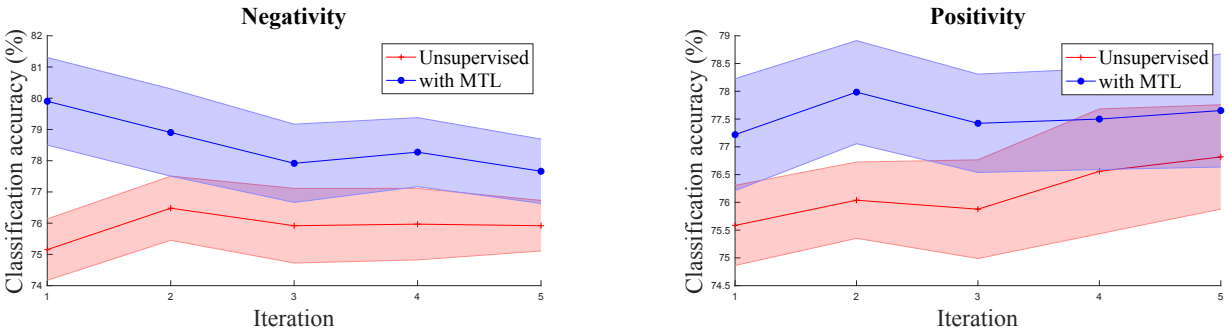
The improvement of our model over the next best performing model across all behaviors is statistically significant with $p < 0.006$.

Table 4.3: Inter-annotator agreement (Krippendorff's alpha) of estimated behavior ratings using different incorporation methods

Method	Model	Acceptance	Blame	Humor	Negativity	Positivity	Sadness
	Human	0.790	0.828	0.584	0.829	0.695	0.623
Random injection	Random ratings	0.387	0.443	0.161	0.522	0.384	0.274
	InferSent [52]	**0.790**	**0.828**	0.455	**0.829**	**0.695**	0.455
	Gensen[54]	0.736	0.773	0.452	0.772	0.649	0.460
	Universal Sentence Encoder [55]	0.742	0.773	**0.457**	0.778	0.643	**0.472**
	Conversation model [14]	0.722	0.757	0.442	0.782	0.644	0.462
	+ Online MTL (proposed)	0.735	0.773	0.450	0.787	0.645	0.468
Worst-annotation-out	Random ratings	0.341	0.405	0.127	0.521	0.392	0.304
	InferSent [52]	0.790	**0.829**	**0.584**	0.829	0.695	0.563
	Gensen [54]	0.804	0.820	0.565	0.844	**0.731**	0.559
	Universal Sentence Encoder [55]	**0.814**	0.818	0.575	0.846	0.726	0.574
	Conversation model [14]	0.786	0.796	0.568	0.856	0.725	0.572
	+ Online MTL (proposed)	0.801	0.815	0.567	**0.861**	0.727	**0.578**

No statistical significant differences were found between pairs of models, however all models have significant improvement over randomly generated ratings.

To factor out the influence of hyper-parameters and randomness of training we analyzed the performance of all the *seq2seq* models in our hyper-parameter search space. For each model configuration, five intermediate checkpoints from training were randomly selected. Sentence embeddings were then extracted from these individual models and applied to the behavior classification task. We then compared the performance of models with and without multitask learning. The standard error plot of the performance in *Positivity* and *Negativity* recognition is shown in Figure 4.2. We observed that the addition of the multitask learning objective collectively increased performance in the final task for most behaviors. This shows that the addition of online transfer learning through multitask to unsupervised sentence embeddings does indeed provide an advantage in performance.



A B

Figure 4.2: Standard error plot of classification accuracy on *Negativity* and *Positivity* for various model hyper-parameter configurations across multiple iterations

4.5.2 Results on IEMOCAP

We evaluated the performance of emotion recognition on IEMOCAP using weighted accuracy (WA) which avoids inflation due to imbalanced number of labels in each class. This is also equivalent to the macro-average of recall scores per class. In addition to general purpose embeddings we also compared with other works that only used IEMOCAP transcripts [76], [79], [80] . It should be noted that there is no official consensus on train/test split or evaluation procedure in IEMOCAP,

and while we made every effort to be consistent with past work (in terms of label classes, number of utterances used, and cross-validation scheme) the results may not be exactly comparable.

The results of emotion recognition on IEMOCAP are shown in Table 4.4. We observed that the addition of online MTL improved the accuracy of conversation model embeddings by an absolute value of 8.02%, which is more than 14% relative improvement. When comparing among our own implementations we observed that the highest accuracy was obtained using embeddings from the Universal Sentence Encoder which had a weighted accuracy of 64.83%. The system trained using our sentence embeddings offered a close second by less than one percent with 63.84% accuracy. Statistical analysis using McNemar's test showed that the improvement of the best system over our proposed embeddings was not significant. However, we observed significant improvement from our model over embeddings from InferSent with p-value < 0.02. Given the considerably smaller amount of pre-training data required and the simpler structure of our proposed MTL system this similarity in performance to Universal Sentence Encoder and advantage over other embeddings is notable.

Table 4.4: Weighted Accuracy of Emotion Recognition on IEMOCAP

Method	WA (%)
Lex-eVector [79]	57.40
E-vector + MCNN [76]	59.63
mLRF [80]	63.80
InferSent [52] + DNN	62.60
GenSen [54] + DNN	60.62
Universal Sentence Encoder [55] + DNN	**64.83**
Conversation Model [14] + DNN	55.82
+ Online MTL (proposed) + DNN	63.84

4.6 Conclusion

In this chapter we explored the benefits of introducing additional objectives to unsupervised contextual learning of sentence embeddings. We found empirical evidence that supports the hypothesis

that MTL can increase the affective deftness of unsupervised sentence embeddings, even when the multitask labels are generated online using a naive, knowledge driven, approach.

Our proposed model has the benefit of not requiring additional effort in generating or collecting data for multitask training. This allows learning from large-scale corpora in an unsupervised manner while simultaneously applying transfer learning. In contrast to general purpose sentence embeddings, our model for learning sentence representations is less complex and requires less training effort, while at the same time yields similar or higher performance in our target task. We have shown that there are benefits in adopting guided unsupervised learning during embedding pre-training instead of overemphasis on universal applications.

While we do expect that further improvements can be obtained through better labels for the multitask objective, that would entail additional effort in system design and label generation while not undermining our conclusions. In addition, we also expect that multitask labels that are too domain-specific (*e.g.* focusing on a specific way or definition of affective expression) may actually hinder the performance of unsupervised embeddings.

Chapter 5

Multimodal Approaches to Modeling Behavior

In this chapter I explore various multimodal approaches for modeling human behavior. I propose two types of models that combine acoustic features with spoken language for the selected task of emotion recognition. I then show that unsupervised pretraining on spoken language modeling enables transfer learning of mutlimodal fusion improving the performance of emotion recognition over other systems.

5.1 Introduction

Acoustic and visual elements in human communication, such as vocal intonation and facial expressions, incorporate semantic information and paralinguistic cues conveying intent and affect [81]. For this reason many multimodal systems have been proposed which integrate information from multiple modalities to improve natural language understanding. This effort has many applications such as in video summarization [82], [83], dialogue systems [84], [85], and emotion and sentiment analysis [86]–[89].

The study of multimodal fusion in affective systems is an especially prevalent and important topic. This follows from the fact that human behavioral expression is fundamentally a multifaceted phenomenon that manifests over multiple modalities [90] and can be more accurately identified through multimodal models [91]. Another factor is the importance of affective information as an

ingredient in a variety of downstream tasks such as in language modeling [92], dialogue system design [93], [94], and video summarization [95].

Many multimodal systems for recognition of sentiment, emotion, and behaviors have been proposed in prior work, including recent neural network based approaches. In feature-level fusion, Tzirakis et al. [96] combined auditory and visual modalities by extracting features using convolutional neural networks (CNN) on each modality which then were concatenated as input to an LSTM network. Hazarika et al. [97] proposed the use of a self-attention mechanism to assign scores for weighted combination of modalities. Other works have applied multimodal integration using late fusion methods [98], [99].

For deeper integration between modalities many have proposed the use of multimodal neural architectures. Lee et al. [100] have proposed the use of an attention matrix calculated from speech and text features to selectively focus on specific regions of the audio feature space. The memory fusion network was introduced by Zadeh et al. [101] which accounted for intra- and inter-modal dependencies across time. Akhtar et al. [102] have proposed a contextual inter-modal attention network that leverages sentiment and emotion labels in a multi-task learning framework.

The strength of deep models arises from the ability to learn meaningful representations of, and association between, features from multiple modalities. This is learned implicitly by the model in the course of training [11]. In this work we propose a model to explicitly learn informative joint representations of speech and text. This is achieved by modeling the dynamic relations between lexical content and acoustic paralinguistics through a language modeling task on spoken language. We augment a bidirectional language model (biLM) with word-aligned acoustic features and optimize the model first using large-scale text corpora, and then followed by speech recordings. We evaluate the effectiveness of representations extracted from this model in encoding multimodal information on the task of emotion recognition on the emotion datasets IEMOCAP [71], MSP-IMPROV [103], and CMU-MOSEI [104].

5.2 Related Work

Lexical representations such as ELMo [105] and BERT [106] have recently been shown to model word semantics and syntax with greater efficacy. This is achieved through contextualized learning on large-scale language corpora which allows internal states of the model to capture both the complex characteristics of word use as well as polysemy due to different contexts. The integration of these word embeddings into downstream models has improved the state of the art in many NLP tasks through their rich representation of language use.

To learn representations from multimodal data Hsu et al. [107] proposed the use of variational autoencoders to encode inter- and intra-modal factors into separate latent variables. Later, Tsai et al. [108] factorized representations into multimodal discriminative and modality-specific generative factors using inference and generative networks. Recent work by Rahman et al. [109] has concurrently proposed the infusion of multimodal information into the BERT model. There the authors have combined the generative capabilities of the BERT model with a sentiment prediction task to allow the model to implicitly learn rich multimodal representations through a joint generative-discriminative objective.

In this chapter we propose to explicitly learn multimodal representations of spoken words by augmenting the biLM and BERT models with acoustic information. This is motivated from how humans integrate acoustic characteristics in speech to interpret the meaning of lexical content from a speaker. Our work differs from prior work in that we do not include or target any discriminative objectives and instead rely on generative tasks to learn meaningful multimodal representations. We adopt the ELMo and BERT architecture for its use of a language modeling task and explore methods of injecting acoustic information during language understanding. We show how these model can be easily trained with large-scale unlabeled data and also demonstrate the usefulness of the resulting multimodal embeddings in an example task of emotion recognition.

5.3 Multimodal Embeddings from Language Models

We extract multimodal embeddings through the use of a bidirectional language model (biLM) infused with acoustic information. The biLM comprises stacked layers of bidirectional LSTMs which operate over lexical and audio embeddings. The lexical and audio embeddings are calculated from respective convolutional layers and combined using a sigmoid-gating function. Multimodal embeddings are then computed using a linear function over the internal states of the recurrent layers. The architecture of the mutlimodal biLM is shown in Figure 5.1.

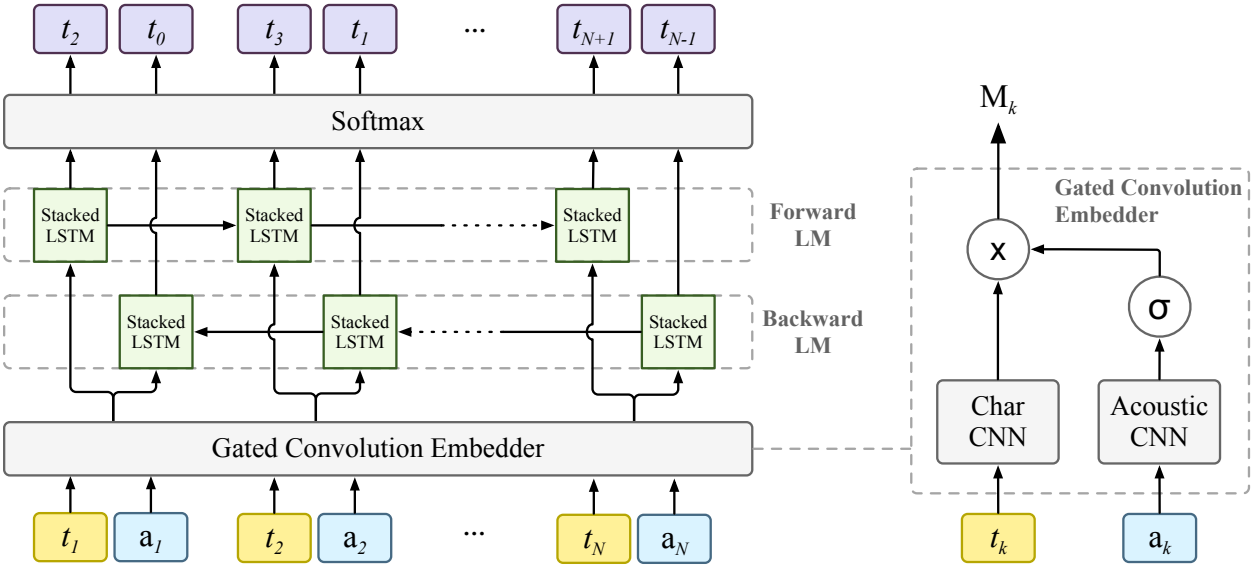


Figure 5.1: Architecture of the multimodal bidirectional language model

5.3.1 Bidirectional language model

A language model (LM) computes the probability distribution of a sequence of words by approximating it as the product of conditional probabilities of each word given previous words. This has been implemented using neural networks in many prior work yielding state of the art results [110]. In this work we applied the biLM model used in ELMo, which is based on the character-level RNN-LM [111].

The biLM is composed of a forward and backward LM each implemented by a multi-layer LSTM. The forward LM predicts the probability distribution of the next token given past context while the backward LM predicts the probability distribution of the previous token given future context. Each LM operates on the same input, which is a token embedding of the current token calculated through a character-level convolutional neural network (CharCNN) [112]. A softmax layer is used to estimate token probabilities from the output of the two-layer LSTM in the LMs. The parameters of the softmax layer are shared between the LMs in both directions.

Different from ELMo, our input to the biLM includes acoustic features in addition to word tokens. Now the forward LM aims to model, at each time step, the conditional probability of the next token t_{k+1} given the current token t_k, acoustic features $\mathbf{a}_k$, and previous internal states of the stacked LSTM $\vec{\mathbf{s}}_{k-1}$:

$$P(t_{k+1} \mid t_k, \mathbf{a}_k, \vec{\mathbf{s}}_{k-1}) \tag{5.1}$$

The backward LM operates similarly but predicts the previous token t_{k-1} given the current token t_k, acoustic features $\mathbf{a}_k$, and internal states resulting from future context $\overleftarrow{\mathbf{s}}_{k+1}$. Details of the acoustic features are given in Section 5.5.1.

5.3.2 Acoustic convolution layers

To integrate paralinguistic information into the language model, time-aligned acoustic features of each word are provided in adjunct to word tokens. We add additional convolutional layers at the input of the biLM to compute acoustic embeddings from the acoustic features. The convolutional layers provide a feature transformation of the acoustic features which are then combined with token embeddings using a gating function.

Due to the varying duration of spoken words, acoustic features are zero-padded to a fixed frame size before being passed to the CNN. This is similar to the use of a maximum number of characters per word in the CharCNN. The acoustic CNN is implemented by series of 1-D convolution layers each followed by a max-pooling layer. The final feature map is then projected to the same

dimension size as token embeddings to allow for element-wise combination. The architecture of the acoustic CNN is shown in Figure 5.2.

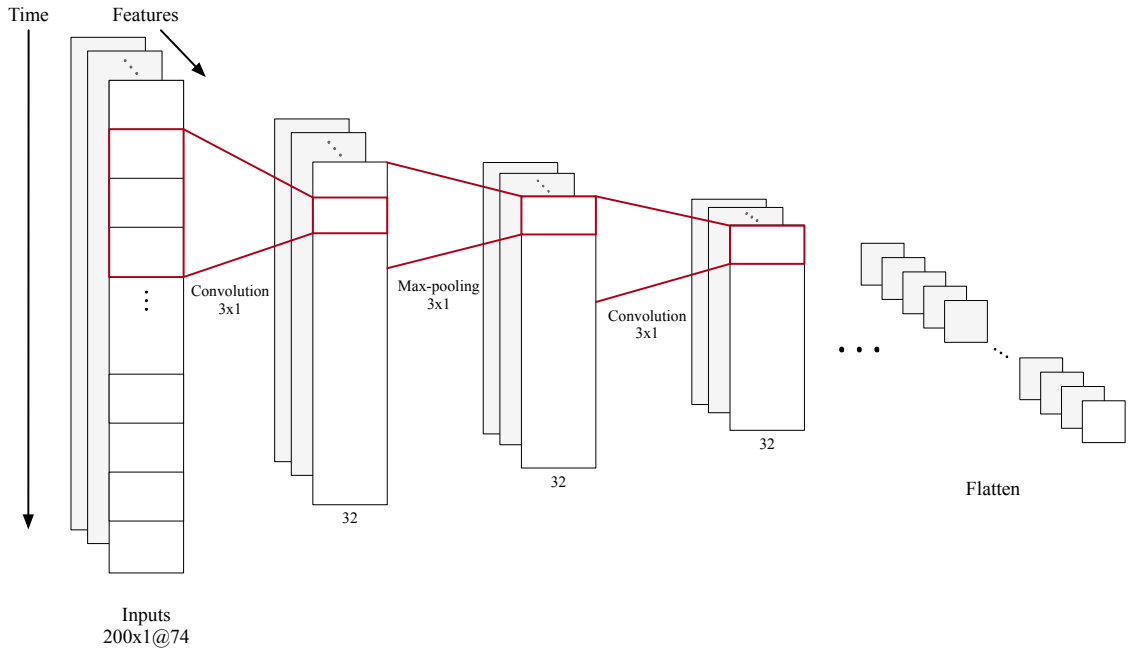


Figure 5.2: Architecture of the acoustic CNN

5.3.3 Extracting Multimodal Embeddings

We combine token and acoustic embeddings using a sigmoid gating function:

$$\mathbf{M}_k = \mathbf{U}(t_k) \odot \sigma(\mathbf{V}(\mathbf{a}_k)) \tag{5.2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the embeddings calculated from the token and corresponding acoustic features, respectively, σ is the sigmoid function, and $\odot$ represents element-wise multiplication. The sigmoid gate is a useful mechanism in language modeling [113] as it allows the network to select relevant features in the token embedding. In our case it serves to modify semantic meaning of words through scaling of the token embedding based on acoustic information. The embeddings after the gated sigmoid function are considered to be multimodal and are used as input to both the forward and backward LM.

Word embeddings are extracted for use in downstream models in a similar fashion to ELMo. That is, we define each word vector as a task-specific weighted sum of all LSTM outputs as well as the input token embedding $\mathbf{M}_k$. We additionally average over all word vectors in a sentence to form sentence embeddings for use in downstream models. A similar approach of obtaining sentence embeddings from a weighted average of word vectors was shown in [114] to be surprisingly effective in many NLP tasks.

The final multimodal ELMo (M-ELMo) sentence embedding is given as

$$\textbf{M-ELMo} = \gamma \frac{1}{N} \sum_{k=1}^{N} \sum_{j=0}^{L} c_j \mathbf{h}_{k,j} \tag{5.3}$$

where $\mathbf{h}_{k,j}$ are the concatenated outputs of LSTMs in both directions at the j^{th} layer for the k^{th} token and $j = 0$ corresponds to the input to the LSTM. Values $\{c_j\}$ are softmax-normalized weights for each layer and γ is a scalar value, all of which are task-specific and tunable parameters in the downstream model. An architectural overview of the downstream model for speaker emotion recognition is shown in Figure 5.3.

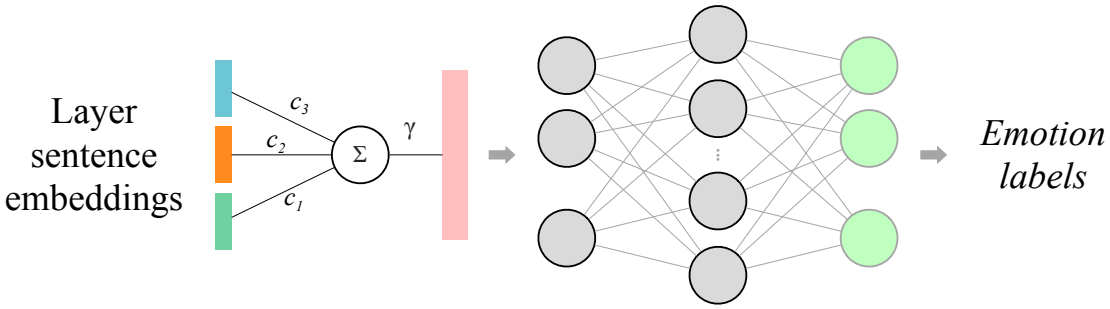


Figure 5.3: DNN model for emotion recognition using the multimodal embeddings

5.4 Multimodal Transformers

Recent work has shown the benefits of using attention in the form of Transformers for natural language processing applications [115]. One such model, BERT [106], demonstrates the importance of pre-training bidirectional language representations using a language modeling-based task. A

significant contribution in their work is the use of Transformer blocks in their task of masked language modeling as well as the addition of multitask pre-training followed by fine-tuning. In this section I describe our method of extending our previous multimodal ELMo model to a Transformer based implementation and the addition of multitask learning.

5.4.1 Masked language modeling using attention layers

Attention layers combine information from all time steps of an input sequence concurrently irrespective of position. A major advantage of this mechanism is that this allows model layers to be processed in a single step which allows for efficient parallel processing. This is opposed to the biLM which requires unfolding of the recurrent structure followed by back-propagation through time. An effect of this method is that prediction of the next token as derived in Equation 5.1 is no longer applicable since the attention weights operate with the entire view of the input sequence. To overcome masked language modeling was applied as in [106], however in our case we additionally include acoustic information as another sequence type. The pretraining task thus becomes prediction of a masked-out word given surrounding words and acoustic features, as shown in Figure 5.4.

To implement the multimodal Transformer we replaced the bidirectional LSTMs described in Section 5.3.1 with Transformer layers. Now instead of recurrent pathways operating across the sequence the attention layers combine information from the previous layer in a single pass. A consequence of this is that tokens are now without order meaning there is little need for the modalities to be aligned. With that in mind we used separate CNNs for each modality and passed the resulting embeddings to the Transformer layers without the use of a gated-sigmoid function as in mutlimodal ELMo. To maintain token order and modality information we added a modality and positional encodings after respective convolutional layers for each modality. The arhcitecture of the multimodal Transformer is shown in Figure 5.5.

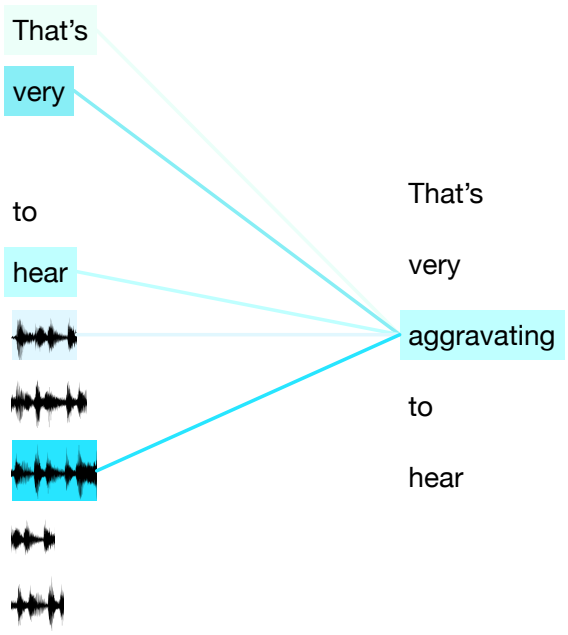


Figure 5.4: Example of multimodal masked language modeling

5.5 Experimental Setup

5.5.1 Features

Since a CharCNN is used as the lexical embedder, input words to the biLM are first transformed into a character map and padded to a fixed length. The character-level representation of each word is then given as a $c \times l_c$ matrix, where c is the dimension size of the character embedding and l_c is the maximum number of characters in a word. For the Transformer model the original word tokens are used.

We used acoustic features extracted using COVAREP (v1.4.2) [116] similar to [102], [104]. There are 74 features in total and include, among others, pitch, voiced/unvoiced segment features, mel-frequency cepstral coefficients, glottal flow parameters, peak slope parameters, and harmonic model parameters.

The acoustic features are aligned with word timings to provide acoustic information for each word. Since the time duration varies between words we pad the number of acoustic frames per token to a fixed length. Thus, word-aligned acoustic features are given as a $d \times l_a$ matrix, where

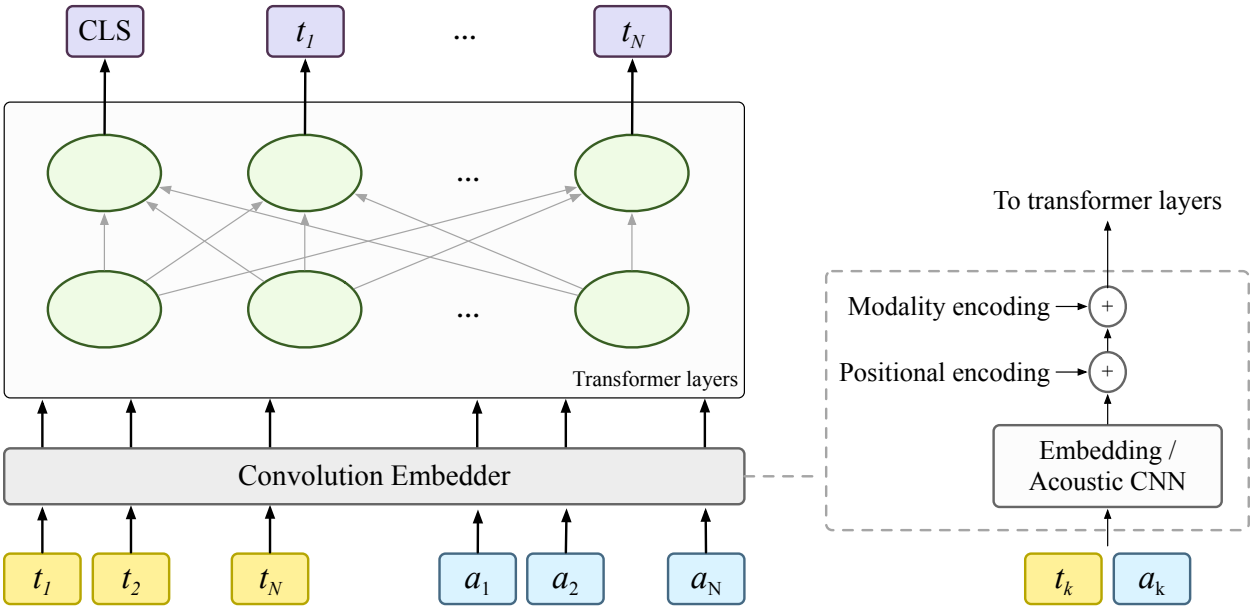


Figure 5.5: Architecture of the multimodal Transformer model

d is the number of acoustic features and l_a is the maximum number of acoustic frames in a word. In our experiments we used a maximum frame length of 2 seconds per word. This corresponds to more than 99.9% of all words in the dataset. We assume any truncated words to be unrepresentative of conventional articulation during conversations (*e.g.* purposely drawn-out words) which may require specific modeling outside the scope of normal interactions.

5.5.2 Multimodal biLM architecture parameters

For the lexical and recurrent components of the multimodal biLM we used the same model architecture as the final model in [105]. This model comprises a character CNN with 2048 character n-gram convolutional filters followed by a two-layer biLSTM ($L = 2$) with 4096 units and a projection size of 512 dimensions.

The architecture of the acoustic CNN is inspired by keyword spotting CNNs proposed in [117], however we applied 1-D convolution since our acoustic features include non-spatial categories. We also used a smaller kernel size in the time dimension to model acoustic variations at a finer scale.

The acoustic CNN comprises three 1-D convolutional layers using kernels of size 3 and a stride of 1. Each layer is followed by a max-pooling function over three frames.

5.5.3 Pre-training the multimodal biLM

The multimodal biLM is pre-trained in two stages. In the first stage the lexical components of the biLM are optimized prior to the inclusion of acoustic features. This is achieved by training on a text corpus and fixing the acoustic input as zero. We use the 1 Billion Word Language Model Benchmark [118] and train the biLM for 10 epochs. After training, the model achieves perplexities of around 35 which is similar to values reported in pretrained models from [105].

In the second stage of pre-training we optimize the biLM using the multimodal dataset CMU-MOSEI (described in Section 5.5.6). In our experiments we use text and audio which are not in the testing split of the dataset to train the biLM. In terms of word count CMU-MOSEI contains around 447K words which is much smaller than the 1-billion word LM benchmark. Therefore, to prevent over-fitting we reduce the learning rate used in the previous stage by a factor of 10 and train for an additional 5 epochs. After pretraining, the we extract multimodal sentence embeddings for use in downstream models.

5.5.4 Multimodal Transformer architecture parameters

For the multimodal Transformer we applied a configuration based on the `bert-base-uncased` model from [106]. The model is built using 12 Transformer layers with 12 attention heads in each layer. The size of the hidden layers is 768 dimensions. For the convolutional embedder the same acoustic CNN structure was used as in Section 5.5.2. For the lexical tokens WordPiece embeddings [119] were used to generate the token embeddings. We use positional encodings representing up to 512 positions and 2 types of modality encodings.

5.5.5 Multitask learning on the multimodal Transformer

The first output token of the Transformer layers is defined as a pooled output of the input sequence **CLS** while the remaining are defined as sequence outputs $\{t_i\}$. By applying additional layers on top of these two outputs we can train the mutlimodal Transformers using multiple tasks. In this thesis we apply linear layers on the **CLS** and $\{t_i\}$ outputs for the task of emotion recognition and masked language modeling, respectively.

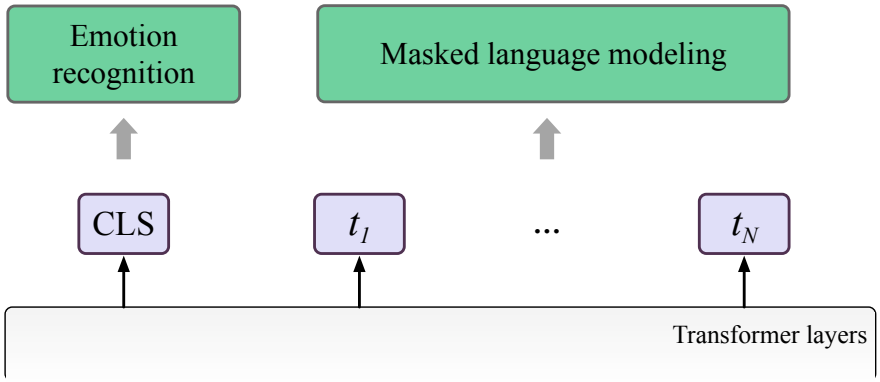


Figure 5.6: Training the multimodal Transformer model with multitask

5.5.6 Emotion recognition as a downstream task

In our experiments we adopt emotion recognition as the downstream task and evaluate on the datasets IEMOCAP, MSP-IMPROV, and CMU-MOSEI.

CMU-MOSEI contains 23,453 single-speaker video segments from YouTube which have been manually transcribed and annotated for sentiment and emotion. Emotions are annotated on a [0,3] Likert scale and include categories such as *happiness*, *sadness*, *anger*, *fear*, *disgust*, and *surprise*. We binarize these annotations to arrive at class labels by predicting the presence of emotions, *i.e.* any emotion with a rating greater than one. Since video segments have ratings for all emotions this becomes a multi-label classification task.

IEMOCAP is a multimodal emotion dataset consisting of 10 actors displaying emotion in scripted and improvised hypothetical scenarios. This dataset was described in Section 4.4.1.3

however following more recent work [120] we evaluate on 7 emotion categories: *angry* (1103), *excite* (1041), *happy* (595), *sad* (1084), *frustrated* (1849), *surprise* (107), and *neutral* (1708) for a total of 7487 utterances. We report results from 10-fold cross validation where in each fold one speaker is used as the test set while their conversation partners and remaining speakers are used as the development and training set, respectively.

MSP-IMPROV is another collection of emotional audiovisual recordings performed by actors. However in this dataset the authors strove to increase naturalness by using target sentences within improvised conversational scenarios. Additional segments of natural interaction of actors during breaks were also included. We evaluate on the emotion categories *angry* (789), *happy* (2603), *sad* (882), *neutral* (3340), for a total of 7714 utterances. The dataset consists of 6 sessions between male and female pairs which leads to 12-fold cross validation.

We trained the network using data from the training split provided in each dataset and validated using the validation split. We also used the validation split as a development set in choosing hyper-parameters of the network.

5.5.7 Evaluation methods

We evaluated the emotion recognition model using weighted accuracy (WA) and F1 score on each emotion. Weighted accuracy [104] refers to the macro-average recall value of a multi-class problem. We also averaged the metrics across all emotions to obtain an average WA and F1 score. F1 scores are weighted by support to account for label imbalance. The downstream model was trained for 30 epochs using binary cross-entropy loss for each individual class. The best model was selected based on the average WA and F1 across all emotions using the validation set. The final model was a neural network with two hidden layers using *Tanh* activation functions.

Due to the lack of work on CMU-MOSEI which focuses on text and audio only, we compared with two recent state of the art emotion recognition models that additionally consider the visual modality. Specifically, these are the graph Memory Fusion Network (Graph-MFN) [104] and the contextual inter-modal attention framework (CIM-Att) [102]. To match learning conditions, we

compared with the single task learning (STL) model of [102] where only emotion labels are used in training. As a baseline we also compared to various single modality models, including a neural network model using sentence embeddings from a fine-tuned ELMo model.

Table 5.1: Emotion recognition results on CMU-MOSEI test set for various multimodal models

	Anger		Disgust		Fear		Happy		Sad		Surprise		**Average**	
Method	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	**WA**	**F1**
(Single modality)														
Acoustic [104]	56.4	71.9	60.9	72.4	62.7	89.8	61.5	61.4	62.0	69.2	54.3	85.4	59.6	75.0
Lexical [104]	56.6	71.8	64.0	72.6	58.8	89.8	54.0	47.0	54.0	61.2	54.3	85.3	57.0	71.3
ELMo + NN‡	64.4	75.4	73.6	82.4	61.8	86.0	65.4	65.1	60.1	71.8	62.5	84.7	64.6±0.3	77.6±0.3
(A + L + V)														
Graph-MFN [104]	62.6	72.8	69.1	76.6	62.0	89.9	66.3	66.3	60.4	66.9	53.7	85.5	62.3	76.3
CIM-Att-STL [102]	64.5	75.6	72.2	81.0	51.5	87.7	61.6	59.3	65.4	67.3	53.0	86.5	61.3	76.2
(A + L)														
CIM-Att-STL [102]	-	-	-	-	-	-	-	-	-	-	-	-	59.6	76.8
M-ELMo + NN	63.9	75.7	72.3	81.7	57.4	84.8	67.2	66.6	61.2	72.1	61.4	85.0	65.0	77.6
M-BERT	-	-	-	-	-	-	-	-	-	-	-	-	**66.6**	-

Modalities: acoustic (A), lexical (L), visual (V).

Table 5.2: Emotion recognition results on IEMOCAP

Model	**Modality**	**UA (%)**
MDRE [86]	Audio + Lexical	53.6
MHA [121]	Audio + Lexical	55.5
ELMo + NN	Lexical	49.32
M-ELMo + NN	Audio + Lexical	52.3

Table 5.3: Emotion recognition results on MSP-IMPROV

Model	**Modality**	**UAR**
[103]	Audio	41.4
[122]	Audio	52.6
[123]	Audio	44.4
[124]	Audio	52.43
ELMo + NN	Lexical	50.70
M-ELMo + NN	Audio + Lexical	53.94

5.6 Results and Analysis

The results of emotion recognition on CMU-MOSEI are shown in Table 5.1. Our feedforward neural network using multimodal embeddings shows improvement in terms of average WA and F1 over all emotions at 65.0% and 77.6%, respectively. While the mutlimodal Transformer model achieved a weighted accuracy of 66.6%. The improvement of M-ELMo over ELMo embeddings is significant with p-value < 0.05 under McNemar's test. Surprisingly, a neural network using ELMo embeddings led to higher performance than other advanced models using multiple modalities. This result demonstrates the effectiveness of unsupervised contextualized embeddings over other methods such as GloVe [13], which were used in [102] and [104].

By observing the mixing weights $\{c_j\}$ for each layer in ELMo and M-ELMo we can highlight the importance of syntactic (lower layers) versus semantic (higher layers) information in each model [105]. An analysis of the distribution of mixing weights is shown in Figure 5.7. We observed that models using M-ELMo focused more on the lower layers compared to those using ELMo. This could be due to the importance of semantics over syntax in text-based emotion recognition, whereas in multimodal configurations paralinguistic information is more dominant in lower layers which can aid in improving recognition ability.

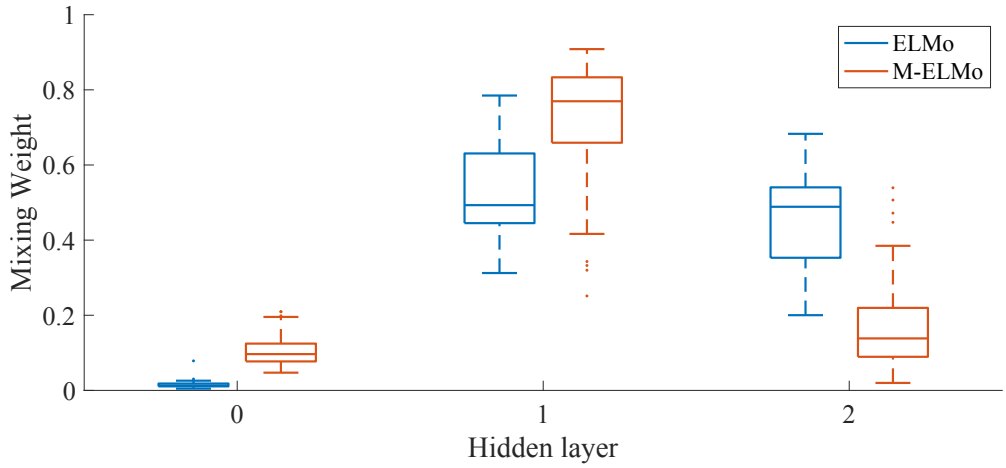


Figure 5.7: Distribution of scalar weight for each layer calculated over 50 runs.

5.7 Conclusion

In this chapter I proposed a method for extending ELMo and BERT embeddings to include acoustic information. Convolutional layers were used on acoustic features to calculate acoustic embeddings. These were then combined with token embeddings using a sigmoid gating function the ELMo model and with attention layers in the BERT model. The multimodal ELMo model was trained on a language modeling task, first with a text corpus followed by inclusion of audio from a multimodal dataset. And the mutlimodal BERT model was trained using multitask on emotion recognition and masked language modeling. I then showed the effectiveness of sentence embeddings extracted from this multimodal biLM on emotion recognition as well as the direct improvements in multimodal BERT. The results are promising especially given that our downstream model using a neural network with two hidden layers outperform state of the art architectures. This demonstrates that such multimodal approaches are effective in capturing inter- and intra-modal dynamics of lexical content and paralinguistic cues in spoken language.

Chapter 6

Summary, Future Work, and Closing Words

6.1 Summary

6.1.1 Behavior annotation using recurrent networks

In the first part of this thesis I presented a neural network framework for rating human behaviors in recorded sessions from couples therapy. The neural network was composed of recurrent layers to handle sequences of words obtained from a sliding window operating over transcripts of the therapy sessions. Each observation window, or frame, covers only a portion of words within the session, however behaviors ratings take into account behaviors exhibited throughout the entire interaction. The labels are therefore considered coarse-grained with respect to the frames and not always representative of the ground-truth. This constitutes as a weakly supervised learning problem. I tackle the issue of weak labels by directly assigning the session-level rating as targets for all frames contained in respective sessions. A fusion layer using a Support Vector Regressor is then applied on the median of frame-level scores in a session to map back to a predicted behavior rating. Evaluating on behavior annotation of couples therapy sessions, I show that such a framework can generate behavior ratings with high inter-annotator agreement to trained human annotators. In addition, experiments show improvement of predicted ratings over outlier human annotators. The experimental results are a promising step towards more advanced automated behavior annotation

systems and showcases the applicability of deep models to behavior signal processing for improved performance.

6.1.2 Unsupervised learning of deep sentence embeddings

In the next part of this thesis I presented multiple methods to further improve the previously described framework using unsupervised learning for domain transfer. The first approach addressed the issue of limited context within frames in the previous model by expanding the observation window to cover entire sentences instead of multiple words. To improve convergence and increase generalizability of the model sentence embeddings from a neural conversation model were applied as input features. The conversation model offers an unsupervised method for encoding sentences into representations which maintain information on conversational content, which were proved experimentally to be beneficial in behavior annotation.

In the second approach I explored methods of gaining domain relevance in unsupervised sentence embeddings. This was achieved through the inclusion of an additional task of sentiment analysis during learning of sentence embeddings using a multitask framework. To show application of this technique in an unsupervised scenario, labels for the task of sentiment classification were generated online using a naive look-up table approach. I then experimentally showed that such online multitask unsupervised sentence embeddings showed improved performance in behavior recognition and annotation on our in-domain data.

6.1.3 Multimodal embeddings

In the final portion of this thesis I explored methods of aggregating information from speech-based modalities for behavior signal processing. This entailed the combination of acoustic features and lexical content extracted from the spoken interactions. To achieve this I proposed a RNN-LM based multimodal model for fusing the two modalities. The bidirectional RNN-LM model included additional convolutional layers at the input to process acoustic feature inputs. Acoustic embeddings generated from these convolutional layers were then combined with lexical embeddings through

a sigmoid gating function. The main novelty of my contribution was to show that by pre-training this model on a task of spoken language modeling, the multimodal RNN-LM could be trained to combine embeddings from multiple modalities in a way that was suited for the downstream task of speaker emotion recognition.

6.2 Future Work

6.2.1 Advanced methods of modality fusion

The multimodal fusion methods described in the latter part of this dissertation combine modality embeddings at the initial stage of the model, either through sigmoid-gating or attention. This method of early fusion attempts to model spoken words as a joint representation of acoustic and lexical meaning. However, it may be beneficial if we allow the model to form higher levels of representation from each modality before joint modeling of the embeddings. This way the fusion of information is based on a more processed transformation of the input. There are various ways that this can be achieved.

One method is through the addition of heads in the attention layers. In the multi-head attention formulation multiple attention units operate on the input and the outputs from all the attention units are concatenated to form the output. By adding additional attention heads and restricting the input to a certain modality through masking we can allow the model to form high levels of abstraction based on only on a single modality. The outputs from these single-modal attention heads can be combined in layer layers of the Transformer. A multimodal attention path can be maintained in conjunction with the single-modal path which gradually incorporates information at each layer. An example architecture of individual attention heads per modality is shown in Figure 6.1.

6.2.2 Incorporating additional modalities and tasks

In this thesis I have focused primarily on speech-based modalities but many behavioral cues are expressed through visual means, such as facial expressions, gestures, and overall body language

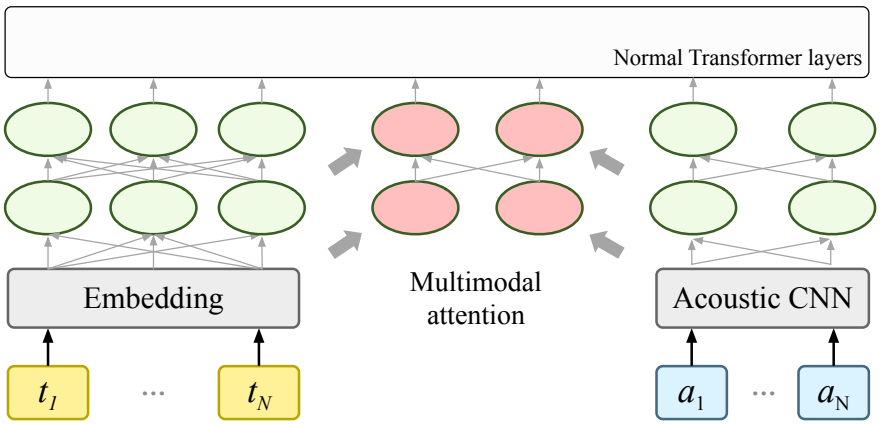


Figure 6.1: Individual attention heads per modality

(posture, orientation, *etc*). Computer vision technologies for facial recognition and pose detection *etc*. are currently quite mature and many well-performing implementations have been proposed. However, study into multimodal fusion of all these modalities for affective computing or behavior annotation is an ongoing endeavor. A possible approach is the exploration of unsupervised pretraining methods for learning how to interpret each modality individually as well as in conjunction with others. Data containing combinations of these modalities may be abundant, however the implementation of efficient pretraining procedures with or without labels, as well as design of fusion models for various downstream applications is a challenging future direction.

6.3 Closing Words

Throughout this dissertation I have focused on developing neural models that strive to recognize and annotate human behavior with performance and capabilities exceeding that of human annotators and observers. To this end I explored many representation learning techniques that take advantage of readily available out-of-domain data sources through unsupervised methods. I believe that by leveraging large-scale datasets neural models can be trained to efficiently encode not only contextual but also behavioral information from both natural language and paralinguistic cues. I have shown that such behavioral information encoded in unsupervised embeddings can be used effectively for behavior annotation in Couples Therapy and speaker emotion recognition. I

acknowledge that many applications of behavior signal processing exist and whether such methods can be applied from theory to real-world implementations for the benefit of users is a feat I hope to witness.

References

[1] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[2] H. Badr, "New frontiers in couple-based interventions in cancer care: Refining the prescription for spousal communication," *Acta Oncologica*, vol. 56, no. 2, pp. 139–145, 2017.

[3] B. R. Baucom, P. Georgiou, C. J. Bryan, E. L. Garland, F. Leifker, A. May, *et al.*, "The promise and the challenge of technology-facilitated methods for assessing behavioral and cognitive markers of risk for suicide among us army national guard personnel," *International journal of environmental research and public health*, vol. 14, no. 4, p. 361, 2017.

[4] R. F. Baumeister, K. D. Vohs, C. N. DeWall, and L. Zhang, "How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation," *Personality and Social Psychology Review*, vol. 11, no. 2, pp. 167–203, 2007, PMID: 18453461. DOI: 10.1177/1088868307301033.

[5] C. Heavey, D. Gill, and A. Christensen, "Couples interaction rating system 2 (cirs2)," *University of California, Los Angeles*, vol. 7, 2002.

[6] G. Margolin, P. H. Oliver, E. B. Gordis, H. G. O'hearn, A. M. Medina, C. M. Ghosh, *et al.*, "The nuts and bolts of behavioral observation of marital and family interaction," English, *Clinical child and family psychology review*, vol. 1, no. 4, pp. 195–213, 1998.

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

[8] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling.," in *Proceedings of Interspeech*, 2012, pp. 194–197.

[9] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[10] S.-Y. Tseng, S. N. Chakravarthula, B. Baucom, and P. Georgiou, "Couples behavior modeling and annotation using low-resource LSTM language models," in *Proceedings of Interspeech*, San Francisco, CA, 2016.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *In Proceedings of Workshop at ICLR*, 2013.

[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[14] S.-Y. Tseng, B. Baucom, and P. Georgiou, "Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings," in *Proceedings of Interspeech*, 2017.

[15] S.-Y. Tseng, B. Baucom, and P. Georgiou, "Unsupervised online multitask learning of behavioral sentence embeddings," *PeerJ Computer Science*, vol. 5, e200, 2019.

[16] P. G. Georgiou, M. P. Black, A. Lammert, B. Baucom, and S. S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" In *Proceedings of Affective Computing and Intelligent Interaction*, Memphis, TN, 2011.

[17] S. N. Chakravarthula, R. Gupta, B. Baucom, and P. Georgiou, "A language-based generative model framework for behavioral analysis of couples' therapy," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[18] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, *Innovations in Machine Learning: Theory and Applications - Neural Probabilistic Language Models*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, ch. Neural Probabilistic Language Models, pp. 137–186. DOI: 10.1007/3-540-33486-6_6.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2013, pp. 3111–3119.

[20] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[21] C. Heavey, D. Gill, and A. Christensen, "Couples interaction rating system 2 (cirs2)," *University of California, Los Angeles*, vol. 7, 2002.

[22] J. Jones and A. Christensen, "Couples interaction study: Social support interaction rating system," University of California, Los Angeles, Technical manual, 1998.

[23] M. P. Black, A. Katsamanis, C.-C. Lee, A. Lammert, B. R. Baucom, A. Christensen, *et al.*, "Automatic classification of married couples' behavior using audio features," in *Proceedings of InterSpeech*, 2010.

[24] M. Black, A. Katsamanis, B. Baucom, C. Lee, A. Lammert, A. Christensen, *et al.*, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, 2012.

[25] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[26] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations.," in *HLT-NAACL*, 2013, pp. 746–751.

[27] *Google news corpus*, `http://www.statmt.org/wmt14/training-monolingual-news-crawl/`, 2014.

[28] "General psychotherapy corpus," in *Alexander Street Press*, `https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series`, 2011.

[29] "Google word2vec toolkit," `https://code.google.com/archive/p/word2vec/`, 2015.

[30] S. N. Chakravarthula, R. Gupta, B. Baucom, and P. Georgiou, "A language-based generative model framework for behavioral analysis of couples' therapy," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2015.

[31] B. Xiao, D. Can, J. Gibson, Z. Imel, D. Atkins, P. Georgiou, *et al.*, "Behavioral coding of therapist language in addiction counseling using recurrent neural networks," in *Proceedings of Interspeech*, San Francisco, CA, 2016.

[32] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar, "A comparison of natural language processing methods for automated coding of motivational interviewing," *Journal of substance abuse treatment*, vol. 65, pp. 43–50, 2016.

[33] M. Boden, "A guide to recurrent neural networks and backpropagation," *the Dallas project*, 2002.

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

[35] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents.," in *ICML*, vol. 14, 2014, pp. 1188–1196.

[36] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, *et al.*, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 4, pp. 694–707, 2016.

[37] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[38] J. Tiedemann, "News from opus-a collection of multilingual parallel corpora with tools and interfaces," in *Recent advances in natural language processing*, vol. 5, 2009, pp. 237–248.

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

[40] J. Jones and A. Christensen, "Couples interaction study: Social support interaction rating system," University of California, Los Angeles, Technical manual, 1998.

[41] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[42] C. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69–78.

[43] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2015, pp. 959–962.

[44] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2249–2255.

[45] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[46] S. Venugopalan, L. A. Hendricks, R. Mooney, and K. Saenko, "Improving lstm-based video description with linguistic knowledge mined from text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1961–1966.

[47] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., 2014, pp. 3104–3112.

[48] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, *et al.*, "Skip-thought vectors," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 3294–3302.

[49] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, *et al.*, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 4, pp. 694–707, 2016.

[50] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 528–540. DOI: 10.18653/v1/N18-1049.

[51] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *International Conference on Learning Representations*, 2016.

[52] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[53] B. MacCartney and C. D. Manning, "Natural logic and natural language inference," in *Computing Meaning*, Springer, 2014, pp. 129–147.

[54] S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal, "Learning general purpose distributed sentence representations via large scale multi-task learning," in *International Conference on Learning Representations*, 2018.

[55] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[56] D. Klein and C. D. Manning, *The unsupervised learning of natural language structure*. Stanford University Stanford, CA, 2005.

[57] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1367–1377.

[58] J. Jurgovsky, M. Granitzer, and C. Seifert, "Evaluating memory efficiency and robustness of word embeddings," in *European conference on information retrieval*, Springer, 2016, pp. 200–211.

[59] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 160–167.

[60] S. J. Hwang and L. Sigal, "A unified semantic embedding: Relating taxonomies and attributes," in *Advances in Neural Information Processing Systems*, 2014, pp. 271–279.

[61] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2014, pp. 165–180.

[62] J. Yu and J. Jiang, "Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 236–246.

[63] M. Rei, "Semi-supervised multitask learning for sequence labeling," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 2121–2130.

[64] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

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

[66] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[67] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[68] H.-Y. Huang, C. Zhu, Y. Shen, and W. Chen, "Fusionnet: Fusing via fully-aware attention with application to machine comprehension," in *International Conference on Learning Representations*, 2018.

[69] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[70] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[71] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[72] J. Tiedemann, "News from OPUS - A collection of multilingual parallel corpora with tools and interfaces," in *Recent Advances in Natural Language Processing*, vol. V, Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia, 2009, pp. 237–248.

[73] M. P. Black, A. Katsamanis, B. Baucom, C.-C. Lee, A. Lammert, A. Christensen, *et al.*, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, pp. 1–21, 2013. DOI: doi:10.1016/j.specom.2011.12.003.

[74] S. N. Chakravarthula, R. Gupta, B. Baucom, and P. Georgiou, "A language-based generative model framework for behavioral analysis of couples' therapy," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2015.

[75] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[76] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *Proc. Interspeech 2018*, pp. 247–251, 2018.

[77] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[78] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1, no. 1, pp. 77–89, 2007.

[79] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, IEEE, 2015, pp. 4749–4753.

[80] K. W. Gamage, V. Sethu, and E. Ambikairajah, "Salience based lexical features for emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, IEEE, 2017, pp. 5830–5834.

[81] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," *Journal of Voice*, vol. 25, pp. 25–34, 2011.

[82] F. Nihei, Y. I. Nakano, and Y. Takase, "Fusing verbal and nonverbal information for extractive meeting summarization," in *Proceedings of the Group Interaction Frontiers in Technology*, ACM, 2018, p. 9.

[83] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for How2 videos," in *ACL*, 2019, pp. 6587–6596.

[84] L. Liao, Y. Ma, X. He, R. Hong, and T.-s. Chua, "Knowledge-aware multimodal dialogue systems," in *ACM-MM*, ACM, 2018, pp. 801–809.

[85] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, *et al.*, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in *ICASSP*, IEEE, 2019, pp. 2352–2356.

[86] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 112–118.

[87] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *SIGKDD*, ACM, 2018, pp. 350–358.

[88] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *EMNLP*, 2018, pp. 150–161.

[89] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, *et al.*, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the International Conference on Multimodal Interfaces*, [Ten Year Technical Impact Award, 2014 ICMI], 2004, pp. 205–211. DOI: 10.1145/1027933.1027968.

[90] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, pp. 1203–1233, 2013.

[91] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[92] P. G. Shivakumar, S.-Y. Tseng, P. Georgiou, and S. Narayanan, "Behavior gated language models," *arXiv preprint*, 2019.

[93] J. Pittermann, A. Pittermann, and W. Minker, "Emotion recognition and adaptation in spoken dialogue systems," *International Journal of Speech Technology*, vol. 13, no. 1, pp. 49–60, 2010.

[94] D. Bertero, F. B. Siddique, C.-S. Wu, Y. Wan, R. H. Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *EMNLP*, 2016, pp. 1042–1047.

[95] A. Singhal, P. Kumar, R. Saini, P. P. Roy, D. P. Dogra, and B.-G. Kim, "Summarization of videos by analyzing affective state of the user through crowdsource," *Cognitive Systems Research*, vol. 52, pp. 917–930, 2018.

[96] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[97] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *MIPR*, IEEE, 2018, pp. 196–201.

[98] S.-Y. Tseng, H. Li, B. Baucom, and P. Georgiou, "'Honey, I learned to talk': Multimodal fusion for behavior analysis," in *ICMI*, ACM, 2018. DOI: 10.1145/3242969.3242996.

[99] N. Blanchard, D. Moreira, A. Bharati, and W. J. Scheirer, "Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities," *arXiv preprint*, 2018.

[100] C. W. Lee, K. Y. Song, J. Jeong, and W. Y. Choi, "Convolutional attention networks for multimodal emotion recognition from speech and text data," *ACL*, 2018.

[101] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *AAAI*, 2018.

[102] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *NAACL-HLT*, 2019.

[103] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[104] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *ACL*, 2018.

[105] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, *et al.*, "Deep contextualized word representations," in *NAACL-HLT*, 2018, pp. 2227–2237.

[106] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[107] W.-N. Hsu and J. Glass, "Disentangling by partitioning: A representation learning framework for multimodal sensory data," *arXiv preprint*, 2018.

[108] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *ICLR*, 2019.

[109] W. Rahman, M. K. Hasan, A. Zadeh, L.-P. Morency, and M. E. Hoque, "M-BERT: Injecting multimodal information in the BERT structure," *arXiv preprint*, 2019.

[110] C. Gong, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu, "Frage: Frequency-agnostic word representation," in *NeurIPS*, 2018.

[111] R. Józefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint*, 2016.

[112] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[113] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 933–941.

[114] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[115] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[116] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies," in *ICASSP*, IEEE, 2014.

[117] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[118] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, *et al.*, "One billion word benchmark for measuring progress in statistical language modeling," in *Interspeech*, 2014.

[119] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[120] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 3362–3366.

[121] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2822–2826.

[122] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 2741–2745.

[123] J. Gideon, M. McInnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[124] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Interspeech*, 2019, pp. 3920–3924.