Through the Eyes of Viewers: A Comment-Enhanced Media Content Representation for TED Talks Impression Recognition

Huan-Yu Chen*, Yun-Shao Lin* and Chi-Chun Lee*
* Department of Electrical Engineering, National Tsing Hua University, Taiwan
* MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

***Abstract*—Developing computational frameworks for personalized content query and recommendation has sparked numerous research into automatic indexing and retrieval of multimedia data. Assessing viewer impression as an appropriate index of media content is especially important as it links directly to the audience preferences toward media content. Most of the existing machine learning frameworks rely on modeling the media contents solely without considering the potential usefulness of user feedback in order to assess the viewer impressions. In this work, we develop a cross-modal network that projects the multimodal media content through the viewer's comment space in order to learn a joint (content and viewer) embedding space to perform viewer impression recognition. Specifically, we gather a large corpus of TED talks including viewer's online comments for each of the presentation video. Our proposed cross-modal projection network achieves 80.8%, 79.5%, and 80.8% of unweighted average recall (UAR) in binary classification tasks for three different viewer impression ratings (i.e., inspiring, persuasive, and funny, respectively). Our experiments demonstrate intuitively that online user comments reflect the viewer impression the most, but an interesting finding shows that it is important to project the content's information into the user comment space, i.e., through the eyes of the comment, in order to obtain an improved recognition accuracy as compared to simply concatenating content and comment features directly.**

***Index Terms*—cross-modal projection, viewer impressions, TED talks**

I. Introduction

The ability to automatically predict viewer impression of multimedia contents can benefit both viewers and publishers. With the rapid growth of internet carrying huge amount of media content, it is becoming increasingly difficult for users to find the desired information. A recent industry report [1] shows that an average person spends nearly an hour a day just on searching for contents on media platforms, and this number is still increasing. The report also indicates that seven out of ten consumers state that an universal search feature would be helpful. The intelligent search and content recommendation feature has largely been developed conventionally through indexing contents using a variety of media's meta attributes, e.g., media types, content genre, language, topic, etc. These attributes mostly are content-*production* related (base on what is being generated), recent work has demonstrated that having a model with capability to predict viewer impression can advance such an intelligent indexing and content queries for media recommendation and search services (base on what is being perceived) [2].

Developing methods to automatically predict viewer impression of videos has been a prevalent topic especially in recent years. Many prior works focus on modeling the media contents for automatic impression recognition. For example, Brezeale et al. use subtitles and visual descriptors (including color, texture, objects, and motion) to cluster the content, and then predict user preferences using the sequence of clusters with Hidden Markov Model (HMM) [3]. Yoon et al. propose to use static audio-visual features with temporal sentiment flow to predict "interestingness" of a video [4]. Trzcinski et al. design a Long-term Recurrent Convolutional Network (LRCN) aiming to predict online video popularity only with visual cues [5]. Nojavanasghari et al. perform "persuasiveness" prediction using multimodal features with a deep multimodal fusion strategy [6]. In this work, our goal is to also predict the viewer impression on media content; given the large heterogeneity exists in the type of media data, we focus first on the video genre of single speaker presentation, i.e., the TED talks.

TED talk videos includes a particular characteristics, i.e., a single speaker sharing information/messages often without background music or special effects. Speakers utilize different speaking styles and strategies that attempt to induce and arouse various feelings, opinions, and ideas, and this effect is important especially in TED talks because speakers tend to share knowledge and take a particular stance on a topic. For example, to influence others' beliefs and behaviors, a persuasive speech would stir the hearers' emotion with credible speakers stating truth [7]. A funny or humor speech expects the viewers to gain a sense of relief, incongruity, or even superiority [8] and need to be carried out often that is highly tailored toward audience's viewpoint [9]. The interpretation, i.e., the resulting viewer impression, of the same content can vary depending on each audience's viewpoint.

In this work, our core technical idea is to extract the usefulness of media content by constraining it with respect to the audience viewpoint in order to perform viewer impression recognition task. Specifically, online comments can be imagined as a concrete realization of audience viewpoint after he/she watches the media content. Several previous works have already examined the use of online comments in predicting

the impressions of videos for content recommendation. For example, Pappas et al. improve video recommendation by combining video contents, metadata, and semantic analysis of comments [10], [11]. Siddiquie et al. build an automatic classification framework on politically persuasive online videos based on audio-visual contents and sentiment scores of viewer's comments [12]. Mishne et al. [13] and Oghina et al. [14] predict movie rating and box office with comments and reviews on social media. In terms of TED talks, most of the works concentrate on analyzing verbal/non-verbal behaviors of the speakers to predict the viewer impressions without considering the comments as additional information source[15], [2], [16], [17]. Most of these previous literature simply extract features (semantically or sentiment-related) from comments and concatenate directly with the content-based descriptors in order to perform viewer impression recognition. We propose to construct the relation between TED talk contents and its associated online comments with a joint space cross-modal projection network learned with a distance constraint, which imposes similar samples to become proximal to one another.

We use 1618 videos of TED Talks and focus on predicting three different viewer impression ratings as binary classification tasks. We choose "inspiring", "persuasive", and "funny" as the three target impression ratings gathered from the official TED website. Our proposed framework leverages the information of both speech and lexical modality of the TED talk speaker and project these features through the online user comment space, and by using both content and comment joint space embedding, it achieves 80.8%, 79.5%, and 80.8% on three binary classification tasks for the ratings of inspiring, persuasive, and funny, respectively. The rest of this paper is organized as follows: section 2 illustrates the details in our proposed framework along with dataset and feature extraction; section 3 describes the experiment results; section 4 summarizes the conclusion and future work.

II. Research Methodology

A. TED Talks Dataset

TED is an organization that hosts conferences and speakers under the slogan of "ideas worth spreading". The talks at TED include a wide range of topics such as technology, science, culture, politics, and academy. TED website is an online library that archive these talks from TED conferences. A huge community is built around this website. Viewers can comment on the talks, and rate the talks with fourteen different keywords indicating their impressions of the talks. Each viewer can vote up to three of fourteen keywords, i.e., beautiful, confusing, courageous, fascinating, funny, informative, ingenious, inspiring, jaw-dropping, long-winded, obnoxious, ok, persuasive and unconvincing to describe the impression to videos, instead of simple like and dislike.

In this work, we select the most viewed 1800 talks on the TED website. We collect video, transcript, comments, and ratings of these talks on the TED website along with meta attributes like video length, tags, and description. We only use top level comments as the direct reaction of the viewers to the videos, rather than a reply to another comment. Those talks that are not delivered in English are removed. Some contents that are not mainly a speech such as music or magic performance are also discarded. This reduces to a total of 1618 talks remained in our dataset. We crawl the raw vote counts on all fourteen keywords from the TED website. Similar to previous work [15], we focus on predicting "Inspiring", "Persuasive", and "Funny" ratings. We define a binary classification problem on the above three impression ratings, i.e., to predict if a given video is above or below median of the given keywords in the whole dataset. For each task, we give positive label if one has rating counts larger than median rating counts of all videos, otherwise negative if the rating count is smaller. The distribution of rating counts for each of three keywords is shown in Fig. 1.

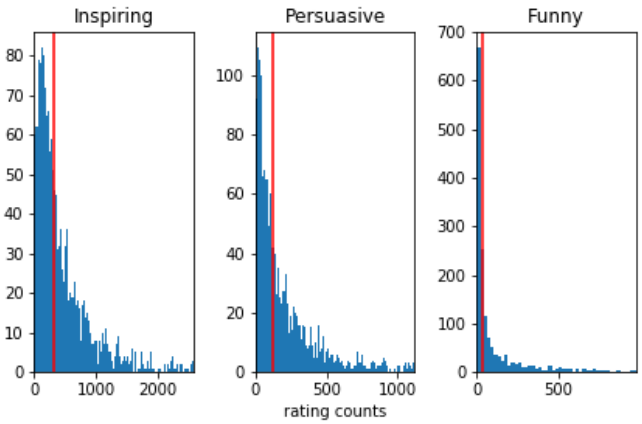

Fig. 1. The histogram of rating counts for each word. The keywords are inspiring, persuasive, and funny from left to right, respectively. The median value which separates positive and negative label is shown as a vertical red line.

B. Content and Comment Features

We extract audio and text features from the talks delivered by the speakers as representation of contents. For the audio features, we first perform segmentation using voice activity detection (VAD) to obtain the sentences of speech. Then, we compute the ComParE [18] feature set using the openSMILE[19] toolkit. The ComParE feature set includes 6373 dimensions of acoustic features based on the statistical functionals of acoustic low-level descriptors (LLDs), including features such as MFCC, pitch, intensity, and spectral analysis calculated at a framerate of 10ms. The average of feature values over all sentences is used to represent the entire speech. The textual features are obtained using pre-trained bidirectional skip-thought sentence embedding for each sentences of the transcript. An embedding of 2400 dimension is obtained for each line of the transcript. Similarly, we also take average of embedding of the entire talks as the textual feature.

Aside from the speech and language contents of talks from

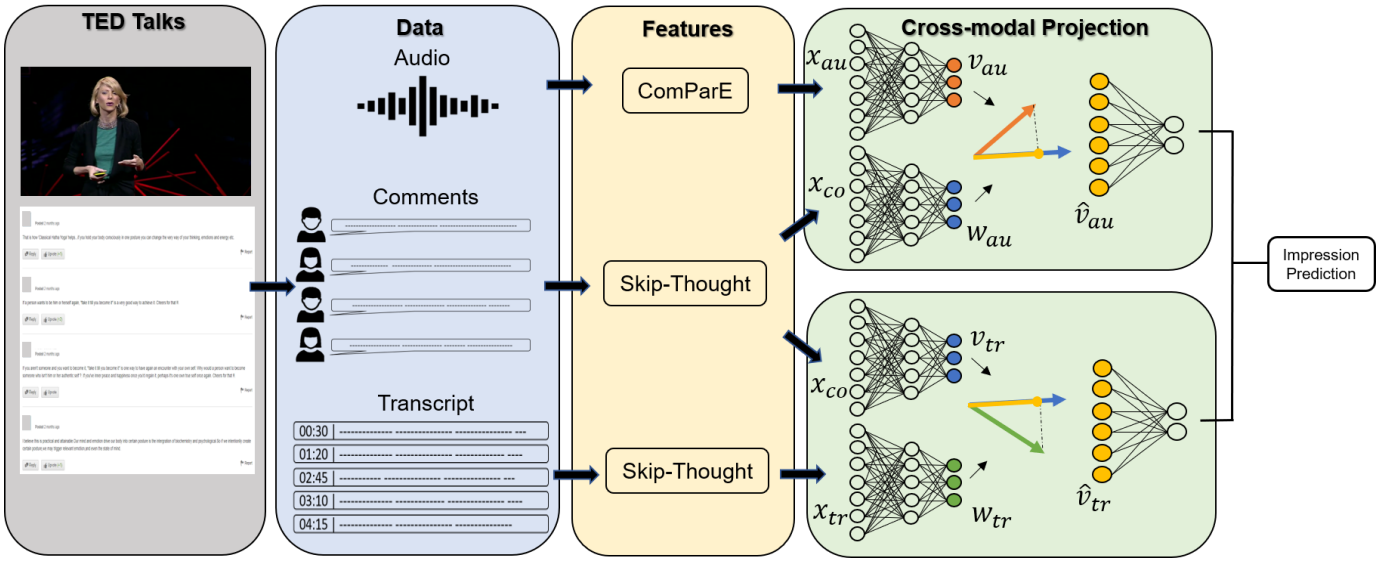

Fig. 2. Detail of our proposed approach. We extract features from audio, transcript, and comments of TED talks, and then we apply cross-modal network to project content space onto comment space. Lastly, two different modalities are fused by averaging the decision scores.

the speaker in the TED talks, the comments from the viewers of the TED website are seen as another information source. We take all the comments of each talk, and transform them into sentence embedding. Again, the average of embedding over all comments is used as a feature from the viewers for each talk. The final mean encoded features are denoted as x_{co}, x_{au}, and x_{tr} for comments, audio, and transcript, respectively.

C. Projection of Media Content onto Viewer Comments

In order to represent the media contents from viewers' perspective, our objective is to learn a joint embedding space based on comment space. A pair of a content vector v_i and a comment vector w_i is used in predicting whether the talk is appealing. Inspired by Zhang et al. [20], who use cross-modal projection to learn a better matching between image-text modalities, we propose to project content representation onto viewer comment space through a deep embedding projection network. We define a projection loss to be used as a constraint in learning the joint space. Specifically, given n samples of talks as a mini-batch, pairs between representation of content and comment are constructed as $\{(v_i, w_j), y_{i,j}\}_{j=1}^{n}$, where i, j are the indices of the samples in the mini-batch, and $y_{i,j} = 1$ implies the two representations are matched, which means they are both from the positive label or both from the negative label. The probability of matching content vector v_i and comment vector w_j is defined as

$$p_{i,j} = \frac{exp(v_i^T \bar{w}_j)}{\sum_{k=1}^{n} exp(v_i^T \bar{w}_k)} \quad s.t. \quad \bar{w}_j = \frac{w_j}{\|w_j\|}$$

where $\bar{w}_j$ means the normalized comment vector, and $v_i^T \bar{w}_j$ is the scalar projection of content vector onto comment vector. A larger scalar projection indicates that the two vectors are more similar. In case of multiple matches inside a mini-batch, the summation of the true matching probability should equal to one. The normalized true matching probability would be

$$q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{n} y_{i,k}}$$

To learn both embedding feature vectors, the matching loss used is the KL divergence from q_i to p_i

$$L_i = \sum_{j=1}^{n} p_{i,j} log \frac{p_{i,j}}{q_{i,j} + \epsilon}$$

where ϵ is a small positive value to avoid numerical problems. The matching loss of the mini-batch is the summation over all samples.

$$L_m = \sum_{i=1}^{n} L_i$$

Following [20], we further perform binary classification with norm-softmax by applying a weight matrix on the projection vector through minimizing the negative log likelihood

$$L_c = \frac{1}{N} \sum_i -log \frac{exp(W_{y_i}^T \hat{v}_i)}{\sum_j exp(W_j^T \hat{v}_i)}$$

$$s.t. \quad \|W_j\| = 1, \quad \hat{v}_i = v_i^T \bar{w}_i \bar{w}_i$$

where W_{y_i} and W_j represent the y_i-th and j-th column of weight matrix W. This constraint encourages feature $\hat{v}_i$ to be more condensed along the weight vector since the weights are normalized to the same length. The final projection matching loss and projection classification loss are combined as the following to be our complete loss function used in the training of our network:

$$L = L_m + L_c$$

TABLE I
SUMMARY OF THE VIEWER IMPRESSION RECOGNITION: COMMENT, TRANSCRIPT, AUDIO ON TOP OF THE TABLE REPRESENTING THE DIFFERENT MODALITY USED IN THE FRAMEWORK. THE METRICS SHOWN ARE ALL ACCURACY CORRESPONDING TO THREE TASKS OF DIFFERENT RATINGS.

	Comment	Transcript			Audio			Multimodal	
Keyword	comment ft	trans. ft	trans.+comm.	proj.	audio ft	audio+comm.	proj.	trans+audio	proj. fusion
Inspiring	72.7%	67.6%	72.9%	80.6%	61.6%	61.6%	78.6%	60.8%	**80.8%**
Persuasive	73.5%	71.8%	74.2%	78.6%	62.5%	62.4%	78.1%	62.2%	**79.5%**
Funny	75.5%	76.3%	80.0%	78.9%	65.6%	65.6%	77.9%	65.6%	**80.8%**

III. EXPERIMENT SETUP AND RESULTS

A. Experimental Setup

In this work, we perform classification on three different viewer impression ratings: inspiring, persuasive, and funny. We project the audio and transcript modalities from content space onto comment space, and then we perform decision score fusion to obtain the final multimodal recognition result. The architecture of our projection network includes three dense feed forward layers with number of nodes as $\{1024, 512, 256\}$. The final embedding dimension of the joint space is 256. We apply 70% dropout to each layer. The learning rate is set as 0.00001 and batch size is defined as 32. Leaky ReLu with alpha being 0.2 is used as the activation function in each layer. Our experiment is carried out using 5-fold cross validation, and the accuracy is used for evaluation of model performance.

We compare our proposed model with the following features in this work:

- Viewer Comment
  - comment ft: x_{co} Using original embedding features extracted from comment data
- Talk Transcript
  - transcript ft: x_{tr} Using original embedding features extracted from transcript data
  - trans.+comm.: $concat(x_{co}, x_{tr})$ Concatenating of original embedding features from comment and transcript
  - proj $\hat{v}_{tr}$: Projecting transcript embedding vector onto comment embedding vector to derive the learned joint space
- Talk Audio
  - audio ft: x_{au} Using original ComPareE acoustic features extracted from audio data
  - audio+comm: $concat(x_{co}, x_{au})$ Concatenating comment embedding and ComPareE acoustic features from comment and audio respectively
  - proj: $\hat{v}_{au}$ Projecting audio features onto comment embedding to derive the learned joint space
- Multimodal Fusion
  - trans+audio: Concatenating original transcript embedding features and acoustic features from transcript and audio
  - Fusion: Performing decision-level fusion between projected joint space of transcript-to-comment and audio-to-comment representations

For the original features and concatenation between features, we use linear SVM as classifier. Our proposed projection network performs classification directly at the final output of our joint space projection network. Table 1 lists the summary of our experiment results. For inspiring classification, there are several observations that can be made by examining the accuracy obtained using features from single modality. Performance of viewer comment features outperforms the content based transcript and audio counterpart by +5% and +11%, respectively. This shows an intuitive result that the comments have a stronger and direct relation to viewer impression rating. We further observe the effect of projection by comparing direct comment-content concatenation and projected joint space for the same modality. In the transcript modality, by simply combining comment features and transcript features, we achieve a boost of 5.3%, and by applying our proposed projection network, it obtains 13% improvement. Joint space learning, i.e., projecting transcript through comment space embedding, indeed gives better representation at cross modal fusion learning in predicting viewer impressions for TED talks. In the audio modality, the results by direct concatenation is limited due to larger dimension of audio features, which seem to dominate the predictions. Through our projection network, we could observe a significant performance gain of 17%, which also mitigates the issue of the dimension mismatch for the original features. Finally, we perform fusion by directly averaging decision score from projection of transcript and projection of audio, which gives 80.8% as our best performance on the rating of inspiring. Similar trend, though smaller boost, also occurs on the rating of persuasive. Our project network achieves 6.8% and 15.6% related improvement on transcript and audio modality, respectively. The final fusion results obtained is 79.5%. Funny, on the other hand, is the only rating that by using the original embedding features of transcript give the best accuracy. However, projected audio-comment representation improves over audio concatenating comment features, and the final fusion still results in the best accuracy for Funny rating, i.e., 80.8% accuracy.

B. Effect of Projection

The learned joint embedding space of audio and comment modalities are visualized in figure 3 using t-SNE. Four sample points are enlarged in the figure to examine the differences between the three feature sets used in this work. We observe that most of the samples are distributed quite uniformly when examining the original audio embedding, while comment

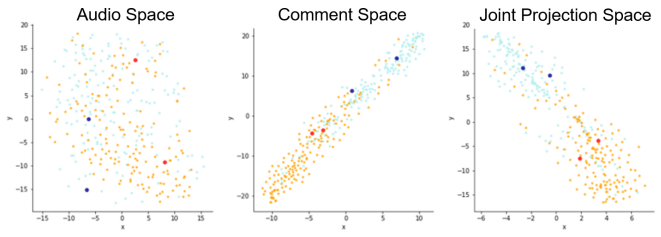

Fig. 3. The learned embedding of audio modality, comment modality, and the projection of audio onto comment, respectively. Dots in blue are positive samples, and dots in red are negative ones.

TABLE II
THE AVERAGE EUCLIDEAN DISTANCE BETWEEN DIFFERENT GROUPS OR WITHIN THE SAME GROUP ON KEYWORD "INSPIRING". x INDICATES THE ORIGINAL FEATURE AND v IS THE EMBEDDING VECTOR UNDER THAT MODALITY. $\hat{v}_{au}$ IS THE PROJECTION FROM AUDIO VECTOR ONTO COMMENT VECTOR.

	Comment		Audio		
	x_{co}	v_{co}	x_{au}	v_{au}	$\hat{v}_{au}$
within high samples	0.225	0.030	112.41	15.36	6.867
within low samples	0.337	0.051	106.82	14.77	6.870
btw high and low samples	0.291	0.054	110.36	15.53	8.954

embedding is much more concentrated. The projection from audio onto comment essentially condense the representation of the original content acoustic features onto the constraint of the online comment space. We further compute the average euclidean distance (shown in table 2). The distance within the same group indicates the compactness of the embedding, and the distance between different group implies the discriminative performance. Comparing the original feature space and the learned embedding space, the samples are much more compact in the latter. When observing the learned embedding vector from two modalities, high scoring samples are more concentrated for the comment modality, while low scoring samples are more concentrated on the audio modality. After projection, the average distance for both classes are about the same. The projection space combines the best of these two modalities.

IV. CONCLUSIONS AND FUTURE WORKS

In this work, we propose a deep joint projection network that learns a comment-enhanced multimodal content embedding space to predict view impression on TED talks. Transcript and audio are used as contents representation from TED talks speakers and comments are used as viewer perspective. Our proposed network learns a joint embedding space by projecting contents onto comments. We consider this projection as viewing talks from the audience perspective. The model performance on three different viewer impression shows that projection improves the recognition accuracy, and it outperforms most of the recent works on predicting viewer impressions on TED talks. To best of our knowledge, this is one of the first works that jointly considers viewers' perspective on content (all of which are data-driven, i.e., by using speeh acoustics and lexical contents of transcripts and online comments) on such automatic recognition tasks on multimedia data. For future work, we plan to extend to other modalities e.g. facial expression, visual features from contents to improve the robustness of the framework.

REFERENCES

[1] ERICSSON CONSUMERLAB, "Tv and media 2017," 2017.
[2] Fasih Haider, Fahim A Salim, Saturnino Luz, Carl Vogel, Owen Conlan, and Nick Campbell, "Visual, laughter, applause and spoken expression features for predicting engagement within ted talks," *Interspeech*, 2017.
[3] D. Brezeale and D. J. Cook, "Learning video preferences using visual features and closed captions," *IEEE Multimedia*, pp. 39–47, 2009.
[4] Sejong Yoon and Vladimir Pavlovic, "Sentiment flow for video interestingness prediction," *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia*, pp. 29–34, 2014.
[5] Tomasz Trzcinski, Pawe l Andruszkiewicz, Tomasz Bochenski, and Przemys law Rokita, "Recurrent neural networks for online video popularity prediction," *ISMIS: Foundations of Intelligent Systems*, pp. 146–153, 2017.
[6] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltruaitis, and Louis-Philippe Morency, "Deep multimodal fusion for persuasiveness prediction," *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 284–288, 2016.
[7] Colin Higgins and Robyn Walker, "Ethos, logos, pathos: Strategies of persuasion in social/environmental reports," *Accounting Forum*, vol. 36, pp. 194, 2012.
[8] Patti M. Valkenburg Moniek Buijzen, "Developing a typology of humor in audiovisual media," *MEDIA PSYCHOLOGY*, vol. 6, pp. 147, 2004.
[9] John C. Meyer, "Humor as a double-edged sword: Four functions of humor in communication," *Communication Theory*, p. 310, 2000.
[10] Nikolaos Pappas and Andrei Popescu-Belis, "Combining content with user preferences for ted lecture recommendation," *11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 47–52, 2013.
[11] Nikolaos Pappas and Andrei Popescu-Belis, "Sentiment analysis of user comments for one-class collaborative filtering over ted talks," *36th Intl ACM SIGIR Conf. on Research and Development in Information Retrieval.*, pp. 773–776, 2013.
[12] Behjat Siddiquie, Dave Chisholm, and Ajay Divakaran, "Exploiting multimodal affect and semantics to identify politically persuasive web videos," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 203–210, 2015.
[13] Gilad Mishne and Natalie Glance, "Predicting movie sales from blogger sentiment," *AAAI Spring Symposium*, pp. 155–158, 2016.
[14] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke, "Predicting imdb movie ratings using social media," *Advances in Information Retrieval*, pp. 503–507, 2012.
[15] Ailbhe Cullen and Naomi Harte, "Thin slicing to predict viewer impressions of ted talks," *The 14th International Conference on Auditory-Visual Speech Processing*, 2017.
[16] TJ Tsai, "Are you ted talk material? comparing prosody in professors and ted speakers," *Interspeech, ISCA*, p. 2534 2538, 2015.
[17] Fahim A. Salim, Killian Levacher, Owen Conlan, and Nick Campbell, "Examining multimodal characteristics of video to understand user engagement," *in Conf. on User Modelling, Adaptation, and Personalisation (UMAP)*, 2015.
[18] Bjorn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," *Interspeech*, 2016.
[19] Florian Eyben, Felix Weninger, Florian Gross, and Bjrn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," *In Proc. ACM Multimedia (MM), Barcelona, Spain*, pp. 835–838, October 2013.
[20] Ying Zhang and Huchuan Lu, "Deep cross-modal projection learning for image-text matching," in *ECCV*, 2018.