M2P2: Multimodal Persuasion Prediction using Adaptive Fusion

Chongyang Bai, Haipeng Chen, Srijan Kumar, Jure Leskovec, and V.S. Subrahmanian

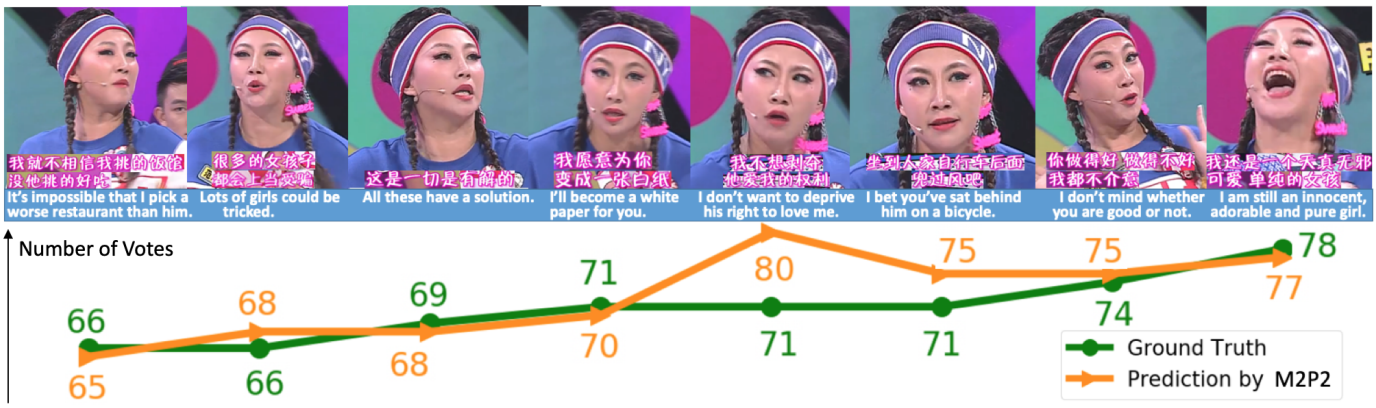

Fig. 1: Real-time prediction of debate persuasiveness (number of votes) using our proposed model M2P2. The debate is from a Chinese debate TV show, Qipashuo. M2P2 closely predicts the ground truth number of votes.

***Abstract*—Identifying persuasive speakers in an adversarial environment is a critical task. In a national election, politicians would like to have persuasive speakers campaign on their behalf. When a company faces adverse publicity, they would like to engage persuasive advocates for their position in the presence of adversaries who are critical of them. Debates represent a common platform for these forms of adversarial persuasion. This paper solves two problems: the Debate Outcome Prediction (DOP) problem predicts who wins a debate while the Intensity of Persuasion Prediction (IPP) problem predicts the change in the number of votes before and after a speaker speaks. Though DOP has been previously studied, we are the first to study IPP. Past studies on DOP fail to leverage two important aspects of multimodal data: 1) multiple modalities are often semantically aligned, and 2) different modalities may provide diverse information for prediction. Our M2P2 (Multimodal Persuasion Prediction) framework is the first to use multimodal (acoustic, visual, language) data to solve the IPP problem. To leverage the alignment of different modalities while maintaining the diversity of the cues they provide, M2P2 devises a novel adaptive fusion learning framework which fuses embeddings obtained from two modules – an *alignment* module that extracts shared information between modalities and a *heterogeneity* module that learns the weights of different modalities with guidance from three separately trained unimodal reference models. We test M2P2 on the popular IQ2US dataset designed for DOP. We also introduce a new dataset called QPS (from Qipashuo, a popular Chinese debate TV show) for IPP. M2P2 significantly outperforms 4 recent baselines on both datasets.**

***Index Terms*—Multimodal learning, Persuasion, Adaptive fusion**

C. Bai is with the Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA. H. Chen is with the Department of Computer Science, Harvard University, Boston, MA 02138, USA. S. Kumar is with the College of Computing at Georgia Institute of Technology, Atlanta, GA 30332, USA. J. Leskovec is with the Department of Computer Science at Stanford University, Stanford, CA 94305, USA. V.S. Subrahmanian is with the Department of Computer Science and the Roberta Buffett Institute of Global Affairs at Northwestern University, Evanston, IL 60208, USA.

Corresponding author: Professor V.S. Subrahmanian.

Code and datasets can be found at https://snap.stanford.edu/persuasion.

I. Introduction

Controversial topics (e.g. foreign policy, immigration, national debt, privacy issues) engender much debate amongst academics, businesses, and politicians. Speakers who are persuasive often win such debates. Given videos of discussions between two participants, the goal of this paper is to provide a fully automated system to solve two persuasion related problems. The Debate Outcome Prediction problem (DOP) tries to determine which of two teams "wins" a debate. Suppose the two teams are A and B and suppose bef_A, bef_B denote the number of voters for A and B's positions respectively before the debate and aft_A, aft_B denote the same after the debate. Hence, $bef_A + bef_B < n$ and $aft_A + aft_B < n$ where n is the total number of voters in the audience. In the DOP problem, we say that team A (resp. team B) wins the debate if $aft_A - bef_A > aft_B - bef_B$ (resp. $<$). We say a speaker is a winner if s/he belongs to the winning team. The Intensity of Persuasion Problem (IPP) tries to predict the increase (or decrease) in the number of votes of each speaker (as opposed to a team). We use the same notation as before but assuming we have two speakers S_1, S_2. The intensity of speaker X's persuasiveness is $\frac{aft_X - bef_X}{n}$ for $X \in \{S_1, S_2\}$. It is clear that both these problems are important. In a business meeting, it might be important to win (DOP), but in other situations, peeling away support for an opponent might be important (IPP). *The more support a speaker can peel away from the opponent, the more persuasive s/he is.*

Fig. 2: *In multimodal content, the modalities are semantically aligned.* This example shows a case where the visual modality (facial expressions) and the language modality (the content of the speech) are closely aligned.

Solving DOP and IPP using video data alone can pose many challenges. In this paper, we test our M2P2 algorithm against two datasets, the IQ2US dataset[1] from a popular US debate TV show and the QPS dataset from the popular Chinese TV show Qipashuo[2]. Real-world videos such as these come with three broad properties: (i) as we can see in Figure 3b, the detected language can be very noisy — this must be accounted for, (ii) as we can see from Figure 3a, there can be considerable noise in the video modality as well — for instance, a man's face might be shown in the video while a woman is speaking and these kinds of audio-video mismatches must be addressed, (iii) but in some cases — as shown in Figure 2, the modalities might be nicely aligned where the audio, language, and video modalities are all correct and the speaker's speech and visual signals are aligned. The problem of identifying these types of mismatches poses a major challenge in building a single model to predict both DOP and IPP.

Though we are not the first to take on the DOP problem, we are the first to solve IPP. DOP has been addressed by [1], [2], [3] who use multimodal sequence data to predict who will win a debate. However, these efforts do not address all the three challenges described above. To the best of our knowledge, there is no existing dataset that addresses the IPP problem and there are no algorithms to solve the IPP problem. *In this paper, we develop a novel algorithm called M2P2 and show that M2P2 improves upon past solutions to DOP by 2%–3.7% accuracy (statistically significant with a p-value below 0.05) and beats adaptations of past work on DOP to the IPP case by over 25% MSE (statistically significant with $p < 0.01$).* Figure 1 shows a sample of how our M2P2 framework predicts speaker persuasiveness at interim points during a debate from the QPS dataset — the reader can readily see that the M2P2 prediction of number of votes (orange line) closely matches the ground truth (green line).

When all three modalities (audio, video, language) agree, then that "common" information must be correctly captured by a predictive model. In this case, we say that the modalities are *aligned*. However, there can be cases where some modalities suggest one thing, while the other(s) suggest something different. In this case, we say the modalities are *heterogeneous*. Our solution, M2P2, captures both aspects and also learns how to weight the two aspects in order to maximize prediction accuracy. M2P2 first leverages the Transformer encoder structure [4] to project the three modalities into three latent spaces. To combine the information from the latent spaces, the model then devises two major modules: *alignment* and *heterogeneity*.

[1]https://www.intelligencesquaredus.org
[2]https://www.imdb.com/title/tt4397792/

(a) *There are cases where the visual modality is noisy, while the language modality is clean.* In 4 consecutive frames when the woman is speaking, the face of a man appears (see frames 2 and 3) and the man's face is incorrectly assumed to be the woman's. The language modality, however, is correct.

(b) *There are cases where the language modality can be noisy, while the visual modality is clean.* In the video frame (the right side of the figure), the subtitles extracted by the OCR system (the left side) are incorrect due to the milk ads shown. Moreover, the font and color of subtitles vary from video to video, so it is difficult to automatically separate these subtitles from other texts.

Fig. 3: *Individual modalities can be noisy.* Here we show examples where the visual or the language modality are wrong. M2P2 learns to down-weight the noisy modalities.

The *alignment* module learns to highlight the shared, aligned information across modalities. It enforces an alignment loss in the loss function as a regularization term during training. This ensures that there is relatively little discrepancy between the latent embeddings of different modalities when they are aligned.

The *heterogeneity* module first learns the weights of modality-specific information and applies weighted fusion to harden the model against noisy modalities (cf. Figure 3). M2P2 uses a novel interactive training procedure to learn the weights from three separately trained reference models, each corresponding to a single modality. Intuitively, a modality with smaller unimodal loss should be assigned a higher weight in the multimodal model. Finally, the outputs of both modules are combined with the debate meta-data for persuasion prediction.

We evaluate M2P2 on the IQ2US and QPS datasets. IQ2US was first used by [1] to evaluate the DOP problem. The IQ2US dataset only has the final debate outcomes, without any labels about how persuasive each speaker is during the debate. Hence, IQ2US cannot be used to evaluate IPP. To this end, we created a new dataset *QPS*, based on an extremely popular Chinese entertainment debate TV show called Qipashuo[2]. In QPS, the audience provides real-time votes before and after each speaker in order to gauge how persuasive the speaker is. QPS therefore provides a direct measure of each speaker's persuasiveness for training and evaluation. We use the IQ2US dataset for the DOP problem and the QPS dataset for IPP

problem. M2P2 outperforms baselines based on four recent papers [1], [2], [3], [5] which were originally designed to predict debate outcomes (or other related problem scenarios). We also conduct ablation studies and visualize our results to show the effectiveness of different novel components in M2P2.

To summarize, we make the following contributions:

- To the best of our knowledge, M2P2 is the first to solve the IPP problem.
- We design a novel adaptive fusion learning framework to solve the IPP and DOP problems.
- We curate a new dataset QPS from the well-known Chinese debate TV show Qipashuo. QPS will be a strong benchmark for future work on persuasion prediction as well as multimodal learning.
- M2P2 outperforms reasonable baselines adapting recent papers [1], [2], [3], [5] in IPP and DOP problems — and these improvements are statistically significant.

II. Related Work

Unimodal persuasion prediction. There has been some work on using a single modality for predicting persuasion. [6], [7], [8], [9] explored the linguistic modality by studying style, context, semantic features and argument-level dynamics in English transcripts to solve DOP. For the visual modality, Joo et al. [10] defined nine visual intents related to persuasion (e.g. dominance, trustworthiness) and trained SVMs to predict them and persuasion using hand-crafted features. Huang et al. [11] improved these results by fine-tuning pre-trained CNNs. In the case of audio, MFCC features and LSTM were used by [12] to solve DOP.

Multimodal persuasion prediction. Brilman et al. [1] solved DOP by extracting facial emotions, voice pitch and word category related features and then training separate SVMs for each modality. The overall prediction for DOP was obtained through a majority vote by the three models. Nojavanasghari et al. [2] solved DOP by first building a Multi-Layer Perceptron (MLP) for each modality, then concatenating the predicted probabilities,and sending them as input to yet another MLP. Because both methods use simple aggregate feature values (e.g. mean, median), they ignore the dynamics of features over time. As a result, these two approaches do not work well with short video clips, and do not leverage temporal dynamics. To address this problem, Santos et al. [3] used an LSTM to take each time-step into account, but their feature-level multimodal fusion considers all modalities to be equally important — thus ignoring the noise, heterogeneity, and alignment properties.

M2P2 is the first to address the Intensity of Persuasion Prediction problem (IPP). Moreover, M2P2 captures temporal dynamics via a multi-headed attention mechanism that: (i) learns the importance of different modalities at different times in long video sequences, and (ii) thus learns better representations of multiple modalities. Moreover, M2P2 is the first to capture both alignment and heterogeneity — hence addressing noise. With these innovations, M2P2 performs well in both IPP and DOP.

General Multimodal Learning. A body of multimodal learning methods defines constraints between modalities in a latent space to capture their inter-relationships. Andrew et al. [13] extended Canonical Correlation Analysis by deep neural networks to maximize inter-modal correlations. Song et al. [14] further proposed to maximize the correlation of the residual matrices of multimodal features. Such correlation constraints have since been used in human action recognition [15], emotion recognition [16] and video captioning [17]. In addition to capturing the shared relationship, [18], [19], [5] tried to extract the individual component of each modality through low-rank estimation. [20], [21] trained cross-modal encoders to reconstruct a modality from another modality. While these efforts provide important insights for creating multimodal embeddings, they do not show how to combine the learned embeddings for accurate prediction.

A second body of work explores architectures for fusing embeddings from modalities. Zadeh et al. [22] introduced bimodal and trimodal tensors via cross products to express inter-modal features. Mai et al. [23] further proposed to combine local and global interaction learning for time-dependent multimodal fusion. As cross products significantly increase the dimensionality of the feature space, [24], [25], [26] introduced bilinear pooling techniques to learn compact representations. Although these methods explicitly model inter-modal relationships, they introduce additional features that require larger networks to be learned for subsequent prediction tasks. In contrast, attention-based fusion [27], [28] learns the weighted sum of multimodal embeddings taking the prediction task into account. However, they require huge amounts of data to learn the optimal attention weights. In order to capture long-term dependencies, M2P2 uses the Transformer encoder [4], [29] to learn latent embeddings for modalities. On one hand, inspired by the first class of work, M2P2 uses a shared projector and enforces high correlation among the encoded embeddings. On the other hand, M2P2 computes a weighted concatenation of latent unimodal embeddings, where the weights are guided by the persuasiveness loss of each embedding through interactive training. These two innovations lead to a compact embedding that can be learned with a small dataset.

III. The M2P2 Framework

Figure 4 shows an overview of our M2P2 architecture with a brief description of its major components. Note that the key novelties of this paper are the two novel modules (i.e., the alignment module and the heterogeneity module shaded in yellow in Figure 4) that constitute the adaptive fusion framework (Section III-C) [3].

A. Generating Primary Input Embeddings

Given a video clip, we respectively represent the acoustic, visual and language input as $X_A \in \mathcal{R}^{T_A}$, $X_V \in \mathcal{R}^{(H\times W\times C)\times T_V}$, $X_L \in \mathcal{R}^{D\times T_L}$, where T_A, T_V, T_L are respectively the lengths of the audio signal, face sequence, and word sequence. H, W, C are the height, width and the number of channels of each image, and D is the length of

[3]Our proposed adaptive fusion framework has the potential of being broadly utilized in other multimodal learning tasks. We leave that exploration for future work.

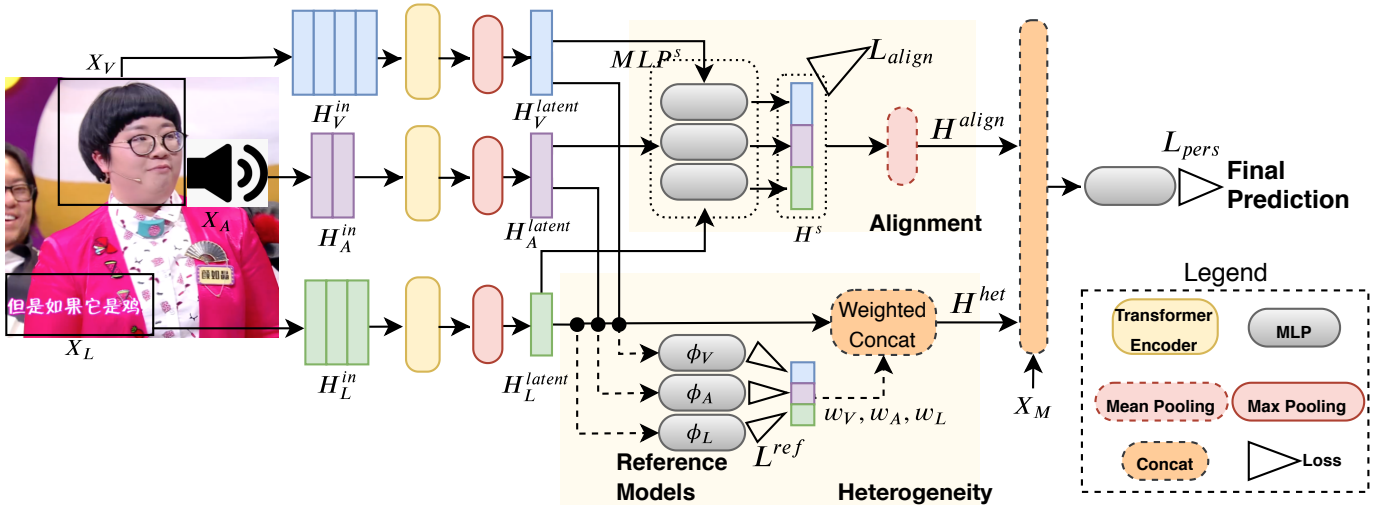

Fig. 4: M2P2 architecture. First, audio, face and language sequences are extracted from a video clip and fed to three separate modules to get primary input embeddings. Second, each of these embeddings is fed to a Transformer encoder [4] followed by a max pooling layer, which yields the latent embeddings. Third, the latent embeddings are fed to the alignment and heterogeneity modules to generate the embeddings H^{align} and H^{het}. Last, we concatenate H^{align}, H^{het} and the debate meta-data X_M which is fed to an MLP for persuasiveness prediction. The latent embeddings interact with two procedures alternately: optimize the alignment loss L_{align} and persuasiveness loss L_{pers}, and learn weights through 3 reference models.

our dictionary of words. In addition, we also use two debate meta-data features: the number of votes before a speech and the length of the speech. We generically denote the debate meta-data as a vector $X_M \in \mathcal{R}^{d_M}$, where $d_M = 2$.

We first extract features from the three modalities, then add a fully-connected (FC) layer for each modality to obtain low dimensional primary input embeddings. The generated primary input embeddings are depicted as multi-dimensional bars (as a symbol of vector sequences) in Figure 4. Here we describe the detailed feature extraction components.

Feature extraction from the acoustic modality. For each audio clip, we use Covarep [30] to extract MFCCs[4], Glottal source parameters, pitch-related features, and features using the Summation of Residual Harmonics method [31]. These features capture human voice characteristics from different perspectives and are all shown to be relevant to emotions [32]. These 73 dimensional features are averaged over every half second.

Feature extraction from the visual modality. Since the speakers in both datasets can be highly dynamic and occluded, we capture only their faces as Brilman et al. [1] did to reduce noisy input. The details of face detection and recognition are in Section IV. Given each facial image, we use the VGG19 architecture [33] pre-trained on the Facial Emotion Recognition FER2013 dataset[5] and extract the 512 dimensional output before the last FC layer as the face features.

Feature extraction from the language modality. We use the Jieba[6] Chinese text segmentation library to segment Chinese sentences (utterances) into words. We use the Tencent Chinese embedding corpus [34] to extract 200 dimensional word embeddings. In the case of English, we extract 64 dimensional Glove word embeddings [35] trained from all transcripts from the IQ2US debates.

All features are passed to a learnable FC+ReLU layer which converts the initial features into *primary input embeddings*. The primary input embeddings thus obtained for each of the three modalities are respectively $H_A^{in} \in \mathcal{R}^{d_{in} \times T'_A}, H_V^{in} \in \mathcal{R}^{d_{in} \times T'_V}, H_L^{in} \in \mathcal{R}^{d_{in} \times T'_L}$, where $d_{in} = 16$ is the row-dimension of the primary input embeddings, which is same across different modalities. T'_A, T'_V, T'_L denote the sequence lengths of the modalities, where $T'_V = T_V, T'_L = T_L$. Note that in our primary input embeddings, the timestamps of the acoustic, visual, and language modality respectively represent a short time window, a frame, and a word.

B. Generating Compact Latent Embeddings of Modalities with Transformers

To get a compact representation of the primary input embeddings for each modality, we aggregate the sequence of features into a single representation vector using one Transformer encoder per modality. Transformer encoders have been shown to outperform many other deep architectures, including RNNs, GRUs, and LSTMs in many sequential data processing tasks in computer vision [36] and natural language processing [37]. The multi-head self-attention mechanism of Transformer better memorizes the long-term temporal dynamics [4].

With the Transformer encoder, the primary input embedding $H_m^{in}, m \in \{A, V, L\}$ of each modality is respectively transformed into a representation as:

$$H_m^{trans} = \text{TransformerEncoder}(H_m^{in}), \quad (1)$$

[4]The energy-related 0th coefficient is excluded

[5]https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/overview

[6]https://github.com/fxsjy/jieba

where $H_m^{trans} \in \mathcal{R}^{d_{trans} \times T'_m}$, and $d_{trans} = 16$ is the dimension of the latent space after the Transformer encoder.

To convert arbitrary length time sequences into standardized latent embedding vectors $H_m^{latent} \in \mathcal{R}^{d_{trans} \times 1}$, we additionally use a max pooling layer:

$$H_m^{latent} = \text{MaxPool}(H_m^{trans}). \tag{2}$$

H_m^{latent} intuitively captures the maximum activation over the time sequence along each dimension of d_{trans}.

C. Balancing Shared and Heterogeneous Information with Adaptive Fusion

As mentioned earlier, there are two conflicting aspects of multimodal data. First, data from different modalities within the same time frames may sometimes be highly aligned (i.e., have shared information). Second, different modalities may sometimes contain diverse cues which may not be equally important for prediction. To balance the aligned and heterogeneous multimodal information, we propose a novel adaptive fusion framework consisting of two key modules: an *alignment* module and a *heterogeneity* module (shaded in yellow in Figure 4).

1) Alignment Module: To extract information shared across different modalities, we first use a *shared* multi-layer perceptron (MLP^s) to project the latent embeddings of each modality $m = A, V, L$ into the same latent space:

$$H_m^s = \text{MLP}^s(H_m^{latent}) \tag{3}$$

Here, $H_m^s \in \mathcal{R}^{d_s}$, where $d_s = 16$ is the dimension of the shared projection space. MLP^s is shown as three rounded grey boxes in Figure 4.

Inspired by existing multimodal representation learning work [13], [38] and the success of domain adaptation techniques [39], [40], [41], [42], for each pair of modalities $\{m, n\}$, we use a cosine loss term $l_{cos}(m, n)$ and a domain adaptation loss term $l_{da}(m, n)$ to measure the alignment of m, n in the shared projection space:

$$\mathcal{L}_{align} = \sum_{\{m,n\} \subset \{A,V,L\}} l_{cos}(m, n) + l_{da}(m, n) \tag{4}$$

The cosine loss term measures the similarity between embeddings of modalities m, n of one sample:

$$l_{cos}(m, n) = 1 - cos(H_m^s, H_n^s)$$

and the domain adaptation loss is one deep coral loss [39], which measures the distance between the correlation matrices of modalities m, n:

$$l_{da}(m, n) = \frac{1}{4d_s^2} \|C_m^s - C_n^s\|_F^2$$

where C_m and C_n are correlation matrices of the embeddings in the shared space for modalities m, n respectively. The cosine loss aligns the modalities in sample basis whereas the domain adaptation loss aligns the two modalities in a distribution level.

During training, the alignment loss $\mathcal{L}_{align}$ will be added to the entire prediction loss function as a regularization term to penalize lack of alignment between the 3 modalities in the projected space.

After the shared MLP layer, the regularized embeddings H_m^s are in the same latent space. We apply mean pooling to average the three embeddings:

$$H^{align} = \text{MeanPool}(H_A^s, H_V^s, H_L^s) \ , \tag{5}$$

$H^{align} \in \mathcal{R}^{d_s}$ now contains shared information from all modalities.

2) Heterogeneity Module: Another key observation discussed in Section I is that different modalities may contain diverse information, and therefore make unequal contributions to the final prediction of persuasiveness (e.g., due to the noisy data from certain modalities as shown in Figure 3). We therefore propose a novel heterogeneity module which utilizes an interactive training procedure (Algorithm 1) to learn weights for different modalities.

Intuitively, the importance of each modality should be inversely proportional to the "error" caused by the modality. To estimate this error term, we create three unimodal MLP reference models (represented as dashed arrows and rounded grey boxes at the central bottom of Figure 4) parameterized by ϕ_A, ϕ_V, ϕ_L for the acoustic, visual, and language modalities respectively. Each unimodal MLP takes the compact latent embedding H_m^{latent} generated by the Transformer encoder as input and generates a unimodal prediction $\hat{Y}_m^{ref}$ for each modality $m = A, V, L$:

$$\hat{Y}_m^{ref} = \text{MLP}_m^{ref}(\phi_m; H_m^{latent}) \ . \tag{6}$$

We use T_{val} to denote the validation set, $Y_{val} \in \mathcal{R}^{|T_{val}|}$ are the labels, and $\hat{Y}_{m,val}^{ref} \in \mathcal{R}^{|T_{val}|}$ are the predictions made by the unimodal reference model for modality m. The reference models (ϕ_m's) are updated using the following Mean Squared Error (MSE) loss alone:

$$\mathcal{L}_m^{ref} = \frac{\left\| Y_{val} - \hat{Y}_{m,val}^{ref} \right\|_2^2}{|T_{val}|} \tag{7}$$

After several epochs of training ϕ_m's, we are able to obtain a converged MSE loss of each reference model. We then use the updated reference model to estimate the prediction errors by $\mathcal{L}_m^{ref}$. $\mathcal{L}_m^{ref}$ is used to guide the weights w_m of latent embeddings H_m^{latent} ($m = A, V, L$) to be concatenated in the heterogeneity module:

$$H^{het} = w_A H_A^{latent} \oplus w_V H_V^{latent} \oplus w_L H_L^{latent}. \tag{8}$$

w_A, w_V, w_L are scalars incrementally updated over epochs:

$$w_m = \alpha w_m + (1 - \alpha)\tilde{w}_m, \tag{9}$$

where $\alpha \in (0, 1)$ controls the rate of update, and $\tilde{w}_m$ is obtained using the following softmax function of the reference model validation losses:

$$\tilde{w}_m = \frac{\exp\{-\beta \mathcal{L}_m^{ref}\}}{\sum_{m'=A,V,L} \exp\{-\beta \mathcal{L}_{m'}^{ref}\}}, \forall m = A, V, L \tag{10}$$

$\beta > 0$ is a scaling factor. Since $\sum_m \tilde{w}_m = 1$, combining Equation (9), it is guaranteed that $\sum_m w_m = 1$.

3) Adaptive Fusion with Interactive Training: The representations obtained from the alignment module (H^{align}) and the heterogeneity module (H^{het}) are then concatenated together with the debate meta-data X_M and fed into a final MLP layer to make the final prediction $\hat{Y}$:

$$\hat{Y} = f(\theta; X_A, X_V, X_L, X_M) = \text{MLP}(H^{align} \oplus H^{het} \oplus X_M) \tag{11}$$

where θ is the set of parameters of the M2P2 model excluding the reference model parameters ϕ_m. The modality weights in H^{het} are adapted from the losses of the unimodal models. Together with H^{align}, the representations from both modules are learned adaptively through an interactive training process.

To train the M2P2 model, we have two loss terms: a novel alignment loss $\mathcal{L}_{align}$, and a persuasiveness loss term $\mathcal{L}_{pers}$. In the case of the IPP problem, $\mathcal{L}_{pers}$ is the MSE loss. In the case of DOP, we use cross-entropy loss for the binary classification. The total loss function is a weighted combination:

$$\mathcal{L}_{final} = \mathcal{L}_{pers} + \gamma \mathcal{L}_{align}, \tag{12}$$

where γ is a weight factor.

The entire training proceeds in a master-slave manner, as shown in Algorithm 1. In each epoch of the master training procedure (Lines 4 to 14), we use the total loss function in Equation (12) to update the parameters θ of the main M2P2 components. The weights w_A, w_V, w_L of the 3 modalities are obtained using reference models ϕ_m, and their losses $\mathcal{L}_m^{ref}$ are then updated in the slave procedure. In each epoch of the slave procedure (Lines 8 to 10), we take the latent embeddings from the master procedure as input and update the reference models with the loss function in Equation (7). We then obtain the weights w_A, w_V, w_L of different modalities in the heterogeneity module.

Algorithm 1: M2P2 interactive training procedure

Input: Training dataset T, validation datset T_{val}; Number of epochs n and N
Output: Multi-modality model $f(\theta; X_A, X_V, X_L, X_M)$, modality weights w_m $(\forall m = A, V, L)$

1 Initialize three unimodal reference models $\phi_m (\forall m = A, V, L)$ and θ;
2 Initialize $w_A = w_V = w_L = 1/3$;
3 % Master Procedure Start
4 **for** *epoch=1,...,N* **do**
5 Update θ with loss function Equation (12);
6 Get latent embeddings $H_m^{latent}, \forall m = A, V, L$;
7 % Slave Procedure Start
8 **for** *epoch=1,...,n* **do**
9 Update $\phi_m, \forall m = A, V, L$ with loss function in Equation (7);
10 **end**
11 % Slave Procedure End
12 Get reference model losses $\mathcal{L}_m^{ref}, \forall m = A, V, L$;
13 Update modality importance weights $w_m, \forall m = A, V, L$ using Equations (9)-(10);
14 **end**
15 % Master Procedure End
16 **return** θ, $w_m (\forall m = A, V, L)$

IV. Datasets

We describe our two datasets below.

A. QPS Dataset

We created the QPS dataset by getting videos[7] *from the popular Chinese TV debate show Qipashuo.* In each episode of the TV show, 100 audience members initially vote 'for' or 'against' a given debate topic. Debaters from 'for' and 'against' teams speak alternately, and the audience can change their votes anytime. In general, there are 6–10 speech turns. Final votes are turned in after the last speaker. The winner is the team which has more votes at the end than at the beginning. For example, if the initial and final 'for' vs. 'against' votes are 30:70 and 40:60, respectively, then the 'for' team wins because they increased their votes from 30 to 40 (even though they still have fewer votes than the "against" team). In total, we collected videos of 21 Qipashuo episodes with 205 speaking clips spanning a total of 582 minutes.

We extracted the transcripts from the video subtitles. To sufficiently preprocess the videos for subtitle extraction, we took the following steps. First, we sampled 2 frames per second and binarize the images with a threshold 0.6, which can avoid the influence from various colors of subtitles in videos. Second, we cropped the subtitles by a fixed bounding box since the position of subtitles is fixed in all the videos. Third, we clustered the binarized images into buckets such that any two binarized images in the same bucket are identical on 90% or more pixels. We then randomly selected one of these images to represent the cluster. This helps reduce noise (e.g. from advertisements displayed on the image). Finally, the surviving binary images were fed into an OCR API to get accurate transcripts. We used Baidu's off-the-shelf pre-trained OCR API[8], so no extra data is needed for training.

If we take each speaking clip as a train/test instance, there would be a total of 205 data points. This paucity of information poses a huge challenge for machine learning. We therefore segment each speaking clip into clips of 50 utterances each according to the transcript we extract above. Note that 50 is the smallest number of utterances in any speaking clip of our dataset. Moreover, note that these "sub-clips" of 50 utterances yield a temporal sequence whose temporal dynamics can be important. The labels are shared for segments extracted from the same clip. This trick yields 2297 such segments which are used as train/test instances in our evaluation.

As the speakers are highly dynamic and often occluded, we only use speakers' faces as the visual input. We extract 2 frames per second from videos and use Dlib[9] for face detection and recognition. The recognition is based on one pre-annotated profile for each speaker and is only needed for training. To

[7] An example can be found in https://youtu.be/P5ehhs0hpFI.

[8] https://ai.baidu.com/tech/ocr

[9] http://dlib.net

further reduce false positives (i.e., extracting the face of the non-speakers), we first use the model from [43] to remove faces in the image that are not speaking, and then use the method from [44] for face tracking.

B. IQ2US

We also evaluate M2P2 on the benchmark IQ2US TV debate dataset used by [1], [12], [6], [7], [8]. This dataset was originally collected by [1]. The audience can only vote at the beginning and at the end of the debate, and the winner is determined in the same way as in Qipashuo. Note that we cannot use the same set of videos as [1], since they were interested in predicting the result of the whole debate, which doesn't require the transcripts to be aligned within shorter clips. Of the 100 episodes we collected, only 58 had transcripts that were correctly aligned with the visual modality at the minute level. Finally, we get 852 one-minute single-speaker clip instances from the 58 episodes — 428 of them belong to the winning side. As transcripts are available in the IQ2US data, no pre-processing is required for the language modality in this dataset. For the visual modality, we use the same procedures as in the QPS dataset to extract the face image sequences of the speakers. Since there are no intermediate votes in IQ2US, we only predict the debate outcome (i.e. whether a single-speaker clip instance belongs to the winning team).

V. Experimental Evaluations

Our experiments assess the performance of M2P2 on the DOP and IPP tasks. Specifically:

1) (IPP) We predict the change of votes after a speech by a debater — this is done on the QPS dataset. Note that the number of votes are scaled by the total number of audience members and hence is guaranteed to lie in the [0,1] interval. Hence, the change of votes always lies in the [-1,1] interval.
2) (DOP) We predict whether a clip in which a debater is speaking is part of the winning team of the debate — this is done on the IQ2US dataset.

In addition, we also conduct an ablation study that assesses the contributions of different components of M2P2. Moreover, we assess the importance of different modalities as well as time frames using the QPS dataset. Finally, we test the sensitivity of the optimal hyper-parameters used in the model..

A. Experimental Settings

QPS uses a 10-fold rolling window prediction. Specifically, we construct 10 sequences of consecutive episodes of the show. For instance, if $E_1, \ldots, E_k$ represent the set of all QPS episodes, then one sequence would be $Seq_k = E_1, \ldots, E_k$, another would be $Seq_{k-1} = E_1, \ldots, E_{k-1}$. For any such sequence $Seq_i = E_1, \ldots, E_i$, we set E_i as the test episode (i.e. the episode on which we make predictions). We learn a model from the first $i-3$ episodes $E_1, \ldots, E_{i-3}$ and identify the best parameters for our model by using episodes E_{i-2}, E_{i-1} as the validation set. As the same subject can occur in multiple episodes of QPS, in order to avoid information leakage from training to test data, we do not train a model from E_i to predict $E_{j,j<i}, \forall i, j$.

For IQ2US, 10-fold cross validation is used since a debater can only appear in one episode. The initial vote score and speaking length features are normalized to $[0, 1]$.

Denote FC*n* as a fully-connected layer that outputs *n*-dimension vectors. The MLPs in the reference models and final multimodal prediction model are all configured as FC16+ReLU, FC8+ReLU, and FC1+Sigmoid. The shared MLP in *alignment module* is FC16+ReLU. M2P2 uses Batch Normalization [45] right after each of the FC layers for input embeddings, and uses 0.4 as dropout [46] after all FC16 layers. For the Transformer encoder, we use a single layer with 4 heads, where the input, hidden, and output dimension are all 16. We use the Adam [47] optimizer with a weight decay of 10^{-5}. The numbers of epochs in Algorithm 1 is $N = 200$ and $n = 10$. The learning rate lr, alignment loss weight γ, update scalar α, scaling factor β are finalized by grid search. We ended up using $lr = 0.001, \gamma = 0.1, \alpha = 0.5, \beta = 50$ as these yield the best results on the validation sets.

B. Comparison with Baselines

We compare both tasks with three multimodal persuasion prediction baselines: early fusion + SVM [1], deep multimodal late fusion [2], early fusion + LSTM [3], and a more recent multimodal fusion baseline, DeepCU [5]. Brilman et al. [1] extract audio, visual and linguistic features from IQ2US debate videos and concatenate these features, which are fed into an SVM for classification. Although [1] also solves the DOP problem on the IQ2US dataset, it is different from our work in that (i) the used episodes are different (see Section IV-B and (ii) it uses long video inputs (9–36 minutes) of all debates while we only use a short speaking clip (1 minutes) of a single speaker. Thus, for fair comparison, we implemented their method and ran experiments in our data. Nojavanasghari et al. [2] first feed features of each modality to a neural network to get predictions of the modality, then uses a fusion neural network to combine the modality-based predictions. Santos et al. [3] model the temporal dynamics by using an LSTM on the concatenated features from all modalities. Verma et al. [5] propose a deep model to integrate both common and unique latent information for multimodal sentiment analysis.

In the case of the IPP problem, we adapt the first baseline by modifying it to use an SVM regressor (rather than an SVM classifier). For the other three baselines, we replace the final layers by a regression and use MSE loss to train the models. For fairness, we also allow the baselines to use the two debate meta-data features. The results comparing M2P2 on IPP and DOP with past approaches are shown in Tables I and II, respectively.

IPP Problem. Table I shows the MSE obtained by different approaches in each fold and the average on the QPS dataset. Note that the vote scores are normalized to lie in the $[0, 1]$ interval. The last line of Table I shows the decrease percentage of MSE which is defined as $dec.$ = 1-MSE(M2P2)/MSE(the best baseline). For instance, from the first column of Table I,

Fold	1	2	3	4	5	6	7	8	9	10	Average
Brilman et al. [1] (early fusion)	0.009	0.011	0.016	0.017	0.030	0.018	0.020	0.012	0.013	0.018	0.016
Nojavanasghari et al. [2] (late fusion)	0.007	0.015	0.019	**0.011**	0.027	**0.014**	0.020	0.012	0.020	0.015	0.016
Santos et al. [3] (early fusion)	0.025	0.019	0.018	0.019	0.018	0.017	0.029	0.016	0.024	0.018	0.020
Verma et al. [5]	0.012	0.013	0.016	0.012	0.021	0.016	0.016	0.019	0.012	0.016	0.015
M2P2 without DA loss	0.006	0.010	0.015	0.015	0.020	0.015	0.012	**0.009**	0.009	0.013	0.012
M2P2 (proposed method)	**0.006**	**0.010**	**0.011**	0.014	**0.017**	0.015	**0.012**	0.010	**0.008**	**0.012**	**0.011**
dec. %	14.2	9.1	31.3	-27.3	5.6	-7.1	25.0	16.7	33.3	20.0	26.7

TABLE I: MSE for each test fold of different approaches to solving the Intensity of Persuasion Prediction (IPP) on the QPS Dataset. The last row shows the MSE decrease percentage of M2P2 compared to the best baseline in each fold. DA loss stands for the domain adaptation loss l_{da}. On average, M2P2 achieves a lower MSE than the baselines by at least 26.7%, which is statistically significant with p-val < 0.05. Note that the vote scores we predict are normalized.

Method	DOP (Accuracy)
Brilman et al. [1] (early fusion)	0.614
Nojavanasghari et al. [2] (late fusion)	0.615
Santos et al. [3] (early fusion)	0.598
Verma et al. [5] (DeepCU)	0.622
M2P2 without DA loss	0.635
M2P2 with MDD DA loss	0.629
M2P2 (proposed method)	**0.639**

TABLE II: Prediction accuracy for Debate Outcome Prediction in IQ2US dataset. Our M2P2 is 1.7%–2.5% better than baselines. The DA stands for domain adaptation and the MDD DA loss is employed from [41]. M2P2 improvements over baselines are statistically significant with p-val < 0.05.

Method	MSE
M2P2 without DA loss (l_{da})	0.012
M2P2 without alignment loss	0.018
M2P2 without reference models	0.014
M2P2-LSTM	0.033
M2P2-Acoustic (unimodal)	0.017
M2P2-Visual (unimodal)	0.019
M2P2-Language (unimodal)	0.016
M2P2	**0.011**

TABLE III: Ablation study results. All improvements are statistically significant (p-val < 0.05).

we see that the percentage decrease is $1 - \frac{0.006}{0.007} \approx 0.14$ representing a 14% decrease of MSE generated by M2P2 compared to the best of the baselines. We observe that on average, M2P2 yields a 26.7% decrease of MSE compared with the best baseline which is statistically significant via a Student t-test (p-val < 0.01). In addition, the comparison of the last two methods shows that the domain adaptation loss l_{da} in Equation 4 improves the performance by $1 - \frac{0.011}{0.012} = 8.3\%$. Moreover, M2P2 is more robust and performs better than all baselines in 7 out of 10 folds.

DOP Problem. Table II shows the average prediction accuracy over 10 folds on the DOP problem w.r.t. the IQ2US dataset. When we compare M2P2 (last row) with the four baselines, it is clear that M2P2 achieves 1.7%–2.5% higher average accuracy than the baselines. The improvement is statistically significant (p-val < 0.05). To further investigate the domain adaptation (DA) loss (l_{da} in Equation 4, we evaluate two variations of M2P2: M2P2 without DA loss, and M2P2 with the MDD DA loss [41]. We observe that the proposed method (with deep coral DA loss) achieves the best, although the difference among the three are *not* statistically significant (p-val = 0.08).

Overall, the two experiments make M2P2 the best performing system for both the IPP and the DOP problems.

C. Ablation Study

To measure the contributions of the different components of M2P2, we create four methods, each with one component removed from M2P2 :

- M2P2 without the domain adaptation (DA) loss l_{da}.
- M2P2 without the alignment loss (l_{da} and l_{cos}).
- M2P2 without reference models. The latent embeddings are concatenated by equal weights $1/3$.
- M2P2-LSTM. The Transformer encoder and max pooling layer are replaced by a 1-layer LSTM.
- M2P2-unimodal. We input a single modality without alignment loss and latent embedding concatenation. That is, the latent embedding is directly concatenated with the debate meta-feature and fed to the final MLP.

Table III shows the average MSE obtained on the QPS dataset for both M2P2 and the 7 variations above. First, rows 2,3 and the last row show that if M2P2 does not use the alignment module and reference models in the heterogeneity module, the MSE increases from 0.011 to 0.018 and 0.014 respectively. This is statistically significant (p-val < 0.05) and hence shows the power of both proposed adaptive fusion modules in Section III-C. Second, we observe the power of the Multihead-attention Transformer encoder to handle long sequences, as the M2P2-LSTM model achieves the worst MSE amongst all methods. Third, we observe from rows 4-6 that the language modality is the most important in the prediction task, while the acoustic and visual modalities are less important.

D. Visualization of Prediction

In this experiment, we show (1) the importance of modalities through their learned weights (cf. Equation (8)), and (2) the examples of learned temporal attention weights from different modalities.

Modality weights. We report the modality weights in the heterogeneity module of the trained M2P2 in all folds of QPS dataset. Figure 5 shows box plots for the three modalities.

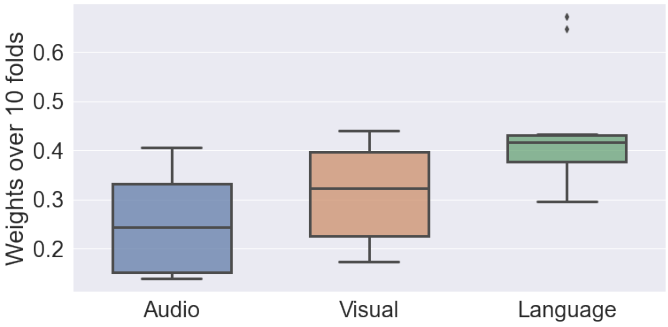

Fig. 5: Modality weights in the heterogeneity module.

Fig. 6: Temporal attention of visual modality – color coded as blue. Darker color implies higher attention weight.

The language modality is the most important and robust over all folds with a median weight of 0.42, while the median weights of acoustic and visual modalities are 0.24 and 0.33 respectively.

Temporal attention weights. We visualize the temporal attention weights of two sample sequences of visual (Figure 6) and language (Figure 7) modalities. For each timestamp t, assume $\alpha_{i,t}^{(h)}$ is the scaled dot-product attention weight from query i to time t in head h, learned by the Transformer Encoder [4]. We calculate the temporal attention weight at time t as

$$a_t = \frac{1}{IH} \sum_{i=1}^{I} \sum_{h=1}^{H} \alpha_{i,t}^{(h)},$$

which represents the amount of attention the model pays to time t, where I and H are the number of queries and heads. In Figure 6 (top), the man's face is not detected correctly in frames 3 and 6 – and we see that M2P2 assigns near-zero attention weights to both frames, suggesting that these frames should be ignored. Moreover, the happy expression in frame 2 gets a high attention weight. The woman below gets high attention weights when she actively talks to someone (frames 2,4,5). In Figure 7, we notice that reasonable keywords like 'wear', 'shackle', 'passive', and 'hold' also get high attention weights. Therefore, our M2P2 captures the meaningful long-range temporal dynamics with the help of Transformer Encoder.

Fig. 7: Temporal attention of language modality – color coded as red. Darker color implies higher attention weight. The original Chinese transcripts are translated to English.

E. Parameter sensitivity analysis

We study the sensitivity of the hyper-parameters around their optimal values obtained from grid search. The update rate α of modality weights in Equation 9, scalar β for Softmax in Equation 10, and weight γ of the alignment loss in Equation 12 are studied. We perturb each of them by $\pm 5\%$ while fixing the other two. We use the modified hyper-parameter value and evaluate the relative change to the original accuracy 0.639 obtained by M2P2 on DOP task (Table II). As a result, the relative change is at most 1.1%, 0.8% and 1.7% for perturbing α, β, γ respectively. α is the most sensitive among the three, implying that the alignment loss is important in training the model. However, all the changes are within a very small range, which indicates that our model is robust.

VI. Discussion

A. Text Encoder Comparison for Linguistic Inputs

In M2P2 , the sequence of word embeddings is used as the sequence input to the Transformer encoder. Another way is to encode each sentence to an embedding and feed the sequence of sentences to the Transformer encoder. We have conducted experiments to compare these two methods. To get English sentence embeddings in IQ2US, we employ the pre-trained Universal Sentence Encoder [48] in TFHub[10]. For Chinese sentences in QPS, we train an LSTM to get 128-dimensional sentence embeddings. We replace word embeddings with sentence embeddings and conduct the experiments in both datasets. As a result, the accuracy in IQ2US is 0.623 (1.4% worse than M2P2) and the mean squared error in QPS is 0.014 (20% worse than M2P2). Thus, the fine-grained word-level embeddings are better than sentence-level embeddings. The word order and semantic meaning is already captured by the word-level embeddings.

B. Heterogeneity Module vs. Attention Mechanism

Intuitively, the heterogeneity module in M2P2 aims to learn the modality-wise importance from data. An alternative is to use the attention mechanism to attend the model to different modalities. However, the attention mechanism introduces extra amount of trainable parameters into M2P2 . In our early experiments, this resulted in worse results due to the small dataset (2297 and 805 data points for QPS and IQ2US resp.). On the contrary, the parameters introduced by heterogeneity module are independent from the rest of M2P2 model, which fuses modalities and achieves better prediction results.

[10]https://tfhub.dev/google/universal-sentence-encoder/1

C. Modality importance

The importance of the different modalities is shown in two experiments: ablation study (Table III) and modality weights (Figure 5). Table III shows that when using a single modality as input, the language modality is the best while the visual modality is the worst. Yet when the three modalities are combined together in M2P2, Figure 5 shows that the audio modality makes the smallest contribution while the language modality is still the strongest.

VII. Limitations and Future Work

First, it is important to note that the adaptive fusion technique in M2P2 can be generalized to other multimodal sequence prediction problems such as video question answering and video sentiment analysis. We leave this exploration for future work.

Second, since the importance of modalities can vary from sample to sample, a future effort could investigate methods to learn sample-specific modality weights (e.g. different attention mechanisms), which might further improve the performance.

Third, prior knowledge is an important factor for humans when judging whether a speaker is persuasive or not. We were unable to assess the impact of prior knowledge in this work as we used publicly available datasets from TV shows. An alternative mechanism would be to conduct such debates ourselves with a studio audience who fill out a pre-debate survey that sheds light on their prior knowledge and then gets their votes periodically during and after the debate. This could be an important experiment to conduct under appropriate IRB protocols.

Last, in real-world persuasion challenges, factors such as logic and the structure of arguments may be more important than the TV shows we have studied where acting and theatrics may be unreasonably important. An important future effort might run IRB-approved experiments involving such persuasion challenges data and use that to predict persuasion in other settings.

VIII. Conclusion

In this paper, we have solved two problems. First, we provide a solution to the Debate Outcome Prediction (DOP) problem that improves on past work by 2%–3.7%. Though these numbers are not huge, they are statistically significant. Second, we are the first to pose and solve the Intensity of Persuasion Prediction (IPP) problem. We show that we are able to beat baselines built on top of past solutions to IPP by 25% on average. Our proposed M2P2 framework leverages both the common and modality-specific information contained in multimodal sequence data (audio, video, language), while learning to focus attention on the meaningful part of the data. Moreover, our newly created QPS dataset provides a valuable new asset for future research — it will be released upon publication of this paper.

Acknowledgment

We gratefully acknowledge the support of NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), IIS-2030477 (RAPID), IIS-2027689 (RAPID); DARPA under No. N660011924033 (MCS); ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, JPMorgan Chase, Docomo, Hitachi, JD.com, KDDI, NVIDIA, Dell, Toshiba, UnitedHealth Group, Adobe, Facebook, Microsoft, and IDEaS Insitute. J. L. is a Chan Zuckerberg Biohub investigator. Special thanks to Xinyu Cong for annotating the QPS dataset and Yuxuan Zhang for translating the Chinese transcripts in the figures of this paper.

References

[1] M. Brilman and S. Scherer, "A multimodal predictive model of successful debaters or how i learned to sway votes," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 149–158.

[2] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 284–288.

[3] P. B. Santos and I. Gurevych, "Multimodal prediction of the audience's impression in political debates," in *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*. ACM, 2018, p. 6.

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

[5] S. Verma, C. Wang, L. Zhu, and W. Liu, "Deepcu: integrating both common and unique latent information for multimodal sentiment analysis," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 3627–3634.

[6] J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil, "Conversational flow in oxford-style debates," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 136–141.

[7] P. Potash and A. Rumshisky, "Towards debate automation: a recurrent model for predicting debate winners," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2465–2475.

[8] L. Wang, N. Beauchamp, S. Shugars, and K. Qin, "Winning on the merits: The joint effects of content and style on debate outcomes," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 219–232, 2017.

[9] I. Habernal and I. Gurevych, "Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1589–1599.

[10] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu, "Visual persuasion: Inferring communicative intents of images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 216–223.

[11] X. Huang and A. Kovashka, "Inferring visual persuasion via body language, setting, and deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 73–79.

[12] P. B. Santos, L. Beinborn, and I. Gurevych, "A domain-agnostic approach for opinion prediction on speech," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 2016, pp. 163–172.

[13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.

[14] G. Song, S. Wang, Q. Huang, and Q. Tian, "Learning feature representation and partial correlation for multimodal multi-label data," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[15] N. E. D. Elmadany, Y. He, and L. Guan, "Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1317–1331, 2019.

[16] G. Aguilar, V. Rozgic, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 991–1002. [Online]. Available: https://www.aclweb.org/anthology/P19-1095

[17] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[18] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust correlated and individual component analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1665–1678, 2015.

[19] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic, "Audio-visual object localization and separation using low-rank and sparsity," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2901–2905.

[20] D. U. Jo, B. Lee, J. Choi, H. Yoo, and J. Y. Choi, "Cross-modal variational auto-encoder with distributed latent spaces and associators," *arXiv preprint arXiv:1905.12867*, 2019.

[21] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1923–1935, 2015.

[22] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114. [Online]. Available: https://www.aclweb.org/anthology/D17-1115

[23] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 122–137, 2020.

[24] J. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B. Zhang, "Hadamard product for low-rank bilinear pooling," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. [Online]. Available: https://openreview.net/forum?id=r1rhWnZkg

[25] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2612–2620.

[26] X. Zhang, X. Gao, W. Lu, L. He, and J. Li, "Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks," *IEEE Transactions on Multimedia*, 2020.

[27] X. Long, C. Gan, G. De Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal keyless attention fusion for video classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter- and intra-modality interactions," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[29] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Florence, Italy: Association for Computational Linguistics, 7 2019.

[30] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[31] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[32] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *Interspeech*, 2016, pp. 3603–3607.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[34] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 175–180. [Online]. Available: https://www.aclweb.org/anthology/N18-2028

[35] 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. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[38] S. H. Dumpala, R. Chakraborty, and S. K. Kopparapu, "Audio-visual fusion for sentiment classification using cross-modal autoencoder," in *32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2019, pp. 1–4.

[39] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[40] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Heterogeneous domain adaptation through progressive alignment," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 5, pp. 1381–1391, 2019. [Online]. Available: https://doi.org/10.1109/TNNLS.2018.2868854

[41] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, "Maximum density divergence for domain adaptation," *CoRR*, vol. abs/2004.12615, 2020. [Online]. Available: https://arxiv.org/abs/2004.12615

[42] J. Li, M. Jing, K. Lu, L. Zhu, and H. T. Shen, "Locality preserving joint transfer for domain adaptation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6103–6115, 2019. [Online]. Available: https://doi.org/10.1109/TIP.2019.2924174

[43] C. Bai, S. Kumar, J. Leskovec, M. Metzger, J. F. Nunamaker, and V. S. Subrahmanian, "Predicting the visual focus of attention in multi-person discussion videos," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 7 2019, pp. 4504–4510. [Online]. Available: https://doi.org/10.24963/ijcai.2019/626

[44] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman, "LAEO-Net: revisiting people Looking At Each Other in videos," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[46] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[48] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 169–174. [Online]. Available: https://www.aclweb.org/anthology/D18-2029

Chongyang Bai is a Ph.D. candidate at Dartmouth College advised by Prof. V.S. Subrahmanian. He obtained a BS in Computational Mathematics and BEng in Computer Science from University of Science and Technology of China in 2016. He was a research intern in Microsoft Research and Google Research in 2016 and 2020 respectively. His research interests are multimodal learning and prediction as well as their applications to human behavioral analysis.

Haipeng Chen is a postdoc in the Computer Science Department, Harvard University. Before that, he was a postdoc in the CS Department at Dartmouth College and obtained the PhD from Interdisciplinary Graduate School, Nanyang Technological University, Singapore in 2018. His research lies in the general areas of Artificial Intelligence, including machine learning (reinforcement learning in particular), data mining, and algorithmic game theory, as well as their applications towards social good. He was winner for the 2017 Microsoft Malmo Collaborative AI Challenge, and runner-up for the Innovation Demonstration Award of IJCAI'19. He has published multiple papers in top tier conferences such as AAAI, IJCAI, UAI, KDD, ICDM. He serves as program committee member for top tier AI conferences such as Neurips, ICLR, AAAI, IJCAI and AAMAS.

Srijan Kumar is an Assistant Professor at CSE, College of Computing at Georgia Institute of Technology. His research area is data science, machine learning, and network analytics. He has won several awards including Facebook Research Award, Adobe Faculty Research Award, ACM SIGKDD Doctoral Dissertation award runner-up, and best paper honorable mention at the World Wide Web Conference. He received his bachelor's degree in computer science from Indian Institute of Technology, masters and Ph.D. degree from the University of Maryland and completed his postdoctoral training from Stanford University.

Jure Leskovec is Associate Professor of Computer Science at Stanford University, Chief Scientist at Pinterest, and investigator at Chan Zuckerberg Biohub. His research focuses on machine learning and data mining with graphs, a general language for describing social, technological and biological systems. Computation over massive data is at the heart of his research and has applications in computer science, social sciences, marketing, and biomedicine. This research has won several awards including a Lagrange Prize, Microsoft Research Faculty Fellowship, the Alfred P. Sloan Fellowship, and numerous best paper and test of time awards. Leskovec received his bachelor's degree in computer science from University of Ljubljana, Slovenia, PhD in machine learning from Carnegie Mellon University and postdoctoral training at Cornell University.

V.S. Subrahmanian is the Walter P. Murphy Professor of Computer Science and Buffett Faculty Fellow in the Buffett Institute for Global Affairs at Northwestern University. He previously served as the Dartmouth College Distinguished Professor in Cybersecurity, Technology, and Society and Director of the Institute for Security, Technology, and Society at Dartmouth and as a Professor of Computer Science at the University of Maryland from 1989-2017 where he also served for 6+ years as Director of the University of Maryland's Institute for Advanced Computer Studies. Prof. Subrahmanian is an expert on big data analytics including methods to analyze text/geospatial/relational/social network data, learn behavioral models from the data, forecast actions, and influence behaviors with applications to cybersecurity and counter-terrorism. He has written eight books, edited fourteen, and published over 300 refereed articles. He is a Fellow of the American Association for the Advancement of Science and the Association for the Advancement of Artificial Intelligence and received numerous other honors and awards. His work has been featured in numerous outlets such as the Baltimore Sun, the Economist, Science, Nature, the Washington Post, American Public Media. He serves on the editorial boards of numerous journals including Science, the Board of Directors of SentiMetrix, Inc., and on the Research Advisory Board of Tata Consultancy Services. He previously served on Board of Directors of the Development Gateway, on DARPA's Executive Advisory Council on Advanced Logistics and as an ad-hoc member of the US Air Force Science Advisory Board.