Multimodal Communication Decoder Network for Human Communication Comprehension

Amir Zadeh[1], **Paul Pu Liang**[1], **Soujanya Poria**[2], **Erik Cambria**[2], **Prateek Vij**[2], **Louis-Philippe Morency**[1]

[1]Language Technologies Institute, Carnegie Mellon University
[2]Nanyang Technological University

{abagherz,pliang,morency}@cs.cmu.edu, {sporia,cambria}@ntu.edu.sg, prateek@sentic.net

Abstract

Human face-to-face communication is a complex multimodal signal. We use words (language modality), gestures (vision modality) and changes in tone (acoustic modality) to convey our intentions. Humans easily process and understand face-to-face communication, however, comprehending this form of communication remains a significant challenge for Artificial Intelligence (AI). To model this form of communication, AI must understand each modality and the interactions between them that shape the communication. In this paper, we present a novel neural architecture for understanding human communication called the Multimodal Communication Decoder Network (MCDN). The main strength of our model comes from discovering interactions between modalities through time using a neural component called the Multi-attention Block (MAB) and storing them in the hybrid memory of a recurrent component called the Long-short Term Hybrid Memory (LSTHM). We perform extensive comparisons on six publicly available datasets for multimodal sentiment analysis, speaker trait recognition and emotion recognition. MCDN achieves new state-of-the-art results in all the datasets.

Introduction

Humans communicate using a highly complex multimodal signal. We employ three modalities in a coordinated manner to convey our intentions to another human: *language modality* (words, phrases and sentences), *vision modality* (gestures and expressions), and *acoustic modality* (changes in vocal tones) (Morency, Mihalcea, and Doshi 2011). Understanding this multimodal communication is trivial for humans; we do it subconsciously in the cerebrum of our brains everyday. However, giving artificial intelligence the capability to understand this form of communication the same way humans do, by incorporating all involved modalities, is a fundamental challenge to the field of Artificial Intelligence (AI). Giving AI the capability to understand human communication could potentially narrow the gap in computers' understanding of humans and opens new horizons for the creation of many intelligent entities.

The coordination between the different modalities in human communication introduces two major types of intra-modality and inter-modality dynamics (Zadeh et al. 2017). Intra-modality dynamics, called α-dynamics in this paper, refer to the dynamics within each modality independent of other modalities. For example, the arrangement of words in a sentence according to the generative grammar of the language (language modality), or the activation of facial muscles for the presentation of a smile (vision modality). Inter-modality dynamics, called β-dynamics in this paper, refer to the dynamics between modalities. The β-dynamics are divided into synchronous and asynchronous categories. An example of synchronous β-dynamics is the simultaneous co-occurrence of a smile with a positive sentence and an example of asynchronous β-dynamics is the delayed occurrence of a laughter after the end of sentence. For machines to understand human communication as outlined in this paper, they must be able to understand these α-dynamics and β-dynamics.

To model these dual dynamics in human communication, we propose a novel deep recurrent neural model called the Multimodal Communication Decoder Network (MCDN). MCDN is distinguishable from previous approaches in that it explicitly accounts for both α-dynamics and β-dynamics in the network architecture and continuously models both dynamics through time. In the MCDN, the α-dynamics within each modality is modeled using a Long-short Term Hybrid Memory (LSTHM) assigned to that modality. The hybrid memory also allows each modality's LSTHM to store important β-dynamics related to that modality. The β-dynamics are discovered at each recurrence time-step using a specific neural component called the Multi-attention Block (MAB). The MAB is capable of simultaneously finding multiple β-dynamics in each recurrence timestep. The MCDN resembles the mechanism of our brains for understanding communication, where different regions independently process and understand different modalities (Kuzmanovic et al. 2012; Sergent and Signoret 1992) – LSTHM in our pipeline – and are connected together using neural links for multimodal information integration (Jiang et al. 2012) – MAB in our pipeline. We benchmark MCDN by evaluating its understanding of different aspects of human communication. These aspects cover sentiment of speech, emotions conveyed by the speaker and various displayed speaker traits. We perform extensive experiments on 16 different attributes related to human communication on public multimodal datasets. Our approach shows state of the art performance in modeling human communication for all the datasets.

Related Work

Modeling multimodal human communication has been studied in past research. Previous approaches can be categorized as follows:

Removing Factor of Time: Studies have removed the time dimensions from β-dynamics (Poria et al. 2017; Pérez-Rosas, Mihalcea, and Morency 2013a; Wöllmer et al. 2013) in order to model co-occurrences of information across the modalities. In these models, each modality is summarized in a representation by collapsing the time dimension, such as averaging the modality information through time (Abburi et al. 2016). While these models are successful in understanding co-occurrences, the lack of alignment is a major flaw as these models cannot deal with multiple contradictory evidences, eg. if a smile and frown happen together in an utterance. Furthermore, these approaches cannot accurately model long sequences since the representation over long periods of time become less informative.

Early Fusion: Approaches have used multimodal input feature concatenation instead of modeling α-dynamics and β-dynamics explicitly. In other words, these approaches rely on generic models (such as Support Vector Machines or deep neural networks) to learn both α-dynamics and β-dynamics without any specific model design. This concatenation technique is known as early fusion. A shortcoming of these models is the lack of detailed modeling for α-dynamics, which in turn affects the modeling of β-dynamics, as well as causing overfitting on input data (Xu, Tao, and Xu 2013). In this paper, we compare to a number of these approaches (Wang et al. 2016; Poria et al. 2016). Often, these early fusion approaches remove the time factor as well (Zadeh et al. 2016; Morency, Mihalcea, and Doshi 2011). We additionally compare to a stronger recurrent baseline that uses early fusion while maintaining the factor of time.

Late Fusion: Late fusion methods learn different models for each modality and combine the outputs using decision voting (Wrtwein and Scherer 2017; Nojavanasghari et al. 2016). While these methods are generally strong in modeling α-dynamics, they have shortcomings for β-dynamics since these inter-modality dynamics are normally more complex than a decision vote. As an example of this shortcoming, if a model is trained for sentiment analysis using the vision modality and predicts negative sentiment, late fusion models have no access to whether this negative sentiment was due to a frowning face or a disgusted face.

Multi-view Learning: Extensions of Hidden Markov Models (Baum and Petrie 1966) and Hidden Conditional Random Fields (Quattoni et al. 2007; Morency, Quattoni, and Darrell 2007) have been proposed for learning from multiple different views (modalities) (Song, Morency, and Davis 2012; 2013). Extensions of LSTMs have also been proposed in a multi-view setting (Rajagopalan 2016).

MCDN is different from the first category since we model both α-dynamics and β-dynamics. It is differs from the second and third category since we explicitly model α-dynamics using a LSTHM for each modality as well as β-dynamics using the MAB. Finally, MCDN is different from the fourth category since it explicitly models α-dynamics and proposes more advanced temporal modeling of β-dynamics.

MCDN Model

In this section we outline our pipeline for human communication comprehension: the Multimodal Communication Decoder Network (MCDN). MCDN has two key components: Long-short Term Hybrid Memory (LSTHM) and Multi-attention Block (MAB). Long-short Term Hybrid Memory (LSTHM) is an extension of the Long-short Term Memory (LSTM) by reformulating the memory component to carry hybrid information. LSTHM is intrinsically designed for multimodal setups and each modality is assigned a unique LSTHM. LSTHM has a hybrid memory that stores α-dynamics of its assigned modality and β-dynamics related to its assigned modality. The component that discovers β-dynamics across different modalities is called the Multi-attention Block (MAB). The MAB first uses information from outputs of all LSTHMs at a timestep to regress coefficients for the outputs according to different β-dynamics. It then weights the output dimensions based on these coefficients and learns a neural β-dynamics code for LSTHMs to modify their hybrid memories. Figure 1 shows the overview of the MCDN. MCDN is differentiable end-to-end which allows the model to be learned efficiently using gradient decent approaches. In the next subsection, we first outline Long-short Term Hybrid Memory. We then proceed to outline Multi-attention Block and how the two components work together.

Long-short Term Hybrid Memory

Long-short Term Memory (LSTM) networks have been among the most successful models in learning from sequential data (Hochreiter and Schmidhuber 1997). The most important component of the LSTM is a memory which stores a dense representation of its input through time (in other words models the α-dynamics of each modality). In the Long-short Term Hybrid Memory (LSTHM) model, we seek to build a memory mechanism for each modality that in addition to storing α-dynamics, is also able to store the β-dynamics related to that modality. This allows the memory to function in a hybrid manner. We extensively discuss the hybrid factor in this section and show how the hybrid factor is derived in the next section where we introduce the Multi-attention Block.

The Long-short Term Hybrid Memory is formulated in Algorithm 1. We assume a set of M modalities to exist in the domain of the problem. Subsequently, M LSTHMs are built in the MCDN pipeline. For each modality $m \in M$, the input to the mth LSTHM is of the form $\mathbf{X}^m = \{x_1^m, x_2^m, x_3^m, \cdots, x_T^m \,; x_t^m \in \mathbb{R}^{d_{in}^m}\}$, where x_t^m is the input at time t and d_{in}^m is the dimensionality of the input of modality m. For example if $m = l$(anguage), we can use word vectors with $d_{in}^l = 300$ at each word step t. d_{mem}^m is the dimensionality of the memory of modality m. σ is the (hard-)sigmoid activation function and $tanh$ is the tangent hyperbolic activation function. $\oplus$ denotes vector concatenation and $\odot$ denotes element-wise multiplication. Similar to the LSTM model, i is the input gate, f is the forget gate, and o is the output gate. $\bar{c}_t^m$ is the proposed update to the hybrid memory $c_t^m \in \mathbb{R}^{d_{mem}^m}$ at time t. $h_t^m \in \mathbb{R}^{d_{mem}^m}$ is the time distributed output of each modality. The neural β-dynamics code z_{t-1} is the output of the Multi-attention Block at the previous time-step and is dis-

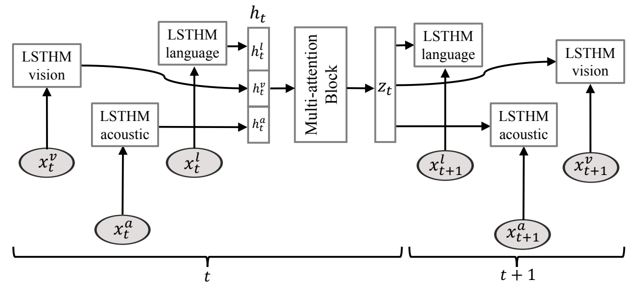

Figure 1: Overview of Multimodal Communication Decoder Network (MCDN) with Long-short Term Hybrid Memory (LSTHM) and Multi-attention Block (MAB) components, for $M = \{l, v, a\}$ representing the language, vision and acoustic modalities respectively.

cussed in detail in next subsection. This neural β-dynamics code z_{t-1} is passed to each of the individual LSTHMs and is the hybrid factor, allowing each individual LSTHM to integrate the relevant information from z_{t-1} to update their individual memory. In other words, z_{t-1} is the channel of communication between LSTHM memories and is the key step that allows each LSTHM to capture and store complex β-dynamics at each time step. To do so, the set of weights W_*^m,U_*^m and V_*^m respectively map the input of LSTHM x_t^m, output of LSTHM h_t^m, and neural β-dynamics code z_{t-1} to each LSTHM memory space using affine transformations.

Multi-attention Block

At each timestamp t, different β-dynamics across the modalities can occur simultaneously. For example, the first one can be the connection between a smile and positive phrase both happening at time t. A second one can be the occurrence of the same smile at time t being connected to an excited voice at time $t-4$, that was carried to time t using the audio LSTHM memory. As we can see, not only do β-dynamics span across all the modalities, they also span across the duration of the sequence, making asynchronous β-dynamics especially difficult to capture. To solve this problem, the Multi-attention Block is a network that can capture multiple different, possibly asynchronous, β-dynamics and encode all of them in a neural β-dynamics code z_t. In the most important step of the Multi-attention Block, different dimensions of LSTHM outputs h_t^m are assigned attention coefficients according to whether or not they form β-dynamics. These attention coefficients will be high if the dimension contributes to formation of a β-dynamics and low if they are irrelevant. The attention coefficient assignment is performed multiple times due to the existence of possibly multiple such β-dynamics across the outputs of LSTHM.

The Multi-attention Block is formulated in Algorithm 1. We assume a maximum of K β-dynamics to be present at each timestamp t. To obtain the K attention coefficients, K softmax distributions are assigned to the concatenated LSTHM memories using a deep neural network $\mathcal{A} : \mathbb{R}^{d_{mem}} \mapsto \mathbb{R}^{K \times d_{mem}}$ with $d_{mem} = \sum_{m \in M} d_{mem}^m$. At each timestep t, the output of LSTHM is the set $\{h_t^m : m \in M, h_t^m \in \mathbb{R}^{d_{mem}^m}\}$. $\mathcal{A}$ takes the concatenation of LSTHM

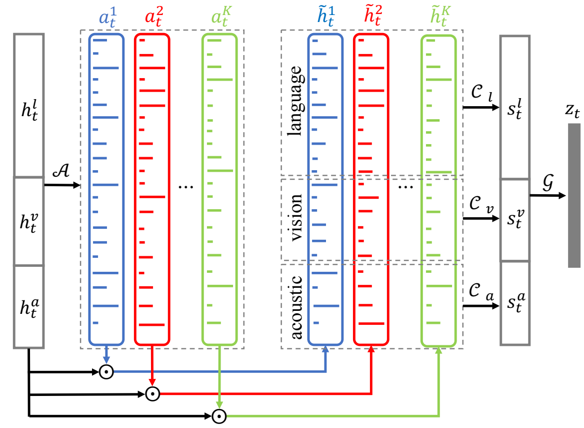

Figure 2: Overview of Multi-attention Block (MAB).

outputs $h_t = \bigoplus_{m \in M} h_t^m, h_t \in \mathbb{R}^{d_{mem}}$ as input and outputs a set of K attentions $\{a_t^k : k \le K,\ a_t^k \in \mathbb{R}^{d_{mem}}\}$ with $a_t = \bigoplus_{k=1}^K a_t^k,\ a_t \in \mathbb{R}^{K \times d_{mem}}$. $\mathcal{A}$ has a softmax layer at top of the network which takes the softmax activation along each one of the K dimensions of its output a_t. As a result, $a_t^k \ge 0, \sum_{i=1}^{d_{mem}} (a_t^k)_i = 1$ which forms a probability distribution over the output dimensions. h_t is then broadcasted (from $\mathbb{R}^{d_{mem}}$ to $\mathbb{R}^{K \times d_{mem}}$) and element-wise multiplied by the a_t to produce attended outputs $\widetilde{h}_t = \{\widetilde{h}_t^k : k \le K, \widetilde{h}_t^k \in \mathbb{R}^{d_{mem}}\}$,

Algorithm 1 Multimodal Communication Decoder Network (MCDN), Long-short Term Hybrid Memory (LSTHM) and Multi-attention Block (MAB) Formulation

procedure MCDN($\mathbf{X}^m$)
  $c_0, h_0, z_0 \leftarrow \mathbf{0}$
  for $t = 1, ..., T$ **do**:
    $h_t \leftarrow$ LSTHM_Step($\bigcup_{m \in M} \{x_t^m\}, z_{t-1}$)
    $z_t \leftarrow$ MAB_Step(h_t)
  return h_T, z_T
procedure LSTHM_STEP($\bigcup_{m \in M} \{x_t^m\}, z_{t-1}$)
  for $m \in M$ **do**: ◁ for all the M modalities
    $i_t^m \leftarrow \sigma(W_i^m\ x_t^m + U_i^m\ h_{t-1}^m + V_i^m\ z_{t-1} + b_i^m)$
    $f_t^m \leftarrow \sigma(W_f^m\ x_t^m + U_f^m\ h_{t-1}^m + V_f^m\ z_{t-1} + b_f^m)$
    $o_t^k \leftarrow \sigma(W_o^m\ x_t^m + U_o^m\ h_{t-1}^m + V_o^m\ z_{t-1} + b_o^m)$
    $\bar{c}_t^m \leftarrow W_{\bar{c}}^m\ x_t^m + U_{\bar{c}}^m\ h_{t-1}^m + V_{\bar{c}}^m\ z_{t-1} + b_{\bar{c}}^m$
    $c_t^m \leftarrow f_t^m \odot c_{t-1}^m + i_t^m \odot tanh(\bar{c}_t^m)$
    $h_t^m \leftarrow o_t^m \odot tanh(c_t^m)$
  $h_t \leftarrow \bigoplus_{m \in M} h_t^m$
  return h_t
procedure MAB_STEP(h_t)
  $a_t \leftarrow \mathcal{A}(h_t; \theta_{\mathcal{A}})$ ◁ K output coefficients
  $\widetilde{h}_t \leftarrow a_t \odot \langle \Uparrow_K h_t \rangle$
  for $m \in M$ **do**: ◁ calculating the β-dynamics
    $s_t^m \leftarrow \mathcal{C}_m(\widetilde{h}_t^m; \theta_{\mathcal{C}_m})$
  $s_t \leftarrow \bigoplus_{m \in M} s_t^m$
  $z_t \leftarrow \mathcal{G}(s_t; \theta_{\mathcal{G}})$
  return z_t

$\widetilde{h}_t \in \mathbb{R}^{K \times d_{mem}}$. $\Uparrow_K$ denotes broadcasting by parameter K.

The first dimension of $\widetilde{h}_t$ contains information needed for the first β-dynamics highlighted using a_t^1, the second dimension of $\widetilde{h}_t$ contains information for the second β-dynamics using a_t^2, and so on until K. $\widetilde{h}_t$ is likely extremely high dimensional and is considered sparse due to presence of dimensions with zero value after element-wise multiplication with attentions. Therefore, $\widetilde{h}_t$ is split into m different parts – one for each modality m – and undergoes dimensionality reduction using $\mathcal{C}_m : \mathbb{R}^{K \times d_{mem}^m} \mapsto \mathbb{R}^{d_{local}^m}, \forall m \in M$. The set of networks $\{\mathcal{C}_m : m \in M\}$ maps the attended outputs of each modality $\widetilde{h}_t^m$ to the same vector space. This dimensionality reduction produces a dense code s_t^m for the K times attended dimensions of each modality. Finally, the set of all M attended modality outputs, $s_t = \bigoplus_{m \in M} s_t^m$, are passed into a deep neural network $\mathcal{G} : \mathbb{R}^{\sum_{m \in M} d_{local}^m} \mapsto \mathbb{R}^{d_{mem}}$ to generate the neural β-dynamics code z_t at time t.

Experimental Methodology

In this paper we benchmark MCDN's understanding of human communication on three tasks: 1) multimodal sentiment analysis. 2) multimodal speaker traits recognition and 3) multimodal emotion recognition. We perform experimentations on six publicly available dataset and compare the performance of MCDN with the performance of state-of-the-art approaches on the same datasets. To ensure generalization of the model, all the datasets of this study are split in train, validation and test sets that include no identical speakers, i.e. all the speakers in the test set are new speakers. All baselines are re-trained on the same train/validation/test splits. To train the MCDN for different tasks, the final outputs h_T and neural β-dynamics z_T are the inputs to another deep neural network that performs classification (categorical cross-entropy loss function) or regression (Mean Squared Error loss function). The code, hyperparameters and instruction on data splits are publicly available at `github.com/hidden-for-blind-review`.

Following is the description of different benchmarks.

Multimodal Sentiment Analysis

CMU-MOSI The CMU-MOSI dataset (Zadeh et al. 2016) is a collection of 2199 opinion video clips. Each opinion video is annotated with sentiment in the range [-3,3]. There are 1284 segments in the train set, 229 in the validation set and 686 in the test set.

ICT-MMMO The ICT-MMMO dataset (Wöllmer et al. 2013) consists of online social review videos that encompass a strong diversity in how people express opinions, annotated at the video level for sentiment. The dataset contains 340 multimodal review videos, of which 220 are used for training, 40 for validation and 80 for testing.

YouTube The YouTube dataset (Morency, Mihalcea, and Doshi 2011) contains 47 videos from the social media web site YouTube were collected that span a wide range of product reviews and opinion videos. Out of 46 videos, 30 are used for training, 5 for validation and 11 for testing.

MOUD To show that MCDN is generalizable to other languages, we perform experimentation on the MOUD dataset (Perez-Rosas, Mihalcea, and Morency 2013b) which consists of product review videos in Spanish. Each video consists of multiple segments labeled to display positive, negative or neutral sentiment. Out of 79 videos in the dataset, 49 are used for training, 10 for validation and 20 for testing.

Multimodal Speaker Trait Recognition

POM Persuasion Opinion Multimodal (POM) dataset (Park et al. 2014) contains movie review videos annotated for the following speaker traits: confidence, passion, dominance, credibility, entertaining, reserved, trusting, relaxed, nervous, humorous and persuasive. 903 videos were split into 600 were for training, 100 for validation and 203 for testing.

Multimodal Emotion Recognition

IEMOCAP The IEMOCAP dataset (Busso et al. 2008) consists of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral) as well as valence, arousal and dominance. The dataset is recorded across 5 sessions with 5 pairs of speakers. To ensure speaker independent learning, the dataset is split at the level of sessions: training is performed on 3 sessions (6 distinct speakers) while validation and testing are each performed on 1 session (2 distinct speakers).

Multimodal Computational Descriptors

All the datasets are in video format and only one speaker is in front of the camera. The descriptors we used for each of the modalities are as follows:

Language All the datasets provide manual transcriptions. We use pre-trained word embeddings (glove.840B.300d) (Pennington, Socher, and Manning 2014) to convert the transcripts of videos into sequence of word vectors. The dimension of the word vectors is 300.

Vision Facet (iMotions 2017) is used to extract a set of features including per-frame basic and advanced emotions and facial action units as indicators of facial muscle movement.

Acoustic We use COVAREP (Degottex et al. 2014) to extract low level acoustic features including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients.

Modality Alignment To reach the same time intervals between different modalities we choose the granularity of the input to be at word level. The words are aligned with audio using P2FA (Yuan and Liberman 2008) to get the exact time. In simple terms $t = i$ means ith word. We treat speech pause as a word with vector values of all zero across dimensions. The vision and acoustic modalities follow the same granularity. We use expected feature values across the entire word for vision and acoustic since they are extracted at a higher frequency (30 Hz for vision and 100 Hz for acoustic).

Comparison Metrics

Different datasets in our experiments have different labels. For binary classification and multiclass classification we use

A^C where C denotes the number of classes, and F1 score. For regression we report Mean Absolute Error MAE and Pearson's correlation r. For all the metrics, higher values denote better performance, except MAE where lower values denote better performance.

Experimental Results

Baseline Models

We compare the performance of our MCDN to the following state-of-the-art models in multimodal sentiment analysis, speaker trait recognition, and emotion recognition. All the baselines are trained for all the benchmark datasets for complete comparison.

TFN (Tensor Fusion Network) (Zadeh et al. 2017) explicitly models α-dynamics and β-dynamics by creating a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions across three modalities. It is the current state-of-the-art for CMU-MOSI dataset.

BC-LSTM (Bidirectional Contextual LSTM) (Poria et al. 2017) is a model for context-dependent sentiment analysis and emotion recognition, currently state-of-the-art on the IEMOCAP and MOUD datasets.

MV-LSTM (Multi-View LSTM) (Rajagopalan 2016) is a recurrent model that designates special regions inside one LSTM to different views of the data.

C-MKL (Convolutional Neural Network (CNN) with Multiple Kernel Learning) (Poria, Cambria, and Gelbukh 2015) is a multimodal sentiment analysis and emotion recognition model which uses a CNN for visual feature extraction and multiple kernel learning for prediction.

THMM (Tri-modal Hidden Markov Model) (Morency, Mihalcea, and Doshi 2011) performs early fusion of the modalities by concatenation and uses a HMM for classification.

SVM (Support Vector Machine) (Cortes and Vapnik 1995) a SVM is trained on the concatenated multimodal features for classification or regression (Zadeh et al. 2016; Perez-Rosas, Mihalcea, and Morency 2013b; Park et al. 2014). To compare to another strong non-neural baseline we use **RF** (Random Forest) (Breiman 2001) using similar multimodal inputs.

SAL-CNN (Selective Additive Learning Convolutional Neural Network) (Wang et al. 2016) is a multimodal sentiment analysis model that attempts to prevent identity-dependent information from being learned by using Gaussian corruption introduced to the neuron outputs.

EF-HCRF: (Hidden Conditional Random Field) (Quattoni et al. 2007) uses a HCRF to learn a set of latent variables conditioned on the concatenated input at each time step. Additionally, we implement the following variations: 1) **EF-LDHCRF** (Latent Discriminative HCRFs) (Morency, Quattoni, and Darrell 2007) are a class of models that learn hidden states in a CRF using a latent code between observed input and hidden output. 2) **MV-HCRF**: Multi-view HCRF (Song, Morency, and Davis 2012) is an extension of the HCRF for Multi-view data, explicitly capturing view-shared and view specific sub-structures. 3) **MV-LDHCRF**: is a variation of the MV-HCRF model that uses LDHCRF instead of HCRF. 4) **EF-HSSHCRF**: (Hierarchical Sequence Summarization HCRF) (Song, Morency, and Davis 2013) is a layered model that uses HCRFs with latent variables to learn hidden spatio-temporal dynamics. 5) **MV-HSSHCRF**: further extends **EF-HSSHCRF** by performing Multi-view hierarchical sequence summary representation. The best performing early fusion model is reported as **EF-HCRF**$_{(\star)}$ while the best multi-view model is reported as **MV-HCRF**$_{(\star)}$, where $\star \in \{\text{h, l, s}\}$ to represent HCRF, LDCRF and HSSCRF respectively [1].

DF (Deep Fusion) (Nojavanasghari et al. 2016) is a deep neural network model that trains one deep model for each modality and performs decision voting on the output of each modality network.

EF-LSTM (Early Fusion LSTM) concatenates the inputs from different modalities at each time-step and uses that as the input to a single LSTM. We also implement

	Binary		Multiclass	Regression	
Method	A^2	F1	A^7	MAE	Corr
Majority	50.2	50.1	17.5	1.864	0.057
RF	56.4	56.3	21.3	-	-
SVM-MD	71.6	72.3	26.5	1.100	0.559
THMM	50.7	45.4	17.8	-	-
SAL-CNN	73.0	-	-	-	-
C-MKL	72.3	72.0	30.2	-	-
EF-HCRF	$65.3_{(h)}$	$65.4_{(h)}$	$24.6_{(l)}$	-	-
MV-HCRF	$65.6_{(s)}$	$65.7_{(s)}$	$24.6_{(l)}$	-	-
DF	72.3	72.1	26.8	1.143	0.518
EF-LSTM	$73.3_{(sb)}$	$73.2_{(sb)}$	$32.4_{(-)}$	$1.023_{(-)}$	$0.622_{(-)}$
MV-LSTM	73.9	74.0	33.2	1.019	0.601
BC-LSTM	73.9	73.9	28.7	1.079	0.581
TFN	74.6	74.5	28.7	1.040	0.587
MCDN (no MAB)	76.5	76.5	30.8	0.998	0.582
MCDN (no $\mathcal{A}$)	$59.3_{(3)}$	$36.0_{(3)}$	$22.0_{(3)}$	$1.438_{(5)}$	$0.060_{(5)}$
MCDN	$\mathbf{77.1}_{(4)}$	$\mathbf{77.0}_{(4)}$	$\mathbf{34.7}_{(3)}$	$\mathbf{0.968}_{(4)}$	$\mathbf{0.625}_{(5)}$
Human	85.7	87.5	53.9	0.710	0.820

Table 1: Sentiment prediction results on CMU-MOSI test set using multimodal methods. Our model outperforms the previous baselines and the best scores are highlighted in bold.

	ICT-MMMO Binary		YouTube Multiclass		MOUD Binary	
Method	A^2	F1	A^3	F1	A^2	F1
Majority	40.0	22.9	42.4	25.2	60.4	45.5
RF	70.0	69.8	49.3	49.2	64.2	63.3
SVM	68.8	68.7	42.4	37.9	60.4	45.5
THMM	53.8	53.0	42.4	27.9	58.5	52.7
C-MKL	80.0	72.4	50.2	50.8	74.0	74.7
EF-HCRF	$81.3_{(l)}$	$79.6_{(l)}$	$45.8_{(l)}$	$45.0_{(l)}$	$54.7_{(h)}$	$54.7_{(h)}$
MV-HCRF	$81.3_{(l)}$	$79.6_{(l)}$	$44.1_{(s)}$	$44.0_{(s)}$	$60.4_{(l)}$	$47.8_{(l)}$
DF	77.5	77.5	45.8	32.0	67.0	67.1
EF-LSTM$_{(\star)}$	$80.0_{(sb)}$	$78.5_{(sb)}$	$44.1_{(-)}$	$43.6_{(-)}$	$67.0_{(-)}$	$64.3_{(-)}$
MV-LSTM	72.5	72.3	45.8	43.3	57.6	48.2
BC-LSTM	70.0	70.1	47.5	47.3	72.6	72.9
TFN	72.5	72.6	47.5	41.0	63.2	61.7
MCDN (no MAB)	82.5	82.4	47.5	42.8	75.5	72.9
MCDN (no $\mathcal{A}$)	$80.0_{(5)}$	$79.1_{(5)}$	$44.1_{(5)}$	$29.3_{(5)}$	$63.2_{(5)}$	$61.9_{(5)}$
MCDN	$\mathbf{86.3}_{(2)}$	$\mathbf{85.9}_{(2)}$	$\mathbf{54.2}_{(6)}$	$\mathbf{52.9}_{(6)}$	$\mathbf{81.1}_{(2)}$	$\mathbf{81.2}_{(2)}$

Table 2: Sentiment prediction results on ICT-MMMO, YouTube and MOUD test sets. Our model outperforms the previous baselines and the best scores are highlighted in bold.

[1] detailed results for each individual model is presented in the Supplementary Material.

Task	Confident	Passionate	Dominant	Credible	Entertaining	Reserved	Trusting	Relaxed	Nervous	Persuasive	Humorous
Method	A^7	A^7	A^7	A^7	A^7	A^5	A^5	A^5	A^5	A^7	A^5
Majority	19.2	20.2	18.2	21.7	19.7	29.6	44.3	39.4	24.1	20.7	6.9
SVM	26.6	20.7	35.0	25.1	31.5	34.0	50.2	49.8	41.4	28.1	36.0
RF	26.6	27.1	26.1	23.2	26.1	34.0	53.2	40.9	36.0	25.6	40.4
THMM	24.1	15.3	29.1	27.6	12.3	22.7	31.0	31.5	27.1	17.2	24.6
DF	25.6	24.1	34.0	26.1	29.6	30.0	53.7	50.2	42.4	26.6	34.5
EF-LSTM	$25.1_{(b)}$	$30.5_{(sb)}$	$36.9_{(s)}$	$29.6_{(b)}$	$33.5_{(b)}$	$33.5_{(sb)}$	$52.7_{(sb)}$	$48.3_{(-)}$	$44.8_{(sb)}$	$25.6_{(sb)}$	$39.4_{(b)}$
MV-LSTM	25.6	28.6	34.5	25.6	29.1	33.0	52.2	50.7	42.4	26.1	38.9
BC-LSTM	26.6	26.6	33.0	27.6	29.6	33.0	52.2	47.3	36.0	27.1	36.5
TFN	24.1	31.0	34.5	24.6	29.1	30.5	38.9	35.5	42.4	27.6	33.0
MCDN (no MAB)	26.1	27.1	35.5	28.1	30.0	32.0	55.2	50.7	42.4	29.1	33.5
MCDN (no $\mathcal{A}$)	$24.6_{(6)}$	$32.0_{(5)}$	$34.0_{(5)}$	$24.6_{(6)}$	$29.6_{(6)}$	$32.5_{(6)}$	$53.2_{(6)}$	$49.3_{(6)}$	$42.4_{(5)}$	$29.6_{(6)}$	$42.4_{(4)}$
MCDN	**29.1**$_{(2)}$	**33.0**$_{(6)}$	**38.4**$_{(6)}$	**31.5**$_{(2)}$	**33.5**$_{(3)}$	**36.9**$_{(1)}$	**55.7**$_{(1)}$	**52.2**$_{(6)}$	**47.3**$_{(5)}$	**31.0**$_{(3)}$	**44.8**$_{(5)}$

Table 3: Speaker personality trait recognition results on POM test set. Our model outperforms the previous baselines and the best scores are highlighted in bold.

the Stacked, (**EF-SLSTM**) Bidirectional (**EF-BLSTM**) and Stacked Bidirectional (**EF-SBLSTM**) LSTMs for stronger baselines. The best performing model is reported as **EF-LSTM**$_{(\star)}$, $\star \in$ {-, s, b, sb} denoting Vanilla, Stacked, Bidirectional and Stacked Bidirectional LSTMs respectively [2].

Majority performs majority voting for classification tasks, and predicts the expected label for regression tasks. This baseline is useful as a lower bound of model performance.

Human performance is calculated for CMU-MOSI dataset which offers annotation results for each human annotator. This is the accuracy of human performance in a one-vs-rest classification/regression.

Finally, **MCDN** indicates our proposed model with K attentions. Additionally, the modified baseline **MCDN (no MAB)** removes the MAB and learns no dense β-dynamics code z. This model can be seen as three disjoint LSTMs and is used to investigate the importance of modeling temporal β-dynamics. The next modified baseline **MCDN (no $\mathcal{A}$)** removes the $\mathcal{A}$ deep network and sets all K attention coefficients $a_t^k = 1$ ($h_t^k = \tilde{h}_t^k$). This comparison will deduce whether the attention coefficients are required in modeling temporal β-dynamics. For MCDN and MCDN (no $\mathcal{A}$), K is treated as a hyperparamter and the best value of K is indicated in parenthesis next to the best reported result.

Results on CMU-MOSI dataset

We summarize the results on the CMU-MOSI dataset in Table 1 and compare to the baselines mentioned previously. We are able to achieve new state-of-the-art results for this dataset in all the metrics using the MCDN. This highlights our model's capability in understanding sentiment aspect of multimodal communication expressed in English.

Results on ICT-MMMO, YouTube, MOUD datasets

We achieve state-of-the-art performance with significant improvement over all the comparison metrics for two whole video English sentiment analysis. Table 2 shows the comparison of our MCDN with state-of-the-art approaches for ICT-MMMO dataset as well as the comparison for YouTube dataset. To assess the generalization of the MCDN to speakers communicating in different languages, we compare with state-of-the-art approaches for sentiment analysis in opinion utterance video clips in Spanish. The final third of Table 2 shows these results. We achieve significant improvement over state-of-the-art approaches in this experiment as well.

[2] specific results across individual models are presented in the Supplementary Material.

Task	Emotions		Valence		Arousal		Dominance	
Method	A^9	F1	MAE	Corr	MAE	Corr	MAE	Corr
Majority	21.2	7.4	2.042	-0.02	1.352	0.01	1.331	0.17
SVM	24.1	18.0	0.251	0.06	0.546	0.54	0.687	0.42
RF	27.3	25.3	-	-	-	-	-	-
THMM	23.5	10.8	-	-	-	-	-	-
C-MKL	34.0	31.1	-	-	-	-	-	-
EF-HCRF	$32.0_{(s)}$	$20.5_{(s)}$	-	-	-	-	-	-
MV-HCRF	$32.0_{(s)}$	$20.5_{(s)}$	-	-	-	-	-	-
DF	26.1	20.0	0.250	-0.04	0.613	0.27	0.726	0.09
EF-LSTM	$34.1_{(s)}$	$32.3_{(s)}$	$0.244_{(-)}$	$0.09_{(-)}$	$0.512_{(b)}$	$0.62_{(-)}$	$0.669_{(s)}$	$0.51_{(sb)}$
MV-LSTM	31.3	26.7	0.257	0.02	0.513	0.62	0.668	**0.52**
BC-LSTM	35.9	34.1	0.248	0.07	0.593	0.40	0.733	0.32
TFN	36.0	34.5	0.251	0.04	0.521	0.55	0.671	0.43
MCDN (no MAB)	31.2	28.0	0.246	0.09	0.509	0.63	0.679	0.44
MCDN (no $\mathcal{A}$)	$23.0_{(3)}$	$10.9_{(3)}$	$0.249_{(5)}$	$0.05_{(5)}$	$0.609_{(4)}$	$0.29_{(4)}$	$0.752_{(4)}$	$0.21_{(5)}$
MCDN	**37.0**$_{(4)}$	**35.9**$_{(4)}$	**0.242**$_{(6)}$	**0.10**$_{(5)}$	**0.497**$_{(3)}$	**0.65**$_{(3)}$	**0.655**$_{(1)}$	$0.50_{(5)}$

Table 4: Emotion recognition results on IEMOCAP test set using multimodal methods. Our model outperforms the previous baselines and the best scores are highlighted in bold.

Results on POM

We experiment on speaker traits recognition based on observed multimodal communicative behaviors. Table 3 shows the performance of the MCDN on POM dataset, where it achieves state-of-the-art accuracies on all 11 speaker trait recognition tasks including persuasiveness and credibility.

Results on IEMOCAP dataset

Our results for multimodal emotion recognition on IEMOCAP dataset are reported in Table 4. Our approach achieves state-of-the-art performance in emotion recognition: both emotion classification as well as continuous emotion regression except for the case of correlation in dominance which our results are competitive but not state-of-the-art.

Discussion

Our experiments indicate outstanding performance of MCDN in modeling human communication. In this section, we aim to better understand different characteristics of our model.

Number of Attentions

In the experiments we report two additional baselines to understand the effect of number of attentions on model performance. Our main research questions (RQ) for these two baselines are RQ1: MCDN (no MAB): whether the β-dynamics are helpful. RQ2: MCDN (no $\mathcal{A}$): whether the attention coefficients are needed. RQ3: MCDN: whether one attention is enough to extract all β-dynamics. RQ4: whether different tasks and datasets require different numbers of attentions.

RQ1: MCDN (no MAB) model only learns simple rules among modalities such as decision voting or simple co-occurrence rules such as Tensor Fusion baseline. Across all datasets, MCDN (no MAB) is outperformed by MCDN. This indicates that continuous modeling of β-dynamics is crucial in understanding human communication.

RQ2: Whether the presence of the coefficients a_t are crucial is an important research question. From the results tables, we notice that the MCDN (no $\mathcal{A}$) baseline severely underperforms compared to MCDN [3]. This supports the importance of the attentions in the MAB. Without these attentions, MCDN is not able to accurately model the β-dynamics.

RQ3: In our experiments the MCDN with only one attention (like conventional attention models) under-performs compared to the models with multiple attentions. One could argue that the models with more attentions have more parameters, and as a result their better performance may not be due to better modeling of β-dynamics, but rather due to more parameters. However we performed extensive grid search on the number of parameters in MCDN with one attention. Increasing the number of parameters further (by increasing dense layers, LSTHM cellsizes etc.) did not improve performance. This indicates that the better performance of MCDN with multiple attentions is not due to the higher number of parameters but rather due to better modeling of β-dynamics.

RQ4: Different tasks and datasets require different number of attentions. This is highly dependent on each dataset's nature and the underlying interconnections between modalities.

Visualization of Attentions

We visually display how each attention is sensitive to different dimensions of LSTHM outputs in Figure 3. Each column of the figure denoted by a^k shows the behavior of the kth attention on a sample video from CMU-MOSI. The left side of a^k is t = 1 and the right side is t = 20, since the sequence has 20 words. The y axis shows what modality the dimension belongs to. Dark blue means high coefficients and red means low coefficients. Our observations (O) are detailed below:

O1: By comparing each of the attentions together, they show diversity on which dimensions they are sensitive to, indicating that each attention is sensitive to different β-dynamics.

[3]a full comparison across different numbers of attentions is attached in the Supplementary Material.

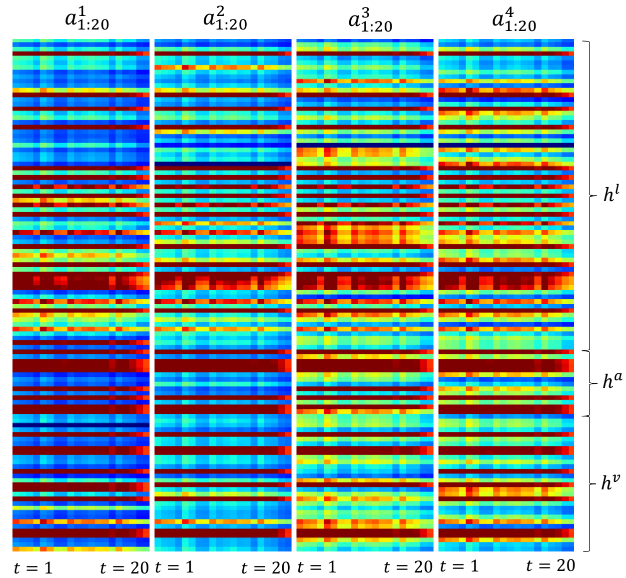

Figure 3: Visualization of the attention units throughout time, obtained using the best model on the CMU-MOSI test set with 4 attentions. Blue indicates activated attentions and red non-activated attentions. Best viewed zoomed in and in color. The learned attentions are diverse and evolve across time.

O2: Some attention coefficients are not active (always red) throughout time. These dimensions carry only α-dynamics needed by that modality and not other modalities. Hence, they are not needed for β-dynamics and will carry no weight in the formation of those structures.

O3: Attentions change their behaviors across time. For some coefficients, these changes are more drastic than the others. We suspect that the less drastic the change in the attention dimension, the higher the chances of that dimension being part of multiple structures thus being active more often.

O4: Some attentions focus on β-dynamics that involve only two modalities. For example, in a^2, the audio modality has no dark blue dimensions, while in a^0 all the modalities have dark blue dimensions. The attentions also seem to have residual effects. a^0 shows activations over a broad set of variables while a^3 shows activation for fewer sets, indicating that attentions could learn to act in a complementary way.

Conclusion

In this paper we modeled multimodal human communication using a novel neural approach called the Multimodal Communication Decoder Network (MCDN). Our approach is designed to model both intra-modality dynamics as well as inter-modality dynamics continuously through time. The intra-modality dynamics are modeled using a Long-short Term Hybrid Memory (LSTHM) for each modality. Various inter-modality dynamics are identified at each time-step using the Multi-attention Block (MAB) which outputs a multimodal neural code for the hybrid memory of LSTHM. MCDN achieves state-of-the-art results in 6 publicly available datasets and across 16 different attributes related to understanding human communication.

References

Abburi, H.; Prasath, R.; Shrivastava, M.; and Gangashetty, S. V. 2016. Multimodal sentiment analysis using deep neural networks. In *International Conference on Mining Intelligence and Knowledge Exploration*, 58–65. Springer.

Baum, L. E., and Petrie, T. 1966. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics* 37(6):1554–1563.

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

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.; Lee, S.; and Narayanan, S. S. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42(4):335–359.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Mach. Learn.* 20(3):273–297.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. Covarepa collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 960–964. IEEE.

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

iMotions. 2017. Facial expression analysis.

Jiang, J.; Dai, B.; Peng, D.; Zhu, C.; Liu, L.; and Lu, C. 2012. Neural synchronization during face-to-face communication. *Journal of Neuroscience* 32(45):16064–16069.

Kuzmanovic, B.; Bente, G.; von Cramon, D. Y.; Schilbach, L.; Tittgemeyer, M.; and Vogeley, K. 2012. Imaging first impressions: Distinct neural processing of verbal and nonverbal social information. *NeuroImage* 60(1):179 – 188.

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

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

Nojavanasghari, B.; Gopinath, D.; Koushik, J.; Baltrušaitis, T.; and Morency, L.-P. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, 284–288. New York, NY, USA: ACM.

Park, S.; Shim, H. S.; Chatterjee, M.; Sagae, K.; and Morency, L.-P. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, 50–57. New York, NY, USA: ACM.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation.

Pérez-Rosas, V.; Mihalcea, R.; and Morency, L.-P. 2013a. Utterance-level multimodal sentiment analysis. In *ACL (1)*, 973–982.

Perez-Rosas, V.; Mihalcea, R.; and Morency, L.-P. 2013b. Utterance-Level Multimodal Sentiment Analysis. In *ACL*.

Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis.

Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*.

Poria, S.; Cambria, E.; and Gelbukh, A. F. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2539–2544.

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

Rajagopalan, Shyam Sundar, L.-P. M. T. B. G. R. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*.

Sergent, J., and Signoret, J.-L. 1992. Processing the facial image - functional and anatomical decomposition of face processing: evidence from prosopagnosia and pet study of normal subjects. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 335(1273):55–62.

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

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

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

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

Wrtwein, T., and Scherer, S. 2017. What really matters – an information gain analysis of questions and reactions in automated ptsd screenings. In *Affective Computing and Intelligent Interaction*.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

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

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

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *Empirical Methods in Natural Language Processing, EMNLP*.